Ontology based semantic similarity for protein interactions

Xiao Luo
University of Houston – Clear Lake
Houston, TX
USA

Hisham Al-Mubaid
Prince Sultan University
Riyadh, 11586
KSA

Saïd Bettayeb
University of Houston – Clear Lake
Houston, TX
USA

Abstract

Ontology structure-based measures, like path length, have been used successfully for semantic similarity in various application domains. In bioinformatics, path length was used with the Gene Ontology (GO), using annotation terms, for gene similarity and clustering. In this paper, we propose to use the path length using GO annotations of proteins as a measure of protein similarity for scoring protein-protein interactions (PPIs). Proteins that interact with each other tend to have similar functions and be involved in similar biological processes compared to proteins that have no interactions. So, with the existence of a reliable well-established ontology, like GO, semantic similarity measures should be able to distinguish between, and rank, fairly well, the interacting and non-interacting protein pairs. The proposed method has been evaluated using datasets of positive and negative protein interactions from human and yeast proteomes. The evaluation results show that this method fares well when used to estimate similarity of interacting and non-interacting proteins.

1. Introduction

Protein-Protein interaction PPI is a very active topic of research in bioinformatics [1, 9, 11, 12, 13, 16]. Large number of projects and publications have been published in the past two decades with fairly interesting and impressive outcomes [12, 13, 16]. Studying and tackling the PPI problem in the context of gene ontology (GO) and using GO annotations of the proteins has opened more possibilities of casting this problem at different levels [7]. GO is considered the most established and structured form of describing gene and protein functions and localization irrespective of the species [7].

In general, interacting proteins are more likely to have higher similarity than non-interacting proteins in terms of their molecular functions and the biological processes that they are involved in. Therefore, an effective similarity measure should be able to distinguish fairly well between interacting and non-interacting protein pairs. For example, proteins that have relatively similar GO annotation terms from the cellular component CC aspect of GO are likely to have interactions and will show relatively higher semantic similarity values [1, 9, 12]. Path length as an ontology structure-based measure has been used successfully as a measure of distance and similarity in various application domains including bioinformatics [3, 5, 10, 17]. In bioinformatics, path length was used in GO, using annotation terms, for gene similarity, gene clustering, and gene disease relationships [17]. In this paper, we present the path length technique using GO annotations of proteins as a measure of semantic similarity between proteins for scoring protein-protein interactions. Graph based methods for semantic similarity have been used and applied in many bioinformatics tasks. However, path length has never been investigated in the context of protein-protein interactions, as to our knowledge. The presented method has been evaluated using datasets of positive and negative interactions from human and yeast protein-proteins interactions PPIs datasets. The evaluation results show that the proposed method fares well as a measure to estimate and assess the protein interactions.

2. Related Work

Gene Ontology: The gene ontology is the primary and most comprehensive source of information for studying and working with gene functions and localizations [10, 17, 1, 3]. GO is a tree-like hierarchical structure, or DAG, where each node represents a term or concept related to gene functions, processes, and locations. So, it is a controlled and structure vocabulary of terms and each term is a node in the DAG. GO is divided into three orthogonal sub ontologies: Molecular function MF, biological processes BP, and cellular component CC. Gene ontology annotations GOA of a protein are terms in the ontology annotated or assigned to that protein. Two proteins are similar if their annotation terms are close. The annotation terms, GOA, of a protein p_i describe the functions, processes, and locations that p_i is involved in.

Similarity measures: a similarity measure is a function, *e.g., sim(),* that quantifies the similarity (or likeness) between two items, *sim*(p_x, p_y), as a numeric value. For example, if a given target protein p_x is more similar to some protein p_i than to p_j then an effective similarity function *sim()* should output that: sim(p_x, p_i) > sim(p_x, p_j). Semantic similarity measures can be roughly divided into two groups: ontology-based measures, information theoretic based measures [3, 10]. Ontology structure based measures are those measures that rely on ontology structure features like path length between node and node depth. On the other hand, information-theoretic based measures, like information content IC, rely on the information revealed by a given node or term in the ontology. Semantic similarity measures have been investigated for long time in different disciplines and applications including natural language processing and bioinformatics [3].

In [5], Al-Mubaid et al. (2007) proposes a semantic similarity approach for similarity of biomedical terms across multiple ontologies and within a unified framework like UMLS [2].

In [2], Pesquita et al. (2009) presents a review study of a number of similarity measures applied to biomedical ontologies. They classify these measures based on various aspects such as edge based versus node based, or pair-wise versus group-wise, and so on [2].

PPI: A huge volume of research has been devoted to studying and investigating PPI from different perspectives in the past two decades [1, 9,11,12,13,16]. Jain and Bader (2010) [1] proposed a new semantic similarity method based on topological clustering semantic similarity for scoring PPIs. Their method performed well on interaction data sets from human and yeast. Their method relies on the fact that different depths of biological knowledge in the various branches of the ontology, viz. GO, may contribute differently into the semantic similarity of the concepts [1]. Several other PPI research studies have used GO as their knowledge resource for protein interaction analysis and prediction [9,11,12,13]

3. Semantic Similarity and PPI

The relationship between semantic similarity measures and protein-protein interactions is straightforward as follows. The semantic similarity value, computed by a semantic similarity measure, for a pair of interacting proteins, in general, is relatively higher than that of a pair of non-interacting proteins. To compute the similarity between proteins we use the GO annotations [1, 9, 12]. In the following we discuss a few semantic similarity measures and then we explain the path length measure.

3.1 Semantic Similarity Measures

The most common and most basic information theoretic based semantic similarity measure is the Resnik measure [14]. Resnik computes the similarity of two terms as the information content IC of their least (or lowest) common ancestor *LCA* as follows:

$$sim(t_1, t_2) = IC(t_i) \text{ ....(1)}$$

where t_i is the least common ancestor LCA of t_1 and t_2; and the IC is computed as follows:

$$IC(t_i)= -\log p(t_i) \ldots\ldots\ldots\ldots\ldots(2)$$

where $P(t_i)$ is the probability of term t_i in the gene ontology annotation (GOA) dataset for this task. In other tasks, the probability of term t_i can be computed differently, for example, from a text corpus. Resnik's measure has been used extensively in the bioinformatics domain [1, 9, 10, 11]. In this task, the probability of a GO term t_x is the number of proteins annotated with the term t_x plus number of proteins annotated with all of its descendants. For example, suppose the term t_x has only two descendants t_i and t_j (i.e., t_x has two child terms that are leaves) and there are 100 proteins annotated with t_x. Also, let there be 200 and 300 proteins annotated with t_i and t_j respectively, and the total number of annotated proteins is10000 in the entire dataset. Then the probability $p(t_x)$ of term t_x *is:* $p(t_x)=(100+200+300)/10000$ = 0.06.

Another important semantic similarity measure is Lin's measure which is also, like Resnik's measure, based on IC [6]:

$$sim(t_1, t_2) = \frac{2*\log(LCA(t_1,t_2))}{\log p(t_1)+\log p(t_2)} \text{ .... (3)}$$

Moreover, Leacock and Chodorow [10] proposed an ontology structure based measure that relies on the distance (*path length*) between nodes and max depth of the ontology structure as follows:

$$sim(t_1, t_2) = -\log \frac{dist(t_1,t_2)}{2*D} \text{ ....(4)}$$

where D is the max depth of the ontology and $dist(t_1, t_2)$ is the shortest path length between t_1 and t_2.

The semantic distance measure of Jiang and Conrath [77] is based on the information content of the nodes and their LCA:

$$Dist(t_1, t_2) = IC(t_1) - IC(t_2) - 2* IC(LCA(t_1, t_2)) \ldots\ldots\ldots(5)$$

where semantic distance *Dist()* can be converted to semantic similarity value using some simple direct mapping function as the semantic similarity is the inverse of semantic distance.

3.2 Path Length Measure

In a given taxonomy or ontology, the basic approach to compute the similarity of two nodes (*concepts*) is using the shortest path length (PL) which is the minimum number of links between the two nodes. Rada et al. (1989) were the first to use path length as a measure of similarity between two concept nodes in a given ontology in the biomedical domain and they applied it to the MeSH ontology [4]; while [5] were the first to apply *PL* as a measure onto the gene ontology GO [5].

In this work, we use *PL* as a similarity method as proposed in our previous work [3, 5, 17] in which the

		iea+	***iea−***
Human	BP	1435	1204
	MF	1421	1268
	CC	1410	1037

(a) Human

		iea+	***iea−***
Yeast	MF	*3753*	*3481*
	CC	4469	4425

(b) Yeast

Table 1: Number of protein-protein interaction pairs used in our evaluation (*positive interactions*) for (a) human and (b) yeast. BP biological process, MF molecular function, and CC cellular component. *iea*+: records with evidence code *iea* are included; *iea*−: records with *iea* evidence code are not included.

similarity *sim*(p_1, p_2) between two proteins p_1 and p_2 is defined as follows:

Sim(p_1,, p_2) = $e^{-f*PL(p_1,p_2)}$ .... (6)

where the path length *PL*(p_1, p_2) between the two proteins p_1, p_2 is computed based on their GO annotation terms, and f is a tuning parameter (*f=0.20* in this evaluation). The path length between two proteins is computed as follows:

$$PL(p_1,p_2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} PL(go_1^i ,go_2^j)}{n* m} \quad \text{............} \ (7)$$

such that: go_1^i and go_2^j are the annotation terms of proteins p_1, and p_2 respectively. Now, the path length PL(go_1, go_2) between two GO terms go_1 and go_2 is computed as follows:

PL(go1, go2) = the minimum path length in the GO graph between the terms go1 and go2

Two proteins p_x, p_y are considered similar if their *Sim()* value, as computed in equation (6), is above certain cutoff value.

3.3 PPI with PL Measure

For a given set of protein-protein interactions PPIs, we compute the similarity of the two proteins in every pair. For a given cutoff value, for example 0.70, if the similarity value of the two proteins in a pair is equal or above the cutoff 0.7 then we consider that these two proteins are similar. The similarity between proteins is computed using GO annotation terms in the three aspects Biological Process BP, Molecular Function MF, and Cellular Component CC. We assume that if two proteins have interaction, then it is highly likely that they exist in similar or same cellular regions, and so, their CC annotation terms should be close thus their semantic similarity using their CC annotation terms should be relatively high. On the other hand, if two proteins have high similarity value based on their CC terms then this indicates that they exist in close areas within the cellular structure and so there is high likelihood that they will interact with each other [1]. Therefore, we expect to observe higher semantic similarity between the proteins that have interactions than randomly selected proteins with no interactions between them.

4. Experiments and Evaluation

The gene ontology downloads and all GO annotations of *yeast* and *human* proteins are downloaded from the official gene ontology site and the SGD website [7, 8]. The datasets of protein interactions PPIs were downloaded from [1] which were extracted from the Database of Interacting Proteins DIP website [1]. The PPI datasets and number of protein pairs in every dataset are presented in Table 1. Since the annotation terms assigned to protein with evidence code *iea* (*inferred from electronic annotation*) are not experimentally proven annotations and many researchers exclude those iea terms, we created two versions of every data set one with *iea* annotations included (we call it *iea*+) and another with excluding all *iea* annotation terms (we call it *iea*−**)**. A dataset of interacting proteins PPIs comprise a set of protein pairs with interaction between proteins in every pairs and thus we call it *positive* PPI dataset. For every positive PPI dataset, we created a dataset of similar size of randomly selected proteins having no interactions between them we call it negative dataset. That is, in the negative set, proteins in every pair are paired randomly with no interactions between them. Table 1 includes number of proteins pairs in every positive PPI dataset. For example, as shown in Table 1, the {human BP iea+}

dataset contains 1435 PPI pairs (pairs of interacting proteins from human proteome) such that each protein in this dataset is annotated with at least one BP term and *iea* terms are included. These protein pairs will be examined for similarity using their BP terms.

Evaluation metrics and settings: for every PPI positive dataset there is also a negative dataset with protein pairs having no interactions. In general, the PPI dataset and the negative dataset are similar size in number protein pairs. For a given cutoff (*threshold*) value t_s , we compute four metrics: true positive TP, true negative TN, false positive FP and false negative FN as follows: TP is number of protein pairs in the positive PPI dataset having similarity equal or greater than t_s; TN: number proteins pairs in the negative set having similarity less than t_s; FP: number of proteins in the negative set having similarity equal or above t_s; and FN: number of proteins in the positive PPI dataset having similarity < t_s. Then precision is: P = $\frac{TP}{TP+FP}$

Recall (also called true positive rate TPR) is:

TPR = R = $\frac{TP}{TP+FN}$

False positive rate: FPR = $\frac{FP}{FP+TN}$

F1-score: $F_1 = \frac{2\,P\,R}{P+R}$

The receiver operating characteristics (ROC) curve represents the relation between FPR and TPR as shown in Figure 1. Then we calculated area under curve (AUC) for every ROC curve. F_1 score, or simply *F-score*, is a commonly used evaluation metric for analyzing the retrieval quality. F-score is basically a harmonic mean of precision and recall; and it is considered more reliable indicator than precision or recall in evaluation. We used these metrics for performance analysis and evaluation of the results of the proposed method.

Experiments and results and discussion: We used six datasets of protein-protein interactions PPIs from human proteome, as shown n Table 1. From yeast proteome, we used four datasets of PPIs as shown in Table 1 (b). We computed protein similarity using GO annotation terms for every protein involved in PPI interactions. We computed the area under ROC curve (AUC) for all experiments and results are in Table 2. In the human PPIs we notice that the best AUC results (0.82 and 0.81) are obtained using the BP terms (Table 2 (a)). The AUC results in the yeast PPIs show that using CC terms produces higher AUC values. The ROC curves for human PPIs experiments are illustrated in Figure 1. Table 3 presents the F-score results for both *human* and *yeast* PPIs using *iea* evidence code (*iea*+) while Table 4 shows the results of *iea*− when records with *iea* codes are excluded . The F-score results of human iea+ experiments with three aspects BP, CC, MF are illustrated in Figure 2. Table 5 contains detailed results of similarity for scoring protein interactions for human proteins using BP terms and excluding *iea* annotations *iea*-. These results, in Table 5, include accuracy, TPR, P, and F-score at every cutoff from 0 to 1.0 with 0.1 step. These results obtained from on 2408 protein pairs where half of them (*1204 protein pairs*) are interacting proteins (*positive PPIs*) and the other half (1204 protein pairs) are non-interacting and randomly selected from human proteins. The best F-score 0.739 was at 0.6 cutoff (Table 5) which makes sense as the similarity value of 0.6 indicates that the two proteins in a PPI pair are fairly similar. Further, at cutoff 0.6 and positive set accuracy is almost 87% which means at 87% of the interacting protein pairs have similarity value of 0.6 of higher. These results support that concept of scoring and ranking protein pairs in support of protein interactions. Moreover, the accuracy of GO annotations of human or yeast proteins have significant impact on the performance of protein similarity.

	GO aspect	***iea+***	***iea−***
Human	BP	0.82	0.81
	MF	0.79	0.76
	CC	0.75	0.78

(a) Human

	GO aspect	***iea+***	***iea−***
Yeast	MF	0.72	0.71
	CC	0.79	0.79

(b) Yeast

Table 2: AUC results of the PL method using (a) Human data, (b) Yeast data

5. Conclusion

We presented an ontology structure based semantic similarity measure to examine the semantic similarity of protein pairs from the *human* and *yeast* PPI networks. It has been shown in the literature that interacting proteins are more similar in terms of their characteristics than randomly selected non-interacting proteins from the proteome. We used the gene ontology as our source of knowledge for measuring protein similarity. We examined the three aspects of GO with and without including the *iea* annotations. The evaluation results are encouraging in the direction of using our similarity measure for scoring and ranking PPI pairs and supporting PPI findings. In the future of this research, we plan to examine other ontology structure based features to augment our measure. For example, we would like to investigate the relative depth of the sub-graph that includes the terms to be measured.

Acknowledgement: We would like to acknowledge the support from PSU for their funding the travel of second author HA for presenting the paper.

iea+		***cutoff***		
		0.6	**0.7**	**0.8**
Human	BP	0.76	0.58	0.29
	MF	0.68	0.55	0.38
	CC	0.67	0.73	0.73
Yeast	BP	0.60	0.52	0.44
	CC	0.67	0.70	0.74

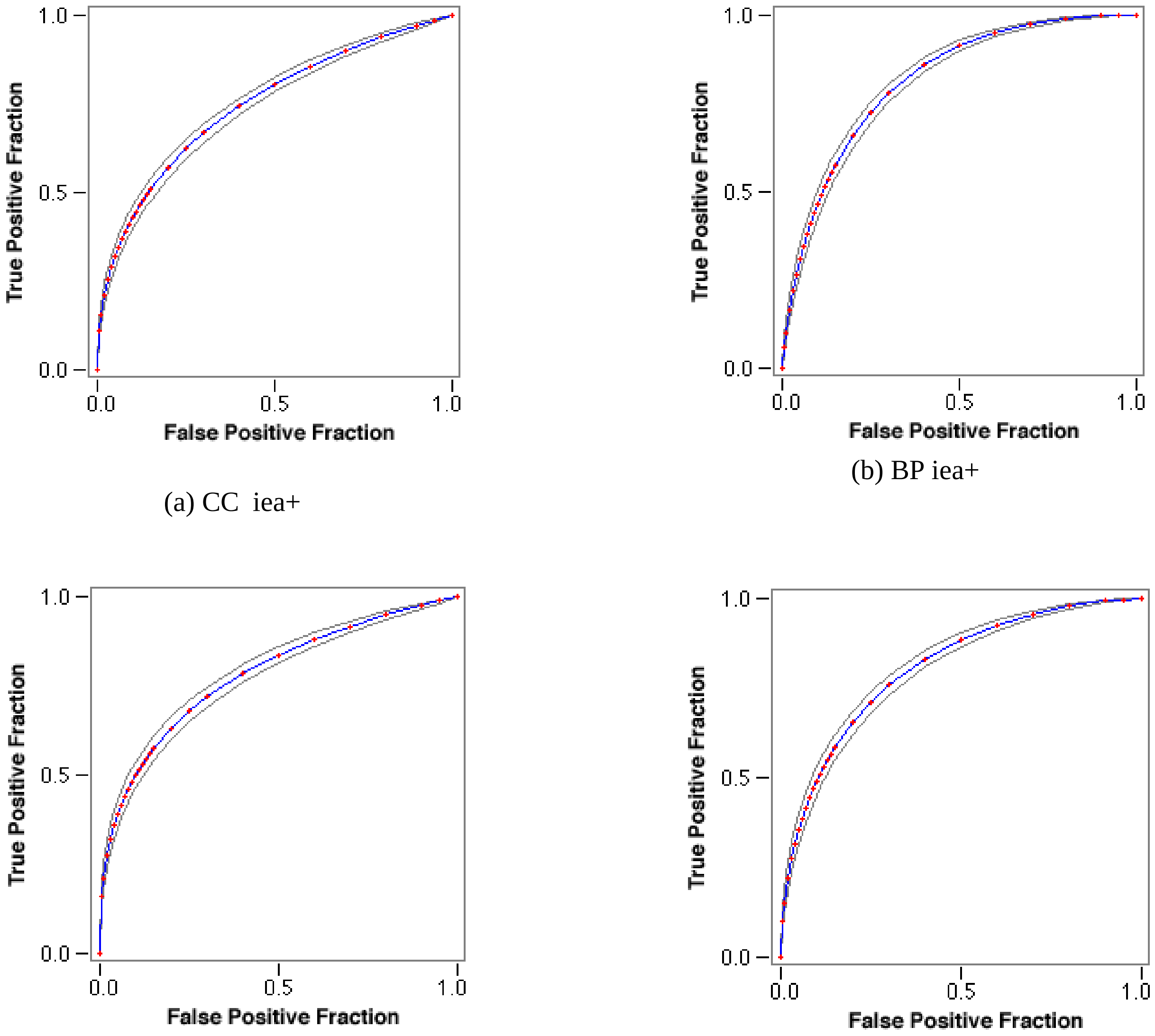
Table 3: F-score results of PL method: iea+

iea−		***cutoff***		
		0.6	**0.7**	**0.8**
Human	BP	0.74	0.57	0.34
	MF	0.68	0.55	0.39
	CC	0.67	0.72	0.74
Yeast	MF	0.60	0.52	0.43
	CC	0.67	0.70	0.74

Table 4: F-score results of PL method: *iea−*

No. of PPI pairs	(1)	(2)	(3)	(4)	(5)	(6)
	cutoff	Pos. Accuracy	Neg. Accuracy	TPR (*recall*)	P	F-Score
Pos = 1204 *Neg = 1204*	0	1.000	0.000	1.000	0.500	0.667
	0.1	1.000	0.000	1.000	0.500	0.667
	0.2	1.000	0.009	1.000	0.502	0.669
	0.3	0.998	0.033	0.998	0.508	0.673
	0.4	0.992	0.085	0.992	0.520	0.682
	0.5	0.982	0.199	0.982	0.551	0.706
	0.6	**0.867**	**0.521**	**0.867**	**0.644**	**0.739**
	0.7	0.422	0.931	0.422	0.860	0.566
	0.8	0.205	0.990	0.205	0.954	0.338
	0.9	0.145	0.998	0.145	0.989	0.253
	1.0	0.120	0.998	0.120	0.986	0.215

Table 5: Detailed results of the experiment *Human BP iea−*. The columns (1) thru (6) are as follows: Column (1) is cutoff, column (2) is the accuracy in positive set (which includes 1204 interacting proteins PPIs), column (3) shows the accuracy of the negative set which includes 1204 non-interacting protein pairs. Columns (4) – (6) show TPR, P, F-score respectively for each cutoff.



(a) CC iea+

(b) BP iea+

(c) CC iea-

(d) BP iea-

Figure 1: Illustration of ROC curves for experiments on human PPI data

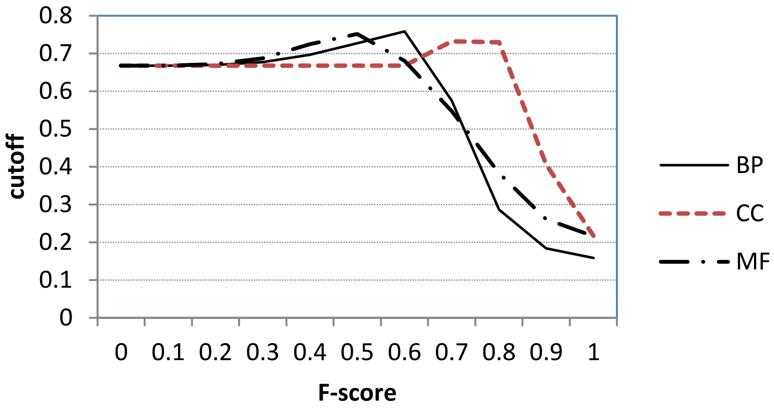


Figure 2: F-score for Human iea+ experiments

References

[1] S. Jain and G. D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. Jain and Bader BMC Bioinformatics, 11:562; 2010.

[2] C. Pesquita, D. Faria1, A.O. Falca, P. Lord, and F.M. Couto1. Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology vol.5, no.7, July 2009.

[3] H. Al Mubaid and A. Nagar. Comparison of four similarity measures based on GO annotations for gene clustering. In proceedings of IEEE Symposium on Computers and Communications ISCC 2008, pp. 531–536, July 2008.

[4] R. Rada, H. Mili, E. Bichnell, and M. Blettner, “Development and application of a metric on semantic nets,” *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 17–30, Jan./Feb. 1989.

[5] Al-Mubaid H. and Nguyen H.A. (2007) “Similarity Computation Using Multiple UMLS Ontologies in a Unified Framework.” Proceedings for the 22nd ACM Symposium on Applied Computing SAC’07, 2007.

[6]. Lin D: An Information-Theoretic Definition of Similarity. Proceedings of the 15th International Conference on Machine Learning Morgan Kaufmann; 1998, 296-304.

[7] Ashburner M et. al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000, 25:25-29.

[8] Saccharomyces Genome Database: http://downloads.yeastgenome.org

[9] Li D, Liu W, Liu Z, Wang J, Liu Q, Zhu Y, He F: PRINCESS, a protein interaction confidence evaluation system with multiple data sources. Mol Cell Proteomics 2008, 7(6):1043-1052.

[10] M. Batet , D. Sanchez, and A. Valls, An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics 44 (2011) 118–125.

[11] Patil A, Nakamura H: Filtering high-throughput protein-protein interaction data using a combination of genomic features. BMC Bioinformatics 2005, 6:100.

[12] Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM: Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 2005, 23(8):951-959.

[13] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droege A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE: A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005, 122(6):957-968.

[14] Resnik, P. (1995). “Using Information Content to Evaluate Semantic Similarity in a Taxonomy.”Proc 14th Int'l Joint Conf Artificial Intelligence.

[15] J. J. Jiang and D.W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proc. of International Conference Research on Computational Linguistics (ROCLING X), 1997.

[16] Xia K, Dong D, Han JDJ: IntNetDB v1.0: an integrated protein-proteininteraction network database generated by a probabilistic model. BMC Bioinformatics 2006, 7:508.

[17] H. Al-Mubaid and A. Nagar, “A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways”, IEEE CBMS, 2008.