CESSM : Collaborative Evaluation of Semantic Similarity Measures

Catia Pesquita[1], Delphine Pessoa, Daniel Faria and Francisco M. Couto
Faculdade de Ciências, Universidade de Lisboa, Portugal
[1] cpesquita@xldb.di.fc.ul.pt

ABSTRACT

The application of semantic similarity measures to proteins annotated with Gene Ontology terms has become a common method in bioinformatics. However, the evaluation of these measures is still challenging, since no common standard of evaluation exists.

We present an online tool for the automated evaluation of GO-based semantic similarity measures, CESSM, that enables the comparison of new measures against previously published ones considering their relation to sequence, *Pfam* and *EC* similarity. The tool also has a collaborative component, by which the authors of published measures can contribute to the enrichment of the evaluation by providing their own results. CESSM is freely available at http://xldb.di.fc.ul.pt/tools/cessm/

BACKGROUND

The creation of the Gene Ontology (GO) [1], a controlled vocabulary for the description of gene product functions, triggered the development of computational methods that take advantage of its structured information. One such method is the application of semantic similarity measures to GO terms, whereby the similarity between two terms is calculated according to their relationship in the ontology. Likewise, semantic similarity measures can also be used to calculate the similarity between gene products, provided they are annotated with GO terms.

Several semantic similarity measures based on GO have been proposed in recent years [2-12], but the evaluation of their performance has been identified as a relevant problem in the field [13]. Various evaluation strategies have been proposed, including the investigation of the relation between the semantic similarity measure and other gene product or protein similarities (such as sequence[2-7], family [12,7] or expression similarity [8,14,15]); and of the feasibility to use semantic similarity measures in such distinct scenarios as the prediction of subnuclear location [16], the ability to characterize human regulatory pathways [17], or the performance in gene clustering [9,10]. This multiplicity of evaluation strategies arises from the lack of a gold standard suitable to this scenario, driving researchers to use diverse data sets, to which they apply distinct evaluation strategies, thus rendering comparison among different works unfeasible.

We present an online tool CESSM (Collaborative Evaluation of Semantic Similarity Measures) for the collaborative and automated evaluation of semantic similarity measures in the context of GO. CESSM allows researchers to compare the performance of their novel semantic similarity measures against several existing ones, using the same protein and annotation dataset and according to three distinct aspects: relation with sequence, *EC* class and *Pfam* family similarities.

METHODS

CESSM provides the user with a list of protein pairs, for which CESSM's database contains semantic, sequence, *EC* class[1] and *Pfam*[2] family similarity data. The user then calculates the similarities between those pairs with his/her own measure, and uploads the results to CESSM. CESSM compares the user's results to the similarity values present in its database, and returns to the user a set of evaluation metrics.

[1] au.expasy.org/enzyme/

[2] pfam.sanger.ac.uk/

The protein pairs set corresponds to UniProt[3] protein pairs characterized by the following:

- both proteins are manually annotated with at least one GO term within all three GO categories (molecular function, biological process and cellular component) with a uniform information content [5] of at least 0.5;
- both proteins have at least one EC class and one Pfam class;
- the proteins BLAST e-values for both directions are below 10 $^{-4}$.

This results in a total of 13,430 protein pairs, composed of 1,039 distinct proteins.

CESSM's database contains data from GO, GOA[4] and UniProt that is used for semantic, sequence, EC class and Pfam family similarity calculations. CESSM's database also stores the data needed to perform the evaluations (the similarities between each protein pair, and the protein pairs themselves) and data about the settings used in each semantic similarity computation.

We currently implement 11 semantic similarity measures: simGIC (GI) [4], simUI (UI) [11], and the average (A) [2], maximum (M) [8] and best-match average (B) [12] combinations of the term similarities by Resnik (R) [18], Lin (L) [19] and Jiang&Conrath (J) [20]. All measures return values between 0 and 1 due to the use of uniform information content [5]. These measures can be applied within each GO category, consider different sets of annotations (for instance, only manually curated ones), and even different sets of ontology relationships (*e.g.* all or just is_a relationships).

CESSM uses three distinct evaluations: correlation with *EC* class similarity; correlation with *Pfam* family similarity and relationship with sequence similarity.

EC class similarity is calculated using the Enzyme Comparison Class (*ECC*) metric proposed by [21]. *ECC* is a value between 0 and 4 that corresponds to the number of *EC* digits two proteins share. For instance, consider two proteins *p1* and *p2*, where *p1* belongs to the *EC* Class *1.1.1.10* and *p2* belongs to class *1.1.2.3*. Their *ECC* would be 2, since they share the first two digits. For proteins with more than one *EC* class, we calculate the maximum *ECC*.

Pfam similarity (*Pfam*) is calculated via Jaccard similarity, where the similarity between two proteins is given by the ratio between the number of *Pfam* families they share and the total number of *Pfam* families they have. This returns a value between 0 and 1.

Sequence similarity (*SeqSim*) is calculated using RRBS [5], which is a relative measure of sequence similarity based on the BLAST bitscores. It takes into account the non-reciprocity of BLAST bitscores and their dependency upon sequence length, and returns a value between 0 and 1.

These similarities are then used to calculate the Pearson's linear correlation between the semantic similarity values and the *Pfam*, *ECC* or *SeqSim* ones.

The relationship between semantic and sequence similarity is further analyzed by plotting two graphs: one of the direct relationship between the user's measure and *SeqSim*, and another of the averaged relationship between the user's and CESSM's semantic similarity measures and *SeqSim*. The latter plot is the result of binning the dataset into 100 intervals of equal size corresponding to averaged values of sequence similarity, over which semantic similarity values are then averaged.

Finally, the general performance of the semantic similarity measures against sequence similarity is also calculated, using, *resolution* [5], a metric that corresponds to the range of the averaged semantic similarity results, and thus reflects the ability of a measure to distinguish between pairs with different levels of sequence similarity.

RESULTS

We implemented an online tool, that allows user's to evaluate the performance of their GO-based semantic similarity measures against several existing measures, using a common dataset and evaluation strategy.

[3] www.uniprot.org/

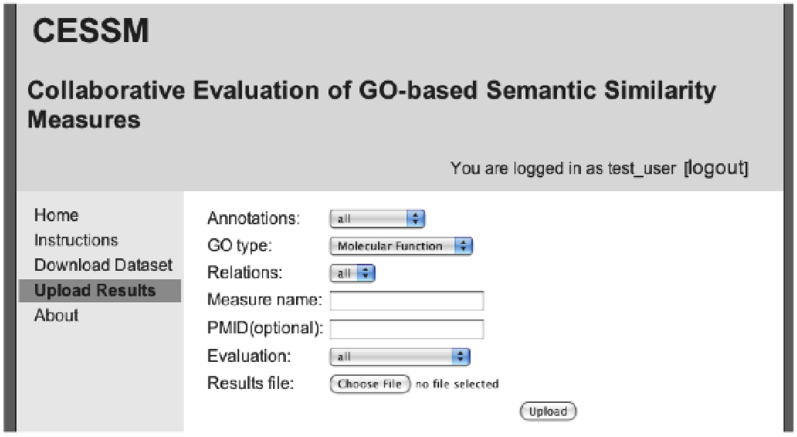
[4] www.ebi.ac.uk/GOA/

CESSM' s user work flow is as follows:

1. 1.CESSM users are requested to download three files:
   1. Gene Ontology file
   2. GOA_UniProt annotations file (same data as used in CESSM's database)
   3. Protein pairs file
2. 2.Using the GO and GOA_UniProt files, users can create a local database (or otherwise parse the data in the files) to support semantic similarity computation in a manner fully comparable to CESSM.
3. Users calculate the semantic similarity between all pairs in the protein pairs file, using the measure they wish to evaluate and considering the following options:
   1. Annotations: all or just manually curated ones.
   2. GO type: molecular function, biological process or cellular component
   3. Ontology relationships: all or just *is_a*.

   Furthermore, similarity values must be bounded between 0 and 1.
4. Users upload the similarity values file, and select the options used for the semantic similarity calculation and the desired evaluation type (Figure 1):
   1. All
   2. based on *ECC* similarity
   3. based on *Pfam* similarity
   4. based on Sequence similarity

   Additionally, a PMID for published measures may be supplied, so that upon manual inspection the submitted results can be included in the database, and be a part of the tool.
5. User downloads a .zip file with the requested evaluation files (Figure 2):
   1. a table with correlation values for the requested metrics (*ECC*, *Pfam* and/or *SeqSim*) against all semantic similarity measures (user and stored ones)
   2. a graph plotting the results for the comparison of the user's measure and sequence similarity
   3. a graph illustrating the averaged relationship (over 100 intervals) between all measures and *SeqSim*
   4. a table with the resolution values for this relationship.



CESSM

Collaborative Evaluation of GO-based Semantic Similarity Measures

You are logged in as test_user [logout]

Annotations: all

GO type: Molecular Function

Relations: all

Measure name:

PMID(optional):

Evaluation: all

Results file: Choose File no file selected

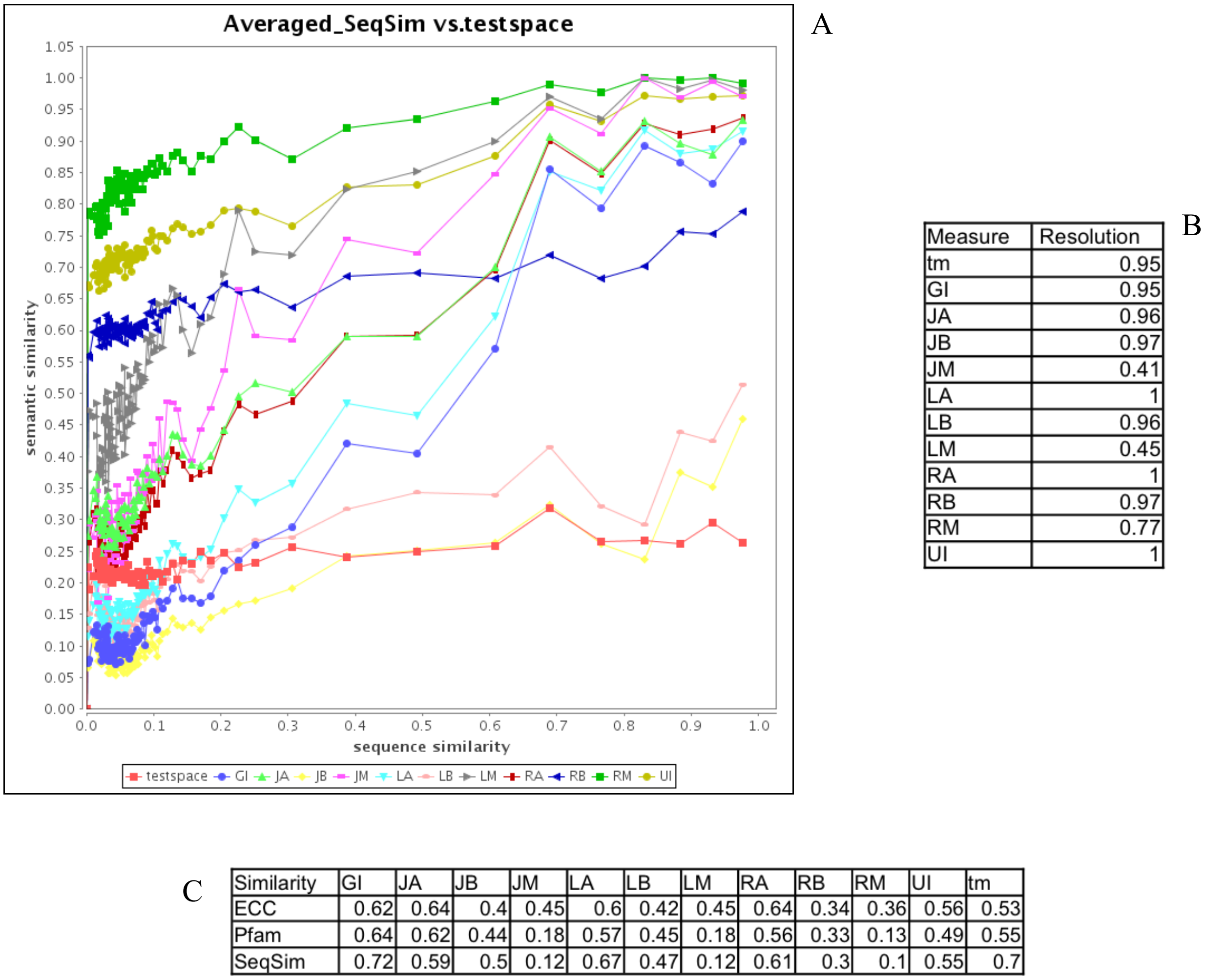
Figure 1: CESSM website. User selects the options he/she used in the calculations and uploads the results file. If the measure has been published the user can provide its PMID or URL.

CONCLUSIONS

CESSM is a common platform for easy evaluation of GO-based semantic similarity measures, rendering comparable results. It also has a collaborative component, since it allows for researchers to contribute results obtained with published measures that upon inspection will be incorporated into the evaluation.

CESSM provides a common dataset of protein pairs, but we do not intend it to be used as a gold standard for protein semantic similarity, simply as a common ground for semantic similarity, based on adequately characterized proteins. CESSM offers the possibility of analyzing the relationship between semantic similarity and several protein similarities based on *EC* class, *Pfam* family and sequence. We hope that their conjugation will give users a better grasp of their measure's overall performance.

Future versions of CESSM will include more options for annotation type (allowing for different combinations of evidence codes) and relationship type (allowing the selection of which relations to use), and web service access, that will improve the communication of results between user and tool and enable the integration of CESSM results into other services or tools.



Measure	Resolution
tm	0.95
GI	0.95
JA	0.96
JB	0.97
JM	0.41
LA	1
LB	0.96
LM	0.45
RA	1
RB	0.97
RM	0.77
UI	1

Similarity	GI	JA	JB	JM	LA	LB	LM	RA	RB	RM	UI	tm
ECC	0.62	0.64	0.4	0.45	0.6	0.42	0.45	0.64	0.34	0.36	0.56	0.53
Pfam	0.64	0.62	0.44	0.18	0.57	0.45	0.18	0.56	0.33	0.13	0.49	0.55
SeqSim	0.72	0.59	0.5	0.12	0.67	0.47	0.12	0.61	0.3	0.1	0.55	0.7

Figure 2: Example of results. A) Averaged behavior of semantic similarity measures against sequence similarity (testspace=user measure). B) Resolution of all CESSM measures and user measure. C) Correlation between all measures and *ECC*, *Pfam* and *SeqSim*.

ACKOWLEDGEMENTS

This work was supported by the Fundação para a Ciência e Tecnologia through the Multiannual Funding Programme, and the grants SFRH/BD/42481/2007 and SFRH/BD/29797/2006. The authors wish to thank Marco A. Alvarez for his help in testing the tool.

REFERENCES

1. Gene Ontology Consortium The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32:D258–D261.
2. Lord P, Stevens R, Brass A, Goble C. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics;19:1275–1283.
3. Mistry M, Pavlidis P (2008) Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 9.
4. Pesquita C, Faria D, Bastos H, Falcão AO, Couto F (2007) Evaluating GO-based Semantic Similarity Measures. ISMB/ECCB 2007 Bio-ontologies SIG
5. Pesquita C, Faria D, Bastos H, Falcao AO, Couto F (2008) Metrics for GO-based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9.
6. Pozo AD, Pazos F, Valencia A (2008) Defining functional distances over Gene Ontology. BMC Bioinformatics 9.
7. Schlicker A, Domingues FS, Rahnenfhrer J, Lengauer T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics.7
8. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A. (2005) Correlation between Gene Expression and GO Semantic Similarity. IEEE/ACM Transactions on Computational Biology and Bioinformatics.
9. Wang JZZ, Du Z, Payattakool R, Yu PSS, Chen CFF (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics
10. Sheehan B, Quigley A, Gaudin B, Dobson S (2008) A relation based measure of semantic similarity for Gene Ontology annotations. BMC Bioinformatics 9.
11. Gentleman(2005) Manual for R.
12. Couto FM, Silva MJ, Coutinho PM. (2007) Measuring semantic similarity between Gene Ontology terms, Data & Knowledge Engineering
13. Pesquita C, Faria D, Falcao AO, Lord P, Couto F (2009) Semantic Similarity in Biomedical Ontologies PLoS Comput Biol.2009 July; 5(7): e1000443
14. H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In Proc. of IEEE 2004 CIBCB.
15. Xu T, Du L, Zhou Y (2008) Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. BMC Bioinformatics 9.
16. Lei Z, Dai Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics. 7
17. Guo X, Liu R, Shriver CD, Hu H, Liebman MN. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics. 2006;22:967–973.
18. 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.
19. Lin D. (1998) An information-theoretic definition of similarity. Proc of the 15th International Conference on Machine Learning.
20. Jiang J, Conrath D. (1997) Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proc of the 10th International Conference on Research on Computational Linguistics.
21. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins, 41:98-107.