CESSM

Collaborative Evaluation of Semantic Similarity Measures

Catia Pesquita, Delphine Pessoa, Daniel Faria and Francisco M. Couto
University of Lisbon, Portugal
cpesquita@xldb.di.fc.ul.pt

Abstract

Motivation: The application of semantic similarity measures to gene products 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.

Results: 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 in terms performance against sequence, Pfam and EC similarity. The tool also has a collaborative component, by which the authors of 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/**

Introduction

The application of semantic similarity measures to GO, and consequently to gene products annotated with GO terms has become an area of blooming interest, with many novel or adapted measures, tools and applications being proposed in the last six years.

Along with the profusion of measures, came a variety of evaluation strategies, including investigating the relation (mostly linear correlation) between the semantic similarity measure and other gene product or protein similarities, such as sequence, family or expression similarity, etc., and also comparison against the performance of other existing semantic similarity measures.

This multiplicity of evaluation strategies arises from the lack of a gold standard suitable to this scenario, driving researchers to use diverse datasets, to which they apply distinct evaluation strategies, thus rendering comparison among different works impossible.

We present an online tool for the collaborative and automated evaluation of semantic similarity measures in the context of GO, CESSM (Collaborative Evaluation of Semantic Similarity Measures).

CESSM allows researchers to compare their novel semantic similarity measures against several previously proposed ones, in three distinct aspects:

- correlation with EC class similarity;
- correlation with Pfam family similarity;
- relationship with sequence similarity.

CESSM

Database

CESSM's database contains data from GO, GOA and UniProt that is used for semantic 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. The current version of GO corresponds to August,2008.

Dataset

The protein pairs dataset corresponds to UniProt protein pairs characterized by the following:

1) both proteins are manually annotated with at least one GO term within all 3 GO types with a uniform IC (Pesquita et al.,2008) of at least 0.5;

2) both proteins have at least one EC class and one Pfam class;

3) the proteins BLAST e-values for both directions are below 10^{-4}.

This results in a total of 13430 protein pairs, composed of 1039 distinct proteins.

With this dataset, we do not aim at providing a gold standard for gene product semantic similarity, but simply to provide a common ground for semantic similarity, based on adequately characterized proteins.

Semantic Similarity Measures

CESSM currently implements 11 semantic similarity measures: simGIC [GI](Pesquita et. al, 2007), simUI [UI](Gentleman, 2005), and the average (Lord et al. 2003) [A], maximum [M](Sevilla et al. 2006) and best-match average [B](Couto et al. 2005) combinations of the term similarities by Resnik [R](1995), Lin [L](1998) and Jiang&Conrath [J](1997).

All measures return values between 0 and 1 due to the use of uniform information content (Pesquita et al, 2008).

Gene Product Similarity Measures

Enzyme Commission Similarity is calaculated using ECC a metric proposed by Devos & Valencia (2000) that returns a value between 0-4, corresponding to the number of EC digits, two proteins share. 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 (Pesquita et. al, 2008), which uses BLAST bitscores and takes into account the reciprocity of BLAST and sequence length, returning a value between 0 and 1.

User experience

1.CESSM users are requested to download three files (Figure 1):

1. Gene Ontology file
2. GOA_UniProt annotations file
3. protein pairs file

2.Using the GO and GOA_UniProt files, 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 manual
2. GO type: molecular function, biological process or cellular component
3.Ontology relationships: all or just is_a.

Similarity values must be bounded between 0 and 1.

Users produce a file containing the protein pairs and their similarity values (Figure 2).

3.Users upload the similarity values file , and select the options used for the semantic similarity calculation and the desired evaluation type (Figure 3):

1. All
2. based on EC similarity
3. based on Pfam similarity
4. based on Sequence similarity

PMID for published measures may be supplied.

4. User receives an e-mail with the results (Figure 4).

Evaluation

CESSM returns to the user three evaluation components: a table with correlation values for the three metrics (ECC, Pfam, and SeqSim) against all semantic similarity measures (user and in-house ones); a graph illustrating the averaged relationship between all measures and SeqSim; and a table with the resolution values for this relationship.

The correlation values correspond to the Pearson's linear correlation between the semantic similarity values and the Pfam, ECC or SeqSim ones.

The data points used to plot the graph (example in Figure 4A) are 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.

The resolution values (Pesquita et al, 2008) correspond 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.

Conclusions

CESSM provides a common platform for easy evaluation of GO-based semantic similarity measures, rendering comparable results.

CESSM 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.

In the immediate future, we will add more settings options for annotation type (allow for different combinations of evidence codes), and relationship type (allow to select which relations to use).

Further on, CESSM will include web services that will enable a better communication of results between user and tool, allowing for the integration of CESSM results (similarities and evaluation results) into other services or tools.

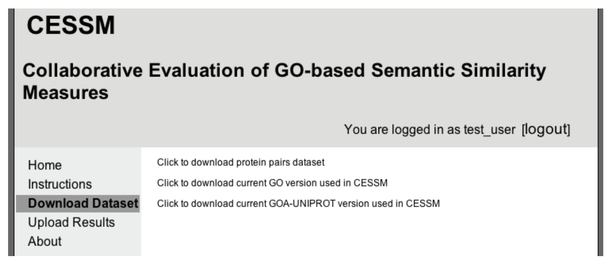


Figure 1: CESSM website. User can download ontology, annotation and protein pairs datasets.

Q9JKJ9	Q9NYL5	0.5
Q9JKJ9	Q60991	0.5
Q9JKJ9	Q64505	0.5
Q9JKJ9	P11511	0.5
Q9JKJ9	P27786	0.5
Q9JKJ9	P00184	0.5
...		

Figure 2: Example of user generated file. Each protein pair is followed by the similarity value calculated by the user.

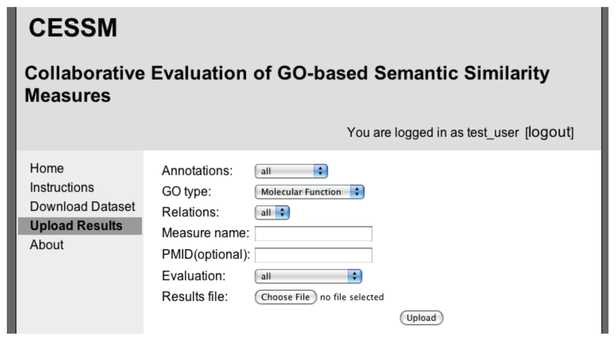
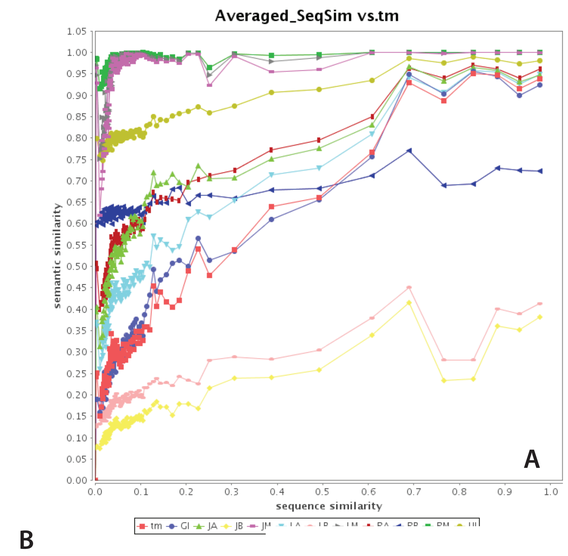


Figure 3: CESSM website. User selects the options he/she used in the calculations and uploads the results file. If the measure has been published the user can provide its PMID or URL.



B

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

C

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 4: Example of results. A) Averaged behaviour of semantic similarity measures against sequence similarity. B) Resolution of all CESSM measures and test measure. C)Corelation between all measures and ECC, Pfam and Sequence Similarity.

Collaborators Needed

Are you the author of a GO-based semantic similarity measure?

Would like to see how it performs against other measures?

Go to: http://xldb.di.fc.ul.pt/tools/cessm and follow the instructions.

Published measures will be included in future CESSM analyses.

If you're still developing your measure, use CESSM results to get fast and comparable results on your measure's performance.

Thank you!