Bioinformatics, 32(9), 2016, 1380–1387
doi: 10.1093/bioinformatics/btv755
Advance Access Publication Date: 26 December 2015
Original Paper

Data and text mining

simDEF: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes

Ahmad Pesaranghader[1,2,*], Stan Matwin[1,2,3], Marina Sokolova[2,4] and Robert G. Beiko[1,*]

[1]Faculty of Computer Science, Dalhousie University, Halifax, NS B3H 4R2, Canada, [2]Institute for Big Data Analytics, Halifax, NS B3H 4R2, Canada, [3]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland and [4]Faculty of Medicine and Faculty of Engineering, University of Ottawa, Ottawa, ON K1H 8M5, Canada

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 7 September 2015; revised on 16 December 2015; accepted on 21 December 2015

Abstract

Motivation: Measures of protein functional similarity are essential tools for function prediction, evaluation of protein–protein interactions (PPIs) and other applications. Several existing methods perform comparisons between proteins based on the semantic similarity of their GO terms; however, these measures are highly sensitive to modifications in the topological structure of GO, tend to be focused on specific analytical tasks and concentrate on the GO terms themselves rather than considering their textual definitions.

Results: We introduce simDEF, an efficient method for measuring semantic similarity of GO terms using their GO definitions, which is based on the Gloss Vector measure commonly used in natural language processing. The simDEF approach builds optimized definition vectors for all relevant GO terms, and expresses the similarity of a pair of proteins as the cosine of the angle between their definition vectors. Relative to existing similarity measures, when validated on a yeast reference database, simDEF improves correlation with sequence homology by up to 50%, shows a correlation improvement $>4\%$ with gene expression in the biological process hierarchy of GO and increases PPI predictability by $>2.5\%$ in F1 score for molecular function hierarchy.

Availability and implementation: Datasets, results and source code are available at http://kiwi.cs.dal.ca/Software/simDEF

Contact: ahmad.pgh@dal.ca or beiko@cs.dal.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene Ontology (GO) (Ashburner *et al.*, 2000) describes the attributes of genes and gene products using a structured vocabulary. Many biomedical databases, such as UniProt (The UniProt Consortium, 2008) and SwissProt (Boeckmann *et al.*, 2003), are annotated by GO terms to communicate semantic meanings of biomedical entities. Computing functional similarity of biomedical entities has been applied to problems such as prediction of protein–protein interaction (PPI) (Pu et al., 2015), gene expression studies (Ovaska, 2015) and homology analysis (Piovesan et al., 2015). Also, in the context of text mining various studies (Jin and Lu, 2010; Jin *et al.*, 2011) have aimed to enhance the literature-based GO annotation of gene products.

There are two main computational models available to measure similarity of terms. *Ontology-based* models take advantage of lexical structures in their estimation of term similarity. Edge-based ontology measures like Wu (Wu and Palmer, 1994) and RSS (Wu *et al.*, 2006) consider the number of edges along the paths that link two GO terms. Node-based measures (which we designate as

information-content-based), such as Resnik (1995), Jiang (Jiang and Conrath, 1997), Lin (1998), Schlicker *et al.* (2006), TCSS (Jain and Bader, 2010), GraSM (Couto *et al.*, 2011) and AIC (Song *et al.*, 2014) compare the properties of the terms augmented with the properties of their ancestors or descendants. IC vectors (Pesaranghader and Muthaiyah, 2013) represent IC values in distributed forms in the computation of semantic similarity. Hybrid measures such as those of Wang *et al.* (2007), Liu *et al.* (2009) and HRSS (Wu *et al.*, 2013) combine node-based and edge-based measures. While these measures first compute semantic similarity of two gene products and then aggregate the results as a single functional similarity value, group-wise measures such as simUI (Falcon and Gentleman, 2007), simGIC (Pesquita *et al.*, 2007) and SORA (Teng *et al.*, 2013) calculate similarity by measuring two sets of GO terms annotating these genes. Huang *et al.* (2007) also proposed a similarity measure where gene functional similarity is based on vector representations of their GO terms. Ontology-based measures suffer from three important limitations: first, they depend on the constantly changing topological structure of GO; second, they use incomplete GO annotations to compute statistical information; and third, they offer no guarantee of generalization to multiple biological tasks.

Distributional-based approaches derive from Firth's idea (1957) that a term is characterized by the company it keeps in its context. Measures following this notion calculate terms' specifications from relevant text data and represent them in a vector space for subsequent computation of their similarities. The Gloss Vector semantic relatedness measure (Pedersen *et al.*, 2004) is a distributional-based approach with a wide application in natural language processing. This measure constructs definitions (glosses) of terms from a predefined thesaurus, and estimates semantic relatedness of two terms as the cosine of the angle between those terms' gloss-vectors. Interpolation of content words of a text corpus into the terms' definition was shown to outperform the direct definition comparison. Gloss vectors offer a new opportunity to exploit the information of GO term definitions and to infer gene functional similarity. Liu *et al.* (2012) successfully applied the Gloss Vector measure to the biomedical domain using MEDLINE as the text corpus and the unified medical language system and WordNet for the construction of extended definitions of medical concepts. The Gloss Vector approach requires a frequency cut-off in selecting the best features describing one term (Pesaranghader *et al.*, 2013, 2014a). We have developed simDEF, an optimized version of the Gloss Vector targeted to analysis of gene functions. Here, by using MEDLINE as the text corpus, we compare the performance of simDEF with other leading approaches, and demonstrate its effectiveness using comparisons based on sequence homology, gene expression and PPI data.

2 Experimental data

2.1 GO and GO annotations

GO comprises three GOs which express different biological attributes: *biological process* (BP) for processes such as metabolism or cell proliferation; *cellular component* (CC) such as the nucleus or cell membrane; and *molecular function* (MF) such as catalytic or binding activities. GO is maintained and constantly updated by a group of curators.

A GO annotation consists of a GO term associated with a specific reference and an evidence code to indicate how a given annotation is supported. Out of all the evidence codes available, Inferred from Electronic Annotation (IEA) is not assigned by a curator and is thus the least reliable so we treat them separately. GO and the required GO annotations were downloaded from the GO website (http://geneontology.org (November 2, 2015).

2.2 MEDLINE abstracts

MEDLINE (https://mbr.nlm.nih.gov/Download/) contains over 20 million citations of biomedical articles from 1966 to the present. The database includes journal articles from medicine, pharmacy, dentistry, nursing, healthcare and covers the literature in biology and biochemistry. For this study, we used MEDLINE 2013 as the corpus to build a first-order word–word co-occurrence matrix for the later computation of second-order co-occurrence (SOC) matrices which are used by simDEF.

2.3 Validation datasets

2.3.1 Sequence homology

We used bitscores from the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997) to create our sequence homology dataset. In the first step, we performed an all-versus-all comparison of proteins in the yeast *Saccharomyces cerevisiae* database (Cherry *et al.*, 1998) with an expectation-value threshold of 0.1. Although this threshold is liberal, the corresponding bitscores associated with *e*-values near this threshold will be very low and have a minimal effect on our analysis. Since a bitscore for query and subject proteins is not symmetrical, we calculate log-reciprocal BLAST score (LRBS) and relative reciprocal BLAST score (RRBS) to express the general sequence similarity of protein pairs. For proteins A and B, the LRBS and RRBS are

$$\mathrm{LRBS}(A,B) = \log\left(\frac{\mathrm{Bitscore}(A,B) + \mathrm{Bitscore}(B,A)}{2}\right), \quad (1)$$

$$\mathrm{RRBS}(A,B) = \frac{\mathrm{Bitscore}(A,B) + \mathrm{Bitscore}(B,A)}{\mathrm{Bitscore}(A,A) + \mathrm{Bitscore}(B,B)}. \quad (2)$$

Finally, after LRBS and RRBS computations, we have a dataset of 20 167 protein pairs from the yeast *S.cerevisiae* database along with their LRBS and RRBS sequence similarity scores. All proteins in the dataset have their own GO annotations from the CC, BP and MF ontologies without considering IEAs.

2.3.2 Gene expression

The gene expression dataset comes from the study by Jain and Bader (2010). In their study, the gene-expression dataset for *S.cerevisiae* was downloaded from GeneMANIA (Warde-Farley *et al.*, 2010) and other microarray experiments. The authors prepared test datasets of 5000 *S.cerevisiae* gene pairs randomly selected from a list of all possible pairs of proteins in their gene expression dataset. This was done independently for CC, BP and MF annotations of gene products. Since in our experiments we mainly consider genes with non-electronic annotations, we used 4800 fitting gene pairs from their study.

2.3.3 Protein–protein interaction

For the PPI experiment, we employed subsets of the yeast PPI dataset from Wu *et al.* (2013). In that study, for each GO, independent gold-standard positive datasets for yeast were built from a core subset of the Database of Interacting Proteins (DIP) (Salwinski *et al.*, 2004). Negative datasets were independently generated by randomly choosing annotated protein pairs in BP, CC and MF, which are absent from a combined dataset of all possible PPIs. Since for different GOs the numbers of generated PPI pairs are different and more importantly many of them do not have GO annotations after excluding

IEA, we selected subsets of 3000 positive and 3000 negative PPIs for each ontology from that study to evaluate our measure against other similarity measures in a PPI prediction task.

3 Methods

Pointwise Mutual Information (PMI) is a measure of association used in information theory. In computational linguistics, the PMI for two given words indicates the likelihood of finding one word in a text document that includes the other word. PMI is formulated as

$$\mathrm{PMI}(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1) \times p(w_2)}, \quad (3)$$

where $p(w_1, w_2)$ is the probability that words w_1 and w_2 co-occur in a document, and $p(w_1)$ and $p(w_2)$ for w_1 and w_2, respectively, are the marginal probabilities of their occurrence in a document. It is expected that rare words are highly associated with and descriptive of each other, yet due to their sparse nature their bigram frequency (i.e. number of times they have been seen next to each other) is small in the corpus. This is the main drawback of the Gloss Vector measure in selection of the best descriptive features.

We use PMI in our proposed measure, simDEF, for statistical elimination of insignificant features (words). simDEF requires procedures for building the co-occurrence matrix from a proper text corpus, constructing extended definitions for GO terms using GO term definitions, and finding words that are appropriate descriptors of that GO term. simDEF comprises six steps (Fig. 1) (see Supplementary Material for pseudocode of the steps laid out in simDEF).

3.1 Step 1—counting bigrams and building the first-order co-occurrence matrix

After discarding punctuation, changing all characters to lowercase, and removing stop words (a pre-defined list of 204 non-informative words like *a* and *the*) from the MEDLINE corpus, a list of bigrams and their frequencies for all the content words is constructed. A window size of 2 is used for extraction of bigrams. This window size controls how close two words can appear in bigrams. Stemming was found to reduce accuracy and was not adopted in simDEF. Then, by ignoring the order of occurrence in a bigram, we transform it from a bigram list to a co-occurrence list. Finally, we construct the first-order co-occurrence matrix, which is symmetric and sparse and represents the contextual information of MEDLINE words. Cell values in the first-order matrix represent how many times the word associated with its row is seen in this corpus alongside the word associated with its column.

3.2 Step 2—definition construction of GO terms and then building BP, CC, and MF definition matrices

In this step, we construct an extended definition for every term in GO. From the theoretical perspective, definition extension of parent GO terms (i.e. broader concepts) with their children's definitions (i.e. narrower and more specific definitions) adds more specific information. Although child GO terms may contain contradictory information, this information may nonetheless provide essential context when calculating functional similarity with other genes (which may in turn be augmented with conflicting information). From the practical perspective, we examined all the combinations of definition extension considering GO relationships such as *is_a*, *has_part*, *part_of*, *regulates*, *siblings* and *synonyms*. What is represented in Figure 2 yielded the best results in our experiments conducted in this study. Improvement in the results using relationships such as *part_of* and *regulates* indicates that besides the similarity, simDEF accounts for relatedness as well. See Supplementary material for more in-depth explanation of why definition extension can be beneficial.

Each GO term has an identifier, a representative name, a GO definition, a namespace defining the sub-ontology of the GO term and other information such as its relationship to the other GO terms. For example, GO:0001104 has the representative name 'RNA polymerase II transcription cofactor activity' and belongs to the MF hierarchy. This GO term has the definition 'Interacting selectively and non-covalently with an RNA polymerase II (RNAP II) regulatory transcription factor and also with the RNAP II basal transcription machinery in order to modulate transcription. Cofactors generally do not bind DNA, but rather mediate PPIs between regulatory transcription factors and the basal RNAP II transcription machinery.' In order to make this definition even richer we concatenate definitions of its direct parents (i.e. GO:0003712 or 'transcription cofactor activity' and GO:0001076 or 'RNA polymerase II transcription factor binding transcription factor activity') and direct children (i.e. GO:0001105 or 'RNA polymerase II transcription co-activator activity' and GO:0001106 or 'RNA polymerase II transcription co-repressor activity') to its definition. We also add this GO term's representative name to this extended definition considering this name as part of its own definition. This process is done for

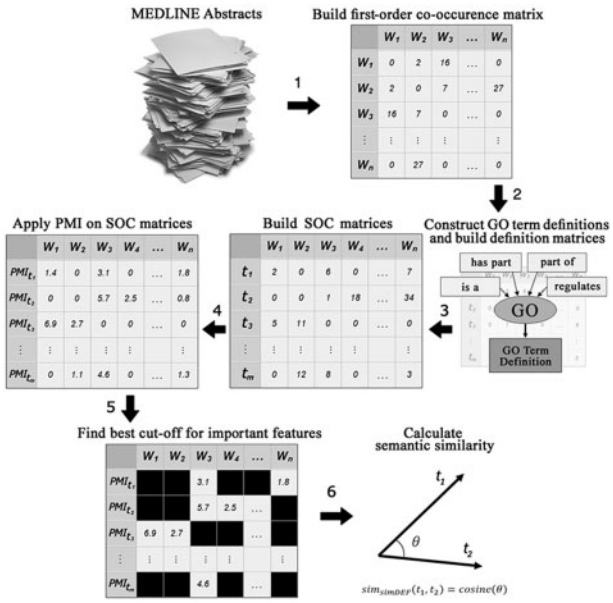

Fig. 1. Computation of the simDEF semantic similarity measure

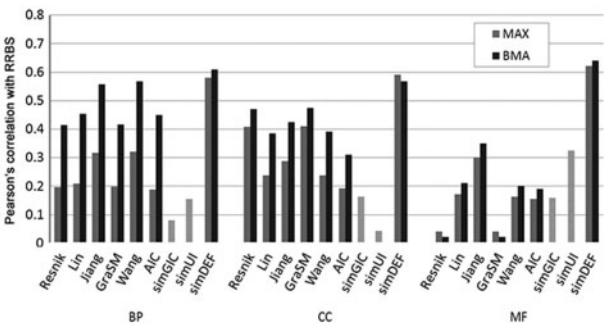

Fig. 2. Pearson's correlation between semantic measures and RRBS (IEA−)

all GO terms in BP, CC and MF. Now we see that for each word in the definition of a GO term we have an associated first-order co-occurrence vector calculated in Step 1. After changing all characters to lowercase and removing punctuation and stop words from these extended definitions we store them in different matrices for three different sub-ontologies. In these matrices, the value of a cell represents how many times the word associated with its column appears in the definition of the GO term associated with its row.

3.3 Step 3—building SOC matrices

To build the SOC vector for a GO term we sum the first-order co-occurrence vectors from the words in the constructed definition of that GO term (i.e. compute the centroid), and then normalize the result vector by the number of words in the definition. We do this process separately for each GO. The results are three different matrices for BP, CC and MF; each row again represents a GO term and features are the words. We have three SOC matrices for BP, CC and MF at the end of this step.

3.4 Step 4—PMI on SOC matrices (PMI-on-SOC matrices)

In our similarity measure, PMI-on-SOC matrices replace a conventional approach of low- and high-frequency cut-offs for detection of insignificant features or words in the Gloss Vector measure. We statistically measure the level of association between GO terms and their describing features in the SOC matrices and then apply a cut-off threshold on this level in the next step. Following (3), here, PMI(t_i, w_j) measures the level of association between GO term i and feature j to discover how descriptive the word j of that GO term is. PMI is biased toward low-frequency words and consequently tends to favor them by assigning them a higher degree of importance (Pesaranghader *et al.*, 2013); in order to resolve this weakness, we employ the add-one technique. Before applying PMI on a matrix, all the elements of the matrix are incremented by 1 unit.

3.5 Step 5—removing insignificant features from the PMI-on-SOC vectors

Defining a PMI threshold allows us to skip those words which provide low information for GO terms in their constructed PMI-on-SOC vectors. By using the available dataset in an iterative way, we gradually increase the threshold of PMI cut-off from zero and then evaluate the results generated by simDEF. Depending on the biomedical task, for a chosen cut-off threshold, criteria such as Pearson's correlation or AUC (see Section 4.3) can be used for the performance evaluation of the estimated similarity results. In general, as cut-off thresholds increase we tend to get better results until a point where performance starts to drop rapidly. Therefore, by recording this curve for different performance results and cut-off points we try to find the optimal cut-off point in order to keep only those informative features describing one GO term. Also, to avoid this interval being sensitive to the choice of dataset, we use 5-fold cross validation to predict the extent to which the threshold will generalize to an independent dataset. This cut-off selection is done separately for the three constructed PMI-on-SOC matrices of the BP, CC and MF ontologies.

3.6 Step 6—calculating semantic similarity

In this final step, the semantic similarity among GO terms is estimated. The cosine of the angle between optimized PMI-on-SOC vectors of two GO terms will indicate the degree of similarity for those terms. For the final usage of the measure, the last produced matrix is loaded into memory and used for measuring similarity between GO terms. In these matrices, each row stores the calculated optimized definition vector of its associated GO term.

As, in most cases, gene products are annotated with more than one GO term in the same ontology hierarchy (BP, CC or MF), there are several methods to measure the functional similarity of gene products based on the semantic similarity of these GO terms. MAX and *AVE* define functional similarity between gene products as the maximum or average semantic similarity values, respectively, over the GO terms annotating the genes. MAX has been shown to be more useful for a PPI task (Wu *et al.*, 2006). If T_A and T_B are the sets of GO terms which annotate proteins A and B, respectively, the MAX for their functional similarity measurement is achieved by

$$\mathrm{sim}_{\mathrm{MAX}}(A, B) = \max_{t_1 \in T_A, t_2 \in T_B} (\mathrm{sim}(t_1, t_2)). \quad (4)$$

The best-match average (BMA) method is found to be the best for evaluation of semantic similarity measures and the correlation of its results with sequence homology and gene expression data (Pesquita *et al.*, 2008). BMA for two gene products A and B with n and m GO annotations is given by

$$\mathrm{sim}_{\mathrm{BMA}}(A, B) = \frac{1}{2}\left(\frac{1}{n}\sum_{t_1 \in T_A} \max_{t_1, t_2 \in T_B} (\mathrm{sim}(t_1, t_2)) + \frac{1}{m}\sum_{t_2 \in T_B} \max_{t_1 \in T_A, t_2} (\mathrm{sim}(t_1, t_2))\right). \quad (5)$$

Consider that in these formulae T_A of different ontologies would be different (likewise for T_B). Therefore, we will achieve three different protein functional similarity values for three different gene ontologies.

MAX and BMA measure similarity between two gene products by combining semantic similarities between their terms. Semantic similarity estimation was used to evaluate the Resnik, Lin, Jiang, GraSM, Wang, AIC and simDEF measures (see Supplementary ma terial for definitions and formulas). In contrast, groupwise measures like simGIC and simUI are functional similarity measures by nature and do not rely on combining similarities between individual terms to assess gene product similarity, but calculate it directly by their annotation sets. By employing GO annotations for the previous measures and MEDLINE for the simDEF as the needed corpora, we implemented these measures as appropriate, and reported results alongside the best cut-off point for feature removal in each task.

4 Results

4.1 Correlation with sequence similarity

Several authors have compared the performance of different semantic similarity measures by testing how well these measures correlate with sequence similarity. Various studies (Lord *et al.*, 2003) showed that the more similar two sequences are the more similar their ontological annotations will be.

To evaluate the semantic similarity measures, we used two distinct sequence similarity measures: LRBS and RRBS with the formulae of (1) and (2). LRBS is similar to the sequence similarity measure used previously by Lord, but compensates for the fact that BLAST scores are not symmetric. RRBS, suggested by Joshi and Xu (2007), is another indicator of functional similarity acting like the sequence identity percentage by taking amino acid substitutions into account. Figure 3 shows the degree of correlation between LRBS and the functional similarity estimations calculated by semantic measures of 20 167 protein pairs (without IEAs included).

In all cases, whether we use MAX or BMA, simDEF correlates with sequence similarity better than the other IC-based measures. The high correlation between simDEF and LRBS in the MF ontology is notable as it is more than the twice of the second best measure's result (Jiang). Supplementary Table S1 shows the exact numerical results of this experiment (with and without IEAs).

The other metric used for sequence similarity measurement is RRBS which is not directly affected by sequence length (unlike LRBS). We assessed whether the dependency on sequence length affects the outcome of the evaluation. Figure 2 shows the degree of correlation between the similarity estimations calculated by semantic measures and RRBS. RRBS, like LRBS, shows the highest degree of correlation with simDEF among the similarity measures.

In general, measures of functional similarity correlate more with LRBS sequence similarity than RRBS. We also observe among IC-based measures tested here that no single measure is superior to all others in the BP, CC and MF ontologies, which suggests task-dependency of these measures. AIC, the latest variant of IC-based measures, does not offer any improvement over the earlier measures. The Wang topological measure of similarity works only slightly better than the IC-based measures in the RRBS sequence similarity comparison of BP. The correlation results for LRBS and RRBS also demonstrate that BMA is the appropriate metric for functional similarity measurement of proteins from BP and CC points of view when we use IC-based measures while for simDEF in CC it is reverse. The difference between results generated by BMA and MAX for simDEF is typically small, whereas other pairwise semantic similarity measures tend to show larger discrepancies. Supplementary Table S2 shows the exact results of this experiment.

4.2 Correlation with gene expression

Correlation with gene expression is another desirable criterion (e.g. Pesaranghader *et al.*, 2014b). Since genes involved in the same process tend to exhibit similar expression patterns, we could expect good semantic similarity estimations calculated on the BP ontology to be correlated with the expression similarity (Yang *et al.*, 2012). For our experiments, the evaluation is done against the available standard reference of 4800 gene expression values. Here, we report Pearson's correlation between gene expression data and the results from simGIC, simUI and BMA of pairwise measures. We focus on the BMA criterion as it always gave higher correlations. Pearson's correlation between gene expression and semantic measures for CC, BP and MF ontologies with and without IEAs considered are shown in Table 1.

The highest correlations in all cases are seen with the CC ontology, followed by BP and MF. Although the difference in correlation coefficients is not as striking as in the homology example, simDEF outperforms the next best method, GraSM, by 4% on the BP ontology and 1–2% on the CC ontology. GraSM has the best correlation for MF, 1–2% better than simDEF, which was also outperformed by the Resnik. Correlation coefficients were generally higher for datasets with IEAs, suggesting that electronic annotations have some value when investigating gene-expression profiles.

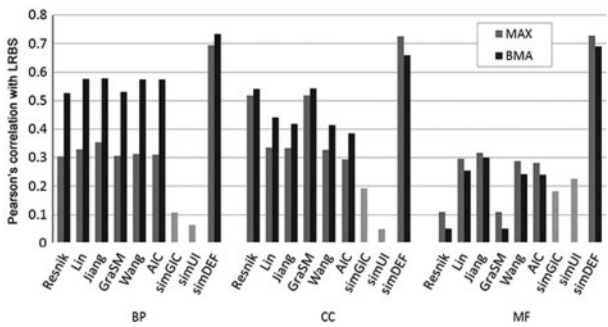

Fig. 3. Pearson's correlation between semantic measures and LRBS (IEA−)

Wang *et al.* (2004) and Sevilla et al. (2005) showed that the correlation between gene expression and semantic similarity was negligible when semantic similarity values were low, but the two measures were highly related when semantic similarity was high. Xu *et al.* (2008) further showed a linear relationship for gene pairs with high levels of expression correlation. We examined this trend by comparing Resnik against simDEF for variable numbers of the highest correlated genes. For this purpose, after sorting gene expression data from the highest to the lowest values, we measured correlation of these data with Resnik and simDEF as we go from the top correlated expressions to the bottom. Figure 4 demonstrates the trend of change for this test.

Considering other studies' findings and our result demonstrated in Figure 4, we see that by being more focused on high-correlated gene expression pairs the overal correlation between functional similarity and gene expression increase only when we take the BP ontology into account. For CC and MF the reverse is true. The other important point learned for BP is that by employing simDEF as semantic measure, when we ignore electronic annotation we get better correlation with highly-correlated gene expression data while this is not true for Resnik. Moreover, we observe that for BP and CC simDEF works better than Resnik no matter which subset we consider. Nevertheless, this issue does not hold for MF and we only get better results from simDEF when we focus on higher correlated genes in terms of their expression.

4.3 Comparison with PPIs

Semantic similarity can also be used as an indicator for the plausibility of putative PPIs, as proteins that interact in the cell *in vivo* are expected to participate in similar cellular locations and BPs. Like other studies (Jain and Bader, 2010; Wu *et al.*, 2013), we formulated this as a classification problem and checked how well the different semantic similarity measures perform for predicting true PPIs. For this purpose, the MAX and BMA results are directly interpreted as the classification probability of 'Interaction' and 'Not Interaction'. The higher this value is, the higher the probability of interaction will be. We applied this approach to a dataset of 6000 PPI pairs for each GO while half of the data have positive labels (due to experimentally confirmed PPIs) and the other half have negative labels.

Table 1. Pearson's correlation of semantic measures for three GOs using BMA against gene expression data (IEA+ and IEA−)

Semantic measure	Including IEA			Excluding IEA		
	BP	CC	MF	BP	CC	MF
Resnik	0.2659	0.4562	0.2514	0.2593	0.4426	0.2231
Lin	0.2541	0.3864	0.2155	0.2567	0.3842	0.2075
Jiang	0.2022	0.3217	0.1566	0.1757	0.2845	0.1708
GraSM	0.2677	0.4542	**0.2516**	0.2624	0.4395	**0.2252**
Wang	0.1911	0.3013	0.1306	0.1638	0.2805	0.1672
AIC	0.2466	0.3735	0.2149	0.2439	0.3593	0.2078
simGIC	0.0812	0.1542	0.1204	0.0667	0.1328	0.1422
simUI	0.1272	0.2418	0.0654	0.0628	0.0773	0.0455
simDEF	**0.3098**	**0.4649**	0.2325	**0.3071**	**0.4559**	0.2166

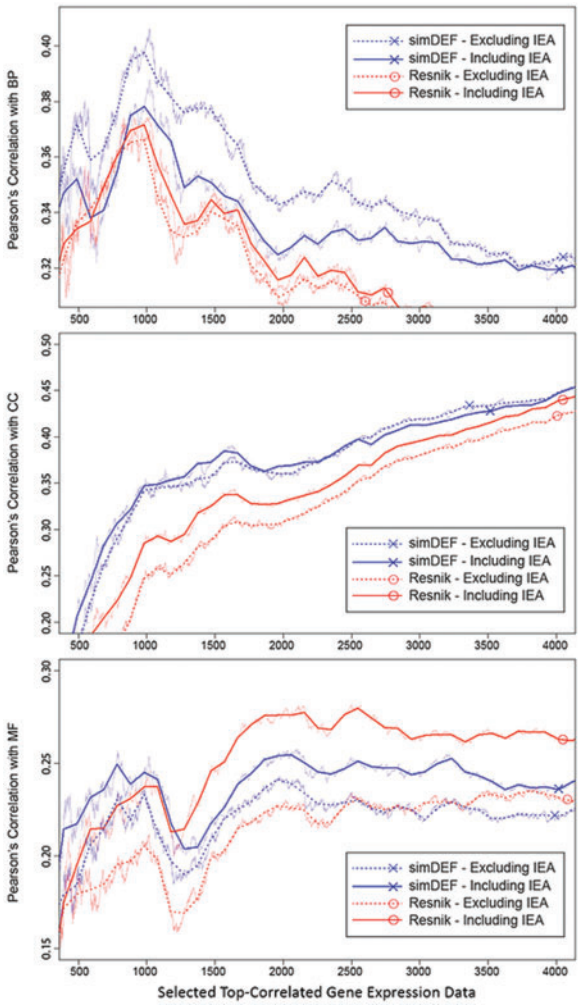

Fig. 4. Relationship of gene expression correlation and semantic similarity in three GO ontologies. $X = 500$ means that only the 500 most highly correlated gene pairs were considered when generating the correlation scores

In our evaluation, the results of prediction were investigated by receiver operating characteristic (ROC) curves, with area under the curve (AUC) as the main accuracy criterion. Here, we report only MAX since, as we expected from the previous studies, for all the cases MAX predicted better results compared with BMA. Table 2 shows the values of AUC for different semantic measures, including a hybrid measure that uses the average of the simDEF and Resnik values as the probability of interaction.

As in the gene expression case, we found that including IEA records from GO improved the accuracy (in this case, the AUC). simDEF gave the highest accuracy when the BP and MF ontologies were used, while Resnik and GraSM performed best for CC. The hybrid classifier's AUC results are represented in the last row of Table 2. This result shows that simDEF is useful on all three ontologies, whether alone or as a complement to the Resnik measure. We believe different approaches of simDEF and IC-based semantic measures in similarity estimation is the main reason for this improvement. With consideration of IEA, the ROC for CC ontology shown in Figure 5 represents that the combination of Resnik and simDEF benefits from the results of simDEF and Resnik both.

ROC is not always the only best approach to evaluate a classifier's performance in a PPI task (Jain and Bader, 2010; Wu *et al.*, 2013). Therefore, in our second experiment, by keeping the feature

Table 2. AUC of the semantic similarity measures for three GOs using MAX in the PPI task on the yeast dataset (IEA+ and IEA−)

Semantic measure	Including IEA			Excluding IEA		
	BP	CC	MF	BP	CC	MF
Resnik	0.8961	**0.8658**	0.7969	0.8685	**0.8525**	0.7429
Lin	0.8856	0.7588	0.7814	0.8629	0.7805	0.7419
Jiang	0.8719	0.7555	0.7613	0.8541	0.7467	0.7621
GraSM	0.8965	**0.8658**	0.7969	0.8691	0.8488	0.7413
Wang	0.8687	0.7835	0.7612	0.8483	0.7507	0.7496
AIC	0.8812	0.7623	0.7802	0.8613	0.7727	0.7427
simGIC	0.8014	0.8003	0.7025	0.7415	0.7673	0.6634
simUI	0.7999	0.7364	0.6921	0.7413	0.7098	0.6705
simDEF	**0.9086**	0.7742	**0.8202**	**0.9059**	0.8001	**0.8115**
simDEF + Res	**0.9264**	**0.8809**	**0.8306**	0.9039	**0.8564**	0.8073

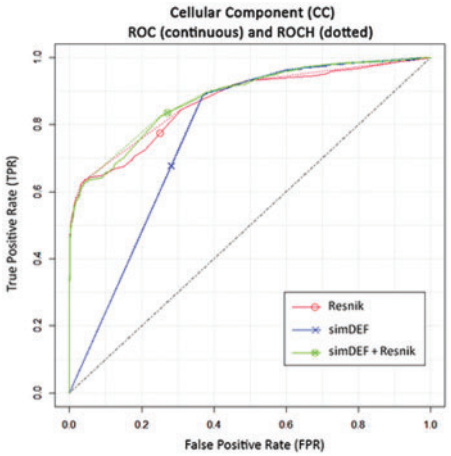

Fig. 5. ROC evaluation of the simDEF, Resnik and the hybrid measure of them by MAX for the PPI task at different classification cut-offs based on the yeast dataset using CC ontology (IEA+)

Table 3. F1-score of the simDEF, Resnik and the hybrid measure by MAX for the PPI task (IEA+)

Semantic measure	Mean of F1-score			Max of F1-score		
	BP	CC	MF	BP	CC	MF
Resnik	0.5973	0.5719	0.4699	0.8416	0.7815	0.7264
simDEF	**0.8154**	**0.7591**	**0.7084**	**0.8483**	**0.7889**	**0.7521**
simDEF + Res	0.6318	0.6686	0.5921	**0.8519**	**0.7962**	**0.7546**

The mean and maximum F1 scores are shown for all three GOs.

cut-off point of simDEF as before, considering Resnik as the baseline measure, and including IEA in the evaluation, we calculated different F1-scores for different classification cut-off points in the simDEF, Resnik and hybrid measures. Then, we compared the calculated mean and maximum F1-score values. While the mean and maximum F1-scores can be indicators of one classifier's performance in the detection of positive interactions (similar to AUC), maximum F1-score also helps in selection of the best classification cut-off point of a classifier having its ROC curve. The mean and maximum F1-score results are shown in Table 3.

The simDEF prediction of PPIs based on the F1-score is always better than the results achieved by Resnik. Even though Resnik gave the best AUC for the CC ontology, the simDEF mean F1-score is

considerably higher than that of Resnik, while the maximum scores differ by $< 1\%$. For the other two ontologies the improved mean of the F1-scores in the simDEF measure against Resnik is notable. For MF the difference between max F1-score in the hybrid measure is $>2.5\%$ compared with Resnik's F1-score itself. We also see this improvement in the result is due more to simDEF than to the Resnik measure.

5 Discussion

Our approach to similarity estimation based on shared context makes intuitive sense, as concepts which share closely related attributes in their representation should exhibit high levels of similarity. We have shown that implementing these ideas via the Gloss Vector representation yields improved effectiveness across the majority of ontologies and problem types. For the yeast database, simDEF increases the correlation of semantic similarity with sequence homology by 50%, yields an increase of $>4\%$ in correlation with gene expression on the BP ontology, and improves the PPI prediction F1-score by $>2.5\%$ on the MF ontology.

A key advantage of simDEF in comparison with IC-based measures is its reduced dependency on annotation data, and the GO structure. New GO terms typically do not have rich annotation information, which can influence the IC calculation of all GO terms as they depend on the *root* frequency which itself depends on all GO term frequencies. In contrast, simDEF needs to access only the direct parents and children of one GO term to expand that GO term's definition.

In future work, simDEF can be evaluated against Enzyme Commission (EC) and protein family (Pfam) similarities. Gene clustering and orthologous protein distinguishing tasks present yet another opportunity for simDEF performance evaluation. simDEF needs to be tested on the other species than *S.cerevisiae* as well. Moreover, other statistical measures of association, such as Chi-square and log-lokelihood, can be examined in replacement of PMI for further improvement of simDEF. More in-depth studies can also find out if using larger window sizes of bigrams or even tri-grams in the word extraction of MEDLINE abstracts would improve the achieved results. Also, current advancement in deep neural networks for the low-dimensional yet more accurate representation of GO terms leaves room for further investigation of semantic similarity measures in the distributional model.

Acknowledgements

The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada for this research.

Funding

S.M. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (grants RGPN 2450-2011 and CREATE 448111), and of Poland's National Scientific Center (grant no. NCN DEC2013/09/B/ST6/01549). M.S. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (grant no. 210812-151799-2001). S.M. and R.G.B. were supported by the Canada Research Chairs program.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Cherry,J.M. *et al.* (1998) SGD: Saccharomyces genome database. *Nucleic Acids Res.*, **26**, 73–79.

Couto,F.M. *et al.* (2011) Disjunctive shared information between ontology concepts: application to Gene Ontology. *J. Biomed. Semant.*, **2**, 5.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Firth,J.R. (1957) *A Synopsis of Linguistic Theory 1930–1955, Volume 1952–59*. The Philological Society, London.

Huang,D.W. *et al.* (2007) David gene functional classification tool: a novel biological module centric algorithm to functionally analyze large gene list. *Genome Biol.*, **8**, R183.

Jain,S. and Bader,G.D. (2010) An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, **11**, 562.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. *ArXiv Prepr*, Cmp-Lg9709008.

Jin,B. and Lu,X. (2010) Identifying informative subsets of the Gene Ontology with information bottleneck methods. *Bioinformatics*, **26**, 2445–2451.

Jin,B. *et al.* (2011). Mapping annotations with textual evidence using an scLDA model. In: *AMIA Annual Symposium Proceedings, Vol. 2011*, p. 834. American Medical Informatics Association.

Joshi,T. and Xu,D. (2007) Quantitative assessment of relationship between sequence similarity and function similarity. *BMC Genomics*, **8**, 222.

Lin,D. (1998) An information-theoretic definition of similarity. In: *Icml*, Madison, Wisconsin, USA, pp. 296–304.

Liu,M. *et al.* (2009) An weighted ontology-based semantic similarity algorithm for web service. *Expert Syst. Appl.*, **36**, 12480–12490.

Liu,Y. *et al.* (2012) Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (ACM)*, Miami, FL, USA, pp. 363–372.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Ovaska,K. (2015) Using Semantic Similarities and csbl. go for Analyzing Microarray Data. *Methods Mol. Biol.*, **10**, 1–12.

Pedersen,T. *et al.* (2004) WordNet:: similarity: measuring the relatedness of concepts. In: *Demonstration Papers at Hlt-Naacl 2004*, *Association for Computational Linguistics*, Boston, Massachusetts, USA, pp. 38–41.

Pesaranghader,A. and Muthaiyah,S. (2013) Definition-based information content vectors for semantic similarity measurement. In: Noah,S.A. *et al.* (eds.), *Soft Computing Applications and Intelligent Systems*. Springer, Berlin, pp. 268–282.

Pesaranghader,A. *et al.* (2013) Improving gloss vector semantic relatedness measure by integrating pointwise mutual information: optimizing second-order co-occurrence vectors computed from biomedical corpus and UMLS. In: *IEEE International Conference on Informatics and Creative Multimedia (ICICM) 2013*, Kuala Lumpur, Malaysia, pp. 196–201.

Pesaranghader,A. *et al.* (2014a) Word sense disambiguation for biomedical text mining using definition-based semantic relatedness and similarity measures. *Int. J. Biosci. Biochem. Bioinforma.*, **4**, 280–283.

Pesaranghader,A. *et al.* (2014b) Gene functional similarity analysis by definition-based semantic similarity measurement of GO terms. In: Sokolova,M. and van Beek,P. (eds.), *Advances in Artificial Intelligence*. Springer, Dordrecht, pp. 203–214.

Pesquita,C. *et al.* (2007) Evaluating GO-based semantic similarity measures. In: *Proceedings of the 10th Annual Bio-Ontologies Meeting,* Vienna, Austria, p. 38.

Pesquita,C. *et al.* (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9**, S4.

Piovesan,D. *et al.* (2015) INGA: protein function prediction combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res*,**43**, W134–W140.

Pu,S. *et al*. (2015) Extracting high confidence protein interactions from affinity purification data: at the crossroads. *J. Proteomics*, **118**, 63–80.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy. *ArXiv Prepr.* Cmp-Lg9511007.

Salwinski,L. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.

Schlicker,A. *et al*. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, **7**, 302.

Sevilla,J.L. *et al*. (2005) Correlation between gene expression and GO semantic similarity. *IEEEACM Trans. Comput. Biol. Bioinforma* **2**, 330–338.

Song,X. *et al*. (2014) Measure the semantic similarity of go terms using aggregate information content. *IEEE ACM Trans. Comput. Biol. Bioinforma*, **TCBB 11**, 468–476.

Teng,Z. *et al*. (2013) Measuring gene functional similarity based on group-wise comparison of GO terms. *Bioinformatics*, **29**, 1424–1432.

The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res*., **36**, D190–D195.

Wang,H. *et al*. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In: *IEEE Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04.*, La Jolla, CA, USA, pp. 25–31.

Wang,J.Z. *et al*. (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Warde-Farley,D. *et al*. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*., **38**, W214–W220.

Wu,X. *et al*. (2006) Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res*., **34**, 2137–2150.

Wu,X. *et al*. (2013) Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and IC-based hybrid method, *PLoS One*, **8**: e66745.

Wu,Z. and Palmer,M. (1994) Verbs semantics and lexical selection. In: *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, Stroudsburg, PA, USA, pp. 133–138.

Xu,T. *et al*. (2008) Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. *BMC Bioinformatics*, **9**, 472.

Yang,H. *et al*. (2012) Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, **28**, 1383–1389.