Predicting False Positives of Protein-Protein Interaction Data by Semantic Similarity Measures[§]

George Montañez[1] and Young-Rae Cho[*,2]

[1]*Machine Learning Department, Carnegie Melon University, Pittsburgh, PA 15213, USA*

[2]*Bioinformatics Program, Department of Computer Science, Baylor University, Waco, TX 76798, USA*

Abstract: Recent technical advances in identifying protein-protein interactions (PPIs) have generated the genomic-wide interaction data, collectively collectively referred to as the interactome. These interaction data give an insight into the underlying mechanisms of biological processes. However, the PPI data determined by experimental and computational methods include an extremely large number of false positives which are not confirmed to occur in vivo. Filtering PPI data is thus a critical preprocessing step to improve analysis accuracy. Integrating Gene Ontology (GO) data is proposed in this article to assess reliability of the PPIs. We evaluate the performance of various semantic similarity measures in terms of functional consistency. Protein pairs with high semantic similarity are considered highly likely to share common functions, and therefore, are more likely to interact. We also propose a combined method of semantic similarity to apply to predicting false positive PPIs. The experimental results show that the combined hybrid method has better performance than the individual semantic similarity classifiers. The proposed classifier predicted that 58.6% of the S. cerevisiae PPIs from the BioGRID database are false positives.

Keywords: Gene Ontology, protein-protein interactions, semantic similarity.

1. INTRODUCTION

Proteins interact with each other for biochemical stability and functionality, building protein complexes as larger functional units. PPIs therefore play a key role in biological processes within a cell. Recently, high-throughput experimental techniques, such as yeast two-hybrid system [1,2,3,4], mass spectrometry [5,6] and synthetic lethality screening [7], have made remarkable advances in identifying PPIs on a genome-wide scale, collectively referred to as the interactome. Since the evidence of interactions provides insights into the underlying mechanisms of biological processes, the availability of a large amount of PPI data has introduced a new paradigm towards functional characterization of proteins on a system level [8,9].

Over the past few years, systematic analysis of the interactome by theoretical and empirical studies has been in the spotlight in the field of bioinformatics [10,11,12]. In particular, a wide range of computational approaches have been applied to the protein interaction networks for functional knowledge discovery, for instance, function prediction of uncharacterized genes or proteins [13,14,15], functional module detection [16,17,18], and signaling pathway identification [19,20]. Although the automated methods are scalable and robust, their accuracy is limited because of unreliability of interaction data. The PPIs determined by large-scale experimental and computational approaches include an extremely large number of false positives, i.e., a significantly large fraction of the putative interactions detected must be considered spurious because they cannot be confirmed to occur *in vivo* [21,22,23]. Filtering PPI data is thus a critical preprocessing step to improve analysis accuracy when handling interactome. The erroneous interaction data can be curated by other resources which are used to judge the level of functional associations of interacting protein pairs, such as gene expression profiles [24,25].

A recent study [26] has suggested the integration of GO data to assess the validity of PPIs through measuring semantic similarity of interacting proteins. GO [27] is a repository of biological ontologies and annotations of genes and gene products. Although the annotation data on GO are created by the published evidence resulted from mostly unreliable high-throughput experiments, they are frequently used as a benchmark for functional characterization because of their comprehensive information.

Functional similarity between proteins can be quantified by semantic similarity, a function that returns a numerical value reflecting closeness in meaning between two ontological terms annotating the proteins [28]. Since an interaction of a protein pair is interpreted as their strong functional association, one can measure the reliability of protein-protein interactions using semantic similarity: proteins with higher semantic similarity are more likely to interact with each other than those with low semantic similarity. Therefore, absent of true information identifying which proteins actually interact, semantic similarity can be an indirect indicator of such interactions.

*Address correspondence to this author at the Bioinformatics Program, Department of Computer Science, Baylor University, One Bear Place # 97356, Waco, TX 76798, USA; Tel: 1-254-710-3385; Fax: 1-254-710-3889; E-mails: young-rae_cho@baylor.edu, ycho21@yahoo.com

[§]Part of information included in this article has been previously published in Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2012.

1574-8936/13 $58.00+.00

In this article, we assess reliability of PPIs determined experimentally and computationally. The performance of existing semantic similarity measures is analyzed in terms of functional consistency, including the combinations of the measures which achieve improved performance over the previous methods. These semantic similarity measures are applied to identify false positive PPIs in current *S. cerevisiae* PPI databases. The experimental results show that the combined hybrid method has better performance than the individual semantic similarity classifiers. The proposed combined classifier predicted that 58.6% of the *S. cerevisiae* PPIs from the BioGRID database [29] are false positives.

2. METHODS

2.1. Gene Ontology (GO)

An ontology is a formal way of representing knowledge which is described by concepts and their relationships [30]. As a collaborative effort to specify bio-ontologies, GO addresses the need for consistent descriptions of genes and gene products across species [31]. It provides a collection of well-defined biological concepts, called GO terms, spanning three domains: biological processes, molecular functions and cellular components. GO is structured as a Directed Acyclic Graph (DAG) by specifying general-to-specific relationships, such as "is-a" and "part-of", between parent and child terms.

As another intriguing feature, GO maintains annotations for genes and gene products to their most specific GO terms, called direct annotations. Because of general-to-specific relationships in the ontology structure, a gene that is annotated to a specific term is also annotated to all its parent terms on the paths towards the root. These are called inferred annotations. Considering both direct and inferred annotations, we can quantify the specificity of a GO term by the proportion of the number of annotated genes on the term to the total number of annotated genes in the ontology. Suppose G_i and G_j are the sets of genes annotated to the GO term t_i and t_j, respectively, and t_i is a parent term of t_j. The size of G_i, $|G_i|$, is always greater than or equals to $|G_j|$.

Note that a gene can be annotated to multiple GO terms. Suppose a gene x is annotated to m different GO terms. $G_i(x)$ denotes a set of genes annotated to the GO term t_i whose annotation includes x, where $1 \le i \le m$. In the same way, suppose n different GO terms have the annotations including both x and y, where $n \le m$. $G_j(x,y)$ denotes a set of genes annotated to the GO term G_j whose annotation includes both x and y, where $1 \le j \le n$. The minimum size of $G_i(x)$, $\min_i |G_i(x)|$, is then less than or equal to $\min_j |G_j(x,y)|$.

2.2. Semantic Similarity Measures

Semantic similarity measures are the functions computing the level of similarity in meaning between terms within an ontology. A variety of semantic similarity measures have been proposed previously [32,33,34]. They can be grouped into four broad categories: *path length-based methods* (or called *edge-based methods*), *information content-based methods* (or called *annotation-based methods*), *common term-based methods* (or called *node-based methods*) and *hybrid methods*. Path length-based methods calculate the path length between terms in an ontology as their similarity. Information content-based methods use an information-theoretic measure based on the notion of term likelihood to assign higher values to terms that have higher specificity. Common term-based methods consider the number of shared ancestor terms in an ontology to assign a similarity value. Hybrid methods incorporate aspects of two different categories. The semantic similarity measures in these four categories are summarized in Table **1**.

2.2.1. Path Length-Based Methods (Edge-Based Method)

Path length-based methods calculate semantic similarity by measuring the shortest path length between two terms. The path length can be normalized with the maximum depth of the ontology, which represents the longest path length out of all shortest paths from the root to leaf nodes.

$$sim_{path}(C_1, C_2) = -\log\left(\frac{length(C_1, C_2)}{2 \times depth}\right)$$

where $length(C_1,C_2)$ is the shortest path length between two terms C_1 and C_2 in an ontology.

The semantic similarity is also measured by the depth to the most specific common ancestor (SCA) of two terms, i.e. the shortest path length from the root to SCA [35]. The longer the path length to SCA from two terms is, the more similar they are in meaning. Wu and Palmer [36] normalized the depth to the SCA by the average depth to the terms. This normalized measure is used to adjust the similarity distorted through the depths of the terms of interest.

$$sim_{Wu}(C_1, C_2) = \frac{2 \times length(C_{root}, C_{sca})}{length(C_{root}, C_{sca}) + length(C_{root}, C_{sca}) + 2 \times length(C_{root}, C_{sca})}$$

where C_{root} denotes the root term and C_{sca} is the most specific common ancestor term of C_1 and C_2.

To compute functional similarity between two proteins, we take into consideration semantic similarity between pairwise combinations of the terms having direct annotations of the proteins. These path length-based methods are applicable to the well-balanced ontology in which each edge between two terms represents the same quantity of specificity. However, according to published results, new terms are added resulting in complex relationships between terms which lead to inconsistent specificity of edges in GO. Therefore, path length-based methods are not suitable for measuring semantic similarity from GO.

2.2.2. Information Content-Based Methods (Annotation-Based Method)

Self-information in Information Theory is a measure of the information content associated with the outcome of a random variable. The amount of self-information contained in a probabilistic event c depends on the probability $P(c)$ of the event. More specifically, the smaller the probability of the event is, the larger the self-information to be received is when the event indeed occurs. The information content of a term C in an ontology is then defined as the negative log likelihood of C, $-\log P(C)$. In the application to GO, the likelihood of a term $P(C)$ can be calculated by the ratio of

Table 1. Summary of Semantic Similarity Measures in Four Categories. SCA Denotes the Most Specific Common Ancestor of Two Terms of Interest in GO

Category/Method	Description
Path length (edge-based) methods	
Path length	Path length between two terms
Normalized path length	Normalized path length between two terms with depth of GO
Depth to SCA of two terms [35]	Depth of SCA of two terms
Normalized depth to SCA [36]	Normalized depth of SCA with average depth of two terms
Information content-based methods	
Resnik [37]	Information content of SCA of two terms
Lin [38]	Normalized Resnik's method by information contents of two terms
Jiang and Conrath [39]	Sum of differences of information contents between SCA and two terms
Common terms (node-based) methods	
Term overlap (TO) [41]	The number of common ancestors of two terms
NTO [41]	Normalized TO method with the smaller set of ancestors of two terms
simUI/DTO [35]	Normalized TO method with the union set of ancestors of two terms
Hybrid methods	
Wang [42]	Combined method of TO with normalized depth
IntelliGO [43]	Combined method of information content with normalized depth
simGIC [44]	Combined method of simUI with information contents

the number of annotated genes on the term C to the total number of annotated genes in the ontology.

The information content-based semantic similarity is measured by commonality of two terms, i.e. more common information the two terms share, more similar they are. Resnik [37] used the information content of the SCA that subsumes two terms C_1 and C_2.

$$sim_{\mathrm{Re}snik}(C_1, C_2) = -\log P(C_{sca})$$

Lin [38] considered not only commonality but a difference between terms by normalizing the Resnik's semantic similarity measure with the average of the individual information contents of C_1 and C_2.

$$sim_{Lin}(C_1, C_2) = \frac{2 \times \log P(C_{sca})}{\log P(C_1) + \log P(C_2)}$$

Jiang and Conrath [39] used the differences of information contents between C_1 and C_{sca} and between C_2 and C_{sca} to measure the semantic distance between C_1 and C_2.

$$dist_{Jiang}(C_1, C_2) = 2 \times \log P(C_{sca}) - \log P(C_1) - \log P(C_2)$$

The semantic similarity between C_1 and C_2 can then be calculated by inversing their semantic similarity.

$$sim_{Jiang}(C_1, C_2) = \frac{1}{1 + dist_{Jiang}(C_1, C_2)}$$

Note that all methods in the path length-based and information content-based categories measure semantic similarity between two GO terms. We however aim at quantifying functional similarity between two proteins which might be annotated to multiple GO terms. We therefore apply three different ways of aggregating semantic similarity values between pairwise combinations of the terms having annotations of the two proteins. Suppose S_1 and S_2 are the sets of GO terms having direct annotations of protein g_1 and protein g_2, respectively. At first, in order to compute functional similarity between two proteins g_1 and g_2, we can select the maximum semantic similarity value among all similarity values of term pairs from S_1 and S_2.

$$sim_{MAX}(g_1, g_2) = \max_{C_1 \in S_1, C_2 \in S_2} sim(C_1, C_2)$$

Next, the average sematic similarity value of all possible pairwise combinations of the terms from S_1 and S_2 can be

Finally, by combining the two used as the functional similarity of g_1 and g_2.

$$sim_{AVG}(g_1, g_2) = \frac{1}{|S_1| \times |S_2|} \sum_{C_1 \in S_1, C_2 \in S_2} sim(C_1, C_2)$$

methods above, the best-match average (BMA) approach computes the average of all pairwise best-matches [40].

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

2.2.3. Common Term-Based Methods (Node-Based Method)

Common term-based methods calculate semantic similarity by measuring the overlap between two sets of

terms, not between two terms. The methods in this category are therefore applied directly to estimating functional similarity between two annotating proteins. As a general measure of this category, the Term Overlap (TO) method counts the GO terms having direct and inferred annotations of both protein g_1 and protein g_2. More common GO terms which g_1 and g_2 are annotated to, higher functional similarity they have. Suppose S_1 and S_2 are the sets of GO terms having both direct and inferred annotations of g_1 and g_2, respectively.

$$sim_{TO}(g_1, g_2) = |S_1 \cap S_2|$$

This approach can be normalized by the union of the two sets of GO terms [35] or by the smaller set of them [41].

$$sim_{UI}(g_1, g_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

$$sim_{NTO}(g_1, g_2) = \frac{|S_1 \cap S_2|}{\min(|S_1|, |S_2|)}$$

The Direct Term Overlap (DTO) measure uses the formula of sim_{UI} (g_1, g_2) with the sets of GO terms having only direct annotations of g_1 and g_2, instead of S_1 and S_2, respectively.

2.2.4. Hybrid Methods

The approaches from different categories can be combined to compute semantic similarity. For example, Wang *et al.* [42] proposed a semantic similarity measure that integrates the Normalized Term Overlap (NTO) with the concept of the normalized depth to the most specific terms in an ontology. IntelliGO [43] is a vector representation model that combines the normalized depth with information contents as weights. However, as discussed, the path length-based approaches do not fit in the GO applications because of complex relationships between terms.

SimGIC [44] integrates the information theoretic measures with term overlaps. It calculates the sum of the information contents in the intersection of S_1 and S_2 divided by the sum of the information contents in the union of them, where S_1 and S_2 are the sets of GO terms having both direct and inferred annotations of g_1 and g_2.

$$sim_{GIC}(g_1, g_2) = \frac{\sum_{C_a \in S_1 \cap S_2} \log P(C_a)}{\sum_{C_b \in S_1 \cup S_2} \log P(C_b)}$$

where $P(C)$ is the likelihood of the term C, i.e. the ratio of the number of annotated genes on the term C to the total number of annotated genes in the ontology.

As another way of integrating the measures from two different categories, we apply a linear combination. For example, we can combine the Resnik's information content-based method with the DTO method in common term-based approaches such as

$$sim_{LC}(g_1, g_2) = \alpha \cdot sim_{Resnik-MAX}(g_1, g_2) + (1-\alpha) \cdot sim_{DTO}(g_1, g_2)$$

where α is a weighting parameter used to assign relative weight to the contributions from both similarity measures. This linear combination (LC) method takes advantage of two orthogonal sources of information: direct annotation term information and the information content of the most specific common term. By considering two distinct sources of information, a more accurate picture of semantic similarity is attained. Since the path length-based methods suffer from the inconsistency of term specificity represented by each edge in GO as discussed previously, we did not choose any measure from that category.

2.3. Classification of PPIs

The false positive interactions can be identified by evaluating how dissimilar each interacting protein pair is semantically. We thus adopt the semantic similarity measures discussed in the previous section. The semantic similarity scores were then subjected to a variable threshold. When the score for an interacting protein pair exceeds the threshold, the corresponding PPI is classified as a true (positive) interaction. Otherwise, it is classified as a false (negative) interaction.

In addition to the semantic similarity classifiers, we propose an additional 'voting' scheme of the combined hybrid method, which only outputs a positive classification when the Resnik-MAX measure exceeds the threshold and the score from DTO is above the median DTO value for the data set. Mathematically, this voting classifier is formulated as follows:

$$C(g_1, g_2) = (sim_{Resnik-MAX}(g_1, g_2) > \theta) \wedge (sim_{DTO}(g_1, g_2) > \beta)$$

where θ is the threshold parameter and β is the median DTO semantic similarity score of the data set. The output of $C(g_1,g_2)$ is restricted to the set {0,1} (binary output) due to the nature of logical conjunction. This method was developed to further reduce the number of identifying false PPIs over most threshold values.

3. EXPERIMENTAL RESULTS

3.1. Evaluation of Semantic Similarity

To compare the performance of the semantic similarity measures, we assessed general correlation of semantic similarity with functional consistency. We downloaded the genome-wide PPI data set of *S. cerevisiae* from the BioGRID database [29] and selected 10,000 interacting protein pairs uniformly at random. The semantic similarity scores have been calculated for each pair using all methods listed in Table **1**.

As a reference ground-truth data set, we used manually curated functional categorizations (FunCat) from the MIPS database [45]. Since the functional categories are hierarchically distributed, we extracted the functional descriptions and their annotations on the third level from the root of the hierarchy. We then computed functional consistency from the FunCat data by taking the number of shared functions for a protein pair divided by the size of the union of their function sets (i.e., the jaccard index). Pearson correlation is then calculated between each semantic similarity score and the functional consistency.

Table **2** lists the Pearson correlation results for the tested semantic similarity measures. We observed that the

combined methods in the hybrid category, such as simGIC and LC, achieved higher correlation with the functional consistency than the other measures. In particular, the linear combination (LC) method of DTO and Resnik-MAX using an α weighting of 0.15 shows the best correlation.

Table 2. Pearson Correlation Results for Semantic Similarity Measures with Functional Consistency on MIPS Functional Categorizations

Semantic Similarity Measures	Pearson Correlation
Resnik-MAX	0.3774
Resnik-BMA	0.5286
Lin-MAX	0.2448
Lin-BMA	0.5162
DTO	0.7683
NTO	0.6726
simGIC	0.7703
LC ($\alpha = 0.10$)	0.7733
LC ($\alpha = 0.15$)	**0.7742**
LC ($\alpha = 0.25$)	0.7715
LC ($\alpha = 0.50$)	0.7215
LC ($\alpha = 0.75$)	0.5815

Fig. (**1**) graphically shows the correlation between the semantic similarity from various measures and the functional consistency. The semantic similarity values for each method were binned and the average functional consistency was taken for each bin. As can be seen, two hybrid methods, simGIC and LC, and a common term-based method, DTO, measured the semantic similarity which have strongly positive correlations with the functional consistency on the MIPS functional categorizations because their plots are close to the diagonal line.

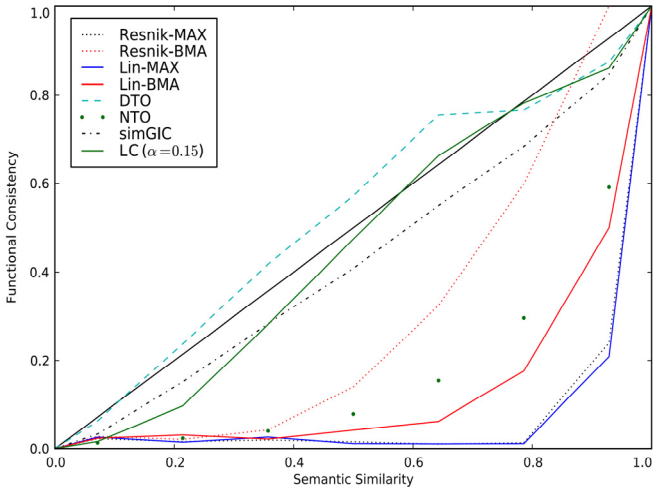

Fig. (1). Correlation plots between semantic similarity from various measures and functional consistency on MIPS functional categorizations. Semantic similarity measured by two hybrid methods (simGIC and LC) and a common term-based method (DTO) has strongly positive correlations with functional consistency because their plots are close to the diagonal line.

3.2. Identification of False PPIs

The genome-wide PPI data of several model organisms are publicly available in a number of open databases, for example, BioGRID [29], IntAct [46], MINT [47], MIPS [48], STRING [49] and DIP [50]. Because they were mostly generated by high-throughput experimental and computational methods, we presume that they contain a significant number of false positives. To test false positive identification, we calculated the semantic similarity using the measures discussed previously for 10,000 PPIs randomly selected from the BioGRID database. All methods were implemented for one hundred different thresholds ranging from 0.00 to 0.99.

To compare the performance of false positive identification, we used as ground-truth any non-empty intersection of functions for two interacting proteins within the MIPS functional categorizations. When a protein pair share at least one functional categorization, they are assumed to interact with each other. Accuracy was then calculated as the number of correct classifications divided by the total number of classifications.

Of the 10,000 PPIs assessed, a majority of them (5,554) are expected to be false interactions as measured by the MIPS ground-truth data set. These interacting protein pairs have no shared functional categorizations, and therefore, are labeled as negative examples. Table **3** shows the classification accuracy for the tested semantic similarity classifiers. The most accurate method for PPI classification is the LC classifier of DTO and Resnik-MAX measures, using an α value of 0.90, which achieves a maximum accuracy of 0.82 over the data set. Equally important is the area under curve, which gives an indication of how accurate the various methods are over all thresholds. The combined voting method achieves the largest area under curve, with a value of 0.76. In addition to this result, it also achieves the second best maximum accuracy, behind the LC classifier with $\alpha = 0.90$ and $\alpha = 0.75$. The combined hybrid methods collectively achieve the best performance on the classification task, with the voting method performing well for almost all thresholds.

Lin's method has the worst performance on the classification task, with the lowest maximum accuracy of all methods tested. DTO appears to trade good performance over many thresholds (area under curve) for maximum classification accuracy, as does NTO. The simGIC measure achieves fairly good performance, with the second best area under curve performance. Since it is also a hybrid method combining the information contents with term overlaps, similar to the linear combination method that achieves the best performance, this provides additional evidence for the performance advantages of using common term-based methods in combination with information content-based methods.

Fig. (**2**) plots the accuracy curves for the LC classifier using several different α weighting values. As expected, the curves begin similar to DTO when the α value is low, since it places more weight on semantic similarity values given by the DTO measure. At $\alpha = 1.0$, the curve is identical to that of

Table 3. Classification Accuracy for Semantic Similarity Classifiers

Classifier	Maximum Accuracy	Area Under Curve
Resnik-MAX	0.8087	0.5348
Resnik-BMA	0.7671	0.5989
Lin-MAX	0.6478	0.4970
Lin-BMA	0.7528	0.5686
DTO	0.7573	0.6519
NTO	0.7636	0.6348
simGIC	0.7892	0.6689
LC ($\alpha = 0.10$)	0.7670	0.6393
LC ($\alpha = 0.15$)	0.7723	0.6336
LC ($\alpha = 0.25$)	0.7810	0.6221
LC ($\alpha = 0.50$)	0.8020	0.5932
LC ($\alpha = 0.75$)	0.8135	0.5643
LC ($\alpha = 0.90$)	**0.8163**	0.5469
Voting	0.8114	**0.7606**

the Resnik-MAX measure. The classifier achieves the maximum accuracy over the tested data set when the α weighting is near 0.9.

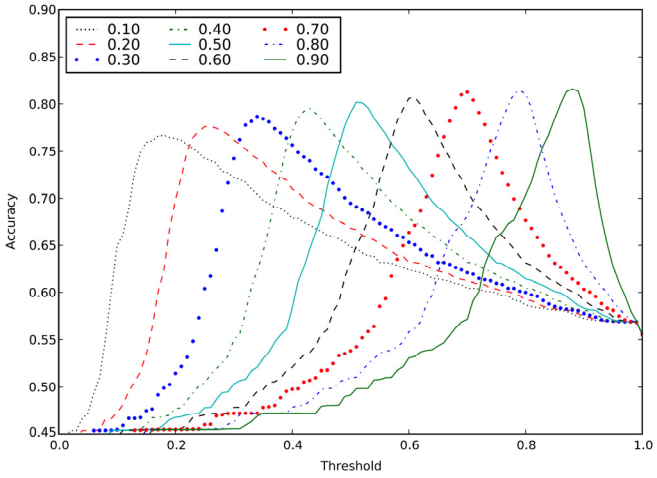

Fig. (2). Classification accuracy of the linear combination (LC) classifier for nine different weighting values of α. The LC classifier achieves the maximum accuracy over the tested data set when the α weighting value is near 0.9.

Fig. (**3**) shows the classification accuracy results for DTO, Resnik-MAX and the combined voting classifiers. Different from the DTO and Resnik-MAX measures, the voting classifier is able to achieve high classification accuracy across all threshold values. By forcing both sub-classifiers to agree on a positive classification, false positives are avoided, leading to higher accuracy given the large percentage of negatively labeled instances in the data set.

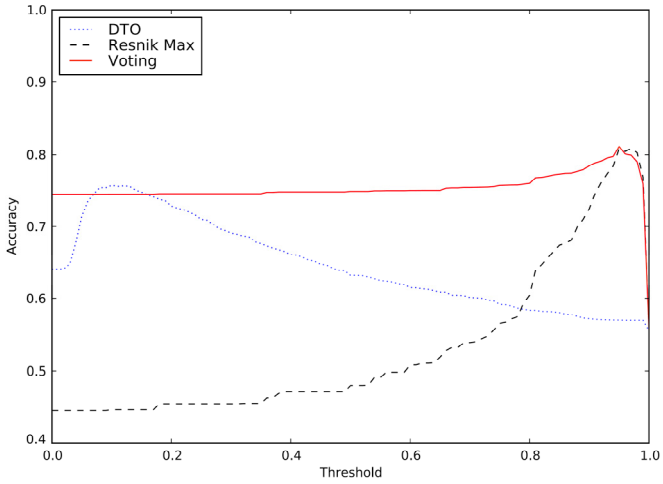

Fig. (3). Classification accuracy over all thresholds for Resnik-MAX, DTO and the Combined Voting classifier. Different from the DTO and Resnik-MAX measures, the voting classifier is able to achieve high classification accuracy across all threshold values.

3.3. Reliability of PPI Data

We extend the classification task to assess the reliability of current PPI data. Using the most accurate parameters for the LC classifier of DTO and Resnik-MAX measures (α = 0.90, threshold = 0.88), we classified all *S. cerevisiae* PPIs in the BioGRID database. As a preprocessing step, we excluded those that lacked corresponding gene annotations within the GO annotation data of *S. cerevisiae*. This resulted in a total of 247,048 interactions, of which 144,677 (58.6%) were classified as false positive interactions.

The PPIs in the BioGRID database have been determined by several different experimental systems. Among the experimental systems, Negative Genetic (0.47%) and Affinity-Capture-MS (0.15%) were the most prevalent in generating false positives. False interactions were most likely to result from genetic experiment types (73%) and high-throughput methods (90%). Table **4** displays an ordered ranking of the experimental systems responsible for the majority of false positive data.

Using the combined semantic similarity classifier, we are able to discover potential false positives existing in PPI data repositories and automate the process of filtering PPI data sets. Given a high accuracy of classification when calibrated against manually curated functional categorization data from the MIPS database (roughly 0.82% accuracy), it is likely that many of the false positive interactions identified by the classifier indeed represent spurious PPIs. Table **5** lists a random sampling of twenty negatively classified PPIs having a zero semantic similarity value as measured by the combined hybrid classifier, which are therefore likely to represent false positive interactions.

4. CONCLUSION

PPIs are crucial resources for functional knowledge discovery. However, as an innate feature, the PPI data sets

Table 4. Experimental System Types and the Proportions of False Positives in the *S. Cerevisiae* PPI Data Set

Experimental System	Number of False Positives	% of Total
Negative Genetic	67,723	0.47
Affinity Capture-MS	21,027	0.15
Positive Genetic	12,078	0.08
Synthetic Growth Defect	11,025	0.08
Synthetic Lethality	6,390	0.04
Two-hybrid	4,847	0.03
Biochemical Activity	4,015	0.03
Affinity Capture-RNA	3,461	0.02
PCA	2,897	0.02
Phenotypic Enhancement	2,485	0.02
Phenotypic Suppression	2,385	0.02
Affinity Capture-Western	1,578	0.01
Synthetic Rescue	1,403	0.01
Dosage Rescue	1,396	0.01
Others	1,967	0.01

Table 5. Twenty PPIs with Zero Valued Semantic Similarity (Likely False PPIs)

Protein A	Protein B	Experimental System
YDR124W	YOR158W	Affinity Capture-MS
YGL122C	YJL107C	Affinity Capture-RNA
YGL122C	YML118W	Affinity Capture-RNA
YJR059W	YER010C	Biochemical Activity
YNL307C	YBR225W	Biochemical Activity
YHR082C	YML083C	Biochemical Activity
YMR216C	OK/SW-cl.3	Biochemical Activity
YOL090W	YGL081W	Negative Genetic
YEL051W	YKL098W	Negative Genetic
YBL015W	YDL118W	Negative Genetic
YDL074C	YMR206W	Negative Genetic
YHR167W	YDR249C	Negative Genetic
YPR078C	YDR488C	Negative Genetic
YGR012W	YLR053C	Negative Genetic
YDR542W	YKL109W	Negative Genetic
YCR091W	YJL147C	Negative Genetic
YNL197C	YOL036W	Negative Genetic
YOR043W	YGR161C	Negative Genetic
YDR388W	YJR083C	Protein-peptide
YMR186W	YER039C-A	Synthetic Growth Defect

include an extremely large number of false positives. Our results indicate that more than 50% of current *S. cerevisiae* PPI data are false positives, determined by mostly high-throughput experimental systems. Identifying the false positive interactions is thus a critical preprocessing step for accurate analysis of PPIs. The work presented in this article focuses on using the ontology structures and annotations from GO to automatically prune false positives from the PPI data sets.

Several semantic similarity methods were assessed for their correlation to manually curated MIPS functional categorizations. A hybrid method by the linear combination was presented that demonstrates performance gains over existing methods. This method takes into account both the maximum information content of the most specific common ancestor as well as the overlap of directly annotated terms in the GO for a pair of genes. Although the individual method, in isolation, is less accurate for classification, it can improve the performance when combined in a majority-vote fashion. It was motivated by the idea that two separate low-accuracy classifiers can become more accurate when combined in a suitable manner. An additional 'voting' variant was also presented that achieves the best overall classification accuracy over a variety of selection thresholds.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation Graduate Research Fellowship (Grant No. 0750271), the Ford Foundation Predoctoral Fellowship program, and the Young Investigator Development Program grant by the Vice Provost for Research at Baylor University.

REFERENCE

[1] Uetz P, Giot L, Cagney G, *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000; 403: 623-627.

[2] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001; 98(8): 4569-4574.

[3] Giot L, Bader JS, Brouwer C, *et al.* A protein interaction map of Drosophila melanogaster. Science 2003; 302: 1727-1736.

[4] Li S, Armstrong CM, Bertin N, *et al.* A map of the interactome network of the metazoan C. elegans. Science 2004; 303: 540-543.

[5] Gavin AC, Bösche M, Krause R, *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002; 415: 141-147.

[6] Ho Y, Gruhler A, Heilbut A, *et al.* Systematic identication of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 2002; 415: 180-183.

[7] Tong AH, Lesage G, Bader GD, *et al.* Global mapping of the yeast genetic interaction network. Science 2004; 303: 808-813.

[8] Rual JF, Venkatesan K, Hao T, *et al.* Towards a proteome-scale map of the human protein-protein interaction network. Nature 2005; 437: 1173-1178.

[9] Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol 2005; 23(5): 561-566.

[10] Stelzl U, Worm U, Lalowski M, *et al.* A human protein-protein interaction network: a resource for annotating the proteome. Cell 2005; 122: 957-968.

[11] Yu H, Braun P, Yildirim MA, *et al.* High-quality binary protein interaction map of the yeast interactome network. Science 2008; 322: 104-110.

[12] Venkatesan K, Rual JF, Vazquez A, *et al.* An empirical framework for binary interactome mapping. Nat Method 2009; 6(1): 83-90.

[13] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol 2007; 3: 88.

[14] Chen X, Liu M, Ward R. Protein function assignment through mining cross-species protein-protein interactions. PLoS One 2008; 3(2): e1562.

[15] Cho Y-R, Zhang A. Predicting function by frequent functional association pattern mining in protein interaction networks. IEEE Trans Inform Technol Biomed (TITB) 2010; 14(1): 30-36.

[16] Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 2003; 100(21): 12123-12128.

[17] Luo F, Yang Y, Chen C-F, Chang R, Zhou J, Scheuermann RH. Modular organization of protein interaction networks. Bioinformatics 2007; 23(2): 207-214.

[18] Song J, Singh M. How and when should interactome-derived clusters be used to predict functional modules and protein function? Bioinformatics 2009; 25(23): 3143-3150.

[19] Scott J, Ideker T, Karp RM, Sharan R. Ecient algorithms for detecting signaling pathways in protein interaction networks. J Comput Biol 2006; 13(2): 133-144.

[20] Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics 2007; 8: 335.

[21] von Mering C, Krause R, Snel B, *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002; 417: 399-403.

[22] Salwinski L, Eisenberg D. Computational methods of analysis of protein-protein interactions. Curr Opin Struct Biol 2003; 13: 377-382.

[23] Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein-protein interaction data? J Mol Biol 2003; 327: 919-923.

[24] Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res 2002; 12: 37-46.

[25] Bader JS, Chaudhuri A, Rothberg JM, Chant J. Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 2004; 22(1): 78-85.

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

[27] The Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 2010; 38: D331-D335.

[28] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 2003; 19(10): 1275-1283.

[29] Stark C, Breitkreutz BJ, Chatr-Aryamontri A, *et al.* The BioGRID interaction database: 2011 update. Nucleic Acids Res 2011; 39: D698-D704.

[30] Bard JBL, Rhee SY. Ontologies in biology: design, applications and future challenges. Nat Rev Genet 2004; 5: 213-222.

[31] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nat Genet 2000; 25: 25-29.

[32] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007; 40: 288-299.

[33] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol 2009; 5(7): e1000443.

[34] Wang J, Zhou X, Zhu J, Zhou C, Guo Z. Revealing and avoiding bias in semantic similarity scores for protein pairs. BMC Bioinformatics 2010; 11: 290.

[35] Guo X, Liu R, Shriver CD, Hu H, Liebman MN. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 2006; 22(8): 967-973.

[36] Wu Z, Palmer M. Verb semantics and lexical selection. Proceedings of 32th Annual Meeting of the Association for Computational Linguistics 1994; 133-138.

[37] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of 14th International Joint Conference on Artificial Intelligence 1995; 448-453.

[38] Lin D. An information-theoretic definition of similarity. Proceedings of 15th International Conference on Machine Learning (ICML) 1998; 296-304.

[39] Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of 10th International Conference on Research in Computational Linguistics 1997.

[40] Tao Y, Sam L, Li J, Friedman C, Lussier YA. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 2007; 23: i529-i538.

[41] Mistry M, Pavlidis P. Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics 2008; 9: 327.

[42] Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007; 23(10): 1274-81.

[43] Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes M-D. IntelliGO: a new vector-based semantic similarity measure including annotation origin. BMC Bioinformatics 2010; 11: 588.

[44] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcao AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 2008; 9(Suppl 5): S4.

[45] Ruepp A, Zollner A, Maier D, *et al.* The FunCat: a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 2004; 32(18): 5539-5545.

[46] Aranda B, Achuthan P, Alam-Faruque Y, *et al.* The IntAct molecular interaction database in 2010. Nucleic Acids Res 2010; 38: D525-D531.

[47] Ceol A, Chatr-aryamontri A, Licata L, *et al.* MINT: the molecular interaction database: 2009 update. Nucleic Acids Res 2010; 38: D532-D539.

[48] Mewes HW, Dietmann S, Frishman D, *et al.* MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res 2008; 36: D196-D201.

[49] von Mering C, Jensen LJ, Kuhn M, *et al.* STRING7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 2007; 35: D358-D362.

[50] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res 2004; 32: D449-D451.

Received: November 30, 2011 Revised: December 30, 2011 Accepted: January 15, 2012