Evaluating GO-based Semantic Similarity Measures

Catia Pesquita[*], Daniel Faria, Hugo Bastos, André O. Falcão and Francisco M. Couto

University of Lisbon, Department of Informatics, Campo Grande, Lisboa – Portugal

ABSTRACT

Motivation: While several efforts have been made in measuring GO-based protein semantic similarity, it is still unclear which are the best approaches to measure it and furthermore whether electronic annotations should be used.

Results: We studied the behaviour of 8 distinct semantic similarity measures as function of sequence similarity with and without electronic annotations. We found that 5 of these measures shared a cumulative normal distribution pattern, which is likely inherent to the relation between functional and sequence similarity. We also present a novel graph-based measure for protein semantic similarity, which produced better results than the other measures studied.

1 INTRODUCTION

Since its foundation, the Gene Ontology (GO) has had a high impact in gene-product annotation, leading to its adoption by an increasing number of sequence databases. This fact, combined with the quality and structure that GO adds to annotation, has enabled its use as a background for functional comparison of gene-products. This type of comparison, called semantic similarity, is usually based on comparing the GO terms to which gene-products are annotated.

To calculate protein semantic similarity, Lord *et al.* (2003[a,b]) used three semantic similarity measures developed for WordNet and based on the notion of information content (*IC*): Resnik´s (1999), Lin´s (1998), and Jiang and Conrath´s (1999). The authors identified a correlation between semantic similarity and sequence similarity, which was stronger in the GO *molecular function* aspect. However, as these three measures were developed for comparing single terms in a hierarchy, some issues arise when applying them to GO-based protein similarity.

One issue is that GO terms can have several disjoint common ancestors. Lord *et al.* (2003[a]) dealt with this by considering only the most informative common ancestor between two terms, whereas Couto *et al.* (2005) proposed the GraSM approach, to account for all disjoint common ancestors.

Another issue is that proteins can be annotated with several GO terms, so computing the semantic similarity between two proteins requires a way of combining the semantic similarity between their terms. To address this, Lord *et al.* (2003[b]) used the arithmetic average of all term pairs, Sevilla *et al.* (2005) opted for their maximum, and Schlicker *et al.* (2006) introduced a composite average where only the best matching term pairs are used.

A different, graph-based approach was proposed by Gentleman (2005), who developed two measures for GO-based protein semantic similarity, both comparing the portion of the GO-graph shared by a pair of proteins.

Despite several studies, it is still unclear which are the best measures and/or approaches to calculate protein semantic similarity, and whether electronic annotations should be used for this purpose or ignored.

In this paper, we investigate the behaviour of several semantic similarity measures as function of sequence similarity, using both the whole annotation space and the subset of non-electronic annotations. We also introduce a novel graph-based measure for protein semantic similarity and compare its performance with that of the other measures.

2 METHODS

2.1 Semantic similarity measures

We used three term semantic similarity measures: Resnik´s (1999), Lin´s (1998), and Jiang and Conrath´s (1999); and combined them with two different approaches to compute protein similarity: the average and the best-match average (*BMA*). The former was applied as described by Lord *et al.* (2003[b]), and the latter was applied as described by Schlicker *et al.* (2006) except that only *molecular function* GO terms are being used. IC and similarity measures were calculated as previously described (Faria *et. al*, 2007).

We also use two graph-based similarity measures: *simUI* (Gentleman, 2005) and the novel *simGIC* (for Graph Information Content). *simUI* calculates similarity as the number of GO terms shared by two proteins divided by the number of GO terms they have together. *simGIC* is an expansion of *simUI* where instead of counting the terms we sum their *IC*. For two proteins *A* and *B* with terms *t*, *simGIC* is given by:

$$simGIC(A,B)=\frac{\sum_{t\in A\cap B} IC(t)}{\sum_{t\in A\cup B} IC(t)} \qquad (1)$$

2.2 Dataset

The full protein dataset used was a subset of 22,067 proteins from the Swiss-Prot database, having at least one *molecular function* GO term of *IC* 0.65 or higher. The goal was to have

[*] To whom correspondence should be addressed.

a dataset that was well characterized functionally but large enough to provide meaningful results.

An all-against-all BLAST search was performed, considering a threshold *e*-value of 10^{-4}. For each protein pair {A,B} with A?B, sequence similarity was defined as:

$$simSeq(A,B) = \log_{10}\left(AVG\left(B_{score}(A,B), B_{score}(B,A)\right)\right) \qquad (2)$$

where B_{score} is BLAST's bit-score (which is not symmetric). For the resulting 618,146 protein pairs, functional semantic similarity was computed with the measures described in 2.1, using *molecular function* GO terms.

A second dataset of proteins with only non-electronic GO annotations was also used. It contained 8,377 proteins which lead to 49,480 protein pairs.

The source data came from the UniProt database (release 2007-02-20), the GO database (release 2007-02) and the GOA-UniProt dataset (release 2007-02).

2.3 Semantic vs. Sequence Similarity

Due to large size and high dispersion of the semantic vs. sequence similarity raw data, discrete intervals of sequence similarity were taken, and average similarity values were calculated for each interval. Intervals had constant size except where the number of protein pairs in an interval was too small (under 200). The procedure was applied to all measures for both datasets.

A cumulative normal distribution curve was fitted to the discrete averaged semantic similarity vs. sequence similarity data. Non-linear regression was done applying the Newton optimization algorithm to solve the least squares method. Besides the normal parameters (mean and standard deviation), two additional parameters were required: a multiplicative scale factor and an additive translation factor (Figure 1).

3 RESULTS AND DISCUSSION

The measures using the average approach (Resnik's, Lin's, and Jiang and Conrath's) were clearly those which performed worse, being the only measures whose behaviour was not monotonically increasing (Figure 1F). This is not unexpected since this approach is biased, penalizing protein pairs which have several distinct functional aspects in common. In fact these measures only became decreasing for high sequence similarity scores, which correspond to protein pairs of larger sequence size, likely to have more than one functional aspect.

The remaining five measures (those with the *BMA* approach, *simUI* and *simGIC*) all showed a crescent behaviour with a similar topology (Figure 1A-E). We found that topology to be well modelled by a scaled cumulative normal distribution (Table 1) despite the higher dispersion visible for the non-electronic dataset, likely due to its smaller size.

What is most striking in the fitted curves is that the parameters for the normal distribution (mean and standard deviation) are nearly identical between measures, within each dataset (Table 1). Considering that these are five distinct measures, one of which (*simUI*) doesn't even rely on the notion of IC, we postulate that a normal distribution curve with these parameters is characteristic of the GO term *molecular function* annotations themselves. What this means is that the ability of *molecular function* GO terms to distinguish different levels of sequence similarity is given by a normal probability density function, which is not altogether surprising. It reflects the fact that sequence pairs with either very low or very high sequence similarity are hard to distinguish functionally, being nearly all unrelated or identical respectively.

Table 1. Regression parameters for the fitted normal distribution curves

	Measure	Regression Parameters				Scaled
		Mean	stdev[1]	scale[2]	trans[3]	Residual[4]
full	simUI	2,2	0,25	0,45	0,48	0,0026
	simGIC	2,2	0,27	0,65	0,26	0,0026
	BMA-R	2,2	0,27	0,51	0,43	0,0029
	BMA-L	2,2	0,27	0,41	0,53	0,0025
	BMA-JC	2,2	0,27	0,25	0,73	0,0029
non-electronic	simUI	2,4	0,31	0,43	0,37	0,0084
	simGIC	2,4	0,30	0,56	0,21	0,0078
	BMA-R	2,4	0,30	0,51	0,21	0,0084
	BMA-L	2,4	0,30	0,46	0,30	0,0091
	BMA-JC	2,4	0,29	0,28	0,58	0,0091

1 – standard deviation; 2 – multiplicative scale factor; 3 – additive translation factor; 4 – average residual by point divided by the scale factor, to dilute scale differences.

By analyzing the regression parameters (Table 1), we see that all these measures are capturing the normal behaviour with similar accuracy (they have similar scaled residuals within each dataset) but with different resolutions, as shown by the different scale factors (see Figure 1).

It is important to note that, while the fitted curves are isomorphic (they are inter-convertible through a linear transformation using the scale and translation factors), the actual semantic similarity measures are not: only their average behaviour is modelled by the curves. One example of this is that all 5 measures produce an equal value (of 1) if two proteins have exactly the same GO terms, of which there are occurrences in several intervals of sequence similarity. Such equality would not be maintained when applying the isomorphism between the measures' curves.

The choice of the best similarity measure therefore should fall to the measure which has the highest resolution, since on average that measure translates differences in annotation to higher differences in semantic similarity, allowing their clearer perception. In this context, the results support the choice of the novel *simGIC* measure, which showed a higher resolution than the other measures with both datasets.

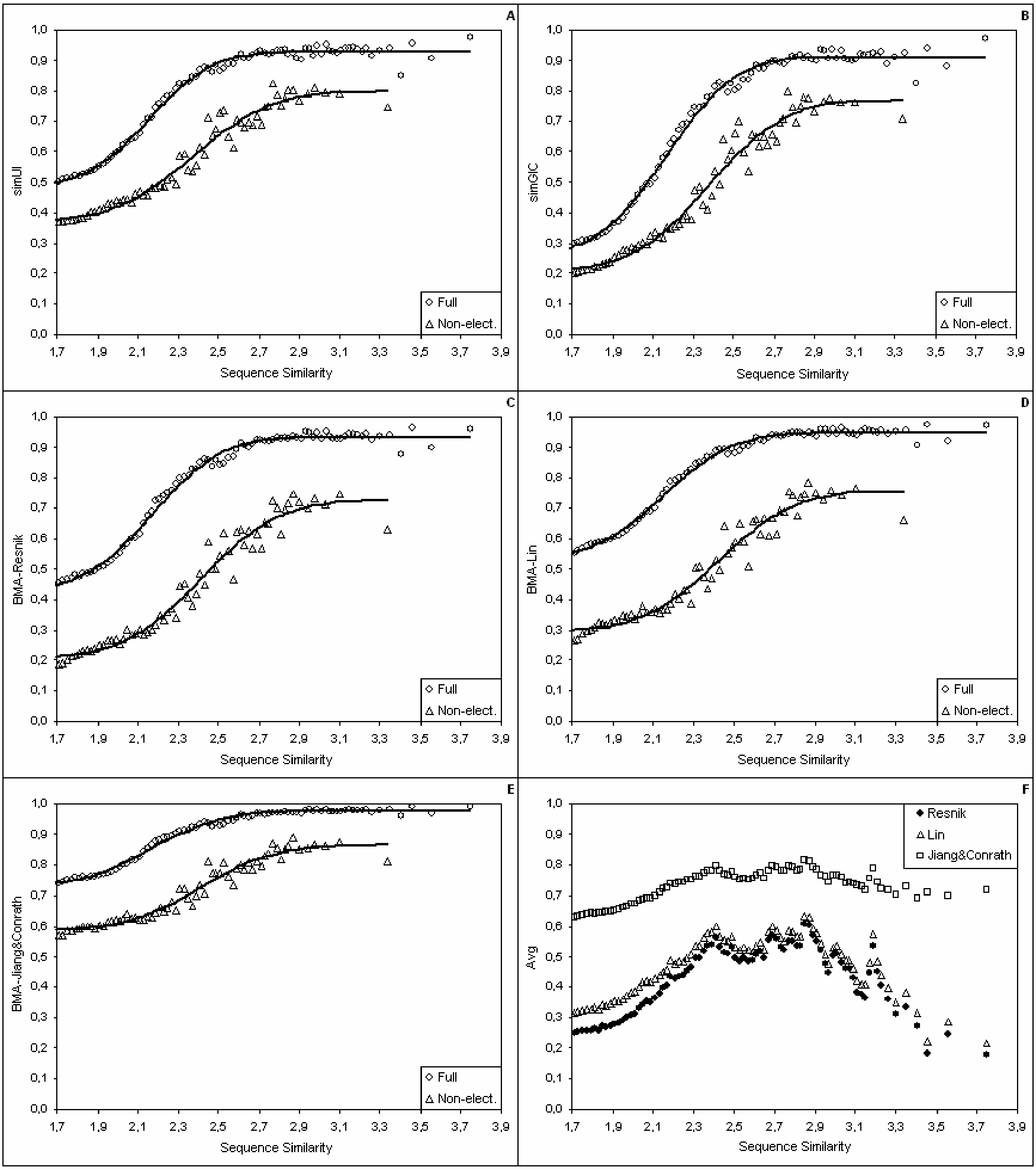

Fig. 1. Semantic similarity vs. sequence similarity for the 8 measures tested, with both full and non-electronic datasets. **A** – *simUI* measure; **B** – simGIC measure; **C** – Resnik's measure with BMA approach; **D** – Lin's measure with BMA approach; **E** – Jiang and Conrath's measure with BMA approach; **F** – Resnik's, Lin's and Jiang and Conrath's measures with the average approach (full dataset only); lines in **A**-**E** correspond to fitted cumulative normal distribution curves. In addition to mean and standard deviation, which determine the inflexion point and width of the curve respectively, two parameters were used to fit the curves: a multiplicative scale parameter to account for the measures not covering the whole 0-1 scale, and an additive translation parameter to account for their minimum value being greater than 0.

However, the only measure which showed a clearly low resolution was Jiang and Conrath's measure. The remaining measures have a resolution not much below that of *simGIC*, with Resnik's measure being second best.

As for the differences in the normal distribution parameters between the two datasets, they reflect the fact that the non-electronic annotation space is different from the full space. For instance, the average number of annotations per protein is smaller in the non-electronic dataset than in the full one (4.8 and 5.5 respectively). Also relevant is the fact that there are much less proteins with non-electronic annotations (8% of the full set), and these could not be representative of the whole protein similarity space.

Despite these differences, the fact remains that the behaviour of the two datasets is similar, which suggests that electronic annotations can not only be reliably used in semantic similarity calculations, but also improve their precision by providing a richer annotation space.

4 CONCLUSIONS

We studied the averaged behaviour of several distinct semantic similarity measures as function of sequence similarity, uncovering an underlying normal distribution-like pattern with constant shape parameters (mean and standard deviation). We postulate that this pattern is characteristic of the variation of functional similarity (as measured by GO *molecular function* annotations) with sequence similarity.

We developed a novel graph-based semantic similarity measure for proteins, which performed better than the remaining measures by translating sequence similarity into a greater coverage of the semantic similarity scale.

We also compared the performance of the similarity measures with and without electronic annotations, concluding that electronic annotations do not significantly affect the behaviour of the similarly measures, and actually increase their precision. While they may lack the reliability of curated annotations, electronic annotations are the present and future of bioinformatics, constituting an increasingly important portion of the annotation space (currently amounting to 97%). What is more, their precision is improving, with values of 91-100% having been reported (Camon *et al*., 2005).

Future work will include comparing semantic similarity with other aspects, such as protein families (Pfam) and Enzyme Commission classes, as well as using a sequence similarity measure independent of sequence length. We will also investigate other semantic similarity measures, such as the GraSM approach.

ACKNOWLEDGEMENTS

This work was partially supported by the Portuguese 'Fundação para a Ciência e Tecnologia' with the grant ref. SFRH/BD/29797/2006.

REFERENCES

Camon, E., Magrane, M., Barrell, D., *et al.* (2004) The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Research* 32 D262.

Camon, E., Barrell, D., Dimmer, E. C., *et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6(Suppl 1):S17.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005) 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.*

Devos, D., and Valencia, A. (2000) Practical limits of function prediction. *Proteins: Structure, Function, and Genetics* **41**, 98–107.

Faria, D., Pesquita, C., Couto, F. M. and Falcao, A. O. (2007) ProteInOn: A Web Tool for Protein Semantic Similarity, *DI/FCUL TR 07-06*, Dpt. Informatics, Univ. Lisbon.

Gentleman, R. (2005) Visualizing and Distances Using GO, Retrieved Jan. 10th, 2007: http://bioconductor.org/packages/2.0/bioc/vignettes/GOstats/inst/doc/GOvis.pdf

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

Jiang, J., and Conrath, D. (1997) Semantic similaritybased on corpus statistics and lexical taxonomy. *In Proc. of the 10th International Conference on Research on Computational Linguistics.*

Lin, D. (1998) An information-theoretic definition of similarity. *In Proc. of the 15th International Conference on Machine Learning.*

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

Lord, P., Stevens, R., Brass, A., and Goble, C. (2003)[b] Semantic similarity measures as tools for exploring the gene ontology. *In Proc. of the 8th Pacific Symposium on Biocomputing.*

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

Schlicker, A., Domingues, F. S., Rahnenfhrer, J., and Lengauer, T. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics* **7**, 302.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martnez-Cruz, L. A., Corrales, F. J., and Rubio, A. (2005) Correlation between gene expression and GO semantic similarity. *In IEEE/ACM Transactions on Computational Biology and Bioinformatics.*

Bodenreider, O., Stevens, R., (2006) Bio-ontologies: current trends and future directions. *Briefings In Bioinformatics.* **7**. No 3. 256-274

Wu, C., Apweiler, R., Bairoch, A., *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* D187–D191.