Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors

Francisco M. Couto
Dep. de Informática
Faculdade de Ciências da Universidade de Lisboa
fcouto@di.fc.ul.pt

Mário J. Silva
Dep. de Informática
Faculdade de Ciências da Universidade de Lisboa
mjs@di.fc.ul.pt

Pedro M. Coutinho
AFMB
CNRS, Marseille
pedro.coutinho@afmb.cnrs-mrs.fr

ABSTRACT

Many bioinformatics applications would benefit from comparing proteins based on their biological role rather than their sequence. In most biological databases, proteins are already annotated with ontology terms. Previous studies identified a correlation between the sequence similarity and the semantic similarity of proteins. The semantic similarity of proteins was computed from their annotated GO terms. However, proteins sharing a biological role do not necessarily have a similar sequence.

This paper introduces our study of the correlation between GO and family similarity. Family similarity overcomes some of the limitations of sequence similarity, thus we obtained a strong correlation between GO and family similarity. Additionally, this paper introduces GraSM, a novel method that uses all the information in the graph structure of the GO, instead of considering it as a hierarchical tree. When calculating the semantic similarity of two concepts, GraSM selects the disjunctive common ancestors rather than only using the most informative common ancestor. GraSM produced a higher family similarity correlation than the original semantic similarity measures.

Categories and Subject Descriptors: I.5.3 [Pattern Recognition]: Clustering - Similarity measures; J.3 [Life and Medical Sciences]: Biology and genetics

General Terms: Algorithms, Experimentation

Keywords: Family Correlation, Gene Ontology, Graph-based Similarity Measure

1. INTRODUCTION

Given the increasing importance of biological ontologies, mechanisms enabling users to measure the similarity between the concepts or, by extension, between the entities annotated with these concepts are required. For example, they can be applied to improve text mining systems [3, 7]. GO (Gene Ontology) has become one of the most important ontologies to annotate proteins. GO provides a structured controlled vocabulary of gene and protein biological roles describing the following aspects: function, process and component.

Many SS (Semantic Similarity) measures applied to ontologies have been proposed. Resnik defined a SS measure based on the information content of the most informative common ancestor [8]. Jiang&Conrath 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 [4]. Lin proposed a SS measure based on the ratio between the information content of the most informative common ancestor and the information content of both concepts [5].

Recently, Lord et al. investigated the effectiveness of the SS measures mentioned above over the GO [6]. The results have shown that GO similarity is correlated with sequence similarity, i.e. they have demonstrated the feasibility of using SS measures in a biological setting. However, the performance of the similarity measures has not been uniform over the different aspects of GO, and it has not been consistent with previous studies using different corpora either [1]. Sequence similarity is not the only kind of structural similarity that can be computed between proteins. Family similarity is also a structural similarity, which is normally based on experimental results about protein domains representing some evolutionarily conserved structure with implications on the protein's biological role.

This paper outlines our main contributions:

- A study of the correlation between GO semantic similarity and Pfam similarity. Pfam is a database of protein families assigned to UniProt proteins. Pfam contains families manually curated and others automatically generated. Since proteins from a same family share biological roles, we measure the effectiveness of a SS measure defined over GO based on its correlation with family similarity.

- 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 SS measures using GO as an hierarchic structure, i.e. they only considered the most informative common ancestor. However, GO is not organized as a tree-like hierarchy but as a DAG (Directed Acyclic Graph), one for each aspect. This permits 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.

2. GRASM

The SS measures mentioned 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. Two common ancestors are disjunctive if there are independent paths from both ancestors to the concept. By

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

Table 1: Correlation coefficients for each aspect of GO and each SS 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%

independent paths we mean those that use at least one concept of the ontology not used by the other paths. Therefore, two disjunctive ancestors of a concept represent two distinct interpretations of a concept. 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 passing through a_2 and a path from a_2 to c not passing through a_1:

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

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 disjunctive ancestors of c_1 or c_2:

$$\begin{aligned} CommonDisjAnc(c_1,c_2) = \{a_1 \mid \\ a_1 \in CommonAnc(c_1,c_2) \wedge \\ \forall a_2 : [(a_2 \in CommonAnc(c_1,c_2)) \wedge (IC(a_1) \leq IC(a_2))] \Rightarrow \\ [(a_1,a_2) \in (DisjAnc(c1) \cup DisjAnc(c2))]\}. \end{aligned}$$

Original SS measures consider the shared information between two concepts c_1 and c_2 as the information content of the most informative common ancestor. GraSM replaces this notion by defining the shared information as the average of the information content of the common disjunctive ancestors:

$$Share_{GraSM}(c_1,c_2) = \overline{\{IC(a) \mid a \in CommonDisjAnc(c_1,c_2)\}}.$$

3. ASSESSMENT

We evaluated the performance of each SS measure based on the correlation between GO and family similarity. We defined the GO similarity between two proteins as the average SS of the GO terms annotated to them. However, since proteins have simultaneous biological roles, for each term annotated to a protein we compared it only with the most similar term annotated to the other protein. We tested the 500 proteins with the largest number of GO annotations from the December 2004 release of Uniprot and GO.

Table 1 presents the correlation coefficients obtained by all SS measures. The results show a strong correlation between GO and family similarity. The correlation coefficients obtained in our study are not directly comparable to the ones obtained by Lord et al., since we are measuring a different correlation using more recent UniProt and GO releases. However, our study obtained a measures' ranking that is preserved in all the aspects of GO and consistent with previous studies using different corpora [1]. In all aspects, Jiang&Conrath's measure have always obtained the strongest correlation, and Lin's measure have always obtained a stronger or equivalent correlation than Resnik's measure. This uniformity and consistency demonstrates that family similarity is more appropriate to validate SS measures than sequence similarity.

GraSM increased the correlation of all the SS measures tested. This shows that using disjunctive ancestors to calculate the shared information of two terms improves the effectiveness of SS measures. The improvement is proportional to the density of each aspect of GO. This was expected, because having more relationships per term increases the probability of having multiple common disjunctive ancestors. 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. Thus, we anticipate that GraSM will improve more its effectiveness in relation to tree-based SS measures as biologic knowledge is added to GO.

4. CONCLUSIONS

By obtaining a SS measures' ranking that is uniform over all the different aspects of GO and consistent with previous studies using different corpora, we have provided a novel and stronger demonstration of the feasibility of SS measures in a biological setting.

By obtaining a higher correlation using disjunctive common ancestors than only using the most informative common ancestor, we have demonstrated the higher effectiveness of GraSM for calculating semantic similarities between GO terms.

All the measures mentioned 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 [2]. FuSSiMeG is available on the Web at:

`http://xldb.fc.ul.pt/rebil/tools/ssm/`

5. REFERENCES

[1] A. Budanitsky and G. Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000), June 2001.

[2] F. Couto, M. Silva, and 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.

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

[4] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *10th International Conference on Research on Computational Linguistics (ROCLING X)*, Taiwan, 1997.

[5] D. Lin. An information-theoretic definition of similarity. In *15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA, 1998.

[6] P. Lord, R. Stevens, A. Brass, and C. Goble. Semantic similarity measures as tools for exploring the Gene Ontology. In *Pacific Symposium on Biocomputing*, pages 601–612, 2003.

[7] D. Rebholz-Schuhmann, H. Kirsch, and F. Couto. Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2):e65, 2005.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.