Measuring Semantic Similarity between Gene Ontology Terms

Francisco M. Couto [a] Mário J. Silva [a] Pedro M. Coutinho [b]

[a] *Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Portugal*

[b] *UMR 6098, CNRS and Universities Aix-Marseille I & II, Marseille, France*

Abstract

Many bioinformatics applications would benefit from comparing proteins based on their biological role rather than their sequence. This manuscript adds two new contributions. First, a study of the correlation between Gene Ontology (GO) terms and family similarity demonstrates that protein families constitute an appropriate baseline for validating GO similarity. Secondly, we introduce GraSM, a novel method that uses all the information in the graph structure of the Gene Ontology, instead of considering it as a hierarchical tree. GraSM gives a consistently higher family similarity correlation on all aspects of GO than the original semantic similarity measures.

Key words: Knowledge manipulation technique, Semantic Similarity, Gene Ontology, Bioinformatics

1 Introduction

The increasing importance of biological ontologies, motivates the development of similarity measures between concepts or, by extension, between entities annotated with these concepts [1]. Many Bioinformatics applications would benefit from using similarity measures to compare proteins based on what they do rather than using sequence similarity, a common technique to compare proteins based on how they are. For example, similarity measures can be applied to improve database querying, filter text-mining results and validate microarray clustering [2–5].

Most research on acquiring semantic properties of concepts has focused on semantic similarity, a research field that aims at calculating how similar two

concepts are based on their semantic properties, normally acquired from corpora [6]. Research on Information Theory developed many semantic similarity measures. Some of them calculate maximum likelihood estimates for each concept using the corpora, and then calculate the similarity between probability distributions. Rada et al. emphasized the use of semantic similarity in ontologies by combining the structure and content of an ontology with statistical information from corpora [7]. Following this approach, many semantic similarity measures applied to ontologies have been proposed. Resnik defined a semantic similarity measure based on the information content of the most informative common ancestor [8]. The information content of a concept is inversely proportional to its frequency in the corpora. Concepts that are frequent in the corpora have low information content. For example, the stop words (such as *the*) that occur almost everywhere in the text normally provide little semantic information. Jiang et al. proposed a semantic distance measure based on the difference between the information content of the concepts and the information content of their most informative common ancestor [9]. Lin proposed a semantic similarity measure based on the ratio between the information content of the most informative common ancestor and the information content of both concepts [10].

Recently, Lort et al. investigated the effectiveness of the above semantic similarity measures over the GO [11]. The GO (semantic) similarity between two proteins was calculated as the semantic similarity of their annotated GO terms. The study compared GO similarity using annotations found in the UniProt/SwissProt database to their sequence similarity [12]. The results showed that GO similarity is correlated with sequence similarity, i.e., they demonstrated the feasibility of using semantic similarity measures in a biological setting. However, the performance of the similarity measures was not uniform over the different aspects of GO, and it was not consistent with previous studies using different corpora either [13]. For example, Resnik's measure achieved the strongest correlation in the *molecular function* aspect and the weakest correlation in the *biological process* aspect. One explanation for the lack of uniformity and consistency can be the significant number of protein pairs with high GO similarity and low sequence similarity [14]. This was expected, because proteins sharing a biological role do not necessarily have a similar sequence [15].

Sequence similarity is not the only kind of structural similarity that can be computed between proteins. Family similarity is a structural similarity of a higher level then sequence similarity. Each family describes a set of related proteins, which can have identical molecular functions, are involved in the same process, or act in the same cellular location. Classifying proteins in families has been a common technique to organize them according to their biological role. For example, the most successful large-scale effort for increasing the coverage of GO annotations within the UniProt database is based on

the exploitation of family annotations [16]. Unlike standard sequence similarity methods, family categorization is normally based on experimental results about protein domains, which represent some evolutionarily conserved structure and have implications on the protein's biological role. Family similarity overcomes some of the limitations of sequence similarity, but the correlation between protein families and semantic similarity has not been studied.

This work extends the research presented above. The contributions of this work are:

- A study of the correlation between semantic similarities on GO and Pfam similarities. Pfam is a database of protein families assigned to UniProt proteins [17]. Pfam contains a mixture of manually curated and automatically generated protein families. Since proteins from same family share biological roles, the effectiveness of a semantic similarity measure defined over GO can be calculated based on its correlation with family similarity.
  By obtaining a uniform and consistent correlation between GO and family similarity over the different aspects of GO, this work provides a novel and stronger demonstration of the feasibility of semantic similarity measures in a biological setting.
- GraSM (Graph-based Similarity Measure), a novel method for incorporating the semantic richness of a graph by selecting disjunctive common ancestors of two concepts. Lord et al. computed the semantic similarity measures using GO as a hierarchic structure, i.e., they only considered the most informative common ancestor. However, GO is not organized as a tree-like hierarchy, but as directed acyclic graphs (DAG), one for each aspect. This enables a more complete and realistic annotation. When all but the most informative common ancestor nodes are ignored, different possible interpretations of the biologic concepts are disregarded. GraSM, on the other hand, selects and uses all the disjunctive common ancestors representing all interpretations.
  By obtaining a higher correlation using disjunctive common ancestors than only using the most informative common ancestor, this work demonstrates the higher effectiveness of GraSM for calculating semantic similarities between GO terms.

The rest of this manuscript is structured as follows: Section 2 provides a brief overview of GO; Section 3 presents state-of-the-art semantic similarity measures; Section 4 describes GraSM in detail; Section 5 explains how to calculate GO and family similarity between proteins; Section 6 presents the experimental evaluation and discusses the obtained results; and Section 7 expresses our main conclusions.

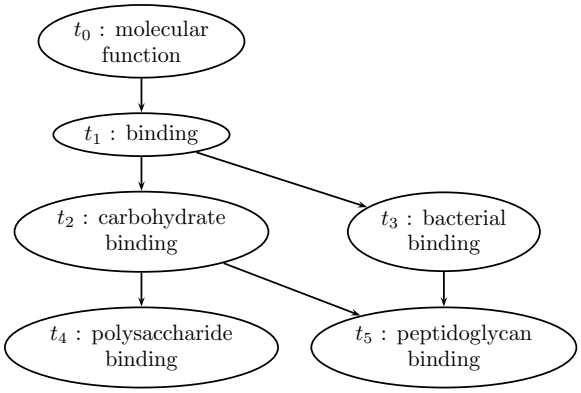


Fig. 1. Subgraph example of GO.

2 Gene Ontology

An ontology is a specification of a conceptualization that describes concepts and relationships used within a community [18]. A popular ontology used in Biology is GO. GO provides a structured controlled vocabulary of gene and protein biological roles, which can be applied to different species [19]. Since the activity or function of a protein can be defined at different levels, GO has three different aspects: *molecular function, biological process* and *cellular component.* Each protein has elementary molecular functions that normally are independent of the environment, such as catalytic or binding activities. Sets of proteins interact and are involved in cellular processes, such as metabolism, signal transduction or RNA processing. Proteins can act in different cellular localizations, such as nucleus or membrane.

GO organizes the terms in three directed acyclic graphs (DAG), one for each aspect. Each node of the graph represents a concept, and the edges represent the links between the concepts (see example in Figure 1). Links can represent two relationship types: *is-a* and *part-of.* GO is a dynamic hierarchy: its content changes every month with the publication of new release. GO is maintained by a group of curators, which add, remove and change terms and their relationships according to requests made by the research community. Any user can request modifications to GO. This prevents GO from becoming outdated and from providing incorrect information.

3 Semantic Similarity Measures

Semantic similarity measures can be used to calculate the similarity of two concepts organized in an ontology. The ontology structure defines the function $Parents(c)$ that, given a concept c, returns the set of more generic concepts directly linked to c. In the case of an ontology organized as a tree, $Parents(c)$

always returns a single concept. On the other hand, in GO, $Parents(c)$ can return more than one term (concept), because each aspect of GO is composed by a set of terms organized as a DAG. Using the function $Parents(c)$ the set of paths between two concepts c_a and c_b can be defined as:

$$\begin{aligned} &Paths(c_a, c_b) = \\ &\{\prec c_1, \ldots, c_n \succ \mid (c_a = c_1) \wedge (c_b = c_n) \wedge \\ &(\forall i : (1 \leq i < n) \wedge (c_i \in Parents(c_{i+1})))\}. \end{aligned} \tag{1}$$

A concept a is an ancestor of a concept c when there is at least one path from a to c:

$$Ancestors(c) = \{a \mid Paths(a, c) \neq \emptyset\}. \tag{2}$$

Note that since $\prec c \succ \in Paths(c, c)$, we have $c \in Ancestors(c)$.

The information content of a concept is inversely proportional to its frequency in a corpus. The frequency of a concept c, $Freq(c)$, can be defined as the number of times that c and all its descendants occur:

$$Freq(c) = \sum\{occur(c_i) \mid c \in Ancestors(c_i)\}. \tag{3}$$

Note that, for each ancestor a of a concept c, we have $Freq(a) \geq Freq(c)$, because the set of descendants of a contains all the descendants of c. An estimate for the occurrence of each GO term is the number of proteins annotated with it.

An estimate for the likelihood of observing an instance of a concept c is:

$$Prob(c) = \frac{Freq(c)}{maxFreq}, \tag{4}$$

where $maxFreq$ is the maximum frequency of all concepts. The $maxFreq$ of each aspect of GO is always equal to the frequency of the maximum term (root) in the DAG. For example, the GO term t =*molecular function* has $Prob(t) = 1$, because all the GO terms in the *molecular function* aspect are descendant of t, and therefore $Freq(t) = maxFreq$.

The information content of a concept c can be defined as the negative logarithm of its probability:

$$IC(c) = -\log(Prob(c)). \tag{5}$$

Note that the information content is monotonic, since it is non-increasing as we descend in the hierarchy.

Semantic similarity measures assume that the similarity between two concepts is related to the extent to which they share information. The common ancestors of two concepts c_1 and c_2 are:

$$CommonAnc(c_1, c_2) = Ancestors(c_1) \cap Ancestors(c_2). \tag{6}$$

Given two concepts c_1 and c_2, their shared information, $Share(c_1, c_2)$, can be defined as the information content of their most informative common ancestor:

$$Share(c_1, c_2) = max\{IC(a) \mid a \in CommonAnc(c_1, c_2)\}. \tag{7}$$

The most informative common ancestor is the one with the largest information content. Note that $Share(c, c) = IC(c)$, because $c \in Ancestors(c)$.

In case of a DAG with multiple roots, two concepts may not have any common ancestor. In these cases the shared information is zero. This never happens in GO, since it has a single root for each aspect.

Given two concepts c_1 and c_2, Resnik defined their semantic similarity as the information content of their most informative common ancestor:

$$Sim_{Resnik}(c_1, c_2) = Share(c_1, c_2). \tag{8}$$

Given two concepts c_1 and c_2, Jiang&Conrath defined their semantic distance as the difference between their information content and the information content of their most informative common ancestor:

$$dist_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times Share(c_1, c_2). \tag{9}$$

Note that Jiang&Conrath's formula measures a distance, the inverse of similarity. A similarity measure based on Jiang&Conrath distance measure can be defined as:

$$Sim_{JC}(c_1, c_2) = \frac{1}{dist_{JC}(c_1, c_2) + 1}. \tag{10}$$

$dist_{JC} + 1$ is used to avoid infinity values, since $dist_{JC}(c, c) = 0$.

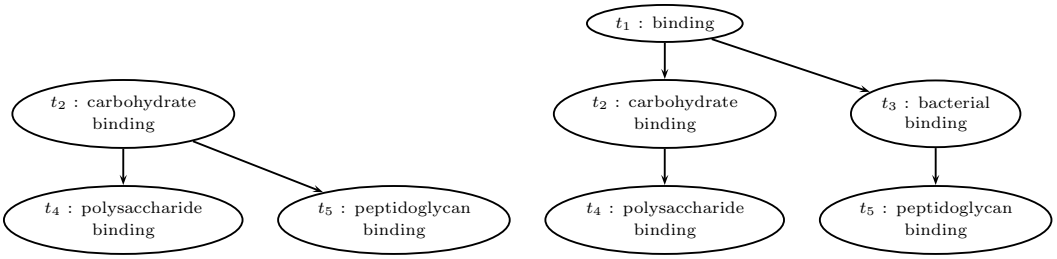


Fig. 2. *Carbohydrate binding* and *bacterial binding* are two common disjunctive ancestors of the terms *peptidoglycan binding* and *polysaccharide binding*, since there are two distinct paths from *peptidoglycan binding* to *carbohydrate binding* and *binding*.

Given two concepts, c_1 and c_2, Lin defined their similarity as the information content of their most informative common ancestor over their information content:

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times Share(c_1, c_2)}{IC(c_1) + IC(c_2)}. \tag{11}$$

4 GraSM

The semantic similarity measures described above only use the most informative common ancestor of both concepts. Therefore, when applied to a DAG, these measures discard other common ancestors even if they are disjunctive ancestors. GraSM assumes that two common ancestors are disjunctive if there are independent paths from both ancestors to the concept [20]. Independent paths mean those that use at least one concept of the ontology not used by the other paths. Two disjunctive ancestors of a concept represent two distinct interpretations of a concept. For example, Figure 2 shows that *carbohydrate binding* and *bacterial binding* are two common disjunctive ancestors of *peptidoglycan binding* and *polysaccharide binding*, since there are two distinct paths from *peptidoglycan binding* to *carbohydrate binding* and *binding*. Thus, the similarity between *peptidoglycan binding* and *polysaccharide binding* is smaller than if *peptidoglycan binding* only had the ancestor *carbohydrate binding*. The similarity is smaller because *peptidoglycan binding* can also be interpreted as *bacterial binding,* which is not an ancestor of *polysaccharide binding.*

Calculating the similarity between two concepts using just the most informative common ancestor only accounts for one of the interpretations. However, similarity measures should also account for other interpretations of both concepts. GraSM selects all the common disjunctive ancestors of two concepts in a DAG to calculate their similarity.

GraSM considers that a_1 and a_2 represent disjunctive ancestors of c if there is

a path from a_1 to c not containing a_2 and a path from a_2 to c not containing a_1:

$$
\begin{aligned}
&DisjAnc(c) = \\
&\{(a_1, a_2) \mid \\
&(\exists p : (p \in Paths(a_1, c)) \wedge (a_2 \notin p)) \wedge \\
&(\exists p : (p \in Paths(a_2, c)) \wedge (a_1 \notin p))\}
\end{aligned}
\tag{12}
$$

Note that if $a_1 \notin Ancestors(a_2)$ and $a_2 \notin Ancestors(a_1)$ then a_1 and a_2 are disjunctive ancestors of c. For example, in Figure 1 $(t_2, t_3) \in DisjAnc(t_5)$. Otherwise, if $a_1 \in Ancestor(a_2)$ it is still possible that a_1 and a_2 represent disjunctive ancestors of c. For example, in Figure 1 $(t_1, t_2) \in DisjAnc(t_5)$ because the path $\prec t_1, t_3, t_5 \succ$ does not pass through t_2, and the path $\prec t_2, t_5 \succ$ does not pass through t_1.

Given two concepts c_1 and c_2, their common disjunctive ancestors are the most informative common ancestor of disjunctive ancestors of c_1 and c_2, i.e., a_1 is a common disjunctive ancestor of c_1 and c_2 if for each ancestor a_2 more informative than a_1, a_1 and a_2 are a disjunctive ancestor of c_1 or c_2:

$$
\begin{aligned}
&CommonDisjAnc(c_1, c_2) = \\
&\{a_1 \mid a_1 \in CommonAnc(c_1, c_2) \wedge \\
&\forall a_2 : [(a_2 \in CommonAnc(c_1, c_2)) \wedge \\
&(IC(a_1) \leq IC(a_2)) \wedge (a_1 \neq a_2)] \Rightarrow \\
&[(a_1, a_2) \in (DisjAnc(c_1) \cup DisjAnc(c_2))]\}
\end{aligned}
\tag{13}
$$

Note that $CommonDisjAnc(c, c) = \{c\}$ because all the ancestors of c are not disjunctive ancestors of c, i.e., $(c, a) \notin DisjAnc(c)$ for all $a \in Ancestors(c)$. In Figure 1, $CommonDisjAnc(t_4, t_5) = \{t_1, t_2\}$ because t_2 is the most informative common ancestor, and t_1 and t_2 are disjunctive ancestors of t_5.

GraSM defines the shared information between c_1 and c_2 as the average of the information content of their common disjunctive ancestors:

$$
\begin{aligned}
&Share_{GraSM}(c_1, c_2) = \\
&\overline{\{IC(a) \mid a \in CommonDisjAnc(c_1, c_2)\}}.
\end{aligned}
\tag{14}
$$

$Share$ can be replaced by $Share_{GraSM}$ yielding three new variants of the semantic similarity measures presented in the previous Section: $Sim_{ResnikGraSM}$, $Sim_{JCGraSM}$ and $Sim_{LinGraSM}$.

4.1 Disjunctive Ancestors Redefinition

Formula 12 can be replaced by the following equivalent formula:

$$
\begin{aligned}
&DisjAnc(c) = \\
&\{(a_1, a_2), (a_2, a_1) \mid IC(a_1) \leq IC(a_2) \\
&(\forall n, m, k\ (|Paths(a_1, c)| = m \wedge |Paths(a_2, c)| = n \wedge \\
&|Paths(a_1, a_2)| = k) \Rightarrow (m > 0 \wedge n > 0 \wedge m > n \times k)\}
\end{aligned}
\tag{15}
$$

The two formulas are equivalent because when $IC(a_1) \leq IC(a_2)$ there are no paths from a_2 to a_1, then any path from a_2 to c does not contain a_1. Additionally, any path from a_1 to c containing a_2 uses one of the k paths from a_1 to a_2 and one of the n paths from a_2 to c, thus there are $n \times k$ independent paths from a_1 to c containing a_2. Therefore, if $m > n \times k$ there is at least one path from a_1 to c not containing a_2.

On the other hand, if exists a path from a_2 to c not containing a_1 and a path a_1 to c not containing a_2, then we have $m > 0 \wedge n > 0$ and assuming $IC(a_1) \leq IC(a_2)$ we have $m > n \times k$ because m includes all the paths from a_1 to c containing a_2 plus the paths (exists at least one) not containing it.

4.2 Example

By assuming a different number of proteins annotated to each GO term of Figure 1, Table 1 shows the information content of these terms. Considering only the subgraph of GO represented in Figure 1 and the values in Table 1, the set of common ancestors of t_4 and t_5 in a descendant order of IC is $\{t_2, t_1, t_0\}$, and the set of common disjunctive ancestors is $CommonDisjAnc(t_4, t_5) = \{t_2, t_1\}$, as described above. Thus, $Share_{GraSM}(t_4, t_5) = \overline{\{IC(t_2), IC(t_1)\}} = 1.5$. The similarity between t_4 and t_5 with and without using GraSM is:

$$
\begin{aligned}
&Sim_{Resnik}(t_4, t_5) \\
&= Share(t_4, t_5) = 2
\end{aligned}
$$

$$
\begin{aligned}
&Sim_{ResnikGraSM}(t_4, t_5) \\
&= Share_{GraSM}(t_4, t_5) = 1.5
\end{aligned}
$$

$$
Sim_{JC}(t_4, t_5)
$$

Table 1
The information content of the GO terms presented in Figure 1 by considering for each term a different number of proteins annotated with it.

GO term	Protein Annotations	$Freq$	$Prob$	IC
t_0	8	16	1	0
t_1	3	8	0.5	1
t_2	2	4	0.25	2
t_3	1	2	0.125	3
t_4	1	1	0.0625	4
t_5	1	1	0.0625	4

$$= \frac{1}{IC(t_4) + IC(t_5) - 2 \times Share(t_4, t_5)}$$
$$= \frac{1}{4 + 4 - 2 \times 2} = 0.25$$

$$Sim_{JCGraSM}(t_4, t_5)$$
$$= \frac{1}{IC(t_4) + IC(t_5) - 2 \times Share_{GraSM}(t_4, t_5)}$$
$$= \frac{1}{4 + 4 - 2 \times 1.5} = 0.2$$

$$Sim_{Lin}(t_4, t_5)$$
$$= \frac{2 \times Share(t_4, t_5)}{IC(t_4) + IC(t_5)}$$
$$= \frac{2 \times 2}{4 + 4} = 0.5$$

$$Sim_{LinGraSM}(t_4, t_5) =$$
$$= \frac{2 \times Share_{GraSM}(t_4, t_5)}{IC(t_4) + IC(t_5)}$$
$$= \frac{2 \times 1.5}{4 + 4} = 0.375$$

If the shared information of one ancestor is high, and then we find another disjunctive common ancestor with lower information content, it seems that finding the additional relationship should increase the similarity rather than

Algorithm 1 $Share_{GraSM}(c_1, c_2)$

1: $Anc = CommonAnc(c_1, c_2)$
2: $CommonDisjAnc = \{\}$
3: **for all** $a \in sortDescByIC(Anc)$ **do**
4: isDisj = true
5: **for all** $cda \in CommonDisjAnc$ **do**
6: $isDisj = isDisj \wedge$
$(DisjAnc(c_1, (cda, a)) \vee DisjAnc(c_2, (cda, a)))$
7: **end for**
8: **if** $isDisj$ **then**
9: $addTo(CommonDisjAnc, a)$
10: **end if**
11: **end for**
12: $shared = 0$
13: **for all** $cda \in CommonDisjAnc$ **do**
14: $shared \mathrel{+}= IC(cda)$
15: **end for**
16: **return** $shared/sizeof(CommonDisjAnc)$

Algorithm 2 $DisjAnc(c, (a_1, a_2))$

Require: $IC(a_1) \leq IC(a_2)$
1: $nPaths = |Paths(a_1, a_2)|$
2: $nPaths_1 = |Paths(a_1, c)|$
3: $nPaths_2 = |Paths(a_2, c)|$
4: **return** $nPaths_1 \geq nPaths \times nPaths_2$

lessen it. However, this is an incorrect intuition. Finding a disjunctive common ancestor means that at least one of the terms has a distinct and more distant interpretation to the other term, which makes the terms less similar. Thus, by taking in account the less informative common ancestor, GraSM provides lower similarities than the original measures.

4.3 Computational Aspect

Algorithm 1 describes a possible implementation of $Share_{GraSM}$. It starts by selecting the common ancestors of both concepts (line 1) and by initializing the list of common disjunctive ancestors as a empty list (line 2). The algorithm selects each common ancestor in descending order of information content (line 3). For each selected ancestor, the algorithm checks if the ancestor is disjunctive to all the common disjunctive ancestors already selected (lines 4 to 7). If the ancestor is disjunctive, it adds it to the list of common disjunctive ancestors (line 9). At the end, the algorithm calculates the average of the information content of all the ancestors in the common disjunctive ancestors list

Table 2
Semantic similarity measures' range and normalization parameters.

	Smallest	Largest	a	b
Resnik	0	$maxIC$	$\frac{1}{maxIC}$	0
Jiang&Conrath	$\frac{1}{2\times maxIC+1}$	1	$1+\frac{1}{2\times maxIC}$	$-\frac{1}{2\times maxIC}$
Lin	0	1	1	0

(lines 12 to 16).

Algorithm 2 describes an efficient technique to check if a pair of ancestors (a_1, a_2) are disjunctive ancestors of a given concept c based on Formula 15. The algorithm checks if the number of paths from a_1 to c is larger than the multiplication of the number of paths from a_1 to a_2 and from a_2 to c.

These implementations show that using GraSM is not prohibitively expensive. In addition to finding the common ancestors, as $Share$, $Share_{GraSM}$ only has to check the list of common ancestors, which is normally much smaller than the depth of the graph. Counting the number of paths is also not time-consuming. For example, in the GO distribution there is a table that stores each path between two GO terms. Therefore $Share_{GraSM}$ has a worst-case performance $O(k^2)$, where k is the maximum number of common ancestors of two terms.

4.4 Normalization

To compare the performance of all measures, Resnik and Jiang&Conrath's measures were normalized to range from 0 to 1 (as Lin's measure), using a linear function $f(x) = a \times x + b$. Table 2 presents the parameters a and b defined according to the largest and smallest value of each measure, where $maxIC$ represents the maximum information content obtained in the respective aspect of GO. The range of each measure was derived from the observation that the information content of a term is always larger than the information content of its ancestors, and the information content of a term ranges from 0 to $maxIC$. The parameters are not affected by using GraSM, since GraSM does not modify the range of the shared information.

Assuming that $maxIC = 10$, the normalized similarities of the previous example are:

$$Sim_{Resnik}(t_4, t_5) = \frac{2}{maxIC} = 0.2$$

$$Sim_{ResnikGraSM}(t_4, t_5) = \frac{1.5}{maxIC} = 0.15$$

$$\begin{aligned} &Sim_{JC}(t_4, t_5) \\ &= (1 + \frac{1}{2 \times maxIC}) \times 0.25 - \frac{1}{2 \times maxIC} \\ &= 0.2125 \end{aligned}$$

$$\begin{aligned} &Sim_{JCGraSM}(t_4, t_5) \\ &= (1 + \frac{1}{2 \times maxIC}) \times 0.2 - \frac{1}{2 \times maxIC} \\ &= 0.16 \end{aligned}$$

The similarities using Lin's measure remain the same after normalization.

5 Protein Similarity

The performance of each measure was evaluated based on the correlation between GO and family similarity. The GO similarity between two proteins is the average similarity of the GO terms annotated to them. Since proteins have simultaneous biological roles, their similarity uses for each term annotated to each protein its similarity to the most similar term annotated to the other protein:

$$GOSim(p_1, p_2) = \frac{GOSim(p_1, Terms(p_2)) + GOSim(p_2, Terms(p_1))}{2} \tag{16}$$

$$GOSim_(p_1, Terms_2) = \overline{\{GOSim(t_1, Terms_2) \mid t_1 \in Terms(p_1)\}} \tag{17}$$

$$GOSim(t_1, Terms_2) = max\{Sim(t_1, t_2) \mid t_2 \in Terms_2\} \tag{18}$$

The function $Terms(p)$ gives all the terms assigned to the protein p in the GOA (Gene Ontology Annotation) database, which provides GO annotations of UniProt proteins [12, 16].

The family similarity between two proteins is the number of families they share:

$$FamSim(p_1, p_2) = |Fam(p_1) \cap Fam(p_2)|. \tag{19}$$

The function $Fam(p)$ gives all the Pfam families assigned to the protein p in UniProt.

The goal is to measure the correlation between GO and family similarity using different semantic similarity measures to calculate the GO similarity. The correlation coefficients were calculated using the following formula:

$$corr(x, y) = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}, \tag{20}$$

where x_i and y_i are the GO and family similarities, respectively.

5.1 Example

Consider two proteins p_a and p_b annotated to $\{t_1, t_2\}$ and $\{t_3, t_4\}$, respectively. Assume $Sim(t_1, t_3) = 0.8$, $Sim(t_1, t_4) = 0.6$, $Sim(t_2, t_4) = 0.4$ and $Sim(t_2, t_3) = 0.2$. Assume also that p_a and p_b belong to the families $\{PF001, PF002\}$ and $\{PF002, PF003\}$, respectively. The GO similarity between p_a and p_b is calculated using Formula 16 as follows:

$$GOSim(p_a, p_b) = \frac{GOSim(p_a, \{t_3, t_4\}) + GOSim(p_b, \{t_1, t_2\})}{2}$$

The similarity between p_a and the terms annotated with p_b is calculated using Formula 17 as follows:

$$GOSim(p_a, \{t_3, t_4\}) = \overline{\{GOSim(t_1, \{t_3, t_4\}), GOSim(t_2, \{t_3, t_4\})\}}$$

The similarity between p_b and the terms annotated with p_a is calculated using Formula 17 as follows:

$$GOSim(p_b, \{t_1, t_2\}) = \overline{\{GOSim(t_3, \{t_1, t_2\}), GOSim(t_4, \{t_1, t_2\})\}}$$

The similarities between each term and each set of terms is calculated using Formula 18 as follows:

$$GOSim(t_1, \{t_3, t_4\}) = \\ max\{Sim(t_1, t_3), Sim(t_1, t_4)\} = Sim(t_1, t_3) = 0.8$$

$$GOSim(t_2, \{t_3, t_4\}) = \\ max\{Sim(t_2, t_3), Sim(t_2, t_4)\} = Sim(t_2, t_4) = 0.4$$

$$GOSim(t_3, \{t_1, t_2\}) = \\ max\{Sim(t_3, t_1), Sim(t_3, t_2)\} = Sim(t_1, t_3) = 0.8$$

$$GOSim(t_4, \{t_1, t_2\}) = \\ max\{Sim(t_4, t_1), Sim(t_4, t_2)\} = Sim(t_1, t_4) = 0.6$$

Using these values in the above formulas we get:

$$GOSim(p_a, \{t_3, t_4\}) = \overline{\{0.8, 0.4\}} = 0.6$$

$$GOSim(p_b, \{t_1, t_2\}) = \overline{\{0.8, 0.6\}} = 0.7$$

$$GOSim(p_a, p_b) = \frac{0.6 + 0.7}{2} = 0.65$$

The family similarity between p_a and p_b is calculated using Formula 19 as follows:

$$FamSim(p_a, p_b) = |\{PF002\}| = 1$$

6 Results

GraSM was evaluated with the 500 proteins with the largest number of GO annotations from the December 2004 releases of UniProt and GO. These proteins were annotated to 234 distinct Pfam families, with an average of 2.5

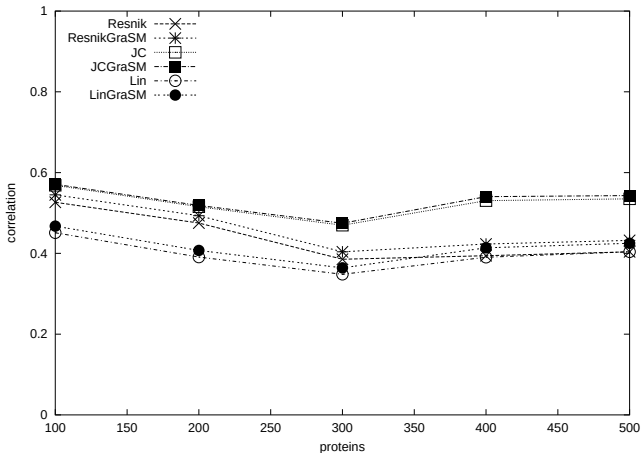
Table 3
Correlation coefficients for each aspect of GO and each semantic similarity measure with and without using GraSM.

	Resnik			Jiang&Conrath			Lin		
	original	GraSM	increase	original	GraSM	increase	original	GraSM	increase
Function	0.404	0.432	6.9%	0.535	0.543	1.5%	0.404	0.426	5.4%
Process	0.246	0.365	48.4%	0.697	0.725	4.0%	0.418	0.526	25.8%
Component	0.216	0.272	25.9%	0.306	0.310	1.3%	0.255	0.279	9.4%

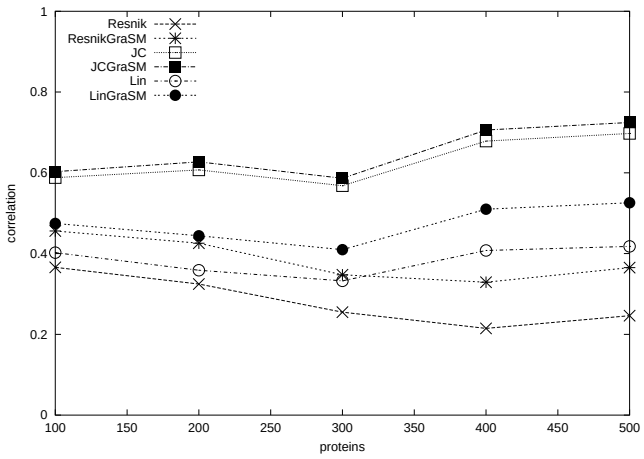
families per protein. Figure 3 presents the correlation for the top-k most annotated proteins, with k ranging from 100 to 500. The average number of GO annotations for the top 100 proteins is 33.3, and 23.8 for the top 500 proteins. The decrease in the number of annotations does not affect the correlation, since there is a stable correlation for the different top-k protein sets. This shows that, on the tested proteins, the number of annotations does not bias the correlation.

Table 3 presents the correlation coefficients obtained by all measures for the top 500 proteins. The results show a strong correlation between GO and family similarity. Having a strong correlation means that GO similarity should increase as we select protein pairs that share more families. Figure 4 shows this behavior by presenting how much GO similarity increases as compared with the GO similarity of protein pairs not sharing any family. The GO similarity increase was calculated as the ratio of the average GO similarity of protein pairs sharing a certain number of families to the average GO similarity of protein pairs not sharing any family. All charts show that GraSM outperforms the original measures without exception.

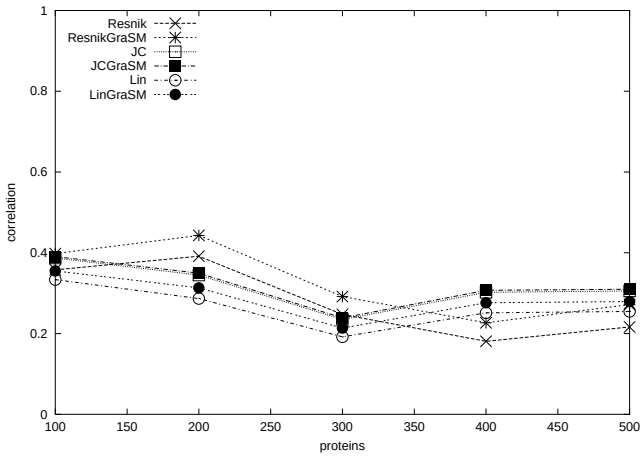
Of the three measures, the one proposed by Jiang&Conrath obtained the strongest correlation in all aspects. Lin's measure obtained a stronger or equivalent correlation than Resnik's measure in all aspects. Resnik's measure assigns an identical similarity to many different pairs of GO terms, since it only uses the information content of the shared ancestor. This explains why Resnik's measure obtained the lowest correlation coefficients. The ranking of the measures is consistent with previous studies using different corpora, and it is uniform over the different aspects of GO [13]. On the contrary, the measure ranking obtained by Lord et al. was neither consistent nor uniform [11]. In their study, there was a different ranking for each aspect: the strongest correlation in the *biological process* aspect was achieved by Jiang&Conrath's measure, in the *cellular component* aspect by Lin's measure, and in the *molecular function* aspect by Resnik's measure. The correlation coefficients obtained in this study are not directly comparable to the ones obtained by Lord et al., since this work is measuring a different correlation using more recent UniProt and GO releases. However, the uniformity and consistency demonstrates that family similarity is more appropriate to validate semantic similarity measures than sequence similarity. This was expected, since family similarity is less error-prone than sequence similarity for creating and validating GO annota-



(a) molecular function

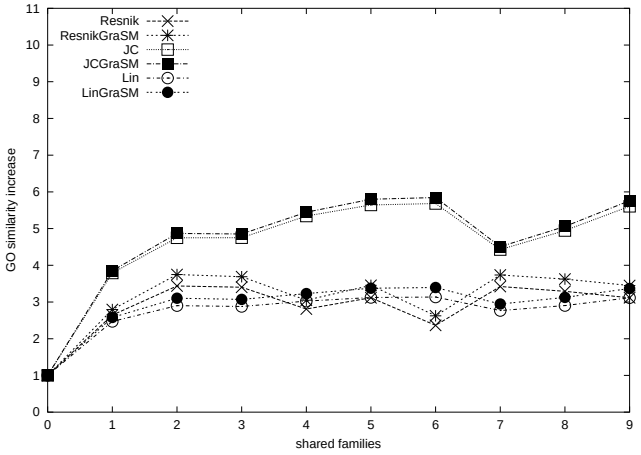


(b) biological process

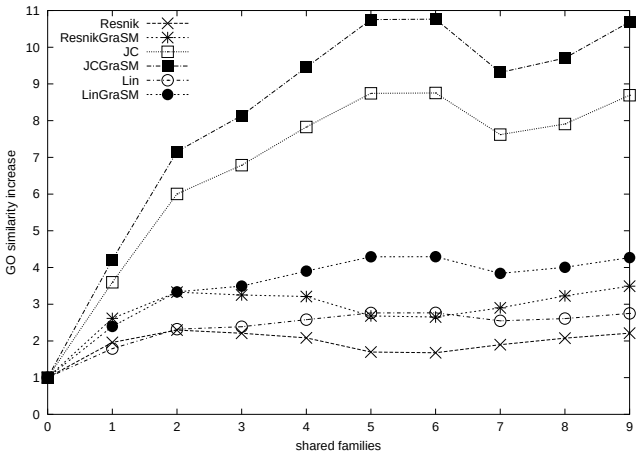


(c) cellular component

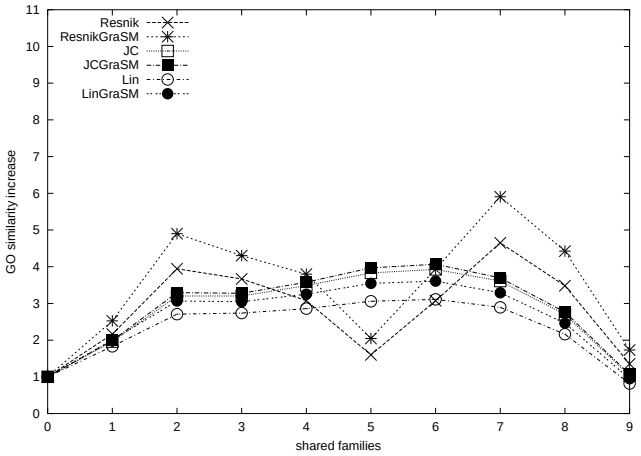
Fig. 3. The charts compare the correlation between GO and family similarity obtained by different semantic similarity measures over the top-k most annotated proteins, with k ranging from 100 to 500.



(a) molecular function

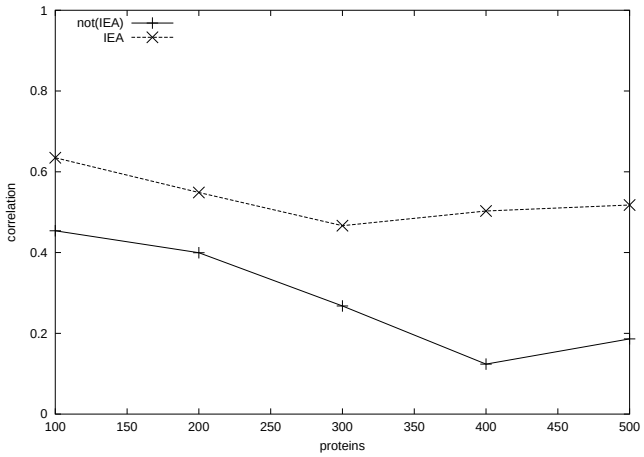


(b) biological process

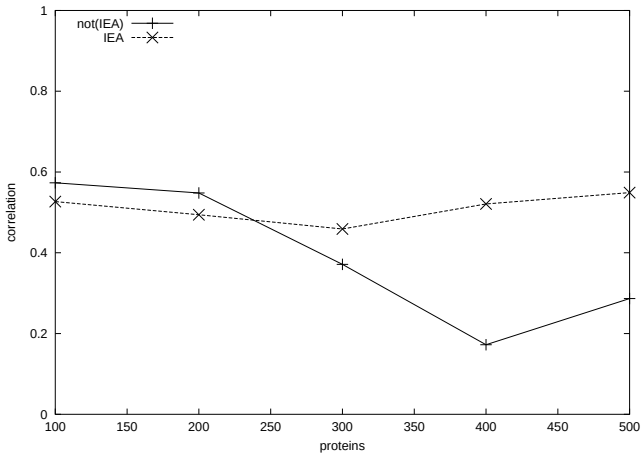


(c) cellular component

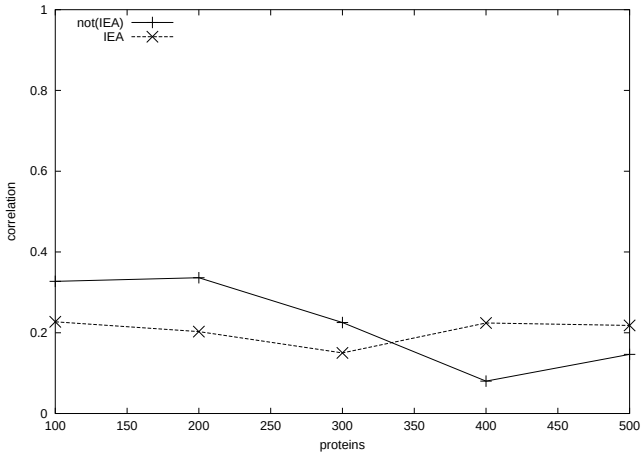
Fig. 4. The charts show the GO similarity increase as we select protein pairs sharing more families. The GO similarity increase is the ratio of the average GO similarity of protein pairs sharing a certain number of families to the average GO similarity of protein pairs not sharing any family.



(a) molecular function



(b) biological process



(c) cellular component

Fig. 5. The charts compare the correlation of automated (IEA) and manual (not(IEA)) annotations using $Sim_{JCGraSM}$ over the top-k most annotated proteins, with k ranging from 100 to 500.

Table 4
Statistics of the version of GO used in the evaluation.

	Function	Process	Component
terms	7437	9055	1489
edges	8664	14356	2157
density	1.16	1.59	1.45

tions [21].

Table 3 shows a higher correlation in the *biological process* aspect, and a lower correlation in the *cellular component* aspect. This was expected, because the *biological process* aspect has the largest number of terms and edges, more than 50% of all GO, and because the *cellular component* aspect only contains 8% of all the terms in GO. Moreover, proteins that share a family can be localized in distinct cellular components. Resnik's measure is the most affected by having such a small amount of terms, since in this aspect a significant number of GO terms pairs have the same common ancestor. This explains the instability of the correlation obtained by Resnik's measure in Figure 4(c).

GraSM increased the correlation of all the semantic similarity measures tested. This shows that using disjunctive ancestors to calculate the shared information of two terms improves the effectiveness of semantic similarity measures. The improvement is proportional to the density (number of edges over the number of terms) of each aspect (see Table 4). This was expected, because having more edges per term increases the probability of having multiple common disjunctive ancestors. For example, the highest improvement (48%) is in the *biological process* aspect, which also has the largest density.

To cope with the large amount of sequences being produced, a significant number of proteins have been functionally characterized by automated tools, which produce less precise and more generic annotations than manual annotations. Figure 5 compares the correlation obtained by automated and manual annotations using $Sim_{JCGraSM}$ over the top-k most annotated proteins, with k ranging from 100 to 500. The correlation of manual annotations is less uniform than the correlation of automated annotations. Manual annotations have a higher quality than automated annotations, but manual curation is a time-consuming task that currently covers less than 5% of UniProt. Thus, the manual annotations are not so well distributed as automated annotations. For example, the set of tested proteins contains proteins manually annotated to more than 60 GO terms and proteins manually annotated to less than 6 GO terms. In addition, most automated annotations are generic and, therefore, closer to family annotations. Thus, the unbalanced distribution and specificity of manual annotations explains why their correlation is less uniform and most of the times smaller than the correlation of automated annotations.

Similarity between the GO terms annotated with:
P16403 (*H12_HUMAN*) and Q8NHM5 (*FXLA_HUMAN*)

P16403's terms	Q8NHM5's terms	Terms' Similarity	Weighted Similarity
function			
GO:0003677 (*DNA binding*)	GO:0003677 (*DNA binding*)	100.0%	17.9%
GO:0003677 (*DNA binding*)	GO:0008270 (*zinc ion binding*)	11.2%	3.4%
process			
GO:0006334 (*nucleosome assembly*)	GO:0006355 (*regulation of transcription, DNA-dependent*)	6.0%	2.6%
GO:0007001 (*chromosome organization and biogenesis (sensu Eukaryota)*)	GO:0006355 (*regulation of transcription, DNA-dependent*)	6.1%	2.1%
GO:0007001 (*chromosome organization and biogenesis (sensu Eukaryota)*)	GO:0000004 (*biological_process unknown*)	5.6%	1.9%
GO:0007001 (*chromosome organization and biogenesis (sensu Eukaryota)*)	GO:0006512 (*ubiquitin cycle*)	4.3%	1.6%
component			
GO:0005634 (*nucleus*)	GO:0005634 (*nucleus*)	100.0%	18.4%
GO:0000786 (*nucleosome*)	GO:0005634 (*nucleus*)	7.0%	2.9%
GO:0005694 (*chromosome*)	GO:0005634 (*nucleus*)	8.7%	2.8%

Fig. 6. Output of FuSSiMeG containing the functional semantic similarity of P42973 and O85465 UniProt proteins using the $Sim_{JCGraSM}$ measure.

GO has two types of edges *is-a* and *part-of*. The results presented in this manuscript were obtained by using both types of edges. Since more than 90% of the edges are *is-a,* using different edge weights has almost no effect on the results. An ontology normally starts by adding the terms and simple relationships to provide a complete coverage of the target domain. Over time, the ontology tends to grow less in the number of terms than in the number of relationships. We believe that GO is not an exception, and therefore the quantity and quality of the relationships will improve. The number of pairs of terms having multiple ancestors will grow, and therefore make GraSM even more effective than tree-based semantic similarity measures.

7 Conclusions

This manuscript shows a strong correlation between GO and family similarity. The correlation is more stable than shown before for sequence similarity.

Jiang&Conrath's measure obtained the best performance with Lin's measure following. The worst performance was obtained by Resnik's measure. This ranking was uniform over the different aspects of GO and consistent with previous studies. This shows that family similarity is an appropriate baseline for validating GO similarity. Therefore, this manuscript provides a novel and stronger demonstration of the feasibility of semantic similarity measures in a biological setting.

This manuscript presents GraSM, a novel measure that incorporates the semantic richness of a graph by selecting disjunctive common ancestors of two concepts. GraSM obtained a higher correlation using disjunctive common ancestors than only using the most informative common ancestor, which demonstrates the effectiveness of GraSM for calculating the similarity between GO terms. It is expected that this improvement will increase over time as GO captures more relationships from the biological domain. GraSM is not specific to GO and can also be applied to other graph-structured taxonomies, such as WordNet [22]. We expect that GraSM will improve the calculation of conceptual similarity between words with multiple senses. For example, when comparing the words *Fluid* with *Fluent* or *Liquid*. Thus, future research may involve the application of GraSM to other taxonomies.

All the semantic similarity measures described in this document were implemented by FuSSiMeG (Functional Semantic Similarity Measure between Gene-Products), which measures the functional similarity between proteins based on the semantic similarity of the GO terms annotated to them [23]. FuSSiMeG is publicly available on the Web (`http://xldb.fc.ul.pt/rebil/tools/ssm/`), affording the similarity calculation on the fly. Figure 6 shows the semantic similarities between the GO terms annotated to two given proteins, which are displayed by FuSSiMeG using the $Sim_{LinGraSM}$ measure. Besides the similarity of the annotated GO terms, their specificity cannot be disregarded when comparing two proteins. For example, both proteins can be annotated with a generic GO term (100% similarity), but this does not mean that they are similar since many other proteins are also annotated to this term. Thus, FuSSiMeG displays the weighted similarity between the GO terms, which divides the semantic similarity by the information content of both terms.

References

[1] R. Stevens, C. Wroe, P. Lord, C. Goble, Handbook on Ontologies, Springer, 2003, Ch. Ontologies in Bioinformatics.

[2] F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, Fatigo: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics 20 (4) (2004) 578–580.

[3] J. Sevilla, V. Segura, A. odhorski, E. Guruceaga, J. Mato, L. Martinez-Cruz, F. Corrales, A. Rubio, Correlation between gene expression and go semantic similarity, IEEE/ACM Transactions Computational Biology Bioinformatics 2 (4) (2005) 330–338.

[4] F. Couto, M. Silva, Advanced Data Mining Techonologies in Bioinformatics, Idea Group Inc., 2006, Ch. Mining the BioLiterature: towards automatic annotation of genes and proteins.

[5] F. Couto, M. Silva, P. Coutinho, Finding genomic ontology terms in text using evidence content, BMC Bioinformatics 6 (S1) (2005) S21.

[6] C. Manning, H. Schütze, Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

[7] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems 19 (1) (1989) 17–30.

[8] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proc. of the 14th International Joint Conference on Artificial Intelligence, 1995.

[9] J. Jiang, D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proc. of the 10th International Conference on Research on Computational Linguistics, 1997.

[10] D. Lin, An information-theoretic definition of similarity, in: Proc. of the 15th International Conference on Machine Learning, 1998.

[11] P. Lord, R. Stevens, A. Brass, C. Goble, Semantic similarity measures as tools for exploring the Gene Ontology, in: Proc. of the 8th Pacific Symposium on Biocomputing, 2003.

[12] R. Apweiler, A. Bairoch, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi, L. Yeh, UniProt: the universal protein knowledgebase, Nucleic Acids Research 32 (Database issue) (2004) D115–D119.

[13] A. Budanitsky, G. Hirst, Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, in: Proc. of the Workshop on WordNet and Other Lexical Resources co-located with the 2nd North American Chapter of the Association for Computational Linguistics, 2001.

[14] P. Lord, R. Stevens, A. Brass, C. Goble, Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, Bioinformatics 19 (10) (2003) 1275–1283.

[15] D. Devos, A. Valencia, Intrinsic errors in genome annotation, Trends Genetics 17 (8) (2001) 429–431.

[16] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, R. Apweiler, The Gene Ontology Annotations (GOA) database: sharing knowledge in UniProt with Gene Ontology, Nucleic Acids Research 32 (2004) 262–166.

[17] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, S. Eddy, The Pfam protein families database, Nucleic Acids Research 32 (Database issue) (2004) D138–D141.

[18] T. Gruber, A translation approach to portable ontologies., Knowledge Acquisition 5 (2) (1993) 199–220.

[19] GO-Consortium, The Gene Ontology (GO) database and informatics resource, Nucleic Acids Research 32 (Database issue) (2004) D258–D261.

[20] F. Couto, M. Silva, P. Coutinho, Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors, in: Proc. of the ACM Conference in Information and Knowledge Management as a short paper, 2005.

[21] T. Doerks, A. Bairoch, P. Bork, Protein annotation: detective work for function prediction, Trends Genetics 14 (6) (1998) 248–250.

[22] C. Fellbaum, WordNet: An Electronic Lexical Database, The MIT Press., 1998.

[23] F. Couto, M. Silva, P. Coutinho, Implementation of a functional semantic similarity measure between gene-products, DI/FCUL TR 03–29, Department of Informatics, University of Lisbon (November 2003).