Journal of Theoretical Biology 401 (2016) 30–37

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Protein–protein interaction inference based on semantic similarity of Gene Ontology terms

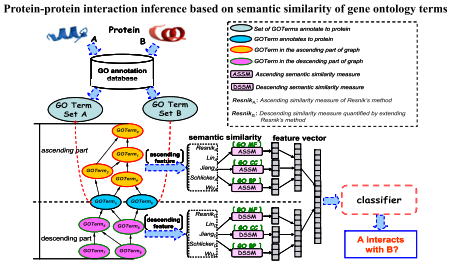
Shu-Bo Zhang [a,*], Qiang-Rong Tang [b]

[a] Department of Computer Science, Guangzhou Maritime Institute, Room 803, Building 88, Dashabei Road, Huangpu District, Guangzhou 510725, PR China
[b] Department of Shipping, Guangzhou Marine Institute, Room 205, Shipping Building, Hongshan No. 3 Road, Huangpu District, Guangzhou 510725, PR China

HIGHLIGHTS

- A GO-driven method to predict protein–protein interaction.
- Deriving similarity measure from the lower part of GO graph.
- Constructing feature vector by combining similarities from both upper and lower parts of the three GO graphs.
- Constructing feature vector by integrating different similarities of various methods.
- Integrated features generally outperform than individual feature.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 10 November 2015
Received in revised form
14 March 2016
Accepted 16 April 2016
Available online 23 April 2016

Keywords:
Protein–protein interaction
Gene Ontology
Ascending similarity
Descending similarity
Feature integration
Support vector machine

ABSTRACT

Identifying protein–protein interactions is important in molecular biology. Experimental methods to this issue have their limitations, and computational approaches have attracted more and more attentions from the biological community. The semantic similarity derived from the Gene Ontology (GO) annotation has been regarded as one of the most powerful indicators for protein interaction. However, conventional methods based on GO similarity fail to take advantage of the specificity of GO terms in the ontology graph. We proposed a GO-based method to predict protein–protein interaction by integrating different kinds of similarity measures derived from the intrinsic structure of GO graph. We extended five existing methods to derive the semantic similarity measures from the descending part of two GO terms in the GO graph, then adopted a feature integration strategy to combines both the ascending and the descending similarity scores derived from the three sub-ontologies to construct various kinds of features to characterize each protein pair. Support vector machines (SVM) were employed as discriminate classifiers, and five-fold cross validation experiments were conducted on both human and yeast protein–protein interaction datasets to evaluate the performance of different kinds of integrated features, the experimental results suggest the best performance of the feature that combines information from both the ascending and the descending parts of the three ontologies. Our method is appealing for effective prediction of protein–protein interaction.

1. Introduction

Identifying protein–protein interactions (PPIs) is a critical problem in molecular biology, because it is fundamental to our understanding of the lifecycle, molecular function, and molecular machinery, such as metabolism and regulatory aspects of the cell organization. Genome-

* Corresponding author.
E-mail addresses: 845996912@qq.com (S.-B. Zhang), tangqiangrong@hotmail.com (Q.-R. Tang).

http://dx.doi.org/10.1016/j.jtbi.2016.04.020
0022-5193/

scale protein interaction networks have been experimentally determined by high-throughput methods such as two-hybrid screens (Ito et al., 2001) and mass spectrometry (Gavin et al., 2006), and tandem affinity purification (Gavin et al., 2002). However, experimental techniques to detect PPIs are time-consuming, labor-intensive and tedious. Moreover, these experiments are often suffer from high false positive and false negative rates, because the data sets that are generated from these techniques are often noisy and incomplete (Von Mering et al., 2002). Due to these reasons, many researchers develop *in silico* approaches to analyze and infer protein–protein interaction.

There have been a variety of computational approaches that can be used to infer protein–protein interactions by exploiting various sources of information including primary structure of proteins (Zhang et al., 2011; Huang et al., 2015; You et al., 2015), three-dimensional structural information (Aloy et al., 2004), genomic information (Tsoka and Ouzounis, 2000), or integration of multiple genomic datasets (Lin et al., 2004), phylogenetic relationships (Ramani and Marcotte, 2003), protein complex data (Martin et al., 2007) and pre-defined domain–domain interactions (Liu et al., 2005), gene expression profiles (Chin et al., 2010), etc. Conventional machine learning techniques are used when there are sufficient known interactions to serve as training data. Despite of this, inferring PPIs is still a challenge in molecular biological domain.

Gene Ontology (GO) is a comprehensive resource across species describing gene and gene product biological properties related to biological process, molecular function, and cellular component. It provides us with promising ways to characterize the functional relationship between pairs of proteins and to infer the interaction between them at functional level. GO contains two components: the GO graph and annotation database. The former is composed of well-defined terms with definite relationships that connect terms in a species-independent manner (Gene Ontology Consortium, 2015). It is structured as a directed acyclic graph (DAG), which contains three orthogonal sub-ontologies: molecular function (MF), cellular compartment (CC) and biological process (BP). The nodes inside the DAG represent certain biological semantic concepts and edges denote the semantic relations between the concepts. The annotation database is composed of GO terms and the corresponding gene products that they annotate to. By exploring the semantic similarity between pairs of GO terms in the DAG, we can infer the semantic relationship between two proteins, which in turn can be used to explore the possibility that they interact with each other. The idea of this lies in the observation that interacting proteins are more likely to participate in the same biological process, and occur in the same cellular component, which will lead to larger semantic similarity score between proteins. Several studies have been carried out for protein–protein interaction prediction, where the similarity score of each protein pair was quantified by the number of GO terms annotating to the two proteins simultaneously (Jansen et al., 2003), and some researchers have recognized the semantic similarity in GO annotation as one of the most powerful indicators for protein interaction (Gene Ontology Consortium, 2015; Miller et al., 2005). However, conventional methods that use Gene Ontology for protein–protein interaction prediction measure the similarity between proteins in terms of the shared GO terms in a controlled vocabulary system (Jansen et al., 2003; Stelzl et al., 2005; Rhodes et al., 2005; Martin et al., 2004). These approaches do not derive similarity measure between pairs of GO terms from the structure of DAG, they are restricted to protein pairs with the same annotations and do not take advantage of the specificity of GO terms. Hence, two proteins may have some common terms that are too general to characterize the semantic relationship among the proteins (Guo et al., 2006).

To overcome the limitation of predicting protein–protein interaction based on GO, some researches adopted another strategy to derive the similarity measure between two proteins for protein–protein interaction prediction, which is the focus of this study. Many similarity measures in this category are established on the intrinsic structure of the GO graph, from which we can derive a subgraph that includes a specific set of GO terms directly annotating to each protein, and all the parents of those terms annotating that protein implicitly. Since this kind of semantic similarity measures can derive more specific information for the similarity measure of a protein pair, they generally produce better results. As a matter of fact, different kinds of measures characterize the similarity or the dissimilarity (based on the distance from one term to the others) of two proteins from different point of views, it may be helpful to integrate various kinds of similarity measures for better prediction capability.

Inspired by previous work on protein–protein interaction prediction based on Gene Ontology, we proposed a new approach to predict protein–protein interaction based on the similarity measures derived from the intrinsic structure of the GO graph (we call the similarity between two GO terms term-based similarity in this study). We firstly extended five existing methods to derive similarity measures from the beneath structure of two terms in the GO graph, then employed a pair-wise strategy with best-match average rule to quantify the semantic similarity measure between two proteins from the term-based similarity scores of two term sets. Subsequently, different kinds of feature vectors were constructed to characterize each protein pair from different aspects, and used as input to train binary support vector machines. Validation experiments were conducted on both human and yeast protein–protein interaction datasets, and the experimental results shown the promising superiority of our method.

2. Related work

Previous researches on GO-based PPI inference focused on two issues: (1) feature retrieval that aims at measuring the semantic similarity between two proteins, and (2) machine learning technique, which develops classification algorithm to learn an efficient classifier. For feature retrieval, most researches measure the semantic relationship of a protein pair from two term sets with elements in each set annotating to one of the two proteins, respectively. The simplest idea is to quantify the ratio of terms shared by two proteins, which was used by He et al. to identified novel interactions in human (He et al., 2009). The limitation of this kind of measure is that it ignores the specificity of the terms in the GO graph. Other measures taking into account the hierarchical structure of GO graph can be categorized into three groups: node-based methods (also called information content-based methods); edge-based methods; and hybrid methods that combine both the node information and the GO graph structure for semantic similarity. Node-based semantic similarity measures are established on the properties of GO terms, and are possibly the most frequently mentioned metrics in the literature (Benabderrahmane et al., 2010), among which the measures proposed by Resnik (1995), Lin (1998) and Jiang and Conrath (1997) are most widely used. Schlicker et al. (2006) developed a relevance similarity to reduce the effect of shallow annotations by combining Resnik's and Lin's measures, where they weighted Lin's measure with a tuning factor $1-p(t)$. Edge-based semantic similarity measures take advantage of the length of the shortest path connecting two terms. Rada et al. (1989) originally introduced a similarity measure of this kind based on the shortest path connecting two terms, and Nagar and Al-Mubaid (2008) applied it on GO terms for the first time. As for hybrid measures, Wu et al. (2013) defined a similarity measure that is composed of three components: the overlap of induced term subgraph, term generality and term distances to their lowest common ancestors. Guo et al. (2006) compared five semantic similarity measures to predict human PPIs and found Resnik's measure to perform best. Jain and Bader (2010) introduced a Topological Clustering Semantic Similarity (TCSS) that

attempts to compensate for the unequal depths of different branches of the GO DAG. Maetschke et al. (2012) introduced an inducer-based prediction algorithm, where they derived information from the GO terms that annotate to protein pairs. Mukhopadhyay et al. (2011) proposed an improved binary differential evolution technique to select GO-based semantic similarity measure for protein–protein interaction prediction.

For the machine learning approach, Jansen et al. (2003) employed a Bayesian network, which used information derived from GO biological process and other resources as features to predict interaction of proteins in yeast. Patil and Nakamura (Patil and Nakamura, 2005) trained a Naive Bayes classifier with input features such as shared GO terms, structure information and sequence homology similarity to infer PPI. Ben-Hur and Noble (2005) and Miller et al. (2005) both used a support vector machine (SVM) and a combination of multiple kernels, information derived from sequence data, GO annotation, network properties and interologs, to predict protein–protein interactions in yeast. Qiu and Noble (2008) also used a multi-kernel SVM but applied it to the prediction of co-complexed protein pairs. Other machine learning algorithms including decision trees (Zhang et al., 2004), random forests (You et al., 2015), AdaBoost strategy (Mei and Zhu, 2014), logistic regression (Lin et al., 2004), and the strategy of summing likelihood ratio scores to predict PPI confidence in human (Rhodes et al., 2005) or yeast (Wu et al., 2006). Among the machine learning techniques used, random forests and support vector machines (SVMs) were found to achieve the best performance (Qi and Noble, 2011).

3. Materials and methods

3.1. Materials

The GO consortium provides publicly available releases of Gene Ontology database and gene annotation dataset. In this study, the GO and the gene annotation datasets released in July 2015 were used to test our approach. The GO database contains 27,985 BP, 3826 CC, and 9954 MF terms. The gene annotation dataset contains 94,980 annotations (49,568 without IEA—Inferred from Electronic Annotation) of 6380 genes for the yeast genome, while 486,639 annotations (340,241 without IEA annotations) of 19,433 genes for the human genome. Both yeast and human protein–protein interaction datasets built by Wu et al. (2013) were used to validate the performance of our method. There are 12,846 and 11,606 human protein pairs, 23,424 and 22,754 yeast protein pairs including and excluding IEA annotations, respectively. The details about human and yeast protein pairs in each datasets are listed in Table 1.

Table 1
The number of positive and negative interaction protein pairs in yeast and human datasets including and excluding IEA.

		Positive samples			Negative samples			Total
		BP	CC	MF	BP	CC	MF	
Yeast	Including IEA	4095	4129	3488	4095	4129	3488	23,424
	Excluding IEA	4061	4113	3203	4061	4113	3203	22,754
Human	Including IEA	2150	2179	2094	2150	2179	2094	12,846
	Excluding IEA	2016	1945	1842	2016	1945	1842	11,606

The protein pairs in the positive datasets are selected from DIP, while those in negative datasets are generated by randomly choosing from iRefWeb, but not identified as interaction pairs.

4. GO-based feature derivation

In this sector, we firstly introduce five existing methods that capture the semantic similarity measures between two GO terms from the ascending part of the GO graph (i.e., the ontology structure above the terms under consideration), and then extend these five methods to quantify the semantic similarity measures of two terms from the descending part (i.e., the ontology structure beneath the terms under consideration). Based on these similarity measures, we subsequently construct the feature vector to characterize each protein pair. The five traditional semantic similarity measures (i.e., Resnik, 1995; Lin's, 1998; Jiang and Conrath's, 1997; Schlicker et al., 2006; Wu's et al., 2013; measures) are referred to as ascending similarity measures, they capture the semantic similarity scores from the upper part of two GO terms, and are defined as follows.

Suppose that t_1 and t_2 are two terms, Resnik's method quantifies the similarity measure of a term pair as the information content (IC) of their most informative common ancestor (MICA),

$$ASIM_{Resnik}(t_1, t_2) = max\{IC(t) | t \in CA(t_1, t_2)\} \quad (1)$$

where $CA(t_1, t_2)$ is the set of common ancestor terms of t_1 and t_2 in the GO graph. $IC(t)$ denotes the IC value of term t, which is defined as $-\log p(t)$, where $p(t)$ is the probability that t occurs in a certain GO annotation corpus, such as human annotation database or yeast annotation database.

Lin's measure is defined as the ratio of information content between the most informative common ancestor and the both terms,

$$ASIM_{Lin}(t_1, t_2) = \frac{2 \times ASIM_{Resnik}(t_1, t_2)}{IC(t_1) + IC(t_2)} \quad (2)$$

Jiang and Conrath's measure is quantified according to the following formula:

$$ASIM_{Jiang}(t_1, t_2) = \frac{1}{1 + Dist_{Jiang}(t_1, t_2)} \quad (3)$$

where $Dist_{Jiang}(t_1, t_2) = IC(t_1) + IC(t_2) - 2 \times ASIM_{Resnik}(t_1, t_2)$.

Schlicker's relevance similarity takes into account the location of the MICA by weighting Lin's measure with a tuning factor $1 - p(t)$,

$$ASIM_{Rel}(t_1, t_2) = \frac{2 \times ASIM_{Resnik}(t_1, t_2)}{IC(t_1) + IC(t_2)} \times (1 - p(t)) \quad (4)$$

where $p(t)$ is the probability that the most informative common ancestor t occurs.

By taking into consideration of both the semantic distance between a term pair and the depth of terms in the GO graph, Wu et al. (2013) defined the their measure as

$$ASIM_{Wu}(t_1, t_2) = \frac{1}{1 + \gamma} \times \frac{\alpha_{IC}}{\alpha_{IC} + \beta_{IC}} \quad (5)$$

where γ is the semantic distance between t_1 and t_2, α_{IC} is the IC value of the MICA, β_{IC} is the average of semantic distance values, which measure the distances from t_1 and t_2 to their corresponding most informative leaf nodes, respectively.

By examining the lower part of a term pair in the GO graph, we observe that a common descendant of two ancestral terms synthesizes the semantics of its parental nodes into new instance or component denoting a new molecular function, biological process or more specific cellular compartment. This means that the common descendants of two GO terms can also be used to characterize their similarity measure. Furthermore, a common descendant closer to two GO terms investigated indicates the larger similarity score of the term pair. Based on this observation, we can derive the similarity measure between two terms from their lower part by taking into account their

closest common child node as we do in the upper part. To this end, we extend the above five existing methods to capture the similarity measures of two terms from their descending part in the GO graph. In order to obtain the same similarity measure scale from the lower part as those from the upper part, we redefine the IC value of a common child term, instead of adopting it original IC value. Suppose IC_{t_1} and IC_{t_2} are the IC values of terms t_1 and t_2, IC_t is the original IC value of their common child term t, we define the new IC value of t as

$$\widehat{IC}(t) = max(IC_{t_1} + IC_{t_2} - IC_t) \tag{6}$$

It should be noted that the IC value defined in formula (6) may not be positive, which is inconsistent with our common knowledge that the information content is positive. To address this issue, we set $\widehat{IC}(t) = 1/IC_t$ if $IC_{t_1} + IC_{t_2} - IC_t \le 0$. Thus, the value of $\widehat{IC}(t)$ conforms to our previous observation that a common descendant closer to the parental term pair has larger similarity score. Then the IC value of the closest descendant, which is corresponding to that of the MICA in the upper part of two terms, can be defined as

$$DSIM_{Resnik}(t) = max\{\widehat{IC}(t), |t \in CD(t_1, t_2)\} \tag{7}$$

where $CD(t_1, t_2)$ is the set of common descendant terms of t_1 and t_2 beneath the GO structure of these two terms. Subsequently, we can define five kinds of descending similarity measures between two terms from their lower part by replacing the $ASIM_{Resnik}$ in formulae (1)–(4) with $DSIM_{Resnik}$, and replacing the α_{IC} with $DSIM_{Resnik}$, γ with the semantic distance between t_1 and t_2 associated with they closest common descendant, and β_{IC} with the average of semantic distance values in formula (5), respectively.

Based on these term-based similarity measures, we can derive protein-based semantic similarity measures, which will serve as feature to characterize protein pairs and identify whether two proteins interact with each other or not. As a protein may be annotated with multiple GO terms, we adopted a pair-wise strategy with best-match average (BMA) rule (Azuaje et al., 2005) to estimate the semantic similarity value between two proteins in this study. By the BMA rule, the similarity measure between proteins p and q is defined as the mean of the following two values: average of maximum similarity scores between each GO term annotated to protein p and those annotated to protein q, and average of maximum similarity scores between each GO term annotated to protein q and those annotated to protein p, which can be quantified by the following formula (Mazandu and Mulder, 2012):

$$BMA(p, q) = \frac{1}{2}\left(\frac{1}{n}\sum_{t\in T_p^A} Sim\left(t, T_q^A\right) + \frac{1}{m}\sum_{t\in T_q^A} Sim\left(t, T_p^A\right)\right) \tag{8}$$

where $Sim\left(t, T_q^A\right) = \max\left\{Sim(t, s): s \in T_q^A\right\}$ is the maximum similarity score between GO term t and those terms annotated to protein q, T_q^A is a set of GO terms in A denoting the biological process (BP), molecular function (MF) or cellular component (CC) sub-ontology annotating protein q, $Sim(t, s)$ is the term-based similarity score between GO terms t and s, and $n = \left|T_p^A\right|$ and $m = \left|T_q^A\right|$ are the numbers of GO terms in these sets.

The similarity score between two proteins can be considered as feature that characterize each protein pair, and we can further build feature vector to represent protein pairs by integrating different kinds of GO-based similarity scores derived from the ascending and the descending parts of the GO structure as

$$F_{asc+des} = [ASIM, DSIM] \tag{9}$$

where $ASIM$ and $DSIM$ denote the ascending and the descending similarity scores of a protein pair, respectively. In order to investigate how the integration of different similarities helps to improve the prediction power, we constructed two kinds of integrated feature vectors; the first one is constructed by concatenating the five ascending similarity scores as

$$F_{intasc} = [ASIM_R, ASIM_L, ASIM_J, ASIM_R, ASIM_W] \tag{10}$$

where $ASIM_R$, $ASIM_L$, $ASIM_J$, $ASIM_R$ and $ASIM_W$ denote the similarity scores of protein pairs derived by Resnik's, Lin's, Jiang's, Schlicker's and Wu's methods, respectively. The second integrated feature vector is constructed by concatenating the five ascending similarity scores and the five descending similarity scores as follows,

$$F_{int(asc+des)} = [ASIM_R, ASIM_L, ASIM_J, ASIM_R, ASIM_W, DSIM_R, DSIM_L, DSIM_J, DSIM_R, DSIM_W] \tag{11}$$

where $DSIM_R$, $DSIM_L$, $DSIM_J$, $DSIM_R$ and $DSIM_W$ denote the descending similarity scores derived from the lower part of GO terms corresponding to the ascending measure of Resnik's, Lin's, Jiang's, Schlicker's and Wu's methods, respectively. Since formula (11) can be used to extract feature vector from each of the sub-ontologies (MF, CC and BP sub-ontologies), we can further obtain a 30-dimensional feature vector to characterize a protein pair by concatenating three feature vectors derived from the three sub-ontologies according to formula (11).

5. Machine learning algorithm

Support vector machine (SVM) is a well known machine learning algorithm, which has been widely used in many fields including pattern recognition, computer vision and bioinformatics. As mentioned above, SVM has been recognized as one of the most effective classifiers for protein–protein interaction, we adopted SVM as the binary classifier to test different GO-based similarity measures for the protein–protein interaction prediction in this study. The radial basis function was selected as the kernel function for its better performance in many cases. The radial basis function is used to quantify the relationship between two pairs of proteins and defined as follows,

$$K(v, u) = exp(-\gamma\|v - u\|^2) \tag{11}$$

where v and u are feature vectors charactering two protein pairs respectively. γ is the parameter of the Gaussian kernel function. As for the ascending part of GO graph, we can derive five similarity score values and build a 5-dimensional feature vector for each protein pair. If we take into account both the ascending and the descending parts of the GO graph, we will obtain a 10-dimensional vector for each protein pair.

The software package libsvm-3.12 (Hsu and Lin, 2002) implemented (in MATLAB) by Chang and Lin was downloaded from http://www.csie.ntu.edu.tw/~cjlin and used to design the binary classifier in this paper. The best regularization parameter C and the kernel parameter γ were determined by a grid search algorithm. For each dataset, the search algorithm performed a five-fold cross validation on each of the following grid points of (C, γ): $[2^{-4},2^{-3},2^{-2}, \cdots,2^{4},2^{5}] \times [2^{-4},2^{-3},2^{-2}, \cdots,2^{4},2^{5}]$. At each point, one fold was used as validation set while the rest as training set. After a reasonable point was identified, we conducted a finer search step for better parameters. The experiments were conducted in MATLAB 2008a on a machine with 2.67 GHz Intel quad core processors and 4 GB of RAM. Various kinds of feature vectors derived from ascending and the descending parts of the three sub-ontologies served as input of the binary SVM classifiers.

6. Implementation of our approach

As we previously described, ten kinds of similarity scores were derived from the GO graph to quantify the relationship between two proteins, and each them was used as a feature to characterize a protein pair. From the view point of pattern recognition, to distinguish interacting protein pairs from non-interacting ones, we should establish some features to characterize each protein pair. To this end, the similarity scores were further assembled into feature vectors. Formula (10) shows how the scores derived from the ascending part of GO graph were assembled into a feature vector $F_{intasc} = [ASIM_R, ASIM_L, ASIM_J, ASIM_R, ASIM_W]$, while formula (11) shows how the scores derived from both the ascending part and the descending part of GO graph were assembled into a feature vector $F_{int(asc+des)} = [ASIM_R, ASIM_L, ASIM_J, ASIM_R, ASIM_W, DSIM_R, DSIM_L, DSIM_J, DSIM_R, DSIM_W]$. These two kinds of feature vectors serve as input of the binary SVM classifier for training the classifier and predicting unknown protein pairs.

The implementation of our approach is demonstrated in Fig. 1, and the pseudo-codes are listed in Algorithm 1. The algorithm contains four stages: 1) map each protein into a GO term set; 2) quantify the similarity scores between protein pairs; 3) construct feature vectors to characterize each pair of proteins; and 4) use feature vectors as input of the SVM classifier to determine whether two proteins interact with each other. In the first stage, the annotation database is queried with each protein, and the GO terms annotating to that protein form a term set. In the second stage, the similarity scores between GO term pairs for different methods are computed from both the ascending and descending part of the GO DAG, and the pair-wise strategy with best-match average (BMA) rule is adopted to estimate the similarity between protein pairs. In third stage, different kinds of similarity scores are combined into a feature vector to characterize each protein pair, and in the last stage, the feature vector serves as input of SVM classifier to predict whether two proteins interact with each other.

Algorithm 1. GOSIMP2P Algorithm Input: protein pair **P1, P2**Output: prediction result **R**Description:

1) load trained SVM classifier
2) load GO graph
3) load annotation database
4) find annotation GO term sets **S1**, **S2** for **P1**, **P2**
5) compute five similarity scores from the ascending part of DAG for **S1**, **S2** $ASIM_R$, $ASIM_L$, $ASIM_J$, $ASIM_R$, $ASIM_W$
6) compute five similarity scores from the descending part of DAG for **S1**, **S2** $DSIM_R$, $DSIM_L$, $DSIM_J$, $DSIM_R$, $DSIM_W$
7) create feature vector from the similarity scores
8) query the SVM classifier with feature vector
9) return prediction result **R**

7. Results

To validate the effectiveness of the proposed method, we adopted five-fold cross validation and employed the overall prediction accuracy as validation measure in this study. In five-fold cross validation, the dataset was randomly partition into five equal subsets, and the validation process was repeated five times. At each time, a subset was retained as validation data and the remaining four subsets were used as training data. The five results of the cross-validation were averaged to produce an overall estimation. The overall prediction accuracy (ACC) is the percentage of correctly discriminated interacting and non-interacting protein pairs and given by:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

where TP, FN , FP and TN denote the numbers of protein pairs that are predicted as true positive, false negative, false positive and true negative, respectively.

In order to identify whether two protein interact with each other or not, we should extract features to characterize each protein pair based on their GO-based similarity scores. To this end, we firstly constructed a GO term set for each protein with each element in the set directly annotating that protein, and then adopted the pair-wise strategy with best-match average (BMA) rule to estimate the semantic similarity value between two proteins. In this study, the similarity scores between proteins were derived from the ascending and the

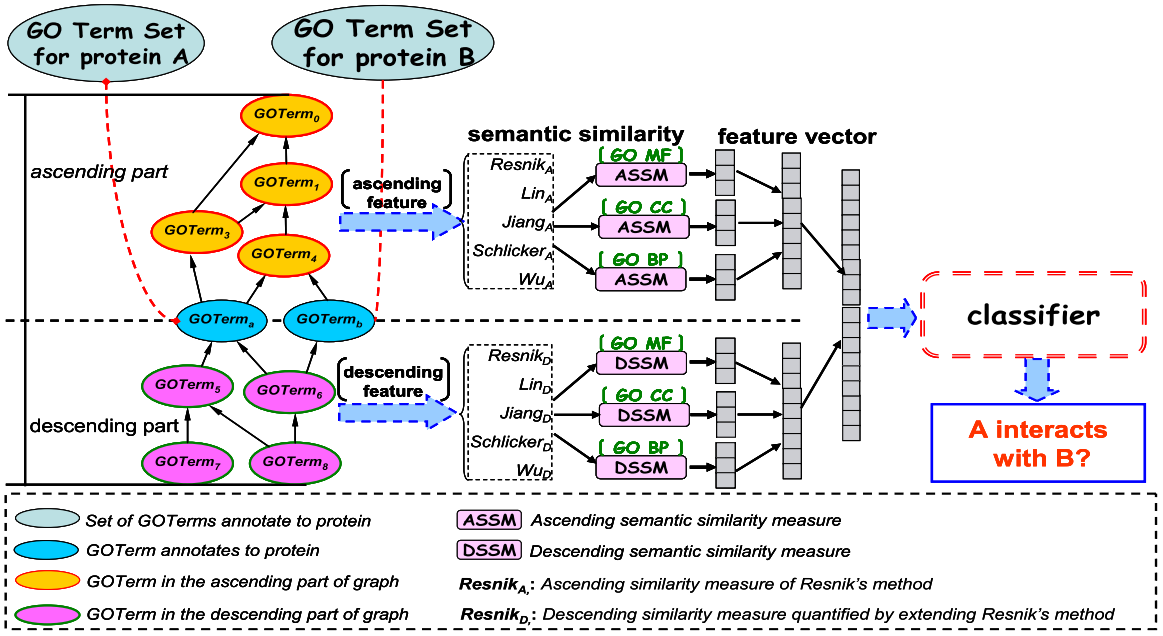


Fig. 1. The architecture of our method. Two GO term sets are firstly assigned to two proteins; then the semantic similarity scores of the two term sets are derived from the upper part and lower part of the GO DAG, and these scores are further used to construct feature vector to characterize each pair of protein; finally, the feature vector serves as input to an SVM classifier.

descending parts of the GO structure by five existing methods and their corresponding extensions, respectively. Different kinds of similarity scores derived from each of the three sub-ontologies: CC, BP, and MF were further used to construct feature vectors according to formulae (10) to (11). All feature vectors that combine different kinds of similarity scores to characterize protein pairs serve as input to train a two-class SVM classifier and used to identify whether two proteins interact with each other.

Tables 2 and 3 list the overall prediction accuracy rates of the features that combine different similarity measures for the three GO aspects including and excluding IEA annotations on human and yeast PPI datasets. In these tables, the increments of prediction accuracy rates produced by the combined feature (integrate ascending with descending similarity scores) against those produced by its corresponding ascending feature are also listed. From these tables, we see that the features that combine both ascending and descending similarity measures generally achieve higher prediction accuracy rates than they corresponding ascending features for different similarity measures, except for Wu's measures on MF and BP sub-ontologies and Schlicker's measure on MF sub-ontology. This implies that the similarity measure derived from the lower part of GO graph contributes to the semantic relationship of two proteins, which better characterizes each protein pair and will in turn contribute to the improvement of the prediction accuracy.

In comparison with the features that contain only one single similarity score, the features that integrate multiple similarity measure scores derived by five different methods achieve higher prediction accuracy rates under all scenarios, no matter if the descending similarity measure is combined with its corresponding ascending similarity measure. These maybe due to the fact that different kinds of similarity measures characterize the relationship between GO terms from distinct aspects, and the combination of these measures can help to descript their semantic relationship of protein pairs.

In order to evaluate the prediction power of features derived from different sub-ontologies, as well as their combination, the five-fold cross validation experiments were also conducted on the features derived from MF, CC and BP sub-ontologies and the integration of these three kinds of features. Figs. 2–5 illustrate the prediction accuracy rates of different methods for the three single sub-ontologies and their combination on human dataset including and excluding IEA annotations (see Figs. S1–S4 for those on yeast dataset including and excluding IEA annotations, respectively). Figs. 2 and 4 show the comparison ACC scores produced by features that only take advantage of information from the ascending part of GO graph including and excluding IEA annotations, respectively. Figs. 3 and 5 show the results produced by features that combine both ascending and descending

Table 2
The ACC scores of different similarity measures for the three sub-ontologies on yeast PPI datasets.

		MF			CC			BP		
		asc	**asc+des**	**↑%**	**asc**	**asc+des**	**↑(%)**	**asc**	**asc+des**	**↑(%)**
Including IEA	Resnik	71.2731	71.8963	0.6232	78.4366	79.8668	1.4302	83.0003	84.9684	1.9681
	Lin	69.1684	69.7831	0.6147	76.0716	76.5241	0.4525	82.0483	83.4828	1.4345
	Jiang	69.1171	69.8386	0.7215	74.4664	77.617	3.1506	82.027	82.9833	0.9563
	Schlicker	70.2741	70.193	−**0.0811**	77.126	77.3864	0.2604	82.8082	83.5852	0.777
	Wu	69.924	70.2912	0.3672	78.0396	79.1325	1.0929	83.8072	83.978	0.1708
	intSIM	**73.5186**	**74.0992**	0.5806	**80.9469**	**83.1839**	2.237	**85.041**	**85.7454**	0.7044
Excluding IEA	Resnik	68.2517	68.9549	0.7032	77.1469	79.4893	2.3424	81.1681	85.1059	3.9378
	Lin	67.72	67.821	0.101	75.5779	77.6479	2.07	81.9636	83.7215	1.7579
	Jiang	66.4982	66.5114	0.0132	73.8551	76.8217	2.9666	82.1482	82.4251	0.2769
	Schlicker	68.2561	68.2957	0.0396	77.2128	78.5137	1.3009	82.2009	83.8094	1.6085
	Wu	66.2345	65.795	−**0.4395**	76.3163	77.1688	0.8525	82.4514	82.6668	0.2154
	intSIM	**70.8359**	**71.1655**	0.3296	**80.9924**	**81.7351**	0.7427	**84.5829**	**85.5322**	0.9493

'↑%' denotes the percentage that the ACC score based on integrated feature larger than that of ascending feature. Minus figures in the table indicates that the ACC values of the former is larger those of the latter. 'asc' denotes feature derived from the semantic similarity measure from the upper part of GO graph, while 'asc+des' denotes feature derived from similarity measure of both upper and lower part of GO graph.

Table 3
The ACC scores of different similarity measures for the three sub-ontologies on human PPI datasets.

		MF			CC			BP		
		asc	**asc+des**	**↑%**	**asc**	**asc+des**	**↑(%)**	**asc**	**asc+des**	**↑(%)**
Including IEA	Resnik (1995)	71.0883	78.1644	7.0761	74.4045	78.5069	4.1024	84.6411	86.2603	1.6192
	Lin (1998)	72.5051	78.2734	5.7683	75.8446	76.4985	0.6539	84.1974	85.7076	1.5102
	Jiang and Conrath (1997)	75.2141	75.8524	0.6383	74.5602	75.3542	0.794	83.6369	84.898	1.2611
	Schlicker et al. (2006)	71.6176	78.2734	6.6558	76.5297	77.4093	0.8796	84.7735	85.7388	0.9653
	Wu et al. (2013)	73.1667	73.0344	**-0.1323**	76.9033	77.3937	0.4904	85.8555	85.7388	−**0.1167**
	intSIM	**78.5692**	**80.3285**	1.7593	**78.5536**	**80.2351**	1.6815	**86.2525**	**86.6962**	0.4437
Excluding IEA	Resnik (1995)	72.1868	79.1229	6.9361	74.3236	78.8471	4.5235	83.3362	85.0164	1.6802
	Lin (1998)	74.9354	79.2521	4.3167	76.1158	77.2014	1.0856	83.1983	84.3185	1.1202
	Jiang and Conrath (1997)	77.503	78.425	0.922	76.2192	77.7012	1.482	83.0174	83.5688	0.5514
	Schlicker et al. (2006)	73.4879	78.7955	5.3076	76.6672	77.7357	1.0685	83.1811	84.4994	1.3183
	Wu et al. (2013)	70.386	70.4377	0.0517	77.4858	78.3991	0.9133	84.4563	84.1289	−**0.3274**
	intSIM	**79.0109**	**79.8725**	0.8616	**79.9414**	**81.5096**	1.5682	**85.0078**	**85.8952**	0.8874

'↑%' denotes the percentage that the ACC score based on integrated feature larger than that of ascending feature. Minus figures in the table indicates that the ACC values of the former is larger those of the latter. 'asc' denotes feature derived from the semantic similarity measure from the upper part of GO graph, while 'asc+des' denotes combined feature derived from similarity measure of both upper and lower part of GO graph.

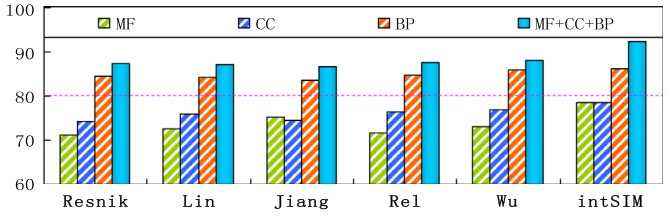


Fig. 2. Comparison of ACC scores of different features derived by different methods from the upper part of various aspects of sub-ontologies on human dataset including IEA annotations.

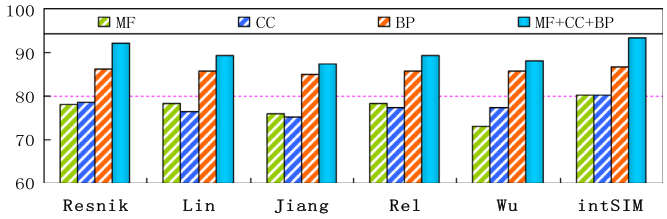


Fig. 3. Comparison of ACC scores of different features derived by different methods from both the upper and the lower parts of various aspects of sub-ontologies on human dataset including IEA annotations.

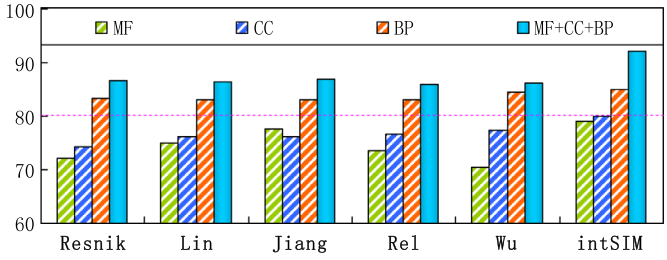


Fig. 4. Comparison of ACC scores of different features derived by different methods from the upper part of various aspects of sub-ontologies on human dataset excluding IEA annotations.

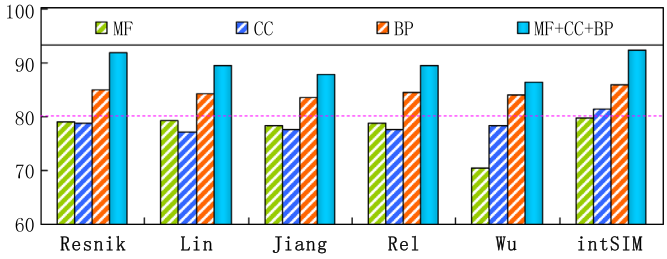


Fig. 5. Comparison of ACC scores of different features derived by different methods from both the upper and the lower parts of various aspects of sub-ontologies on human dataset excluding IEA annotations.

information for different sub-ontologies including and excluding IEA annotations, respectively. From these figures, we see that the features that combine information from MF, CC and BP sub-ontologies consistently perform better than those only taking advantage of information from any individual ontology. This reflects the fact that the three sub-onlotogies can characterize the relationship of GO terms from different aspects, which in turn helps to achieve higher prediction accuracy scores in regard to protein–protein interaction. From another point of view, the features that combine five kinds of similarity measures (denotes as intSIM) all perform somewhat better than each of the individual feature in the four scenarios, no matter if the similarity measures from MF, CC and BP sub-ontologies are integrated with one another.

Furthermore, we observe that the feature that integrates different kinds of similarity scores derived from both the ascending and the descending parts of the three sub-ontologies performs best on both human and yeast datasets. It produced the highest ACC values of 93.39% and 92.34% on the human dataset, and also the highest ACC values of 92.16% and 90.27% on yeast dataset including and excluding IEA annotations, respectively. This suggests that the more information we derive from the GO DAG, the better prediction power of the classifier, which has the following three kinds of information as its input: (1) the information derived by different similarity measures; (2) the information derived from both the ascending and the descending parts of the GO graph; and (3) the information from the three sub-ontologies.

8. Conclusion

In this study, we proposed a new approach to predict protein–protein interaction based on the semantic similarity of GO terms. The similarity values from the upper part of MF, CC and BP sub-ontologies were derived by five existing methods, and those from the lower part of GO graph were also quantified by five extended methods. After the term-based similarity measures were used to measure the semantic similarities of protein pairs, different kinds of protein similarity scores were further assembled into feature vectors to characterize protein pairs and serve as input of machine learning algorithm. SVM was adopted as classifier and five-fold cross validation experiments were conducted on both human and yeast datasets. The experimental results show that the combination of different kinds of similarity scores can generally perform better, in particularly, the feature that takes into account information derived from the ascending and the descending parts of the GO structure in the three sub-ontologies by different methods (five existing methods and their corresponding extensions), produced the highest prediction accuracy rates in all datasets. Consequently, our method can derive more information to characterize the relationship between protein pairs, which in turn performs better for protein–protein interaction inference.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Acknowledgments

The authors thank Dr. Wu Xiaomei for her sharing the PPI datasets. This research was financially supported the Sciences Foundation of Guangzhou Maritime Institute (K31012B09), the Foundation for innovation project in Higher Education of Guangdong, China (A510602 and B510620); the Science and Technology Research and Development Program of Guangzhou (201510010238); and Sciences and Technology Project of Guangdong Province Transportation Hall, China (2012-02-045).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.04.020.

References

Aloy, P., Böttcher, B., Ceulemans, H., et al., 2004. Structure-based assembly of protein complexes in yeast. Science 303 (5666), 2026–2029.

Azuaje, F., Wang, H., Bodenreider, O., 2005. Ontology-driven similarity approaches to supporting gene functional assessment. In: Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies, pp. 9–10.

Benabderrahmane, S., Smail-Tabbone, M., Poch, O., et al., 2010. IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinform. 11 (1), 588.

Ben-Hur, A., Noble, W.S., 2005. Kernel methods for predicting protein–protein interactions. Bioinformatics 21 (suppl 1), i38–i46.

Chin, C.H., Chen, S.H., Ho, C.W., et al., 2010. A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles. BMC Bioinform. 11 (Suppl 1), S25.

Gavin, A.C., Bösche, M., Krause, R., et al., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415 (6868), 141–147.

Gavin, A.C., Aloy, P., Grandi, P., et al., 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440 (7084), 631–636.

Gene Ontology Consortium, 2015. Gene ontology consortium: going forward. Nucleic Acids Res. 43 (D1), D1049–D1056.

Guo, X., Liu, R., Shriver, C.D., et al., 2006. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22 (8), 967–973.

He, M., Wang, Y., Li, W., 2009. PPI finder: a mining tool for human protein-protein interactions. PLoS One 4 (2), e4554.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13 (2), 415–425.

Huang, Q., You, Z., Zhang, X., et al., 2015. Prediction of protein–protein interactions with clustered amino acids and weighted sparse representation. Int. J. Mol. Sci. 16 (5), 10855–10869.

Ito, T., Chiba, T., Ozawa, R., et al., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. 98 (8), 4569–4574.

Jain, S., Bader, G.D., 2010. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinform. 11 (1), 562.

Jansen, R., Yu, H., Greenbaum, D., et al., 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302 (5644), 449–453.

Jiang, J.J., Conrath, D.W., 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics, pp. 19–33.

Lin, N., Wu, B., Jansen, R., et al., 2004. Information assessment on predicting protein-protein interactions. BMC Bioinform. 5 (1), 154.

Lin, D., 1998. An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on Machine Learning, pp. 296–304.

Liu, Y., Liu, N., Zhao, H., 2005. Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. Bioinformatics 21 (15), 3279–3285.

Maetschke, S.R., Simonsen, M., Davis, M.J., et al., 2012. Gene Ontology-driven inference of protein–protein interactions using inducers. Bioinformatics 28 (1), 69–75.

Martin, D., Brun, C., Remy, E., et al., 2004. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. 5 (12), R101.

Martin, S., Mao, Z., Chan, L.S., et al., 2007. Inferring protein-protein interaction networks from protein complex data. Int. J. Bioinform. Res. Appl. 3 (4), 480–492.

Mazandu, G.K., Mulder, N.J., 2012. A topology-based metric for measuring term similarity in the gene ontology. Adv. Bioinform. 2012 (975783), 1–17.

Mei, S., Zhu, H., 2014. AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. PLoS One 9, e110488.

Miller, J.P., Lo, R.S., Ben-Hur, A., et al., 2005. Large-scale identification of yeast integral membrane protein interactions. Proc. Natl. Acad. Sci. 102 (34), 12123–12128.

Mukhopadhyay, A., De, M., Maulik, U., 2011. Selection of GO-Based semantic similarity measures through AMDE for predicting protein-protein interactions, Swarm, Evolutionary, and Memetic Computing. Springer, Berlin Heidelberg, pp. 55–62.

Nagar, A., Al-Mubaid, H., 2008. A new path length measure based on go for gene similarity with evaluation using sgd pathways. In: Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems, CBMS'08, pp. 590–595.

Patil, A., Nakamura, H., 2005. Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinform. 6 (1), 100.

Qi, Y., Noble, W.S., 2011. Protein interaction networks: protein domain interaction and protein function prediction, Handbook of Statistical Bioinformatics. Springer, Berlin Heidelberg, pp. 427–459.

Qiu, J., Noble, W.S., 2008. Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput. Biol. 4 (4), e1000054.

Rada, R., Mili, H., Bicknell, E., et al., 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst., Man Cybern. 19 (1), 17–30.

Ramani, A.K., Marcotte, E.M., 2003. Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol. 327 (1), 273–284.

Resnik, P., 1995. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453.

Rhodes, D.R., Tomlins, S.A., Varambally, S., et al., 2005. Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23 (8), 951–959.

Schlicker, A., Domingues, F.S., Rahnenführer, J., et al., 2006. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform. 7 (1), 302.

Stelzl, U., Worm, U., Lalowski, M., et al., 2005. A human protein-protein interaction network: a resource for annotating the proteome. Cell 122 (2), 957–968.

Tsoka, S., Ouzounis, C.A., 2000. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat. Genet. 26 (2), 141–142.

Von Mering, C., Krause, R., Snel, B., et al., 2002. Comparative assessment of large-scale data sets of protein–protein interactions. Nature 417 (6887), 399–403.

Wu, X., Zhu, L., Guo, J., et al., 2006. Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations. Nucleic Acids Res. 34 (7), 2137–2150.

Wu, X., Pang, E., Lin, K., 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.

You, Z.H., Chan, K.C.C., Hu, P., 2015. Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. PLoS One 10 (5), e0125811.

Zhang, L.V., Wong, S.L., King, O.D., et al., 2004. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinform. 5 (1), 38.

Zhang, Y.N., Pan, X.Y., Huang, Y., et al., 2011. Adaptive compressive learning for prediction of protein–protein interactions from primary sequence. J. Theor. Biol. 283 (1), 44–52.