M-Finder: functional association mining from protein interaction networks weighted by semantic similarity

Young-Rae Cho
Department of Computer Science
Baylor University
Waco, TX, USA
young-rae_cho@baylor.edu

Tak Chien Chiam
Department of Computer Science
Baylor University
Waco, TX, USA
chiam_takchien@baylor.edu

Yanxin Lu
Department of Computer Science
Baylor University
Waco, TX, USA
yanxin_lu@baylor.edu

***Abstract*—Protein-protein interactions (PPIs) play a key role in understanding functional behavior of genes. Discovering association patterns from PPI networks is crucial for functional characterization on a system level. We present a novel approach to discover the functional association pattern of a query gene from the genome-wide PPI networks. This approach consists of two major components. First, we transform the PPI network to a weighted graph representation by measuring semantic similarity. Three enhanced semantic similarity methods are proposed to estimate functional closeness of each interacting pair. Second, we apply a dynamic propagation algorithm to detect the functional association pattern of a gene, represented as a sub-network. The size of the sub-networks is flexibly determined by user-specific parameters. In this paper, we also introduce an interactive web application, called M-Finder, to visualize the functional association pattern of a gene entered by a user. The semantic similarity measures and the dynamic propagation algorithm are embedded in this tool to run on up-to-date PPI networks of model species. M-Finder allows users to carry out further systematic analysis for functional characterization on the genomic scale.**

***Keywords*-protein-protein interactions; protein interaction networks; functional modules; Gene Ontology; semantic similarity**

I. Introduction

Proteins interact with each other to build larger functional units. Functional associations between genes are thus evidenced by PPIs. In recent years, PPI data have been enriched by advanced large-scale experimental and computational techniques. Availability of the interactome, a set of PPIs on a genome-wide scale, has introduced a new paradigm towards functional characterization of genes. The system-level analysis has been performed by theoretical and empirical studies of the interactome [1], [2]. However, the automated high-throughput methods have resulted in decreasing reliability of the PPI data sets [3].

We use Gene Ontology (GO) and its annotation data to assess the validity of PPIs. GO [4] is a repository of biological ontologies and annotations of genes. Although the annotation data on GO have been created by the published evidence resulting from mostly unreliable high-throughput experiments, they are frequently used as a benchmark for functional characterization because of their comprehensive information on the genome level. Semantic similarity [5] is a function that returns a numerical value reflecting closeness in meaning between two GO terms. The semantic similarity of an interacting protein pair can be quantified by the similarity scores between the GO terms annotating the proteins. Since a true PPI is interpreted as a strong functional association of the interacting pair, we can apply the semantic similarity to estimate the reliability of PPIs.

In this study, we present a computational approach to discover the functional association pattern of a gene from the genome-wide PPI networks. This approach consists of two major technical components. First, the PPI networks are weighted by semantic similarity, i.e., the semantic similarity score of each interacting pair is assigned as a weight to the edge of the network. We use three enhanced semantic similarity methods which have higher accuracy in estimating functional closeness between genes than existing methods. Second, given a query gene, our approach explores the functionally associated genes and their links by simulating a functional propagation model on the PPI network. A dynamic propagation algorithm based on repeated random walks is applied for efficient computation. The size of the output sub-networks, as functional association patterns, is flexibly determined by user-specific parameters.

We introduce an interactive web application, called M-Finder, to visualize the functional association pattern of a gene entered by a user. The semantic similarity measures and the dynamic propagation algorithm are embedded in this tool to run on up-to-date PPI networks of model species.

II. Semantic Similarity

A. *Previous semantic similarity measurements*

Various semantic similarity measures have been proposed previously [6], [7], [8], aiming at quantifying functional similarity between genes. They have utilized the GO structure and annotations. The existing semantic similarity measures can be grouped into four broad categories: *edge-based*, *node-based*, *annotation-based*, and *hybrid* methods. Edge-based methods compute the path length between GO terms [9] or

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

the depth to the most specific common ancestor term (SCA) of two GO terms. The depth to SCA can be normalized with the average depth to the two GO terms.

Node-based methods consider the ratio of shared ancestor terms of two GO terms. Suppose we measure the semantic similarity of two genes g_1 and g_2. We can find two GO term sets having both direct and indirect annotations of g_1 and g_2, respectively. The semantic similarity can be computed by the Jaccard index of the GO term sets (called simUI [10]), i.e., the size of the intersection of the sets divided by the size of their union. Considering unbalanced sets, we can use the size of their intersection divided by the size of a smaller set (called NTO [11]).

Annotation-based methods use the notion of information content of GO terms in order to assign higher values to the terms that have higher specificity. The information content of a term C is defined as the negative log likelihood of C, $-\log P(C)$. The likelihood $P(C)$ indicates the specificity of C, calculated by the proportion of the genes annotated to C. Resnik [12] used the information content of SCA of two GO terms. Lin [13] took into consideration normalizing the Resnik's method with the average information content of the two GO terms. Jiang and Conrath [14] proposed measuring the sum of differences of information contents between SCA and the two GO terms.

Hybrid methods incorporate the aspects of two different categories. For example, Wang *et al.* [15] proposed the semantic similarity that integrates the normalized node-based method with the concept of the normalized depth to the terms. SimGIC [16] integrates simUI with information contents. It calculates the sum of information contents in the intersection of two GO term sets having the annotations of two genes, divided by the sum of information contents in their union.

B. New approaches of semantic similarity

Compared to sequence similarity, expression correlations and interaction evidence, many previous studies [16], [17] have shown that the Resnik's method as an annotation-based approach, simUI as a node-based approach, and simGIC as a combined method of the Resnik's and simUI have the best performance. To convert multiple semantic similarity scores between two GO terms C_1 and C_2 to a single functional similarity score between two genes g_1 and g_2, it has been observed that the best-match averaging (BMA) approach [18] to compute the average of pairwise best-matches has the highest accuracy.

$$sim_{\scriptscriptstyle BMA}(g_1, g_2) = \frac{\sum_{C_1 \in S_1} \max_{C_2 \in S_2} sim(C_1,C_2) + \sum_{C_2 \in S_2} \max_{C_1 \in S_1} sim(C_1,C_2)}{|S_1|+|S_2|}.$$

In this study, we propose additional improvements of the semantic similarity measures. Since the Resnik's method computes the information content of SCA of two terms, it focuses on their commonality only, not a difference between them. Although the Lin's method normalizes the Resnik's scores with the average information content of the two individual terms, this normalization process reflects a significant bias because of the shallow annotation problem [8]. We thus normalize the Resnik's semantic similarity of two GO terms with distance between them. Three different approaches, simICNP, simICND and simICS, are introduced to measure the distance.

Suppose we measure the semantic similarity between C_1 and C_2, and C_0 is their SCA. First, as an intuitive way, simICNP uses the information content of SCA normalized with the shortest path length between C_1 and C_2 in the ontology as the distance.

$$sim_{\scriptscriptstyle ICNP}(C_1, C_2) = \frac{-\log P(C_0)}{1 + len(C_1, C_2)}.$$

This method gives a penalty to the Resnik's semantic similarity if C_1 and C_2 are located farther from their SCA in GO. Next, simICND employs the information content of SCA normalized with the difference of information contents from two terms to their SCA, as the Jiang's method uses.

$$sim_{\scriptscriptstyle ICND}(C_1, C_2) = \frac{-\log P(C_0)}{1 - \log P(C_1) - \log P(C_2) + 2 \cdot \log P(C_0)}.$$

This method gives a penalty to the Resnik's semantic similarity if the information contents of C_1 and C_2 are higher than the information content of their SCA.

The last approach explores the information content of the difference of two sets of genes annotated to two GO terms, respectively, whereas the Resnik's method uses the information content of SCA only. Γ_{C_i} denotes the set of genes annotated to the term C_i. Since $|\Gamma_{C_1}| \subseteq |\Gamma_{C_0}|$ and $|\Gamma_{C_2}| \subseteq |\Gamma_{C_0}|$ where $|\Gamma_{C_i}|$ is the size of the set Γ_{C_i}, $|\Gamma_{C_1} \cap \Gamma_{C_2}| \subseteq |\Gamma_{C_0}|$. Moreover, $|\Gamma_{C_1} \cup \Gamma_{C_2}| \subseteq |\Gamma_{C_0}|$. To measure the semantic similarity between C_1 and C_2, we combine three factors: Γ_{C_0}, the union set of Γ_{C_1} and Γ_{C_2}, and the intersection set of Γ_{C_1} and Γ_{C_2}. If $|\Gamma_{C_0}| - |\Gamma_{C_1} \cup \Gamma_{C_2}|$ is small, then C_1 and C_2 are close to C_0, so they are similar each other. $|\Gamma_{C_0}| - |\Gamma_{C_1} \cup \Gamma_{C_2}|$ is clearly bounded by 0 and $|\Gamma_{C_0}|$. Also, the larger $|\Gamma_{C_1} \cap \Gamma_{C_2}|$ is, the more similar C_1 and C_2 are. From these three factors, simICS computes the information content of annotation set difference, i.e., $\alpha \cdot |\Gamma_{C_0}| + \beta \cdot (|\Gamma_{C_0}| - |\Gamma_{C_1} \cup \Gamma_{C_2}|) - \gamma \cdot |\Gamma_{C_1} \cap \Gamma_{C_2}|$. We empirically found this semantic similarity has better performance when $\alpha > \beta = \gamma$. By giving twice more weight to the first factor α than the others, we have $2|\Gamma_{C_0}| + |\Gamma_{C_0}| - |\Gamma_{C_1} \cup \Gamma_{C_2}| - |\Gamma_{C_1} \cap \Gamma_{C_2}| = 3|\Gamma_{C_0}| - |\Gamma_{C_1}| - |\Gamma_{C_2}|$. Therefore,

$$sim_{\scriptscriptstyle ICS}(C_1, C_2) = -\log \frac{3|\Gamma_{C_0}| - |\Gamma_{C_1}| - |\Gamma_{C_2}|}{|\Gamma_{C_r}|},$$

where C_r is the root term in the ontology. Because $(3|\Gamma_{C_0}| - |\Gamma_{C_1}| - |\Gamma_{C_2}|)/|\Gamma_{C_r}|$ ranges from 0 to 3, simICS can produce negative scores.

Table I
THE PEARSON CORRELATION RESULTS BETWEEN SEMANTIC SIMILARITY AND FUNCTIONAL CONSISTENCY OF PPIS.

Category	Semantic Similarity	Correlation
Edge-based	path length	0.699
	normalized depth	0.614
Node-based	NTO	0.571
	simUI	0.679
Annotation-based	Resnik	0.646
	Lin	0.663
	Jiang	0.716
Hybrid	Wang	0.667
	simGIC	0.688
New Approaches	simICNP	**0.726**
	simICND	**0.733**
	simICS	0.678

C. Evaluation of semantic similarity measures

We evaluated the effectiveness of previous semantic similarity measures and three new approaches on the up-to-date PPI data of *S. cerevisiae* from BioGRID [3], which includes 92,906 interactions. The semantic similarity of each interacting pair was computed and compared to their functional consistency as gold standard. For the path length-based methods, annotation-based methods and our three new methods, we applied the BMA method to generate the semantic similarity for each interacting pair.

For functional consistency, we used FunCat data from the MIPS database [19]. Because this small data set has been manually determined and curated, we assumed it is closer to gold standard than GO annotations. The functional catalogues are structured into a relatively well-balanced tree representation. We extracted the functional catalogues and their annotations on the third-level from the root of the hierarchy (the child categories of the root are on the first-level), and measured the functional consistency of each interacting pair by the Jaccard coefficient of two sets of functions that two genes perform, respectively.

We calculated the Pearson correlation between semantic similarity and functional consistency of PPIs. Table I shows the correlation coefficients of nine previous semantic similarity methods and three new approaches. Among nine previous methods, the Jiang's annotation-based method, the path length method, and simGIC as a combination of simUI and the Resnik's method achieved relatively good performance. However, as shown in Table I, simICNP and simICND significantly outperformed those previous methods.

In our approach of functional association mining, these semantic similarity scores are used as edge weights in the range between 0 and 1 for a PPI network. We thus apply the linear transformation of the scores from the Resnik's method and our three new approaches because their scores are not bounded by 0 and 1. We statistically found the upper and lower bounds of the semantic similarity scores and projected them into the range between 0 and 1. All outlier values greater than the upper bound were assigned 1.

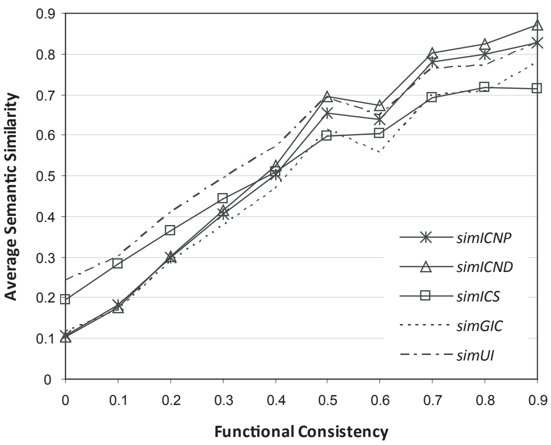

Figure 1. The relationship between semantic similarity scores and functional consistency of the yeast PPIs for each semantic similarity method. According to functional consistency, PPIs were binned into 10 groups and the average semantic similarity was measured for each group.

Figure 1 graphically shows the relationship between semantic similarity and functional consistency of PPIs. We tested five semantic similarity measures: three new approaches and two previous methods, simGIC and simUI. PPIs were binned into 10 groups with the equal range of functional consistency and the average semantic similarity was measured for each group. The ideal case is a diagonal line from 0 to 0.9. When the functional consistency is less than 0.4, simICNP, simICND and simGIC worked better because they produced low semantic similarity scores to the PPIs having low functional consistency. When the functional consistency is higher than 0.7, simICNP, simICND and simUI worked better because of high semantic similarity scores to the PPIs having high functional consistency. When the functional consistency increases from 0.5 to 0.6, the average semantic similarity decreased for all methods except simICS. However, simICNP and simICND show very slight decreases in this range. Overall, simICNP and simICND produce the most suitable semantic similarity scores because, as functional similarity of PPIs increases, their semantic similarity scores have the most likely a linear increase.

III. FUNCTIONAL ASSOCIATION MINING

A. Functional propagation model

A previous study [20] proposed the information propagation model for the application to real-world complex systems which are represented as an undirected and weighted graph $G(V, W)$, i.e., each edge has a weight as its strength. This model was designed to quantify the influence of a source node $v_i \in V$ on the others $v_j \in V, j \neq i$ through connections. A path $p = \langle v_0, v_1, \cdots, v_n \rangle$ is a chain of edges with length n from a source v_0 to a target v_n, and $w_{i(i+1)}$ denotes the weight of the edge between v_i and $v_{(i+1)}$ where each edge weight has a numeric value in the range between 0 and 1. The strength of p in the information propagation model is defined as

$$S(p) = \lambda \cdot w_{0,1} \prod_{i=1}^{n-1} \frac{w_{i(i+1)}}{d_i}. \tag{1}$$

λ is a scale parameter depending on the network structure. d_i is a shape parameter of v_i. It mostly represents the degree of connectivity of v_i. From this formula, it is confirmed that the strength of a path p has inverse relationships with the length of p and the degrees of the nodes on p. Only the degree of the source node does not affect the path strength. The information propagation model then defines the overall impact of v_0 (a source) on v_n (a target) as the sum of path strength for all possible paths from v_0 to v_n including cycling paths.

B. Dynamic propagation algorithm

The propagation model is implemented by the algorithm based upon repeated random walk simulation on an undirected network. This algorithm repeatedly computes the functional impact score $F_s(v_i)$ of a given source node v_s on each node v_i in the network. The strength of a path $\langle v_s, v_1, \cdots, v_n\rangle$ from Equation 1 can be redescribed as

$$S(\langle v_s, v_1, \cdots, v_n\rangle) = S(\langle v_s, v_1, \cdots, v_{n-1}\rangle) \cdot \frac{w_{(n-1)n}}{d_{n-1}}. \tag{2}$$

It indicates that the path strength can be computed in a recursive fashion. Moreover, the functional impact $F_s(v_i)$ of v_s on a node v_i, which is the sum of path strength for all possible paths from v_s to v_i, can be achieved by

$$F_s(v_i) = \sum_{v_j \in N(v_i)} F_s(v_j) \cdot \frac{w_{ji}}{d_j}, \tag{3}$$

where $N(v_i)$ is a set of nodes directly connected to v_i (i.e., a set of neighbors of v_i).

The dynamic propagation algorithm performs the stepwise computation of functional impact scores of a query gene v_s on the others in a PPI network, and finally produces the cumulative functional impact score on each gene. As the first step, it assigns the initial score 1 to v_s and 0 to the others. The next step is to compute the functional impact scores of v_s on its neighboring genes $v_i \in N(v_s)$ by $F_s(v_i) = w_{s,i}$ because $S(\langle v_s, v_i\rangle) = w_{s,i}$ from Equation 1. (We assume λ is 1.) And then, at each step, this algorithm updates the scores on all the genes by Equation 3.

To terminate iterative updating of the functional impact score on each node, we handle two parameters: the minimum path strength threshold and the minimum impact score threshold. The first parameter is to disregard any score of $F_s(v_j) \times \frac{w_{ji}}{d_j}$ on Equation 3 when it is less than the threshold. In this case, $F_s(v_i)$ is calculated by summing all scores from the other neighbors, i.e., $F_s(v_i) = \sum_{v_k \in N(v_i)} F_s(v_k) \cdot \frac{w_{ki}}{d_k}$ where $k \neq j$. The second parameter is to eliminate the score $F_s(v_i)$ when it is less than the threshold. The score is then replaced with 0. It means that the cumulative functional impact score on the node v_i is not updated at this step. This algorithm runs until there are no more updates on cumulative impact scores of any nodes in the network.

This dynamic propagation algorithm is used for mining functional associations (also called a functional module) given a query gene. At first, we build a weighted PPI network by assigning the semantic similarity of each interacting pair to the edge as its weight. We then simulate the propagation of functional information starting from the query gene. The query gene initially has the self-information of 1, and this initial information quantity is propagated through the links of the weighted PPI network. Our approach generates a sub-network consisting of the set of genes, which are functionally associated with the query gene directly or indirectly, and the connections between them. The size of the sub-network is flexibly determined by an additional user-specified parameter, named the minimum association threshold. When the algorithm halts, each gene has a cumulative functional impact score. We then capture the set of genes having the cumulative scores greater than the threshold in order to form a functional sub-network as output. This sub-network reveals the functional association pattern of the query gene.

C. Accuracy of functional associations

We tested the accuracy of this functional association mining approach on the PPI network of *S. cerevisiae*. The weighted PPI network was built using a new semantic similarity measure, simICND, described in the previous section. We randomly selected 1,000 genes from the PPI data set of BioGRID [3] as query genes, and ran the propagation algorithm to generate a sub-network, as a functional association pattern, for each query gene. We used 0.01 and 0.001 for the minimum path strength parameter and the minimum impact score parameter, respectively. We observed that those parameters are not sensitive to the quality of output sub-networks, i.e., increasing (or decreasing) the parameter values does not change the output. However, the minimum association parameter is critical to improve the accuracy of functional association pattern mining. In general, by increasing the minimum association threshold, we can obtain a smaller-size network as output. We thus use the minimum association threshold as a variable parameter.

Figure 2 shows the accuracy improvement of functional association patterns by changing the minimum association parameter value. The sub-network as output for each query gene was compared to the FunCat data from MIPS [19]. The functional categories and their annotations on the first, second and third levels from the root in FunCat were used. We calculated the ratio of genes in the sub-network which are also annotated to each functional category, and then

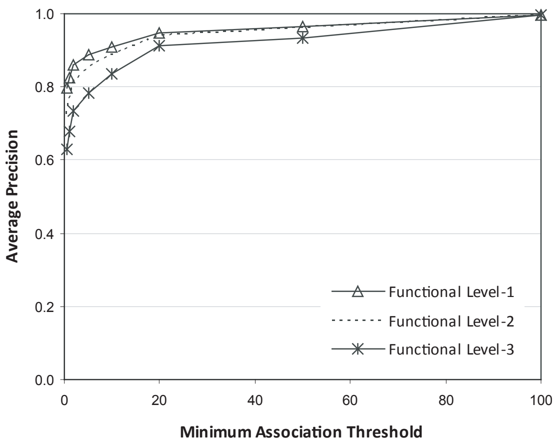

Figure 2. Precision of functional association patterns discovered by the proposed semantic similarity measure and the dynamic propagation algorithm on the yeast PPI network. Precision increases as the minimum association threshold increases. When the output sub-networks are compared to more specific functional categories, precision slightly decreases.

took the maximum ratio among all functional categories as precision of the sub-network. The average precision of all sub-networks generated from 1,000 query genes was plotted in Figure 2. It clearly shows, as the minimum association threshold increases, precision increases because a smaller-sized network including only the genes strongly associated with the query gene is generated. The precision higher than 80% is achieved with any minimum association thresholds when the output sub-networks are compared to the functional categories on the first level. When the sub-networks are compared to more specific functional categories on the second and third levels, their precision slightly decreases. However, it remains higher than 90% with the minimum association thresholds greater than 20.

IV. M-Finder

M-Finder is an interactive web application that enables researchers to discover the functional association pattern of a query gene in several model species. (The current version of M-Finder supports *S. cerevisiae* only.) The primary function of M-Finder is to allow the user to enter the name of a specific query gene (e.g., systematic name or gene symbol) and select a weighting scheme out of three new semantic similarity measures discussed in Section II. This tool then visualizes an undirected sub-network of the genes that are functionally associated with the query gene. The dynamic propagation algorithm is performed on the weighted network which is structured by the most recent version of PPI data from BioGRID, and the resultant sub-network is returned to the client for visualization. The user can specify the minimum association threshold as a parameter. The default parameter value is 20. However, if the user wants to have a larger sub-network, a lower parameter value should be applied. In contrast, for a smaller sub-network, a higher parameter value should be applied.

To visualize the output sub-networks, we used Cytoscape Web [21] which is a Flash-based network library that can easily be embedded on a web page. It models many of network features, including node shapes, visual mappers and layout algorithms, on a stand-alone application. M-Finder utilizes several Cytoscape Web features. Each node is labeled with its official gene name. Callbacks enable specific nodes to be highlighted based on user interaction with other parts of the web page. The visualized sub-network can also be downloaded as any image file format.

As an additional function, M-Finder provides the detailed information of each interactor (gene) and each interaction in the visualized functional association patterns. When clicking any node on the sub-network, the user can obtain not only the accession number, systematic name, official symbol and description of the gene, but also its Gene Ontology information such as GO terms annotating the gene, evidence codes and ontological domains. Similarly, if the user clicks any edge on the sub-network, then the PPI information, such as the interaction type, experimental systems and publications, is displayed. All metadata from BioGRID and Gene Ontology are regularly updated. M-Finder is available online at

`http://bionet.ecs.baylor.edu/mfinder`

V. Conclusion

We presented a novel approach to analyze functional association patterns of a gene of interest from the genome-wide PPI networks. The major components are the semantic similarity measurement of interacting pairs and the functional propagation simulation on a weighted PPI network.

First, measuring semantic similarity is a significant step to estimate functional closeness of each interacting pair. Because most of the PPI data have been produced by high-throughput experimental and computational techniques, the PPI data sets typically include a large amount of false positives, i.e., the interactions in the data sets which do not occur in vivo. Although the proteomic studies have been proceeding actively to determine PPIs on the genome-level, the PPI data sets in many model species still have an extremely large number of false negatives, i.e., the potential interactions that have not been confirmed yet. Weighting PPIs by their semantic similarity can resolve the problems of both false positives and false negatives. Our functional propagation model suggests that a path including more false positive interactions with low weights will have lower strength and deliver a lower chance of being functionally associated. In addition, although a path between two genes is disconnected by false negative interactions, the dynamic propagation algorithm can transfer high functional impact scores through alternative strong paths. The weighted PPI networks, therefore, are very effective in systematic studies for functional characterization.

Second, the dynamic propagation algorithm captures the set of genes functionally associated with a single gene very efficiently. Short running time is one of the critical requirements for the interactive web applications. This algorithm is based on the repeated random walk simulation. It however performs the information propagation starting from a single node only, and prunes an information flow as early as possible if a trivial functional influence is transferred through a link. As a result, it approximates the cumulative functional impact score on each node to improve efficiency. However, the elapsed time might vary depending on the query gene and the parameter value that are entered by a user. The elapsed time is clearly sensitive to parameter value selection since the parameter determines the size of output sub-networks. The connectivity of the selected query gene are also related to the efficiency of MFinder. For example, if the query gene is a hub of the network, then the algorithm searches more neighbors for computation of functional impact scores. Moreover, if the query gene is an intermodule hub, i.e. a hub linked to several different functional modules, then its computational complexity will increase further. As future work, the topological features of the PPI networks can be investigated in order to automatically suggest the best parameter value given a query gene.

References

[1] Yu, H., et al., "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, pp. 104–110, 2008.

[2] Venkatesan, K., et al., "An empirical framework for binary interactome mapping," *Nature Method*, vol. 6, no. 1, pp. 83–90, 2009.

[3] Stark, C., et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, pp. D698–D704, 2011.

[4] The Gene Ontology Consortium, "The Gene Ontology in 2010: extensions and refinements," *Nucleic Acids Research*, vol. 38, pp. D331–D335, 2010.

[5] Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A., "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[6] Pesquita, C., Faria, D., Falcao, A.O., Lord, P. and Couto, F.M., "Semantic similarity in biomedical ontologies," *PLoS Computational Biology*, vol. 5, no. 7, p. e1000443, 2009.

[7] Wang, J., Zhou, X., Zhu, J., Zhou, C. and Guo, Z., "Revealing and avoiding bias in semantic similarity scores for protein pairs," *BMC Bioinformatics*, vol. 11, p. 290, 2010.

[8] Guzzi, P.H., Mina, M., Guerra, C. and Cannataro, M., "Semantic similarity analysis of protein data: assessment with biological features and issues," *Briefings in Bioinformatics*, doi:10.1093/bib/bbr066.

[9] Nagar, A. and Al-Mubaid, H., "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)*, 2008, pp. 590–595.

[10] Guo, X., Liu, R., Shriver, C.D., Hu, H. and Liebman, M.N., "Assessing semantic similarity measures for the characterization of human regulatory pathways," *Bioinformatics*, vol. 22, no. 8, pp. 967–973, 2006.

[11] Mistry, M. and Pavlidis, P., "Gene Ontology term overlap as a measure of gene functional similarity," *BMC Bioinformatics*, vol. 9, p. 327, 2008.

[12] Resnik, P., "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 448–453.

[13] Lin, D, "An information-theoretic definition of similarity," in *Proceedings of 15th International Conference on Machine Learning (ICML)*, 1998, pp. 296–304.

[14] Jiang, J.J. and Conrath, D.W., "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proceedings of 10th International Conference on Research in Computational Linguistics*, 1997.

[15] Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.-F., "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, 2007.

[16] Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E.N., Falcao, A.O. and Couto, F.M., "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, no. Suppl 5, p. S4, 2008.

[17] Jain, S. and Bader, G.D., "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinformatics*, vol. 11, p. 562, 2010.

[18] Tao, Y., Sam, L., Li, J., Friedman, C. and Lussier, Y.A., "Information theory applied to the sparse gene ontology annotation network to predict novel gene function," *Bioinformatics*, vol. 23, pp. i529–i538, 2007.

[19] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H.W., "The FunCat: a functional annotation scheme for systematic classification of proteins from whole genomes," *Nucleic Acids Research*, vol. 32, no. 18, pp. 5539–5545, 2004.

[20] Cho, Y.-R., Shi, L. and Zhang, A., "flowNet: Flow-based approach for efficient analysis of complex biological networks," in *Proceedings of 9th IEEE International Conference on Data Mining (ICDM)*, 2009, pp. 91–100.

[21] Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q. and Bader, G.D., "Cytoscape Web: an interactive web-based network browser," *Bioinformatics*, vol. 26, no. 18, pp. 2347–2348, 2010.