A New Path Based Hybrid Measure for Gene Ontology Similarity

Sanghamitra Bandyopadhyay and Koushik Mallick

Abstract—Gene Ontology (GO) consists of a controlled vocabulary of terms, annotating a gene or gene product, structured in a directed acyclic graph. In the graph, semantic relations connect the terms, that represent the knowledge of functional description and cellular component information of gene products. GO similarity gives us a numerical representation of biological relationship between a gene set, which can be used to infer various biological facts such as protein interaction, structural similarity, gene clustering, etc. Here we introduce a new shortest path based hybrid measure of ontological similarity between two terms which combines both structure of the GO graph and information content of the terms. Here the similarity between two terms t_1 and t_2, referred to as $GOSim_{PBHM}(t_1, t_2)$, has two components; one obtained from the common ancestors of t_1 and t_2. The other from their remaining ancestors. The proposed path based hybrid measure does not suffer from the well-known shallow annotation problem. Its superiority with respect to some other popular measures is established for protein protein interaction prediction, correlation with gene expression and functional classification of genes in a biological pathway. Finally, the proposed measure is utilized to compute the average GO similarity score among the genes that are experimentally validated targets of some microRNAs. Results demonstrate that the targets of a given miRNA have a high degree of similarity in the biological process category of GO.

Index Terms—Gene ontology similarity, semantic similarity, term similarity, information content, protein interaction prediction, functional classification of genes, microRNA

1 Introduction

Gene ontology (GO) is a directed acyclic graph (DAG) of terms that describes semantic relationships in biological functions or cellular locations of a cell. Here nodes represent some biological semantic concepts and edges represent semantic relations. Semantic concepts are composed of three non overlapping ontology categories, namely, cellular component (CC), biological process (BP) and molecular function (MF). Cellular components represent the locations of a gene product within the cell, at the levels of subcellular-structures and macromolecular complexes. Biological Processes are ordered chains of multiple molecular functions or reactions in which a gene product participates (e.g., signal transduction, glucose metabolism, etc.). Molecular functions describe the chemical activity such as catalytic or binding activities of a gene product at the molecular level. However it does not provide details of the context of the pathway, complex, etc. of a molecular function. GO is a hierarchical, tree like structure but has redundant paths. At the top level the nodes represent generic concepts while the leaf nodes represent the most specific concepts. A concept, also called a term, may have multiple ancestors. There are relations among the terms by which we can measure the semantic distance or similarity with respect to a biological activity. A detailed description about GO can be found in [1] and [2].

Discovering functionally related genes and proteins is critical as interacting proteins often participate in the same BP and/or are co-localized in the same CC and/or exhibit the same MF. GO similarity value has been used to predict protein-protein and domain-domain interactions [3], [4], [5]. Sequence and GO similarities are generally correlated [6], [7]. Moreover it is found that structural and domain similarity are, in general, also correlated with GO similarity. Relation between disease and genes can be measured with semantic relations [8]. Hence measuring the GO similarity as accurately as possible is an important task.

To calculate semantic similarity of two gene products, term pairs are generated from directly annotated GO term sets of the two. Then similarity value of each term pair is calculated and a combination scheme is used to compute the similarity between the gene pair. An alternative way of measuring similarity is to use all annotated GO terms of two genes products to create vectors and then to compute similarity among vectors. These are called group wise measures [9]. Different strategies exist for calculating semantic similarity of a GO term pair. Some popular measures are outlined here: (i) Node and corpus annotation statistics based measures, e.g., those by Resnik [10], Lin [11], Jiang and Conrath [12], Schlicker et al. (Relevance) [13], Li et al. (simIC) [14], etc. (ii) Measures dependent on the depth of terms and path length, e.g., Nagar and Hisham [15], Shaohua Zhang et al. [16]. and (iii) Measures that hybridize the path length with all common ancestor terms, e.g., Wang et al. [17]. There are also some measures which mix the annotation statistics with the path length, e.g., intelliGO [18], Shen et al. [19]. Among the group wise measures, the number of all common

- *S. Bandyopadhyay is with Indian Statistical Institute, Machine Intelligence Unit, 203, B.T. Road, Kolkata 700108, West Bengal, India. E-mail: sanghami@isical.ac.in.*
- *K. Mallick is with the CSE Department, RCC Institute of Information Technology, Canal South Road, Beliaghata, Kolkata 700015, West Bengal, India. E-mail: koushikbuie@gmail.com.*

Manuscript received 7 Feb. 2013; revised 27 Sept. 2013; accepted 4 Nov. 2013; date of publication 19 Nov. 2013; date of current version 7 May 2014.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2013.149

ancestor terms of two genes are used, e.g., SimUI [20]. and the measures based on annotation statistics of all common ancestors, e.g., simGIC [21] are very popular. Details about the different similarity measures functions are discussed [2] and [9]. Most of these measures have certain limitations, as discussed in detail in Section 2. To overcome these limitations, in this paper we have proposed a shortest path based hybrid measure (SPBHM) for computing the similarity between two GO terms. It combines both structural and semantic information embedded in the GO graph. Effectiveness of the proposed measure is established through several studies like PPI prediction, correlation analysis and miRNA functional similarity.

2 Related Work

There are many approaches to measure similarity between a GO term pair. An edge based method was introduced by Rada et al. [23], that was based on the simple concept of number of edges traversed to go from one node to another. Although this works well on lexical analysis in medical terms, Resnik [10] has shown that edges in the GO graph do not represent the same uniform distance. This is because the terms at the same level do not have the same specificity for a GO graph. Some terms have many children while some have a few, some GO branches are long and represent deep concepts while some are short representing shallow concepts. Depth of a term does not represent the specificity of a term. Terms at higher level may be more specific than another at a lower level [7]. So an information content (IC) based measure is more effective in this area. Information content of a term is defined as the negative logarithm of the probability of occurrence of the term [10]. It is defined as $IC(t) = -\log(p(t))$, where $p(t)$ is the probability of occurrence of a term t, and is defined as

$$p(t) = \frac{annotation(t) + \sum_{d \in descendent(t)} (annotation(d))}{\sum_{c \in descendent(root)} annotation(c)}. \quad (1)$$

Here $annotation(t)$ is the number of genes annotated by term t, and $descendent(t)$ is the set of all terms that are descendants of t in the GO graph. Note that terms that are more generic (i.e., near the root of the GO) will have higher $p(t)$ values and hence smaller IC values. On the other hand, more specific terms will have smaller $p(t)$ and hence larger IC values. Common well known IC value based measures are proposed by Resnik [10], Lin [11], Jiang [12] and the Relevance measure proposed by Schlicker et al. [13].

According to the traditional IC based measure, the specificity of a term is fully dependent on the number of genes taken in the annotation corpus but it ignores the edge density and graph topology information in the different portions of the GO graph. A new approach for the definition of term specificity is introduced in the field of semantic similarity measure [24]. The new definition of IC value of a term is proportional to the number of descendants and parents and the depth of the term under consideration. In this paper, we are considering measures based on the traditional annotation based IC only.

Resnik defines the similarity between terms t_1 and t_2 as the IC value of the lowest common ancestor term (LCA). For a given pair of terms t_1 and t_2, the LCA is that common ancestor term of these terms which is at the maximum distance from the root of the GO. This LCA is sometimes also called the Most Informative Common Ancestor (MICA) [25]. Therefore, $Sim_{Resnik}(t_1, t_2) = IC(LCA(t_1, t_2))$. According to Resnik, terms are considered to be more similar if the IC value of their LCA is higher. The serious drawback of this measure is that when two different term pairs located in different levels of the GO graph have the same LCA, then they have the same similarity value; this is misleading. In order to overcome this Lin invented another measure defined as follows:

$$Sim_{Lin}(t_1, t_2) = \frac{2 * IC(LCA(t_1, t_2))}{(IC(t_1) + IC(t_2))}. \quad (2)$$

The problem of Lin's measure is that if $IC(t_1) \approx IC(t_2) \approx IC(LCA(t_1, t_2))$, then $Sim_{Lin}(t_1, t_2) \approx 1$ irrespective of whether t_1 and t_2 are very general, or very specific terms. Note that the similarity should be high only in the latter case. This problem is known as shallow annotations of genes [14]. Jiang's [12] measure is similar to Lin's measure, and has the same limitation. Relevance similarity measure proposed by Schlicker et al. [13] has removed these problems by combining both Lin and Resnik similarity. It is defined as follows:

$$Sim_{Rel}(t_1, t_2) = \frac{2 * IC(t)}{(IC(t_1) + IC(t_2))} * (1 - p(t)), \quad (3)$$

where t is the LCA of t_1 and t_2. Here the effect of shallow annotation is reduced by cofactor $(1 - p(t))$. When a term pair is less specific, then their LCA term has higher probability value, which produces a smaller value of the cofactor and vice versa. Now consider the case when the probabilities of LCA terms of two term pairs are either close to 0 or 1. Then the similarity adjustment factor $(1 - p(t))$ has a very small effect according to the Relevance measure, and the two term pairs show almost similar values. This is discussed in detail in Li et al. [14]. Li et al. proposed another measure named simIC [14], with a new cofactor to remove the above problem of the Relevance measure. However, this new measure is able to distinguish well two pairs both of which are highly similar, it fails when the two pairs have lower similar values.

The above mentioned node based measures are sometimes inadequate to describe the distance or similarity of a pair of terms [25]. So edge based measures have become popular. Wang et al. [17] proposed a measure where the semantic value (SV) of a term t_1 is computed as a function of the SVs of t_1 and all its ancestors. They consider $SV(t_1) = 1$, and $SV_{t_1}(t) = max(w_e * SV_{t_1}(t^c)) | t^c \in children(t)$, for all ancestors t of t_1. Here w_e ($w_e < 1$) is the semantic contribution factor of edge e which links the terms t and t^c. They consider that terms further from the annotating term have less semantic contribution as they are more general. The semantic value of a term t_1 is then calculated as $SV(t_1) = \sum_{t \in ancestor(t_1)} SV_{t_1}(t)$. Finally they measured semantic similarity between terms t_1 and t_2 as

$$Sim_{Wang}(t_1, t_2) = \frac{\sum_{t \in ancestor(t_1) \cap ancestor(t_2)} SV_{t_1}(t) + SV_{t_2}(t)}{SV(t_1) + SV(t_2)}. \quad (4)$$

Wang's measure also suffers from the shallow annotation problem. If two terms are the same then they always have similarity value equal to one, regardless of whether they are general or specific terms. Nagar and Hisham [15] have proposed another path length based measure. They define similarity function as $Sim_{Nagar}(t_1, t_2) = e^{-f*pathlength(t_1,t_2)}$. Here $pathlength(t_1, t_2)$ is the length of the shortest path to go from t_1 to t_2 in the GO graph and f is a tuning factor. The function is based on an exponential transfer function that produces a similarity value proportional to the path length between a term pair. Another variant is proposed by Shaohua Zhang et al. [16], where they refine Nagar's model by also considering the depth of the LCA with respect to the root as a contributing factor to the exponential transfer function. One serious drawback of the edge based measures is that, either they consider weight of all the edges as 1 or they consider edges on the same level as having equal weight. As the GO terms on the same level may have different specificities, so edges should also have different weights. Wang, Nagar and Zhang have validated their measures through the functional classification of genes in yeast biological pathway. Shen et al. [19] have proposed a hybrid measure which combines both the structure and node based information. The intuition is that if LCA of a term pair is located at a higher level of a GO graph then their dissimilarity should be larger and vice versa. For this, they assign a weight to each GO term as the reciprocal of its IC value. So the weights will be more for less specific terms appearing at a higher level. The measure finds the path connecting the two terms and their LCA with the smallest sum of weights. This value contributes to the semantic distance between the terms. So if the LCA term is near the root of GO graph then the path will have higher dissimilarity value and vice versa. Note that, here even if two different term pairs have the same LCA, they have different contributions to the dissimilarity value. Finally they normalize the sum of weights by an arctan function, as follows:

$$dist(t_1, t_2) = \frac{arctan\left(\sum_{t_1 \in path_1} \frac{1}{IC(t_1)} + \sum_{t_2 \in path_2} \frac{1}{IC(t_2)}\right)}{\frac{\pi}{2}}, \quad (5)$$

here $path_1$ and $path_2$ are the shortest paths that connect the terms t_1 and t_2, respectively, with the LCA. As LCA appears twice according to the formulae so they did not consider it in the second path of Eq. (5). This distance is converted to similarity value by subtracting it from 1. For validating this measure, the DIP yeast protein interaction data and gene expression data are considered. However, a problem with this measure is that the contributions of common ancestor terms are not considered. It has the same problem as noted for Wang's and other path length based measures, namely, if two terms are the same, then their dissimilarity value is always 0 (similarity is always 1), irrespective of the specificity of the term pair. They have considered the LCA term to contribute to the dissimilar semantic value. This appears to be counter intuitive as the LCA term is common for the two terms under consideration. Note that the weight $\frac{1}{IC(t_1)}$ will produce a very high value for the terms with IC values less than 1. This will produce very high dissimilarity value for all gene pairs which have their LCA near the root regardless of the position of the genes' annotated terms. Another limitation is that although the weighted approach is considered, the contributions of all the terms in the shortest path are the same. Keeping these in mind, we have proposed a shortest path based measure where we consider the contribution of both the similar and dissimilar parts of the GO sub graph corresponding to a given term pair. In our approach we have calculated the individual weights of terms using an exponential transfer function and consider a coefficient which is decreasing proportionally as the path length increases. More details are provided in Section 3.

Validation of the similarity measures is an important issue. Some common approaches are correlation of ontological similarity with sequence similarity [13], [7], with gene expression [26], with Enzyme Commission (EC) number and validation with PPI data [27]. Some authors have also used functional classification of genes in a biological pathway [17], [15] for this purpose. Rather than using a single approach for validating the proposed measure, we establish its effectiveness using (i) PPI prediction efficiency for yeast and human, (ii) correlation with gene expression data of yeast and human, (iii) functional classification of genes in a biological pathway for yeast, (iv) quantifying the GO similarity among the target gene set of a human miRNA and (v) performance validation with the online CESSM [28] tool. Detailed analysis of the validation schemes are provided in Section 4.

3 Proposed Measure

3.1 Measuring Similarity between Two GO Terms

In the previous section, several existing measures have been described, and their limitations have been discussed. In order to overcome these limitations, a new measure called SPBHM has been proposed in this article. Here, we describe this measure in detail, where a weighted shortest path based approach has been adopted. SPBHM combines information content values of GO terms along with their graph structure information. To measure semantic similarity of two GO terms we have considered three shortest paths: one from LCA term to root term, and the other two from the annotated terms to the LCA, but excluding the LCA. While the former path contributes to the similarity component, the latter path contributes to the dissimilarity component. A natural intuition from the GO graph structure is that if the common ancestor terms are more specific, then the GO similarity should be higher than the case when the terms are more general. The converse occurs for the remaining or uncommon terms. That is, if they are more specific in the GO DAG, then the dissimilarity should be lower than if the terms are more general. From this intuition, we have defined the dissimilar component for a GO term pair t_1 and t_2 as

$$\begin{aligned} dis(t_1, t_2) = &\sum_{t_i \in path_{t_1}} \left(1 - e^{-\frac{1}{IC(t_i)}} * W^1_{t_i}\right) \\ &+ \sum_{t_j \in path_{t_2}} \left(1 - e^{-\frac{1}{IC(t_j)}} * W^2_{t_j}\right). \end{aligned} \quad (6)$$

Here $path_{t_1}$ and $path_{t_2}$ are the shortest paths from the terms t_1 and t_2 respectively up to the LCA but excluding the LCA term. Every node i in the subgraph is weighted by the value $(1 - e^{-\frac{1}{IC(t_i)}})$. We refer to the shortest path as the

minimum weight path connecting the annotating term to the immediate descendant of the LCA term. The value of $(1 - e^{-\frac{1}{IC(t_i)}})$ will increase exponentially as the IC value decreases. We define a coefficient $W^1_{t_i}$ for $path_{t_1}$ as $W^1_{t_i} = e^{-\frac{l_i}{L_{t_1}}}$, where L_{t_1} is the total number of nodes on the shortest path from the term t_1 to the immediate descendant of LCA term. The number of nodes starting from the term t_1 to the i_{th} term on the shortest path from t_1 is denoted by l_i. $W^2_{t_j}$ is defined analogously. Note that this coefficient and hence the semantic contribution decrease exponentially as we move up to the LCA term. As is evident from the formulation of $W^1_{t_i}$ and $W^2_{t_j}$, these start with a value of $e^{-\frac{1}{L}}$, and terminate with a value e^{-1}. Therefore if L is large, the weighting coefficients will start at higher values.

The similarity component of SPBHM is written as

$$sim(t_1, t_2) = 2 \times \sum_{t'_i \in path} e^{-\frac{1}{IC(t'_i)}} * W'_{t_i}. \quad (7)$$

Here we consider the shortest path, $path$, as the minimum weighted path connecting the LCA term to the root term. Note that this is the same for both t_1 and t_2 and hence we have multiplication by two in Eq. (7). For the similarity component we consider $e^{-\frac{1}{IC(t'_i)}}$ as the contribution of each node. As stated above, $e^{-\frac{1}{IC(t'_i)}}$ will produce higher value when IC values are more specific. Weight W'_{t_i} is defined as $e^{-\frac{l'_i}{L_{t_{LCA}}}}$, where $L_{t_{LCA}}$ is the total number of nodes on the shortest path from the LCA term to the root term. l'_i is the number of nodes starting from the LCA term to the i_{th} term on the aforementioned shortest path. To normalize $sim(t_1, t_2)$ we use the $\arctan$ function:

$$\frac{\arctan(sim(t_1, t_2))}{\frac{\pi}{2}}. \quad (8)$$

Similarly we normalize the dissimilarity value of Eq. (6) and convert it to a similarity value as follows:

$$1 - \frac{\arctan(dis(t_1, t_2))}{\frac{\pi}{2}}. \quad (9)$$

Finally we get the overall similarity value between t_1 and t_2 by averaging the above Eq. (8) and (9) as bellow:

$$GOsim_{SPBHM}(t_1, t_2) = \left(\frac{\arctan(sim(t_1, t_2))}{\frac{\pi}{2}} + \left(1 - \frac{\arctan(dis(t_1, t_2))}{\frac{\pi}{2}}\right)\right) \Big/ 2. \quad (10)$$

As mentioned above, weights of the shortest paths are used for measuring the similarity of two GO terms. The time complexity of the proposed measure is proportional to that of finding the shortest paths in the ancestor sub-graph of GO terms. Dijkstra's algorithm is used for this purpose, which has time complexity $\mathcal{O}(\log N * E)$. Where N and E is the number of nodes and number of edges respectively in the subgraph of the annotated GO terms. As the number of edges and nodes in a subgraph of a GO term is small compared to the full GO graph, so the required time reasonable. For the purpose of efficient computation, we have precomputed the shortest paths from each annotated GO term to all its ancestors. This avoids repeated calculation of shortest paths for annotated terms during the similarity computation of a long list of gene pairs. Note that the $\arctan$ function has the property of higher resolution for lower values of its input. As the input to the function is the summation of weights of the nodes along the shortest paths, so it shows a better correlation and resolution than many other measures. This is also evident from Fig. 7 in Section 5.6, where the distribution of three different measures are plotted.

3.2 Measuring the Similarity between Genes or Gene Products

Let us consider gene products A and B having the following annotations $anot(A) = t^A_1, t^A_2, t^A_3 \ldots t^A_m$ and $anot(B) = t^B_1, t^B_2, t^B_3 \ldots t^B_n$ with m and n number of GO terms respectively. To compute the semantic similarity between A and B, the proposed measure or any other existing measure is used to compute the similarity values between all the pairs of terms annotating the two gene products. This provides a matrix of dimension $m \times n$, denoted by $SimMat$. Some common methods for aggregating the term pairwise similarities into an overall similarity value between two gene products are discussed here.

Maximum similarity approach $MAX_{\forall(i,j)}(SimMat_{ij})$ considers the maximum value of the matrix as the similarity between A and B. This measure has a problem that it disregards the similarity value of the other term pairs for a multi-functional protein.

In the average (AVG) approach the arithmetic average of similarity values of the all term pairs is taken as follows: $\frac{1}{m*n}\sum_{\forall(i,j)} SimMat_{ij}$. Because of averaging property, this measure sometimes underestimates the true similarity.

Best match average (BMA) is another effective combination method used in [13], which considers all the maximum similarity values of each term of A to all the terms of B and vice versa. Let $Rowscore = \frac{1}{m}\sum_{i=1}^{m} max_{1<=j<=n}(SimMat_{i,j})$ be the average of the maximum similarity values of each term in $anot(A)$ to $anot(B)$. Similarly let $Colscore = \frac{1}{n}\sum_{j=1}^{n} max_{1<=i<=m}(SimMat_{i,j})$ be the average of maximum similarities of each term in $anot(B)$ to $anot(A)$. Now $avg(Rowscore, Colscore)$ is used to obtain the BMA score.

4 Datasets and Validation Approaches of Proposed Measure

4.1 Gene Ontology Data

We have collected annotation data for the species *Saccharomyces Cerevisiae* and *Homo Sapiens* from the Uniprot database dated February 2013. We collected GO graph data from www.geneontology.org. We have considered all relations of GO graph viz., *is_a*, *part_of*, *regulates* etc. In a part of the study, we have also simulated the results on GOSemSim (version 1.12.1) [29] package, an R package from the http://www.bioconductor.org. The results are provided in Table 2 and Table 3 in the supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.149.

4.2 Protein-Protein Interaction Data of Yeast and Human

We collected three yeast protein-protein interaction (PPI) data sets from different databases [30], [31], [32]. We used around 4,837 positive interactions for yeast proteins from DIP (Database of Interacting Proteins) [33] and MIPS (Mammalian Protein-Protein Interaction Database) [34] and the same number of negative instances from Ben-Hur and Noble [30]. The data set in Jain and Davis [31] has around 4,100 PPI data for yeast from the core set of DIP yeast data set (dated 2009). They have generated negative data set of same size as the positive one by selecting random protein pairs which do not occur in iRefWeb [35] (September 2010). The third data set used for analysis is by Yu et al. [32] from where we have taken the high confidence yeast protein interaction data set of 5,621 non-redundant interactions. Yu et al. prepared a negative database for yeast which consists of random protein pairs, filtering out those available in the bioGRID database entries.

Data sets for human were collected from Jain and Davis [31] and Yu et al. [32]. Data set of Jain and Davis [31] has around 1,500 unique human protein interactions while Yu et al. [32] has around 15,000 high confidence interactions from HPRD. The same number of random negative data sets were used. For drawing the ROC plots, the threshold of the similarity values are varied between (0,1). The protein pairs with similarity values greater than the threshold are predicted to be positives, while those below the threshold are predicted to be negatives. Thereafter the true positive and true negative, and false positive and false negative values are computed, and ROC curves are plotted. The area under the curve (AUC) obtained from the ROC plots is used to compare the performance of the proposed measure vis-a-vis the other GO similarity measures. Results are discussed in Section 5.1. We have noticed that presence of same protein pairs in different data sets used in this work are very low (less than 5 percent).

4.3 Gene Expression Data for Yeast and Human

Correlation between gene expression and GO similarity value among the gene products are very important as many gene products that participate in the same biological process or are functionally related, have similar expression profiles [26], [36]. We have downloaded different gene expression data sets for yeast and human. For yeast, Eisen data set [37] which consists of 2,467 genes with 79 experiments, and the Spellman data set [38] containing 6,178 genes with 77 experiments are used. We have collected several gene expression data sets, namely GSE10073, GSE8506 for yeast and GSE20247, GSE9574, GSE5788, GSE13466, GSE20437 for human, from Gene Expression Omnibus of NCBI [39] (http://www.ncbi.nlm.nih.gov/). We measure the Pearson correlation coefficients of gene expression among the gene pairs and also calculated the corresponding GO similarities using SPBHM as well as the measures of Relevance, Resnik, Wang, Jiang, Lin, simIC, Shen, Zhang, simUI and simGIC. All the similarity values lie between [0,1]. Each similarity measure is divided into 100 intervals, and the gene pairs belonging to the same interval are grouped together. Thereafter the average similarity and the average Pearson correlation coefficient are computed for each interval resulting in two vectors $GO^i_{k,avg}$ and $PC^i_{k,avg}$, where $i = 1, 2, \ldots 100$ and $k \in$ [*SPBHM, Relevance, Resnik, Wang, Jiang, Lin, simIC, shen, zhang, TCSS*]. Finally the Pearson correlation coefficient between these two vectors is computed for each similarity measure. This is referred to as the average Pearson correlation of the measure [26]. In general, higher this value, better is the measure.

4.4 Biological Pathways of Yeast

Participating genes in a biological pathway are involved in different molecular functions and some of them are assigned different Enzyme Commission numbers. These EC numbers are assigned to genes according to the subtype of reaction that they catalyze at the molecular level. Classifying the genes according to the molecular function is an important validation for a GO similarity measure. For this purpose we have taken a few pathways from yeast pathway database [40] (http://pathway.yeastgenome.org/). Results are demonstrated for the *Glycolysis* and *tryptophan degradation* pathways. For computing the pairwise gene similarities, we have used the proposed SPBHM and Relevance measures over the MF ontology. We used MAX as the combination approach here for both the similarity measures. The genes are clustered according to each measure using complete linkage hierarchical clustering.

4.5 Human miRNA Target Gene Sets

Yu et al. [22] have calculated the microRNA functional similarity through the analysis of GO similarity of targeted genes. Each miRNA has a certain GO profile [41]. Inspired by Yu et al. [22] we hypothesize that a microRNA often targets similar type of genes with respect to GO. In order to verify this we compute that GO similarity observed among a set of target genes of a miRNA. We have taken experimentally validated human miRNA target genes from miRTarBase [42]. We included those miRNAs having at least two genes in their target set, and where at least 60 percent of genes have significantly enriched GO BP terms (p-value $<$ .001). After this preprocessing step, our data set covers around 70 percent of the total miRNAs in miRTarBase data set. The average GO similarity between the genes that are targets of a particular miRNA is computed using the different measures.

4.6 CESSM Data Set

CESSM [28] is an online tool for performance comparison of a measure. GO graph and annotation information of proteins are included. In this data set there are 13,430 protein pairs collected from different species. GO similarity value of those protein pairs are computed using a given measure. Then the correlation values between the GO similarity and Enzyme Commission number, Pfam domain and sequence similarity are calculated. Resolution of a measure is also calculated by the tool. Higher resolution and correlation value supports the efficiency of a measure. Details are available in [21]. We provided the results obtained for the measures those are now currently discussed in this paper for all three aspects of GO.

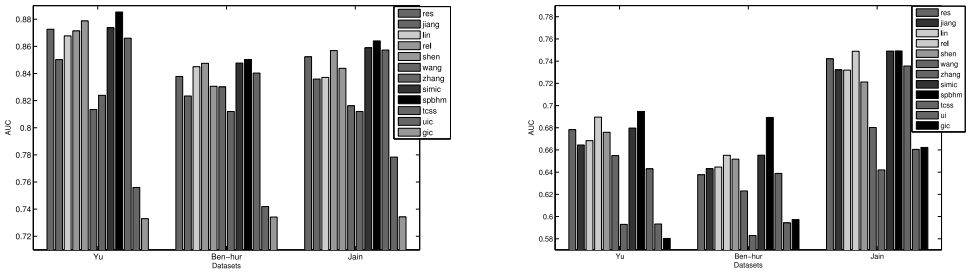


Fig. 1. Barplot of AUCs of different measures on yeast PPI data sets in (a) BP and (b) MF ontology.

5 Results and Discussion

5.1 Protein Protein Interaction of Yeast and Human

First we have investigated the performance of the different GO based similarity measures on three yeast and two human PPI data sets as described in Section 4.2. The GO similarity values of all the protein pairs in the PPI data (both positive and negative) are computed using all the measures, viz, Resnik [10], Jiang and Conrath [12], Lin [11], Relevance (Schlicker) [13], Wang et al. [17], TCSS [31], simUI [21], simGIC [21], Shen et al. [19], Shaohua Zhang [16] and the proposed SPBHM measure. The ROC curve is plotted and the area under the curve is measured. Table 1 in the supplementary material, available online, shows the AUC values corresponding to the different measures in terms of BP, CC and MF ontologies on the three yeast PPI data sets. Here Figs. 1a, 1b and 2 show the summary of results for BP, MF and CC respectively. As can be seen from the figures, the proposed measure performs well for all the three data sets providing the best values in eight out of nine cases. In only one case, the simIC measure performed best, closely followed by the proposed SPBHM. Performances of Relevance measure is also very close to simIC closely followed by the Resnik measure. Many authors have found that Resnik measure provides the best performance [31], [43] for yeast PPI data. However our results show that simIC and Relevance measure are equally good, if not better. On the other hand Jiang, Lin, Wang, simUI, simGIC and Zhang measures performed very poorly for this data sets. Hence these latter measures were not taken into consideration in case of the Human PPI data sets. As mentioned in Section 4.2 we have used two data sets for human. AUC values obtained for the different measures are plotted in Fig. 3, and the detailed results are mentioned in Table 2 of the supplementary file, available online. From Fig. 3 we see that in five out of six cases, SPBHM outperforms Resnik, Relevance, simIC, Shen and TCSS measures. Interestingly, the proposed measure outperforms Relevance, simIC and Resnik with p-values 0.004, 0.01 and 0.002 respectively using wilcoxon signed rank test indicating its potential superiority. Note that irrespective of the measure used, the AUC values are in general larger for BP, showing that it is perhaps the most discriminative among the three categories of gene ontology. Best values are bold faced in the table of supplementary material, available online. In summary, the proposed hybrid measure (SPBHM) and node based measures (simIC, Relevence, Resnik), in general, performed better than the graph structure (Wang, TCSS, Shen), vector (simUI, simGIC) and simple path (Zhang) based measures.

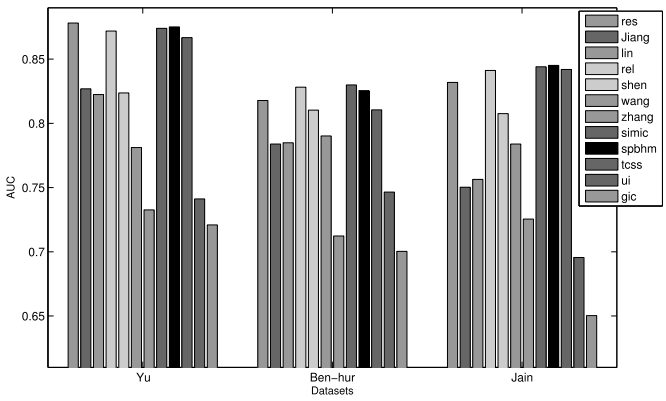
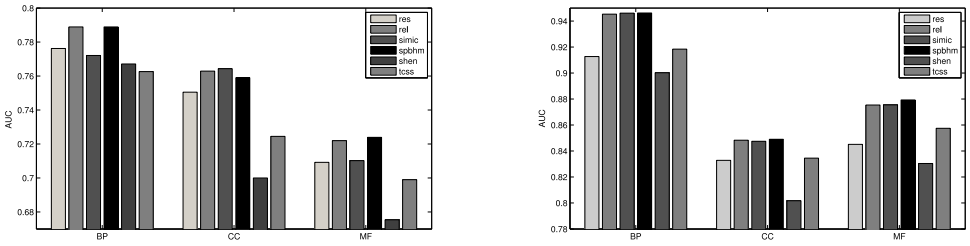


Fig. 2. Barplot of AUCs of different measures on yeast PPI data sets in CC.

5.2 Correlation of Gene Expression with GO Similarity Value

To analyze the correlation with gene expression we used different data sets mentioned in Section 4.3. We have taken around 126,000 gene pairs for yeast data to calculate the correlation. We used BMA approach to aggregate the GO term similarity for all the measures. Detailed results are provided in supplementary material (Tables 5, 6 and 7 for BP, CC and MF respectively), available online. A bar plot in Fig. 4 shows average of the correlation value obtained from the four data set for each ontology. Proposed measure achieved the highest correlation (computed using the approach described in Section 4.3) of 0.81 for BP ontology on Eisen's data set followed by Relevance, Shen, simIC and Lin measures respectively. For other data sets also, SPBHM is found to yield the best performance. For CC similarity, the proposed measure achieves the best performance for Spellman's, Eisen and GSE10073 data sets. On the other data set we find that Resnik measure performed the best, while SPBHM also performed reasonably well. In case of MF similarity, proposed measure achieves the highest correlation for Eisen, Spellman's and GSE10073 data sets (Table 7 of supplementary, available online), while for the other data set SPBHM's performance is close to the best values.

We also have conducted correlation analysis on human data sets as shown in Tables 1, 2 and 3 for BP, CC and MF respectively. In this case we have considered only the



(a) (b)

Fig. 3. Barplot of AUCs of different measures on human PPI data sets for three ontologies in (a) Yu et al.'s and (b) Jain et al's data.

Resnik, Relevance and simIC measures, as these are found to provide competitive performance for the yeast gene expression data as well as the PPI data (Section 5.1). Moreover these two measures are also indicated to be better by many other authors [31], [43]. As can be seen from the tables, SPBHM generally outperforms the other three for most of the data sets. In only a few cases, it is marginally inferior to Resnik, Relevance and simIC measures.

5.3 Functional Classification of Genes in a Biological Pathway

Fourteen genes appear in the *Glycolysis* pathway [40] of yeast. The dendrogram generated by complete linkage hierarchical clustering of these genes using the proposed SPBHM and Relevance measures are shown in Figs. 5a and 5b, respectively. In Table 4 we have listed the gene clusters with EC numbers of individual genes and four clusters obtained using both the measures. We see that SPBHM has clustered the genes such that all the genes in a cluster have similar subtype of EC numbers. Let us consider the second cluster of five genes. As can be seen, SPBHM is able to group together CDC19, PYK2, PGK1, PFK1 and PFK2, all of which have similar molecular functions (EC numbers starting with 2.7 for all of them). In contrast, the cluster obtained using Relevance measure is mixed. Closer analysis reveals that Relevance measure has failed in the first level itself when it assigns high similarity to CDC19, PYK2, ENO1 and ENO2. On the other hand, SPBHM first groups (CDC19 and PYK2), thereafter combining these with PGK1. Note that gene pair CDC19 and PYK2 both have a similar MF GO term *pyruvate kinase activity* (GO:0004743), at level seven with a large IC value. SPBHM measured a similarity value of 0.95 for these two genes. Similarly gene pair PFK1 and PFK2 have a very high GO similarity (they have the same EC number also) and share a common enriched MF term *6-phosphofructokinase activity* (GO:0003872) with a high IC value. So these pairs are placed in same cluster at the first level. Five genes CDC19, PYK2, PGK1, PFK1 and PFK2 are clustered in the second level. All these genes have common ancestor MF GO term *kinase activity* (GO:0016301), a level five GO term with enriched p-value of 4.94e-09, and they have almost common EC number subtypes. The five genes clustered using Relevance measure have a common enriched MF GO term *metal ion binding* (GO:0046872) with a larger p-value of 3.0e-05 as compared to that found by SPBHM. Similar investigation has been conducted using Wang's measure whose performance is found to be quite inferior. The results are omitted here for brevity. The above results establish the superiority of the proposed SPBHM similarity measure.

Clustering results for the *tryptophan degradation* pathway [40] is shown in Table 5 for the two measures. Both of

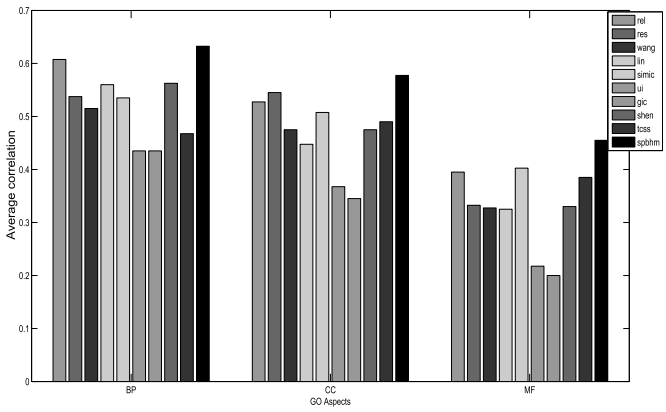


Fig. 4. Correlation with gene expression and GO similarity on yeast data set. Bar plot shows the average value obtained from the four gene expression data set for various measures.

TABLE 1
Correlation of Gene Expression Data [39] of Human with GO BP Similarity (BMA) Score

GSE no	Resnik [10]	Relevance[13]	simIC[14]	SPBHM
GSE20247	0.59	0.54	0.56	**0.61**
GSE9574	0.61	0.49	0.53	**0.63**
GSE5788	0.51	0.53	0.51	**0.59**
GSE13466	**0.52**	0.45	0.49	0.43
GSE20437	0.29	0.32	0.31	**0.52**

TABLE 2
Correlation of Gene Expression Data [39] of Human with GO CC Similarity Score (BMA)

GSE no	Resnik [10]	Relevance [13]	simIC [14]	SPBHM
GSE20247	0.34	0.38	0.39	**0.41**
GSE9574	0.28	0.36	0.35	**0.43**
GSE5788	**0.51**	0.29	0.36	0.37
GSE13466	0.52	0.61	0.56	**0.66**
GSE20437	0.59	**0.70**	0.67	0.67

TABLE 3
Correlation of Gene Expression Data [39] of Human with GO MF Similarity Score (BMA)

GSE no	Resnik[10]	Relevance [13]	simIC [14]	SPBHM
GSE20247	0.38	0.35	0.37	**0.42**
GSE9574	0.28	**0.34**	**0.34**	0.33
GSE5788	0.35	0.48	0.49	**0.52**
GSE13466	0.39	0.51	0.52	**0.52**
GSE20437	0.39	**0.57**	**0.57**	0.54

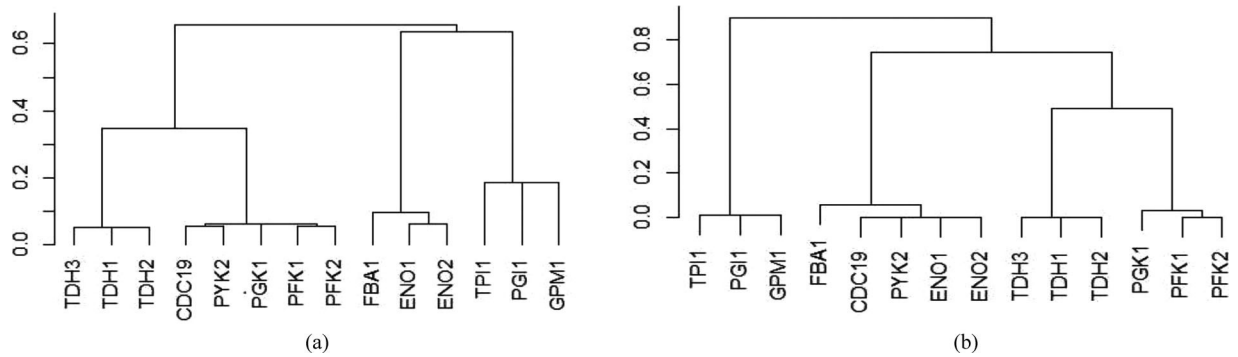


Fig. 5. Dendrogram of genes in the yeast glycolysis pathway based on GO MF distance using (a) the proposed SPBHM and (b) Relevance measure. The vertical axes show the distance (or 1-similarity) at which the clusters are merged.

the measures perform equally for this data. The dendrograms by the two measures are shown in Figs. 6a and 6b respectively. Here genes ADH1, ADH2, ADH3, ADH4, ADH5 and SFA1 have enriched MF term *alcohol dehydrogenase (NAD) activity* (GO:0004022), at level 6 in GO graph with p-value 1.116558e-20. Genes ARO8 and ARO9 also have an enriched MF term *aromatic-amino-acid:2-oxoglutarate aminotransferase activity* (GO:0008793), at level 6 in the GO graph with p-value 4.984313e-08. Genes ARO10, PDC1, PDC5 and PDC6 have a similar MF GO term *pyruvate decarboxylase activity* (GO:0004737) also at level 6 with p-value 1.491544e-14. All three gene sets have similar EC number subtypes. The results imply that a good functional classification is achieved by the proposed one as well as Relevance measure for this data set.

5.4 Semantic Similarity among the Target Gene Set of miRNAs

Here we have measured the average semantic similarity of the experimentally validated targeted gene sets [42] of human miRNAs. After preparation of the data (see Section 4.5), for each targeted gene set we found the enriched GO BP terms and their participation in enriched KEGG pathways. As can be seen from Table 7 in the supplementary material, available online, the gene sets have highly enriched GO terms with very small p-values computed using Clusterprofiler toolbox [44] with hypergeometric test. Many of them participate in common KEGG pathways. Thus it is expected that they should exhibit high average GO BP similarity. Our results also confirmed the same (average GO BP similarity of miRNAs in the range 0.70 to 0.94). In a part of the study, the values of the SPBHM and eight other measures are computed for every pair of genes in the target set of a miRNA. Table 6 shows the correlation between SPBHM and the other eight measures. Results are shown on only 14 miRNAs, though the trend is the same for all others. From the values in Table 6 it is clear that the proposed SPBHM has high correlation with Relevance measure as also with simIC, Jiang, Wang and Lin's measures. The agreement of SPBHM with Resnik's measure appears to be the least. The results indicate that our hypothesis about high GO similarity between target genes of a miRNA is supported not only by SPBHM but also by most of the other widely used measures.

5.5 CESSM Results

In this section a summary of different correlation analysis based on EC number, Pfam domain and sequence similarity, as obtained using the CESSM [28] online tool, are discussed. Twelve measures viz., Resnik, Relevance, simIC, Jiang, Lin, simUI, simGIC, Wang, Shen, TCSS, Zhang and SPBHM are used for comparison. Details are provided in the supplementary material, available online. BMA combination approach is used here for all the measures except for the group wise measures (simUI and simGIC). In the BP ontology, SPBHM performed best for the correlation with EC number similarity and Pfam domain similarity. For sequence similarity, performance of SPBHM is the fourth best after simUI, Resnik and simGIC measures. For CC

TABLE 4
Clustering Result of Yeast Glycolysis Pathway [40] by SPBHM and Relevance Measure

SPBHM threshold 0.84		Relevance threshold 0.90	
Gene name	EC number	Gene name	EC number
TDH1	1.2.1.12	TDH1	1.2.1.12
TDH2	1.2.1.12	TDH2	1.2.1.12
TDH3	1.2.1.12	TDH3	1.2.1.12
CDC19	2.7.1.40	FBA1	4.1.2.13
PYK2	2.7.1.40	CDC19	2.7.1.40
PGK1	2.7.2.3	PYK2	2.7.1.40
PFK1	2.7.1.11	ENO1	4.2.1.11
PFK2	2.7.1.11	ENO2	4.2.1.11
ENO1	4.2.1.11	PGK1	2.7.2.3
ENO2	4.2.1.11	PFK1	2.7.1.11
FBA1	4.1.2.13	PFK2	2.7.1.11
TPI1	5.3.1.1	TPI1	5.3.1.1
PGI1	5.3.1.9	PGI1	5.3.1.9
GPM1	5.4.2.1	GMP1	5.4.2.1

TABLE 5
Clustering Result of Yeast Tryptophan Degradation Pathway [40] by the SPBHM and Relevance Measure

PBHM threshold 0.90		Relevance threshold 0.95	
Gene name	EC number	Gene name	EC number
ADH1	1.1.1.190	ADH1	1.1.1.190
ADH2	1.1.1.190	ADH2	1.1.1.190
ADH3	1.1.1.190	ADH3	1.1.1.190
ADH4	1.1.1.190	ADH4	1.1.1.190
ADH5	1.1.1.190	ADH5	1.1.1.190
SFA1	1.1.1.190	SFA1	1.1.1.190
ARO10	4.1.1.74	ARO10	4.1.1.74
PDC1	4.1.1.74	PDC1	4.1.1.74
PDC5	4.1.1.74	PDC5	4.1.1.74
PDC6	4.1.1.74	PDC6	4.1.1.74
ARO9	2.6.1.28	ARO9	2.6.1.28
ARO8	2.6.1.27	ARO8	2.6.1.27

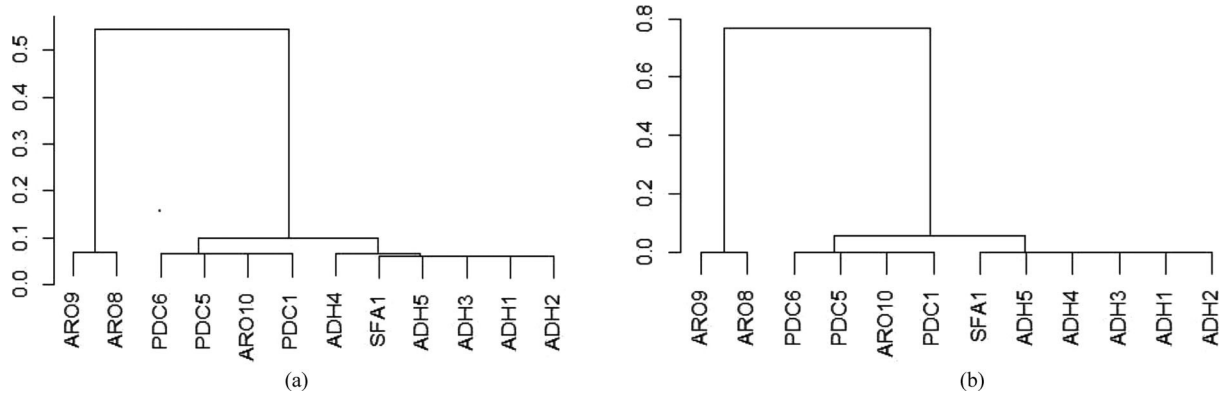


Fig. 6. Dendrogram of genes in the yeast tryptophan degradation pathway based on GO MF distance using (a) the SPBHM and (b) Relevance measure. The vertical axes show the distance (or 1-similarity) at which the clusters are merged.

ontology with EC correlation, Resnik (0.3776) is marginally superior to SPBHM (0.3762) followed by the other measures. For Pfam domain and sequence similarity measure it performed fourth best, being better than relevance, simIC, Lin, Jiang, Wang, TCSS, Zhang and Shen. For MF ontology SPBHM performed better than all eight aforementioned measures for Pfam similarity. In case of EC correlation similarity, SPBHM shows the best correlation performance. In the case of sequence similarity, SPBHM performed better than Shen, simIC and Relevance measures only. In terms of the resolution parameter, SPBHM performed the best for BP and CC, (0.9337 and 0.9798 respectively). We have aggregated the ranks of different measures obtained from the different CESSM results, SPBHM got the rank one followed by simGIC and Resnik measures.

5.6 Further Analysis of SPBHM

In the previous sections we have established the effectiveness of the proposed SPBHM as compared to several well-known measures of ontological similarity. In this section we discuss how well the proposed measure is able to handle the shallow annotation problem. Let us consider three GO terms, namely GO:0046483 *heterocycle metabolic process*, GO:0044237 *cellular metabolic process* and GO:0008152 *metabolic process*. Of these, GO:0046483 is the most specific term (lower down the ontology), while GO:0008152 is the most general (directly connected to the root term). The IC values of the three terms are 3.95, 1.33 and 1.165, respectively, as per the gene annotation database. Now we compute the GO BP similarity of the three terms with respect to themselves using the measures of Lin, Jiang and Wang and also with the proposed SPBHM. While the first three measures are unable to indicate the difference in the specificity of the three terms in the ontology, SPBHM is able to correctly indicate this by providing values 0.86, 0.77 and 0.65, respectively. Resnik and Relevance measures are also able to indicate this difference between the three terms by providing values (0.34, 0.117, 0.103) and (0.98, 0.76 0.688) respectively. We have also seen that simIC and TCSS are also able to provide different similarity scores for these three cases. This example aptly demonstrates that SPBHM is able to tackle the shallow annotation problem quite well.

The resolution of SPBHM is found to be better than the Relevance measure. In fact, genes that are found to be greater than 0.9 similar according to the Relevance measure, are found to have similarity values in the range of 0.78 to 0.99. Thus while the Relevance measure is unable to capture the subtle difference in similarities of many gene pairs SPBHM can easily distinguish such cases. Consider a term pair with nearly equal IC values with their LCA term $IC(t_1) \approx IC(t_2) \approx IC(LCA(t_1, t_2))$. According to the Relevance measure (see Eq. (3)) the similarity value is adjusted by the probability of the LCA term $(1 - p(LCA))$. If the

TABLE 6
GO BP Similarity Analysis of Target Gene Set of Some Human miRNAs [42]

	Pearson Correlation of the proposed SPBHM with							
miRNA id	Resnik [10]	Jiang [12]	Lin [11]	Wang [17]	Relevance [13]	simIC [14]	Shen [19]	TCSS [31]
hsa-miR-20a	0.67	0.82	0.81	0.83	0.94	0.92	0.84	0.83
hsa-miR-137	0.62	0.64	0.67	0.69	0.76	0.75	0.65	0.69
hsa-miR-9	0.76	0.94	0.95	0.93	0.96	0.93	0.91	0.88
hsa-miR-183	0.77	0.86	0.94	0.85	0.92	0.95	0.91	0.85
hsa-miR-31	0.62	0.84	0.89	0.86	0.90	0.88	0.84	0.79
hsa-miR-503	0.88	0.88	0.87	0.88	0.92	0.93	0.78	0.79
hsa-miR-199a-5p	0.54	0.93	0.93	0.93	0.94	0.92	0.74	0.61
hsa-miR-19b	0.81	0.82	0.83	0.88	0.97	0.94	0.85	0.77
hsa-miR-195	0.77	0.89	0.90	0.91	0.92	0.86	0.82	0.79
hsa-miR-520c-3p	0.76	0.93	0.90	0.85	0.95	0.94	0.85	0.76
hsa-let-7c	0.79	0.92	0.91	0.91	0.96	0.95	0.91	0.78
hsa-let-7g	0.77	0.96	0.95	0.97	0.96	0.98	0.93	0.75
hsa-let-7e	0.73	0.89	0.91	0.86	0.92	0.90	0.84	0.70

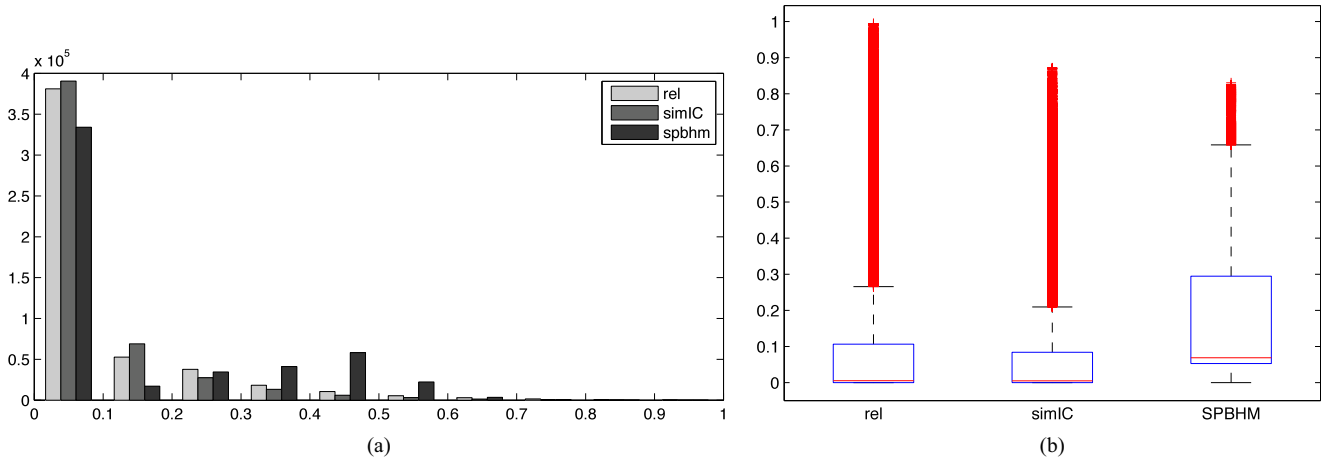


Fig. 7. GO similarity score distribution Histogram (a) and boxplot (b) of Relevance, simIC and SPBHM measures on BP ontology.

LCA terms of two gene pairs have a probability very close to 0 or 1, then their similarity values also become very similar. In contrast, in such cases SPBHM calculates the value by considering the summation of the weighted paths which will be different, depending on the location of the LCA terms. Consider the following human genes: CDK2, CCND2 and ERBB4 with entrez ids 1017, 894 and 2066, respectively. The LCA term *cell division* (GO:0051301) between the gene pair CCND2 and CDK2 (obtained using MAX combination) has 408 annotations with p-value 0.0007. The similarity values of this gene pair using the measures Relevance, Lin, Wang, Jiang, Resnik and SPBHM are 0.999, 1, 1, 1, 0.46 and 0.89 respectively in GO BP ontology. In contrast, the genes CCND2 and ERBB4 has an LCA term *positive regulation of epithelial cell proliferation* (GO:0050679) and is more specific than that in the previous example. This LCA term has 98 annotations with respect to the same background gene set with p-value 4.49e-05. The similarity values of this gene pair using the measures Relevance, Lin, Wang, Jiang, Resnik and SPBHM are 0.999, 1, 1, 1, 0.59 and 0.96 respectively in GO BP ontology. Evidently, the first four measures are unable to capture the fact that although both the gene pairs are highly similar, the similarity in the second case should be higher because the corresponding LCA term is more specific. In contrast, both Resnik and SPBHM (as also simIC and TCSS, results omitted) are able to detect this fact, providing higher similarity values for this gene pair. Several similar examples, including for the measures, have been observed but are omitted here for the sake of brevity.

For better understanding of the unique property of SPBHM, we have shown the distribution chart of pair wise GO BP ontology term similarity scores. There are 2,024 directly annotated terms in the yeast protein corpus. As the number of term pairs generated from this corpus is large we have selected 50 percent of the total terms randomly. Then the similarity value of each pair is calculated using Relevance, simIC, SPBHM measures. As can be seen from Fig. 7a, the histogram plots of the scores follow similar patterns for all the three measures. It is also evident that most of the term pairs lie in a very low score range (0-.1). This is expected since most of the term pairs will have low similarity while only a few will have high similarity scores. While SPBHM measure has distributed the score of the term pairs into different ranges. One reason is that, SPBHM does not assign a score zero to the all term pairs have the root node as LCA. It assigns a score which is proportionally decreasing with the increase of distance from the root to the term pair. The other reason is it has better resolution. Boxplot (Fig. 7b) shows that SPBHM has wider ranges for the distribution of scores.

6 Conclusion

In this paper we have proposed a new GO similarity measure, called SPBHM, integrating the topological property of GO DAG and annotation statistics of GO terms. SPBHM measures similarity between two GO terms as a function of their similarity as well as dissimilarity values along the shortest weighted paths from the root. To measure the dissimilarity value between a term pair, SPBHM uses the aggregation of the contributions of all uncommon ancestor terms along the shortest weighted paths from annotating terms to next descendant of LCA. For measuring the similarity of the term pair, the weighted path from the LCA to the root is considered.

As shown in this article, this measure efficiently handles the shallow annotation problem, as it considers the both the difference of level between the terms and their LCA, as well as the IC values along the path. This is the unique feature of SPBHM, making it different from the other popular path-based measures. Moreover this measure is able to capture subtle differences between highly similar term pairs. It is intuitively clear that although two terms pairs may be highly similar, however from the human perspective it is generally possible to say if one of the two pairs is more similar than the other. SPBHM is able to indicate these finer differences; this is not true for many of the existing measures, as shown in Section 5.6. Extensive experimental results provided in this article demonstrate that the proposed SPBHM performs better when predicting PPIs, has higher correlation with gene expression similarity, and is better able to classify functions of genes in a biological pathway. As a further validation, it is shown that SPBHM correlates well with some other existing measures when used for computing the similarity of genes that are targets of a given miRNA.

As is evident, the performance of the measures is dependent on the properties of the data set. For the case of PPI data set, if the proteins in the corpus have more specific GO term annotations, then the measures that can handle shallow annotations will perform better, e.g., SPBHM, Resnik, Relevance, simIC, Shen and TCSS. The others measures (that cannot handle shallow annotations) do not perform well because they tend to show higher similarity for shallow annotated terms present in the negative data set. In contrast, if the proteins have more general annotations, then neither the above measures that can handle shallow annotations, nor the others that cannot (e.g., Lin, Jiang, Wang, simUI, simGIC), will perform well. In fact, in such situations it might be misleading to use GO similarity scores for PPI prediction.

For the studies related to correlation between GO similarity and other types of similarities (e.g., gene expression and sequence similarity), the choice of the measure depends on the nature of the data. For example, for gene co-expression studies if the data contains highly co-expressed gene pairs, then the measures that have higher resolution in the higher similarity range should be preferred. The converse is true for a data containing gene pairs with low co-expression. The advantage of using SPBHM is that it is not only able to handle shallow annotation problem, and hence performs well for PPI prediction (see Figs. 1a, 1b, 2, 3a and 3b), its resolution is also more for a greater range of similarities (Figs. 7a and 7b and discussion in Section 5.6). Hence it performs well for correlation studies as well.

There is some scope for improving the proposed measure. An exponential and arctan function is used in SPBHM. It is important to see the effectiveness of other transfer functions in this regard. Moreover, instead of using information from only the shortest weighted path, it may be better if all weighted paths are considered. As pointed out in [2], it is important to study the term specificity of SPBHM based other ways of computing IC values. Moreover annotation length bias [2] of SPBHM should also be investigated. These need to be investigated in future.

REFERENCES

[1] L.du Plessis, N. Škunca, and C. Dessimoz, "The What, Where, How and Why of Gene Ontology A Primer for Bioinformaticians," *Briefings in Bioinformatics*, vol. 12, no. 6, pp. 723-735, 2011.

[2] P.H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic Similarity Analysis of Protein Data: Assessment with Biological Features and Issues," *Briefings in Bioinformatics*, vol. 13, no. 5, pp. 569-585, 2012.

[3] S.R. Maetschke, M. Simonsen, M.J. Davis, and M.A. Ragan, "Gene Ontology-Driven Inference of Protein-Protein Interactions Using Inducers," *Oxford Bioinformatics*, vol. 28, pp. 69-75, Nov. 2011.

[4] B. Raghavachari, A. Tasneem, T.M. Przytycka, and R. Jothi, "Domine: A Database of Protein Domain Interactions," *Nucleic Acids Research*, vol. 36, (Database issue), pp. D656-D661, Jan. 2008.

[5] F. Ramirez, A. Schlicker, Y. Assenov, T. Lengauer, and M. Albrecht, "Computational Analysis of Human Protein Interaction Networks," *Proteomics*, vol. 7, no. 15, pp. 2541-2552, Aug. 2007.

[6] D. Zhong-Hui, B. Hughes, L. Reichel, and T. Shi, "The Relationship between Protein Sequences and Their Gene Ontology Functions," *Proc. First Int'l Multi-Symp. Computer and Computational Sciences*, pp. 76-83, 2006.

[7] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship between Sequence and Annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275-1283, Feb. 2003.

[8] L. Franke, H. Bakel, L. Fokkens, E. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes," *Am. J. Human Genetics*, vol. 78, no. 6, pp. 1011-1025, 2006.

[9] N.S. Louis du Plessis and C. Dessimoz, "Semantic Similarity in Biomedical Ontologies," *PLoS Computational Biology*, vol. 5, no. 7, p. e1000443, 2009.

[10] P. Resnik, "Semantic Similarity in a Taxonomy: An Information Based Measure and Its Application to Problems of Ambiguity in Natural Language," *J. Artificial Intelligence Research*, vol. 11, pp. 99-130, Feb. 1999.

[11] D. Lin, "An Information-Theoretic Definition of Similarity," *Proc. 15th Int'l Conf. Machine Learning*, vol. 98, pp. 296-304, 1998.

[12] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," *Proc. Int'l Conf. Research on Computational Linguistics (ROCLING X)*, pp. 19-33, 1997.

[13] A. Schlicker, F.S. Domingues, J. Rahnenführer, and T. Lengauer, "A New Measure for Functional Similarity of Gene Products Based on Gene Ontology," *BMC Bioinformatics*, vol. 7, article 302, Jan. 2006.

[14] B. Li, J.Z. Wang, F.A. Feltus, J. Zhou, and F. Luo, "Effectively Integrating Information Content and Structural Relationship to Improve the Go-Based Similarity Measure between Proteins," *Proc. Int'l Conf. Bioinformatics and Computational Biology (BioComp '10)*, pp. 12-15, July 2010.

[15] A. Nagar and A.-M. Hisham, "A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways," *Proc. 21st IEEE Int'l Symp. Computer-Based Medical Systems (CBMS '08)*, pp. 590-595, 2008.

[16] X.S. Shaohua Zhang, M. Wang, and J. Diao, "A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways," *Proc. WASE Int'l Conf. Information Eng. (WASE '10)*, pp. 85-88, 2010.

[17] J. Wang, Z. Du1, R. Payattakool, P.S. Yu, and C.-F. Chen, "A New Method to Measure the Semantic Similarity of GO Terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274-1281, Mar. 2007.

[18] S. Benabderrahmane, M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes, "Intelligo: A New Vector-Based Semantic Similarity Measure Including Annotation Origin," *BMC Bioinformatics*, vol. 11, no. 1, article 588, 2010.

[19] Y. Shen, S. Zhang, and H.-S. Wong, "A New Method for Measuring the Semantic Similarity on Gene Ontology," *Proc. IEEE Int'l Conf. Bioinformatics and Biomedicine (BIBM '10)*, pp. 533-538, 2010.

[20] R. Gentleman, "Visualizing and Distances Using Go," http://www. bioconductor. org/docs/vignettes. html, 2005.

[21] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto, "Metrics for Go Based Protein Semantic Similarity: A Systematic Evaluation," *BMC Bioinformatics*, vol. 9, no. Suppl. 5, article S4, 2008.

[22] Y. Guangchuang, X. Chuan-Le, B. Xiaochen, L. Chun-Hua, Q. Yide, Z. Sheng, and H. Qing-Yu, "A New Method for Measuring Functional Similarity of MicroRNAs," *J. Integrated OMICS*, vol. 1, no. 1, pp. 49-54, Feb. 2011.

[23] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric On Semantic Nets," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 19, no. 1, pp. 17-30, Jan./Feb. 1989.

[24] X. Wu, E. Pang, K. Lin, and Z.-M. Pei, "Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge-and Ic-Based Hybrid Method," *PloS One*, vol. 8, no. 5, p. e66745, 2013.

[25] C. Pesquita, D. Faria, A. Falco, P. Lord, and F. Couto, "Semantic Similarity in Biomedical Ontologies," *PLoS Computational Biology*, vol. 5, no. 7, p. e1000443, July 2009.

[26] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, "Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship between Sequence and Annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275-1283, 2003.

[27] X. Wu, L. Zhu, G. Jie, D.-Y. Zhang, and K. Lin, "Prediction of Yeast Protein-Protein Interaction Network: Insights from the Gene Ontology and Annotations," *Nucleic Acids Research*, vol. 34, no. 7, pp. 2137-2150, Apr. 2006.

[28] C. Pesquita, D. Faria, and F. Couto, "CESSM: Collaborative Evaluation of Semantic Similarity Measures," *Proc. Jornadas de Bioinformàtica Challenges in Bioinformatics (JB '09)*, 2009.

[29] G. Yu, L. Wang, and Y.H.Y, Q. He, "GOSemSim: An R Package for Measuring Semantic Similarity among GO Terms and Gene Products," *Bioinformatics*, vol. 26, no. 7, pp. 976-978, Apr. 2010.
[30] A. Ben-Hur and W.S. Noble, "Kernel Methods for Predicting Protein-Protein Interactions," *Bioinformatics*, vol. 21, no. 1, pp. i38-i46, 2005.
[31] S. Jain and G.D.B.M.J. Davis, "An Improved Method for Scoring Protein-Protein Interactions Using Semantic Similarity within the Gene Ontology," *BMC Bioinformatics*, vol. 11, article 562, Nov. 2010.
[32] J. Yu, M. Guo, C.J. Needham, Y. Huang, L. Cai, and D.R. Westhead, "Simple Sequence-Based Kernels Do Not Predict Protein-Protein Interactions," *Bioinformatics*, vol. 26, no. 20, pp. 2610-2614, 2010.
[33] I. Xenarios, L. Salwinski, J. Duan, P. Higney, S. Kim, and D. Eisenberg, "DIP: The Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions," *Nucleic Acids Research*, vol. 30, pp. 303-305, 2002.
[34] P. Pagel et al., "The MIPS Mammalian Protein-Protein Interaction Database," *Bioinformatics*, vol. 21, no. 6, pp. 832-834, 2005.
[35] S. Razick, G. Magklaras, and I. Donaldson, "iRefIndex: A Consolidated Protein Interaction Database with Provenance," *BMC Bioinformatics*, vol. 9, article 405, Sept. 2008.
[36] M. Brameier and C. Wiuf, "Co-Clustering and Visualization of Gene Expression Data and Gene Ontology Terms for Saccharomyces Cerevisiae Using Self-Organizing Maps," *J. Biomedical Informatics*, vol. 40, pp. 160-173, 2007.
[37] M. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences USA*, vol. 95, no. 25, pp. 14863-14868, 1998.
[38] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.
[39] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "NCBI GEO: Mining Tens of Millions of Expression Profiles-Database and Tools Update," *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D760-D765, 2007.
[40] K.R. Christie et al., "Saccharomyces Genome Database (sgd) Provides Tools to Identify and Analyze Sequences from Saccharomyces Cerevisiae and Related Sequences from Other Organisms," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D311-D314, 2004.
[41] P. Zotos, G. Papachristoudis, M. Roubelakis, I. Michalopoulos, K. Pappa, N. Anagnou, and S. Kossida, "GOmir: A Stand-Alone Application for Human microRNA Target Analysis and Gene Ontology Clustering," *Proc. Eighth IEEE Int'l Conf. BioInformatics and BioEng. (BIBE '08)*, 2008.
[42] S. Hsu et al., "miRTarBase: A Database Curates Experimentally Validated microRNA-Target Interactions," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D163-D169, 2011.
[43] T. Xu, L. Du, and Y. Zhou, "Evaluation of GO-Based Functional Similarity Measures Using S. Cerevisiae Protein Interaction and Expression Profile Data," *BMC Bioinformatics*, vol. 9, article 472, 2008.
[44] G. Yu, L. Wang, and Y.H.Y, Q. He, "ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters," *A J. Integrative Biology*, vol. 16, no. 5, pp. 284-287, May 2012.

Sanghamitra Bandyopadhyay received the PhD degree in computer science in 1998 from Indian Statistical Institute, Kolkata, India, where she currently serves as a professor. She has received the prestigious S. S. Bhatnagar award in 2010, Humboldt fellowship for experienced researchers, and the Senior Associateship of ICTP, Italy. She is a fellow of the Indian National Academy of Engineering and the National Academy of Science, India. She has co-authored six books and more than 250 research papers. Her research interests include pattern recognition, data mining, evolutionary computing, and bioinformatics.

Koushik Mallick received the ME degree in computer science and engineering from Jadavpur University in 2009. He is working toward the PhD degree from Calcutta University while working at Indian Statistical Institute, Kolkata, India. At present, he is an assistant professor at RCC Institute of Information Technology (RCCIIT), Kolkata.