Assessing Reliability of Protein-Protein Interactions by Gene Ontology Integration

George D. Montañez
Machine Learning Department
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
gmontane@cs.cmu.edu

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

***Abstract*—Recent advances in genome-wide identification of protein-protein interactions (PPIs) have produced an abundance of interaction data which give an insight into functional associations among proteins. However, it is known that the PPI datasets determined by high-throughput experiments or inferred by computational methods include an extremely large number of false positives. Using Gene Ontology (GO) and its annotations, we assess reliability of the PPIs by considering the semantic similarity of interacting proteins. Protein pairs with high semantic similarity are considered highly likely to share common functions, and therefore, are more likely to interact. We analyze the performance of existing semantic similarity measures in terms of functional consistency and propose a combined method that achieves improved performance over existing methods. The semantic similarity measures are applied to identify false positive PPIs. The classification results show that the combined hybrid method has higher accuracy than the other existing measures. Furthermore, the combined hybrid classifier predicts that 59.6% of the *S. cerevisiae* PPIs from the BioGRID database are false positives.**

***Keywords*-protein-protein interactions; Gene Ontology; semantic similarity; direct term overlap;**

I. Introduction

Protein-protein interactions (PPIs) play a key role in biological processes within a cell. Recent high-throughput experimental and computational methods of discovering PPIs have resulted in an increase in raw data indicating potentially shared functions among proteins [1], [2], [3]. Availability of the interactome, a set of PPIs on a genome-wide scale, has thus introduced a new paradigm towards functional characterization of proteins on a system level. The automated methods of interaction inference, however, can result in a significant number of false positives, i.e., a large fraction of the putative interactions detected must be considered spurious because they cannot be confirmed to occur in vivo [4], [5]. These erroneous data can be curated by other resources which describe the level of functional associations of interacting proteins.

Previous research [6] has suggested using Gene Ontology (GO) to assess the validity of PPIs through measurement of the semantic similarity between proteins. GO [7] is a repository of biological ontologies and annotations of genes and gene products. Although the annotation data are based on the published evidence derived from mostly unreliable high-throughput experiments, they are frequently used as a benchmark for functional characterization because of their comprehensiveness.

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 [8]. Since an interaction of a protein pair is interpreted as their strong functional association, one can measure the reliability of proposed PPIs using semantic similarity: proteins with higher semantic similarity are more likely to interact via a PPI than those with low similarity. Therefore, absent of true information identifying which proteins actually interact, semantic similarity can serve as an indirect indicator of such interactions. Although several novel measures of assessing semantic similarity (and, by extension, functional similarity) have been proposed over the past few years [9], [10], [11], recent research suggests that the Resnik's original method [12] based on information content calculation for the most specific common terms remains the most accurate [6].

In assessing the reliability of proposed PPIs using semantic similarity, we make use of GO annotation data experimentally determined and computationally inferred. While using inferred annotation data to prune inferred protein-protein interactions (as is done here) may strike some as circular, the resulting bias is in the direction of confirming the validity of PPIs, such that true interactions are unlikely to be classified as false, at the expense letting some false PPIs go undetected. One can therefore perform valuable pruning, using existing data sources, without a high risk of misclassifying true interactions. While acknowledging the shortcomings of such an approach, this allows leveraging of freely-available GO data to potentially improve the reliability of PPI datasets.

The work presented here is to determine the validity of PPIs in *S. cerevisiae* having the proteins annotated to GO. We first review several semantic similarity measures and introduce a combined semantic similarity method that presents a simple means of improving the performance of existing semantic similarity measures. We then describe two sets of experiments and their results, the first used to assess the reliability of the various similarity measures by comparing their correlation to manually curated data, and the other focused on classification of PPIs using semantic similarity measures. Finally, we briefly

978-1-4673-1191-5/12/$31.00 ©2012 IEEE

discuss the efficiency of the combined similarity measure.

II. Summary of Semantic Similarity Measures

Semantic similarity methods produce a number indicating level of similarity between terms, such as those in the GO. These methods can be grouped into a few broad categories: *path length-based methods*, *information content-based methods*, *common term-based methods* and *hybrid methods*. Path length-based methods compute the path length between terms in an ontology as their similarity. Information content-based methods use the notion of *term likelihood* which defines specificity of terms within an ontology, and convert this into an information measure. Common term-based methods consider the number of shared terms in an ontology to assign a similarity value. Hybrid methods incorporate aspects of these (and possibly other) categories. The semantic similarity measures in these four categories are summarized in Table I.

*a) **Path Length Methods**:* Path length-based methods calculate the semantic similarity between two proteins by measuring the shortest path length in an ontology between the terms that each protein is annotated to. The path length between two terms can be normalized by the depth of the ontology, which represents the longest path length among all shortest paths from the root to leaf nodes.

The semantic similarity is also measured by the shortest path length from the root to the most specific common ancestor (SCA) of the terms that each gene is annotated to [13]. The longer the path length to SCA, the more similar the two terms are in meaning. This method can be normalized by the average depth of the terms that each protein is annotated to.

These path length-based methods are applicable to a well-balanced ontology in which each edge represents the same quantity of specificity. However, since GO has been structured by adding new terms in a random fashion, the path length-based methods are not suitable for measuring semantic similarity from GO.

*b) **Information Content Methods**:* Let

$$p(c) = \frac{|\text{Annotations(c)}|}{|\text{Annotations(Root)}|} \tag{1}$$

where the Annotations function returns a set of all annotating proteins to a term, and Root denotes the root domain of the ontology (i.e. biological process, molecular function or cellular component.)

Resnik's method [12] is representative of the information content-based methods. Within the GO, Resnik's method can be calculated as follows, using the alternative *one minus likelihood* measure discussed in [12] so that all similarity values are mapped to the range $[0, 1]$:

$$Sim_{\text{Resnik}}(C_1, C_2) = 1 - p(\text{SCA}) \tag{2}$$

where SCA is the most specific common ancestor of terms C_1 and C_2.

Lin's method [15] is an information content-based method that takes into account the information content of the individual terms, as well that of the most specific common ancestor. It is defined as:

$$Sim_{\text{Lin}}(C_1, C_2) = \frac{2 \log p(\text{SCA})}{\log p(C_1) + \log p(C_2)} \tag{3}$$

Both Resnik's and Lin's methods are defined in reference to terms, whereas we seek to calculate the semantic similarity between two proteins which may be annotated to multiple terms each. We therefore consider two methods of aggregating annotation term values: MAX and BMA. Suppose S_a and S_b are the sets of terms that proteins a and b are annotated to, respectively. MAX chooses the maximum semantic similarity value between any two terms in S_a and S_b. For BMA, the average of all pairwise best matches between S_a and S_b is used [21].

*c) **Common Term Methods**:* Common term-based methods calculate the semantic similarity between two proteins by measuring the overlap between the term sets each protein is annotated to. This includes terms that a protein is directly annotated to and may include parents of such terms as well, depending on the method.

Indirect Term Overlap (TO) [17] considers the number of common direct and ancestor terms between two annotation sets as a measure of similarity. It is defined as:

$$Sim_{\text{TO}}(g_a, g_b) = |S_a \cap S_b|, \tag{4}$$

where S_a and S_b are the sets of direct and ancestor terms that proteins g_a and g_b are annotated to, respectively.

Indirect Normalized Term Overlap (NTO) [17] considers the number of common direct and ancestor terms between two terms, normalized by the smaller of the two sets. It is defined as:

$$Sim_{\text{NTO}}(g_a, g_b) = \frac{|S_a \cap S_b|}{min(|S_a|, |S_b|)}, \tag{5}$$

where S_a and S_b are again the sets of direct and ancestor terms proteins g_a and g_b are annotated to.

SimUI [13] is similar to Indirect Normalized Term Overlap, but normalization is done using the union of both sets. It is defined as:

$$Sim_{\text{UI}}(g_a, g_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}. \tag{6}$$

In contrast to TO, NTO, and simUI, Direct Term Overlap (DTO) only considers the overlap of terms that proteins are directly annotated to, disregarding ancestors terms. Since annotation data in the GO is only given for the most specific terms of annotation (and we then infer annotation to all parents), it follows that any two proteins sharing a common, maximally specific annotation term are likely to share a common function. DTO is defined as:

$$Sim_{\text{DTO}}(g_a, g_b) = \frac{|D_a \cap D_b|}{|D_a \cup D_b|}, \tag{7}$$

where D_a and D_b are the sets of terms that proteins g_a and g_b are directly annotated to, respectively.

TABLE I
SUMMARY OF SEMANTIC SIMILARITY MEASURES. SCA DENOTES THE MOST SPECIFIC COMMON ANCESTOR OF TWO TERMS OF INTEREST IN GO.

Category	Method	Description
Path Length	Path length	Path length between two terms
	Normalized path length	Normalized path length between two terms by depth of GO
	Depth of SCA [13]	Depth of SCA of two terms
	Normalized depth of SCA [14]	Normalized depth of SCA by average depth of two terms
Information Contents	Resnik [12]	Information content of SCA of two terms
	Lin [15]	Normalized Resnik's method by information contents of two terms
Common Terms	Term overlap (TO) [17]	The number of ancestors of two terms
	NTO [17]	Normalized TO method by the smaller set of ancestors of two terms
	simUI [13]	Normalized TO method by the union set of ancestors of two terms
Hybrid Methods	simGIC [18]	Combined method of simUI with information contents
	Wang [19]	Combined method of TO with normalized depth
	IntelliGO [20]	Combined method of information content with normalized depth
	TCSS [6]	Normalized Resnik's method by clustering GO terms

*d) **Hybrid Methods**:* Hybrid methods combine the approaches from different categories to compute semantic similarity. For example, SimGIC [18] integrates the information theoretic measures with overlap measures. It calculates the sum of information contents in the intersection of S_a and S_b divided by the sum of information contents in the union of them.

$$Sim_{\text{GIC}}(g_a, g_b) = \frac{\sum_{t_1 \in S_a \cap S_b} \log p(t_1)}{\sum_{t_2 \in S_a \cup S_b} \log p(t_2)}, \quad (8)$$

where $p(c)$ follows the definition of Equation 1.

Wang et al. [19] proposed another hybrid method that integrates the term overlap (TO) measure with the concept of the normalized depth to the most specific terms in an ontology. IntelliGO [20] is a vector representation model that combines the normalized depth with information contents as weights. Jain and Bader [6] introduced a novel approach, called TCSS, which applies clustering of similar GO terms to find a subgraph and measures semantic similarity by Resnik's method considering whether two terms are located in the same subgraph. This method attempts to solve the problem of unequal depth in different branches of GO.

We consider a novel combined method that aggregates the semantic similarity as calculated by Resnik's method with that of DTO. It is defined as:

$$Sim_{\text{com}}(g_a, g_b) = \alpha Sim_{\text{Resnik-MAX}}(g_a, g_b) + (1-\alpha) Sim_{\text{DTO}}(g_a, g_b) \quad (9)$$

where α is a weighting parameter used to assign relative weight to the contributions from both similarity measures. The α parameter can be found using a separate training dataset with standard optimization methods such as particle-swarm optimization [22]. For the results presented here, a disjoint training dataset of equal size to the test dataset was created, and α was selected for best performance on the training dataset for each particular task. The trained α values were then used for subsequent experiments on the test dataset. For additional detail, see the Appendix.

Our combined hybrid method takes advantage of two orthogonal sources of information: direct annotation information and the information content from the most specific common ancestor of two terms. 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 use any measure from that category. To form our hybrid measure, we therefore chose DTO, being the best of the common term-based methods, and Resnik-MAX, a standard information content-based method. Our experimental results confirm the performance advantage of using the two orthogonal information sources chosen for the hybrid classifier, in both classification accuracy and correlation with independent data sources.

III. EVALUATION OF SEMANTIC SIMILARITY

A. Correlation with Functional Categorizations

The semantic similarity measures discussed in the previous section can be used to evaluate the reliability of PPIs. In order to compare the performance of the measures, we assessed general correlation of semantic similarity with functional consistency. We first downloaded the genome-wide PPI dataset of *S. cerevisiae* from BioGRID [23] and selected 10,000 interacting protein pairs uniformly at random. The semantic similarity scores were calculated for each pair using all methods.

As a reference ground-truth set, we used manually curated MIPS functional categorizations (FunCat) [24]. Since the MIPS 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 was then calculated between each semantic similarity score and the ground-truth functional consistency.

Table II contains the Pearson correlation results for the tested measures and the MIPS functional categorization values. We found that the combined measure achieved top correlation performance (along with simGIC), using a trained α weighting of 0.15. Other combined methods were also tested, using different combinations of base semantic similarity methods,

TABLE II
CORRELATION BETWEEN SEMANTIC SIMILARITY AND FUNCTIONAL CONSISTENCY FROM MIPS FUNCTIONAL CATEGORIZATIONS.

Semantic Similarity Method	Pearson Correlation
Resnik-MAX	0.3774
Resnik-BMA	0.5286
Lin-MAX	0.2448
Lin-BMA	0.5162
NTO	0.6726
DTO	0.7683
simGIC	0.7703
Combined ($\alpha = 0.15$)	**0.7742**

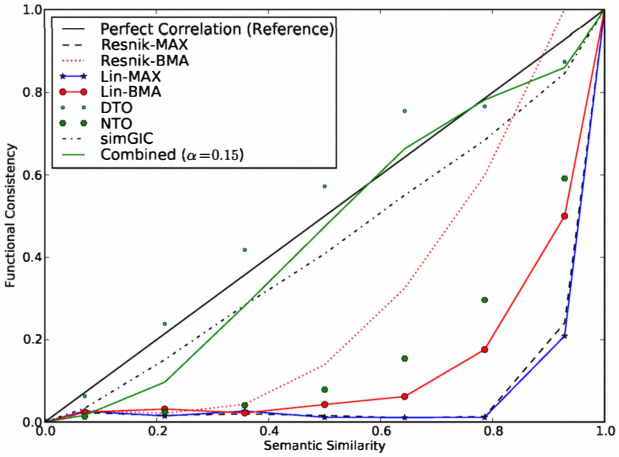

Fig. 1. Correlation between MIPS functional consistency and semantic similarity measures.

and none had performance exceeding that of the DTO/Resnik-MAX combined method (results not shown in Table II).

The second measure with high correlation was simGIC, another hybrid method. It integrates term specificity with the term overlap concept, similar to our combined measure. By including two orthogonal sources of information, the hybrid methods appear to gain more information overall concerning the functional similarity of proteins. Therefore, those combined hybrid methods represent the best choice for evaluating reliability of the PPIs generated from high-throughput experiments.

Figure 1 graphically shows the correlation between MIPS functional consistency and the various semantic similarity measures. The semantic similarity values for each method were binned and the average functional consistency was taken for each bin. As can be seen, the combined method has highest correlation with MIPS functional consistency because its plot is closest to the diagonal line. DTO and simGIC also have fairly good correlation.

B. Verification of True PPIs

The method of using semantic similarity to identify valid protein interaction requires that semantic similarity be higher for proteins that interact than for proteins that do not. To test the veracity of this assumption, we conducted an additional experiment using the manually curated DIP core PPI dataset. The core PPIs have been collected by two forms of curative processes: RNA expression profiles and paralogous verification. An interaction was selected from the full PPI dataset if the putative interacting pair have high cohesiveness of their RNA expression profiles and they have paralogs that also interact. It demonstrated that these methods identified true interactions with high selectivity [25].

TABLE III
AVERAGE SEMANTIC SIMILARITY FOR CORE AND NON-CORE PPIS

Method	Core	Non-Core	P-Value
Resnik-MAX	0.9214	0.8763	2.36E-061
Resnik-BMA	0.5259	0.4626	1.30E-100
Lin-MAX	0.9674	0.9377	9.39E-039
Lin-BMA	0.7338	0.6840	7.35E-059
DTO	0.2388	0.1277	4.61E-179
NTO	0.6763	0.5262	2.27E-282
simGIC	0.3319	0.2054	1.74E-182
Combined ($\alpha = 0.10$)	0.3071	0.2025	3.21E-184
Combined ($\alpha = 0.15$)	0.3412	0.2400	2.51E-186
Combined ($\alpha = 0.25$)	0.4095	0.3148	6.00E-189
Combined ($\alpha = 0.50$)	0.5801	0.5020	1.21E-177
Combined ($\alpha = 0.75$)	0.7508	0.6892	1.86E-127
Combined ($\alpha = 0.80$)	0.7849	0.7266	6.95E-114
Combined ($\alpha = 0.85$)	0.8190	0.7640	4.69E-100
Combined ($\alpha = 0.90$)	0.8532	0.8015	1.97E-086
Combined ($\alpha = 0.95$)	0.8873	0.8389	2.01E-073

For our experiment, two disjoint PPI datasets were created. The first consisted of 5,692 core PPIs and the second consisted of 8,701 non-core PPIs, which were generated by using the BioGRID dataset and removing all PPIs that occurred within the core DIP dataset. Since the core PPI dataset contains interactions that are more likely to be valid, we hypothesized a higher average semantic similarity for this dataset. As can be seen in Table III, the core dataset had average semantic similarity values that were consistently higher than the non-core set, to a statistically significant degree. Significance was measured using a two-tailed, unequal variance t-test [26], with resulting p-values far below 0.05. Therefore, the assumption that higher semantic similarity is associated with true interaction can be taken as a reasonable premise, given the data.

IV. IDENTIFICATION OF FALSE PPIS

A. Classification Method

To identify false positive interactions, we use as ground truth the non-empty intersection of functions for two proteins within the MIPS functional categorizations. When two proteins share a functional categorization, the pair is presumed to interact, which becomes more likely as the functions shared become more specific. Therefore, using functional categorizations at the third level of the hierarchy represents a reasonable starting point for assessing ground truth in the absence of more reliable indicators.

Semantic similarity values by all methods were calculated for the 10,000 PPIs randomly selected from BioGRID. These values were then subjected to a variable threshold. When the similarity value exceeds the threshold, the semantic similarity method classifies the PPI as a positive (true) interaction. Otherwise, the PPI is classified as a false interaction. All methods were tested for one hundred different thresholds

ranging from 0.0 to 0.99. Accuracy was calculated as the number of correct classifications divided by the total number of classifications.

In addition to the semantic similarity methods described in the previous section, we created an additional 'voting' scheme of the combined hybrid classifier, which only outputs a positive classification when the Resnik-MAX measure exceeds the threshold and the DTO value is above the median DTO value of a separate training dataset. Mathematically, the voting classifier is defined as follows:

$$C(g_a, g_b) = (Sim_{\text{Resnik-MAX}}(g_a, g_b) > \theta) \wedge (Sim_{\text{DTO}}(g_a, g_b) > \beta) \quad (10)$$

where θ is the threshold parameter, β is the median DTO semantic similarity value of the training dataset, and $Sim_{\text{Resnik-MAX}}(g_a, g_b)$ is the semantic similarity of genes g_a and g_b using the Resnik-MAX method. The output of $C(g_a, g_b)$ 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 false positive identifications over most threshold values.

B. Classification Accuracy

Of the 10,000 PPIs assessed, a majority of them (5,554) are expected to be false PPIs as measured by the MIPS ground truth dataset. These have no shared functional categorizations, and therefore, are labeled as negative examples. Table IV shows the classification accuracy for the semantic similarity measures. The most accurate methods for PPI classification are the combined (DTO/Resnik-MAX) classifier using a trained α value of 0.83 and the Resnik-MAX classifier, which achieve maximum accuracies of 0.82 and 0.81 over the dataset, respectively. 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, it also achieves a high maximum accuracy, slightly lower than the combined aggregate hybrid classifier. Therefore, the combined hybrid classifier achieves the best performance on the classification task, similar to Resnik-Max but classifying roughly 1% more PPIs correctly, 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. SimGIC achieves fairly good performance, with the second best area under curve performance. Since it is also a hybrid method combining information from common ancestor relationships with term overlap, similar to the combined method that achieves the best performance, this provides additional evidence for the performance advantages of using term overlap methods in combination with information content-based methods.

Figure 2 plots the accuracy curves for the combined hybrid 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 the similarity values given by the DTO measure. At $\alpha = 1.0$, the curve is identical to that of Resnik-MAX, and the classifier achieves maximum accuracy over our dataset when the α weighting is near 0.9.

TABLE IV
CLASSIFICATION ACCURACY FOR SEMANTIC SIMILARITY CLASSIFIERS

Method	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
NTO	0.7636	0.6348
DTO	0.7573	0.6519
simGIC	0.7892	0.6689
Combined ($\alpha = 0.83$)	**0.8157**	0.5552
Voting	0.8134	**0.7606**

Figure 3 shows the classification accuracy results for DTO, Resnik-MAX and the combined voting classifiers. The combined voting classifier is able to achieve high classification accuracy for 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 dataset.

C. Estimating the Percentage of False Positives in PPI Data Repositories

Using the most accurate trained parameters for the combined hybrid classifier ($\alpha = 0.83$, threshold $= 0.82$), we classified all PPIs within the *S. cerevisiae* PPI dataset from BioGRID. 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 PPIs, of which 147,151 (59.6%) were classified as false positive interactions. Of the indicated false positive interactions, Negative Genetic (47%) and Affinity-Capture-MS (15%) were the most prevalent among experimental systems used. False interactions were most likely to result from genetic experiment types (73%) and high-throughput methods (90%). Table V displays an ordered ranking of the experimental systems responsible for the majority of false positive classifications.

Using this classifier, we are able to discover likely false positive interactions within existing data repositories and automate the process of PPI pruning to eliminate false interactions. Our results indicate a high percentage of false positives within current *S. cerevisiae* PPI data, resulting largely from high throughput methods of interaction discovery. Given a high accuracy of classification when calibrated against manually curated MIPS ground-truth data (roughly 82% accuracy), it is likely that many of the false positive interactions identified by the semantic similarity classifier indeed represent spurious protein-protein interactions. Table VI lists a random sampling of twenty negatively classified PPIs with zero semantic similarity as measured by the combined hybrid classifier, which are therefore likely to represent false positive interactions.

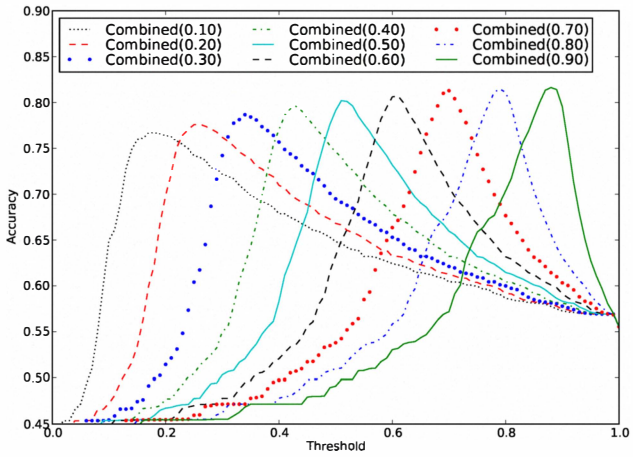

Fig. 2. Classification accuracy of the combined hybrid classifier for nine α weighting values.

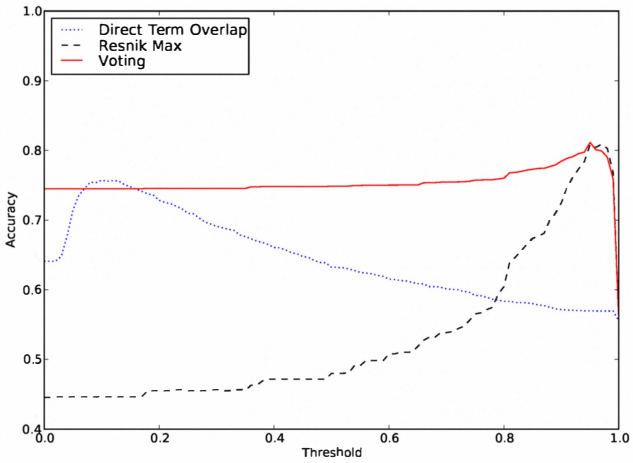

Fig. 3. Classification accuracy over all thresholds for Resnik-MAX, DTO and the Combined Voting classifier.

V. Efficiency

The combined semantic similarity method presented in this paper is a simple modification to Resnik's measure that improves correlation and classification accuracy. Furthermore, the 'voting' extension that produces superior classification accuracy is another efficient extension of the basic Resnik-MAX method. To compute the voting classifier, the Resnik-MAX value is determined and two additional steps are performed:

1) Direct Term Overlap classification: This step runs in time linear to the size of S_a and S_b, which is less than or equal to twice the number of terms in the GO. If n is the number of GO terms, the additional time for this step is $O(n)$.
2) Thresholding and Voting: This step takes constant time.

Therefore, the combined methods are efficient and have a runtime comparable to the Resnik-MAX method.

TABLE V
System Types for False Positive Classification in *S. Cerevisiae* PPI Dataset

Experimental System	Number of False Positives	% of Total
Negative Genetic	68,637	46.6
Affinity Capture-MS	21,294	14.5
Positive Genetic	12,240	8.3
Synthetic Growth Defect	11,296	7.7
Synthetic Lethality	6,755	4.6
Two-hybrid	4,985	3.4
Biochemical Activity	4,144	2.8
Affinity Capture-RNA	3,488	2.4
PCA	2,688	1.8
Phenotypic Enhancement	2,497	1.7
Phenotypic Suppression	2,393	1.6
Affinity Capture-Western	1,668	1.1
Dosage Rescue	1,519	1.0
Synthetic Rescue	1,451	1.0
Others	2,096	1.5

TABLE VI
Sampling of 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

VI. Conclusion

Protein-protein interactions (PPIs) are crucial resources for functional knowledge discovery. However, as an innate feature, the PPI datasets include an extremely large number of false positives. Identifying the false positive interactions is thus a critical preprocessing step for accurate analysis of PPIs. The work presented here focuses on using the ontology structures and annotations from Gene Ontology (GO) to automatically prune false positives from the PPI datasets. Several semantic similarity methods were assessed for their correlation to manually curated MIPS functional categorizations, and a combined hybrid method was presented that demonstrates performance gains over existing methods. An additional 'voting' variant was also developed that achieves the best overall classification accuracy for a variety of selection thresholds.

Our novel method is motivated by the idea that separate low-accuracy classifiers can become accurate when combined in a synergistic manner. This concept underlies popular methods in machine learning such as Boosting [27] and Random

Forests [28], as well as collaborative crowd-sourcing repositories such as Wikipedia. By only classifying a PPI as a positive interaction when two separate classifiers agree on a positive classification, one avoids many false positives, at the cost of missing some true interactions. However, since high-throughput interaction data suffer from an abundance of false positive interactions, the methods presented here have the potential to improve the accuracy of PPI classifications.

APPENDIX
TRAINING OF THE α PARAMETER

To determine the α values for the combined classifier, a separate training set was created by drawing 20,000 PPIs uniformly at random from BioGRID, then retaining 10,000 of those not present in the test dataset. This resulted in two disjoint datasets of equal size (10,000 PPIs in each), one used for training and the other for testing. The combined classifier was then evaluated over the training dataset for the following α values: 0.05, 0.10, 0.15, 0.25, 0.50, 0.60, 0.70, and 0.75-0.99 (in increments of 0.01). The α value with the best performance on the training dataset was chosen for each particular task. The trained α values were then used for subsequent experiments on the test dataset. The results reported in Section IV are over the test dataset, with α values found using the training dataset.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 0750271 and the Ford Foundation Predoctoral Fellowship program. Additional support was provided by a grant from the Young Investigator Development Program (YIDP) by the Vice Provost for Research at Baylor University. The authors would also like to thank Dr. Walter Makous of the University of Rochester and Dr. Larry Wasserman of Carnegie Mellon University for their helpful suggestions concerning this research.

REFERENCES

[1] Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M., "Protein interaction networks from yeast to human," *Current Opinion in Structural Biology*, vol. 14, pp. 292–299, 2004.

[2] Salwinski, L. and Eisenberg, D., "Computational methods of analysis of protein-protein interactions," *Current Opinion in Structural Biology*, vol. 13, pp. 377–382, 2003.

[3] Shoemaker, B.A. and Panchenko, A.R., "Deciphering protein-protein interactions. Part 2. Computational methods to predict protein and domain interaction partners," *PLoS Computational Biology*, vol. 3, no. 4, p. e43, 2007.

[4] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399–403, 2002.

[5] Sprinzak, E., Sattath, S. and Margalit, H., "How reliable are experimental protein-protein interaction data?" *Journal of Molecular Biology*, vol. 327, pp. 919–923, 2003.

[6] 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.

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

[8] 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.

[9] Pedersen, T., Pakhomov, S.V.S., Patwardhan, S. and Chute, C.G., "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of Biomedical Informatics*, vol. 40, pp. 288–299, 2007.

[10] 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.

[11] 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.

[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] 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.

[14] Wu, Z. and Palmer, M., "Verb semantics and lexical selection," in *Proceedings of 32th Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 133–138.

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

[16] 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.

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

[18] 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.

[19] 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, pp. 1274–1281, 2007.

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

[21] 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.

[22] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm Intelligence*, vol. 1, pp. 33–57, 2007.

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

[24] 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.

[25] Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D., "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular and Cellular Proteomics*, vol. 1.5, pp. 349–356, 2002.

[26] Casella, G., Berger, R.L., *Statistical Inference*, ser. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[27] Freund, Y. and Schapire, R.E., "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, 1999.

[28] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, 10.1023/A:1010933404324. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324