# ESTIMATING SUPPORT FOR PROTEIN-PROTEIN INTERACTION DATA WITH APPLICATIONS TO FUNCTION PREDICTION

Erliang Zeng
*Bioinformatics Research Group (BioRG), School of Computing and Information Sciences*
*Florida International University*
*Miami, FL 33199, USA*
*Email: ezeng001@cs.fiu.edu*

Chris Ding
*Department of Computer Science and Engineering, University of Texas at Arlington*
*Arlington, TX 76019, USA*
*Email: CHQDing@uta.edu*

Giri Narasimhan*
*Bioinformatics Research Group (BioRG), School of Computing and Information Sciences*
*Florida International University*
*Miami, FL 33199, USA*
*Email: giri@cs.fiu.edu*

Stephen R. Holbrook†
*Computational Research Division, Lawrence Berkeley National Laboratory*
*Berkeley, CA 94720, USA*
*Email: srholbrook@lbl.gov*

Almost every cellular process requires the interactions of pairs or larger complexes of proteins. High throughput protein-protein interaction (PPI) data have been generated using techniques such as the *yeast two-hybrid systems*, *mass spectrometry method*, and many more. Such data provide us with a new perspective to predict protein functions and to generate protein-protein interaction networks, and many recent algorithms have been developed for this purpose. However, PPI data generated using high throughput techniques contain a large number of false positives. In this paper, we have proposed a novel method to evaluate the support for PPI data based on gene ontology information. If the semantic similarity between genes is computed using gene ontology information and using Resnik's formula, then our results show that we can model the PPI data as a mixture model predicated on the assumption that true protein-protein interactions will have higher support than the false positives in the data. Thus semantic similarity between genes serves as a *metric of support* for PPI data. Taking it one step further, new function prediction approaches are also being proposed with the help of the proposed metric of the support for the PPI data. These new function prediction approaches outperform their conventional counterparts. New evaluation methods are also proposed.

## 1. INTRODUCTION

Protein-protein interactions (PPI) are essential for cellular activities considering the fact that almost every biological function requires the cooperation of many proteins. Recently, many high-throughput methods have been developed to detect pairwise protein-protein interactions. These methods include the yeast two-hybrid approach, mass spectrometry techniques, genetic interactions, mRNA coexpression, and *in silico* methods[1]. Among them, the yeast two-hybrid approach and mass spectrometry techniques aim to detect physical binding between proteins.

The huge amount of protein-protein interaction data provide us with a means to begin elucidating protein function. Functional annotation of proteins is a fundamental problem in the post-genomic era. To date, a large fraction of the pro-

*Corresponding author.
†Corresponding author.

teins have no assigned functions. Even for one of the most well-studied organisms such as *Saccharomyces cerevisiae*, about a quarter of the proteins remain uncharacterized[2].

There are several functional annotation systems. These annotation systems include COGs (Clusters of Orthologous Groups)[3], Funcat (Functional Catalogue)[4] and GO (Gene Ontology)[5]. GO is the most comprehensive system and is widely used. In this paper, we will focus on functional annotations based on GO terms associated with individual genes and proteins.

A lot of previous work has been done on protein function prediction by using the recently available protein-protein interaction data (see review by Sharan *et al.*[2]). The simplest and most direct method for function prediction determines the function of a protein based on the known function of proteins lying in its neighborhood in the PPI network. Schwikowski *et al.*[6] used the so-called majority-voting technique to predict up to three functions that are frequently found among the annotations of its network neighbors. Hishigaki *et al.*[7] approached this problem by also considering the background level of each function across the whole genome. The $\chi^2$-like score was computed for every predicted function. Hua *et al.*[8] proposed to improve the prediction accuracy by investigating the relation between network topology and functional similarity.

In contrast to the local neighborhood approach, several methods have been proposed to predict functions using global optimization. Vazquez *et al.*[7] and Nabieva *et al.*[9] formulated the function prediction problem as a minimum multiway cut problem and provided an approximation algorithm to this NP-hard problem. Vazquez *et al.*[7] used a simulated annealing approach and Nabieva *et al.*[9] applied a integer programming method. Karaoz *et al.*[10] used a similar approach but handled one annotation label at a time. Several probabilistic models were also proposed for this task such as the *Markov random field model* used by Letovsky *et al.*[11] and Deng *et al.*[12], and a statistical model used by Wu *et al.*[13].

Despite some successful applications of the aforementioned algorithms in functional annotation of uncharacterized proteins, many challenges still remain. One of the big challenges is that PPI data has a high degree of noise[1]. Most methods that generate interaction networks or perform functional prediction do not have a preprocessing step to clean the data or filter out the noise. Although some methods include the reliability of experimental sources as suggested by Nabieva *et al.*[14], the reliability estimations are crude and do not consider the variations in the reliability of instances within the same experimental source. Some approaches were proposed to predict protein-protein interaction based on evidence from multi-source data. The evidence score calculated from multi-source data is a type of reliability measure of the protein-protein interaction data. Such approaches include those developed by Jansen *et al.*[15], Bader *et al.*[16], Zhang *et al.*[17], Ben-Hur *et al.*[18], Lee *et al.*[19], Qi *et al.*[20], and many more. Jansen *et al.*[15] combined multiple sources of data using a Bayes classifier. Bader *et al.*[16] developed statistical methods that assign a confidence score to every interaction. Zhang *et al.*[17] predicted co-complexed protein pairs by constructing a decision tree. Ben-Hur *et al.*[18] used kernel methods for predicting protein-protein interactions. Lee *et al.*[19] developed a probabilistic framework to derive numerical likelihoods for interacting protein pairs. Qi *et al.*[20] used a *Mixture-of-Experts* method to predict the set of interacting proteins. The challenges of integrating multi-source data are manily due to the heterogeneity of the data and the effect of a functionally-biased reference set. Another problem is that most multi-source data are unstructured but often correlated.

Another important shortcoming of most function prediction methods is that they do not take all annotations and their relationships into account. Instead, they have either used arbitrarily chosen functional categories from one level of annotation hierarchy or some arbitrarily chosen so-called informative functional categories based on some *ad hoc* thresholds. Such arbitrarily chosen functional categories only cover a small portions of the whole annotation hierarchy, making the predictions less comprehensive and hard to compare. Predicting functions using the entire annotation system hierarchy is necessary and is a main focus of this paper.

In this paper, we propose a method to address the above two problems. We hypothesize that the distribution of similarity values of pairs of proteins can be modeled as a sum of two log-normal distributions (i.e., a mixture model) representing two popu-

lations – one representing pairs of proteins that interact with high support (high confidence), and the other representing pairs that interact with low support (low confidence) (section 2.2). The parameters of the mixture model were then estimated from a large database. This mixture model was then used to differentiate interactions with high confidence from the ones that have low confidence, and was integrated into the function prediction methods. A new evaluation method was also proposed to evaluate the predictions (section 2.4). The new evaluation method captures the similarity between GO terms and reflects the relative hierarchical positions of predicted and true function assignments.

Note that while PPI data involves proteins, GO terms are associated with genes and their products. For the rest of this paper, we will use the terms *genes* and their associated *proteins* interchangeably.

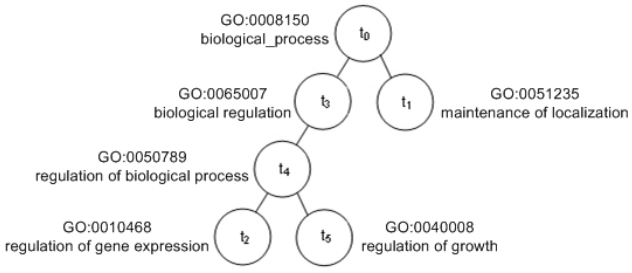


**Fig. 1.** An example showing the hierarchy of sample GO terms.

## 2. METHODS

In this section, we first introduce the concepts of similarity between genes calculated based on gene ontology. Next, we investigate inherent properties of some previously known methods used to calculate such similarity. Then a mixture model is introduced to model the distribution of the similarity values between pairs of genes. Next, we present the new function prediction methods using this mixture model. Finally, we present improved evaluation methods for function prediction.

### 2.1. Similarity between Genes Based on Gene Ontology Data

Suppose that a gene $A$ is associated with the following GO terms $\{t_{a1}, ..., t_{ai}\}$, and that a gene $B$ is associated with the following GO terms $\{t_{b1}, ..., t_{bj}\}$, The similarity between genes $A$ and $B$ based on gene ontology is defined as

$$sim_X(A, B) = \max_{i,j}\{sim_X(t_{ai}, t_{bj})\}. \quad (1)$$

where $sim_X(t_{ai}, t_{bj})$ is the similarity between the GO terms $t_{ai}$ and $t_{bj}$ using method $X$.

Thus, in order to calculate the similarity between genes, we need to calculate the similarity between individual GO terms, for which many methods have been proposed. Below we discuss the methods proposed by Resnik[21], Jiang and Conrath[22], Lin[23], and Schlicker *et al.*[24]. The methods proposed by Resnik, Jiang and Conrath, and Lin have been used in other domain and was introduced to this area by Lord *et al.*[25].

**Resnik**:

$$sim_R(t_1, t_2) = \max_{t \in S(t_1,t_2)} \{IC(t)\} \quad (2)$$

**Jiang-Conrath**:

$$dist_{JC}(t_1, t_2) = \min_{t \in S(t_1,t_2)} \{IC(t_1) + IC(t_2) - 2IC(t)\} \quad (3)$$

**Lin**:

$$sim_L(t_1, t_2) = \max_{t \in S(t_1,t_2)} \left\{ \frac{2IC(t)}{IC(t_1) + IC(t_2)} \right\} \quad (4)$$

**Schlicker**:

$$sim_S(t_1, t_2) = \max_{t \in S(t_1,t_2)} \left\{ \frac{2IC(t)}{IC(t_1) + IC(t_2)}(1 + IC(t)) \right\}. \quad (5)$$

Here $IC(t)$ is the information content of term $t$:

$$IC(t) = -\log(p(t)), \quad (6)$$

where $p(t)$ is defined as $freq(t)/N$, $freq(t)$ is the number of genes associated with term $t$ or with any child term of $t$ in the data set, $N$ is total number of genes in the genome that have at least one GO term associated with them, and $S(t_1, t_2)$ is the set of common subsumers of the terms $t_1$ and $t_2$. Note that the Jiang-Conrath proposal uses the complementary concept of distance instead of similarity.

The basic objective of these methods is to capture the specificity of each GO term and to calculate the similarity between GO terms in a way that reflects their positions in the GO hierarchy. However, as discussed below, we argue that the methods of Lin and Jiang-Conrath are not best suited for this purpose. For example, consider the non-root

terms $t_2$ (GO:0010468) and $t_3$ (GO:0065007) in Figure 1. Then $dist_{JC}(t_2, t_2) = dist_{JC}(t_3, t_3) = 0$, and $sim_L(t_2, t_2) = sim_L(t_3, t_3) = 1$. In other words, the methods of Lin and Jiang-Conrath cannot differentiate between two pairs of genes, one of which is associated with the term $t_2$ (GO:0010468), and the other with $t_3$ (GO:0065007) because it ignores the fact that $t_2$ (GO:0010468, regulation of gene expression) is more specific than $t_3$ (GO:0065007, biological regulation). In contrast, $sim_R(t_2, t_2) = -\log p(t_2) > sim_R(t_3, t_3) = -\log p(t_3)$, if $t_2$ is more specific than $t_3$, thus reflecting the relative positions (and the specificities) of $t_2$ and $t_3$ in the GO hierarchy. For example, in *Saccharomyces cerevisiae*, genes *YCR042C* and *YMR227C* encode TFIID subunits. Both are annotated with GO terms GO:0000114 (G1-specific transcription in mitotic cell cycle) and GO:0006367 (transcription initiation from RNA polymerase II promoter). According to the definition, $sim_L(\text{YCR042C}, \text{YMR227C}) = 1$ and $dist_{JC}(\text{YCR042C}, \text{YMR227C}) = 0$. Now consider another pair of genes *YCR046C* and *YOR063W*, both of which encode components of the ribosomal large subunits, however, one is mitochondrial and the other cytosolic. Both are annotated with the GO term GO:0006412 (translation). Again, according to the definition, $sim_L(\text{YCR046C}, \text{YOR063W}) = 1$ and $dist_{JC}(\text{YCR046C}, \text{YOR063W}) = 0$. Thus, we have

$$\begin{aligned} &sim_L(\text{YCR042C}, \text{YMR227C}) \\ &\quad = sim_L(\text{YCR046C}, \text{YOR063W}) = 1, \end{aligned}$$

and

$$\begin{aligned} &dist_{JC}(\text{YCR042C}, \text{YMR227C}) \\ &\quad = dist_{JC}(\text{YCR046C}, \text{YOR063W}) = 0. \end{aligned}$$

But clearly, the annotations of genes YCR042C and YMR227C are much more specific than the annotations of genes YCR046C and YOR063W. So the similarity between genes YCR042C and YMR227C should be greater than the similarity between genes YCR046C and YOR063W. The similarity between genes calculated by the method of Resnik reflects this fact, since

$$\begin{aligned} &sim_R(\text{YCR042C}, \text{YMR227C}) \\ &\quad = -\log p(\text{GO} : 0000114) = 9.69 \\ &\quad > sim_R(\text{YCR046C}, \text{YOR063W}) \\ &\quad = -\log p(\text{GO} : 0006412) = 4.02. \end{aligned}$$

## 2.2. Mixture Model and Parameter Estimation

The contents of this entire subsection are among the novel contributions of this paper.

As mentioned earlier, PPI data generated using high throughput techniques contain a large number of false positives[1]. Thus the PPI data set contains two groups, one representing true positives and the other representing false positives. However, differentiating the true and false positives in a large PPI data set is a big challenge due to the lack of good quantitative measures. An *ad hoc* threshold can be used for such measures. Our proposed method avoids such choices. Instead, we propose a mixture model to differentiate the two groups in a large PPI data set. One group contains pairs of interacting proteins that have strong support, the other of pairs of interacting proteins that have weak or unknown support. Here we hypothesize that the similarity between genes based on Gene Ontology using the method of Resnik (see Eq.(2)) helps to differentiate between the two groups in the PPI data. We conjecture that the true positives will have higher gene similarity values than the false positives. A mixture model is used to model the distribution of the similarity values (using the Resnik method for similarity of *Biological Process* GO terms). In particular,

$$p(x) = w_1 p_1(x) + w_2 p_2(x), \tag{7}$$

where $p_1(x)$ is the probability density function for the similarity of pairs of genes for pairs of genes with true interactions in the PPI data, and $p_2(x)$ is the probability density function for the similarity of pairs of genes in the false positives; $w_1$ and $w_2$ are the weights for $p_1$ and $p_2$, respectively. Given a large data set, $p_1$, $p_2$, $w_1$, and $w_2$ can be inferred by the maximum likelihood estimation (MLE) method. For our case, we conclude that the similarity of pairs of genes can be modeled as a mixture of two log-normal distributions with probability density functions

$$p_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1 x} exp\left(-\frac{(\log x - \mu_1)^2}{2\sigma_1^2}\right) \tag{8}$$

and

$$p_2(x) = \frac{1}{\sqrt{2\pi}\sigma_2 x} exp\left(-\frac{(\log x - \mu_2)^2}{2\sigma_2^2}\right). \tag{9}$$

After parameter estimation, we can calculate a value $s$ such that for any $x > s$, $p(x \in \text{Group 2}) > p(x \in$

Group 1). This value $s$ is the threshold meant to differentiate the PPI data with high support (Group 2) from those with low support (Group 1). The further away the point is from $s$, the greater is the confidence. Furthermore, the confidence can be measured by computing the p-value since the parameters of distribution are known.

Thus our mixture model suggests a way of differentiating the true positives from the false positives by only looking at the similarity value of pairs of genes (using the method of Resnik in Eq.(2) for similarity of *Biological Process* GO terms), and by using a threshold value specified by the model (Group 1 contains false positives and Group 2 contains the true positives). Note that no *ad hoc* decision are involved.

### 2.3. Function Prediction

The major advantage of the method presented above is that the p-values obtained from the mixture model provide us with a metric of support of a reliability measure for the PPI data set. However, the limitation of our technique is that it can only be applied to pairs of genes with annotations. In order to overcome this limitation, it has been suggested that function prediction should be performed first to predict the functional annotation of unannotated genes. As mentioned earlier, many computational approaches have been developed for this task[2]. However, the prediction methods are prone to high false positives. Schwikowski *et al.*[6] proposed the *Majority-Voting* (MV) algorithm for predicting the functions of an unannotated gene $u$ by the following objective function,

$$\alpha_u = \arg\max_{\alpha} \sum_{v \in N(u), \alpha_v \in A(v)} \delta(\alpha_v, \alpha), \qquad (10)$$

where $N(u)$ is the set of neighbors of $u$, $A(v)$ is the set of annotations associated with gene $v$, $\alpha_i$ is the annotation for gene $i$, $\delta(x, y)$ is a function that equals 1 if $x = y$, and 0 otherwise. In other words, gene $u$ is annotated with the term $\alpha$ associated with the largest number of its neighbors. The main weakness of this conventional majority voting algorithm is that it weights all its neighbors equally, and is prone to errors because of the high degree of false positives in the PPI data set. Using the metric of support proposed in section 2.2, we propose a modified *"Reliable" Majority-Voting* (RMV) algorithm which assigns a functional annotation to an unannotated gene $u$ based on the following objective function

$$\alpha_u = \arg\max_{\alpha} \sum_{v \in N(u), \alpha_v \in A(v)} w_{v,u} \delta(\alpha_v, \alpha), \qquad (11)$$

where $w_{v,u}$ is the reliability of the interaction between genes $v$ and $u$, that is, $w_{v,u} = sim(A(v), \{\alpha\})$.

Another weakness of the conventional MV algorithm is that it only allows exact matches of annotations and will reject even approximate matches of annotations. Here we propose the *Weighted Reliable Majority-Voting* (WRMV) method, a modification of RMV, with the following objective function

$$\alpha_u = \arg\max_{\alpha} \sum_{v \in N(u)} w_{v,u} \big( \max_{\alpha_v \in A(v)} sim(\alpha_v, \alpha) \big), \qquad (12)$$

where $sim(x, y)$ is a function that calculates the similarity between the GO terms $x$ and $y$.

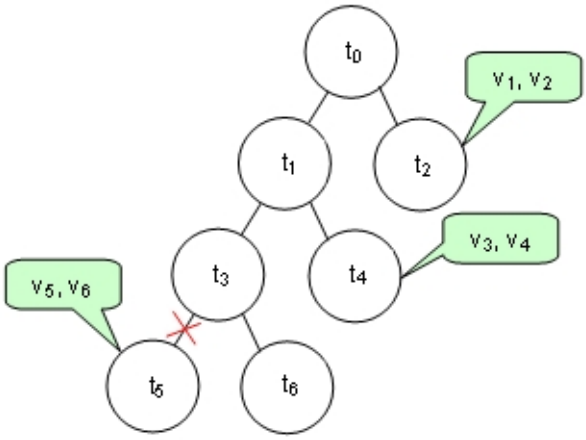


**Fig. 2.** An example showing the hierarchy of GO terms associated with a set of genes. GO term $t_2$ is associated with genes $v_1$ and $v_2$; GO term $t_4$ is associated with genes $v_3$ and $v_4$; GO term $t_5$ is associated with genes $v_5$ and $v_6$.

Note that the aforementioned algorithms only predict one functional annotation term for an uncharacterized gene. But they can be adapted to predict $k$ functional annotation terms for any uncharacterized gene by picking the $k$ best values of $\alpha$ in each case.

The example in Figure 2 illustrates the necessity of considering both the metric of support for the PPI data and the relationships between GO terms during function prediction. Assume we need to predict functions for a protein $u$, whose neighbors in

the interaction network include proteins $v_1$, $v_2$, $v_3$, $v_4$, $v_5$, and $v_6$. As shown in Figure 2, suppose proteins $v_1$ and $v_2$ are annotated with GO term $t_2$, $v_3$ and $v_4$ with GO term $t_4$, and $v_5$ and $v_6$ with GO term $t_5$. According to the MV algorithm, protein $u$ will be assigned all the GO terms $t_2$, $t_4$, and $t_5$, since each of the three terms has equal votes (2 in this case). However, as can be seen from Figure 2, GO term $t_5$ is more specific than GO terms $t_2$ and $t_4$. So GO term $t_5$ should be the most favored as an annotation for protein $u$, assuming that all the PPI data are equally reliable. On the other hand, if the interactions between proteins $u$ and $v_5$ and $v_6$ are less reliable (or false positives), then there is less support for associating protein $u$ with term $t_5$.

Note that the metric of support can also be used to improve other approaches besides the MV algorithm. In this paper, we have employed only local approaches, because as argued by Murali *et al.*[26] methods based on global optimization do not perform better than local approaches based on majority-voting algorithm.

### 2.4. Evaluating the Function Prediction

Several measures are possible in order to evaluate the function prediction methods proposed in section 2.3. For the traditional cross-validation technique, the simplest method to perform an evaluation is to use *precision* and *recall*, defined as follows:

$$Precision = \frac{\sum_i k_i}{\sum_i m_i}, \quad Recall = \frac{\sum_i k_i}{\sum_i n_i}, \qquad (13)$$

where $n_i$ is the number of known functions for the protein $i$, $m_i$ is the number of predicted functions for the protein $i$ when hiding its known annotations, and $k_i$ is the number of matches between known and predicted functions for protein $i$. The conventional method to count the number of matches between the annotated and predicted functions only considers the exact overlap between predicted and known functions, ignoring the structure and relationship between functional attributes. Using again the simple example illustrated in Figure 2, assume that the correct function annotation of a protein $u$ is GO term $t_4$, while term $t_1$ is the only function predicted for it. Then both recall and precision would be reported to be 0 according to the conventional method. However, it overlooks the fact that GO term $t_4$ is quite close to the term $t_1$. Here we introduce a new definition for precision and recall. For a known protein, suppose the known annotated functional terms are $\{t_{o1}, t_{o2}, ..., t_{on}\}$, and the predicted terms are $\{t_{p1}, t_{p2}, ..., t_{pm}\}$. We define the success of the prediction for function $t_{oi}$ as

$$RecallSucess(t_{oi}) = \max_j sim(t_{oi}, t_{pj}),$$

and the success of the predicted function $t_{pj}$ as

$$PrecisionSuccess(t_{pj}) = \max_i sim(t_{oi}, t_{pj}).$$

We define the new precision and recall measures as follows:

$$Precision = \frac{\sum_j PrecisionSuccess(t_{pj})}{\sum_j sim(t_{pj}, t_{pj})}, \qquad (14)$$

$$Recall = \frac{\sum_i RecallSucess(t_{o_i})}{\sum_i sim(t_{o_i}, t_{o_i})}. \qquad (15)$$

## 3. EXPERIMENTAL RESULTS

### 3.1. Data Sets

Function prediction methods based on a protein-protein interaction network can make use of two data sources - the PPI data set and a database of available functional annotations. In this section, we will introduce the two data sources we used in our experiments.

#### 3.1.1. *Gene Ontology*

We used the available functional annotations from the Gene Ontology (GO) database[5]. GO consists of sets of structured vocabularies each organized as a rooted directed acyclic graph (DAG), where every node is associated with a GO term and edges represent either a "IS-A" or a "PART-OF" relationship. Three independent sets of vocabularies are provided: *cellular component*, *molecular function* and *biological process*. Generally, a gene is annotated by one or more GO terms. The terms at the lower levels correspond to more specific function descriptions. If a gene is annotated with a GO term, it is also annotated with the ancestors of that GO term. Thus, the terms at the higher levels have more associated genes. The GO database is constantly being updated; we used version 5.403, and the gene-term associations for *Saccharomyces cerevisiae* from version 1.1344 from SGD.

#### 3.1.2. *Protein-Protein Interaction Data*

Several PPI data sets were used in this paper for our experiments. The first PPI data set was downloaded from the BioGRID database[27]. Henceforth, we will refer to this data set as the *BioGRID* data set. The *confirmation number* for a given pair of proteins is defined as the number of independent confirmations that support that interaction. A pseudo-negative data set was also generated by picking pairs of proteins that were not present in the PPI data set. Thus each pair of proteins in the pseudo-negative data set has a confirmation number of 0. There were 87920 unique interacting pairs in total with confirmation numbers ranging from 0 to 40. This data set is used to estimate the metric of support for pairs of proteins.

Two so-called *gold-standard data sets* (gold-standard positive and gold-standard negative) were used to test the performance of our method. These two gold-standard data sets were hand-crafted by Jansen *et al.*[15]. The gold-standard positives came from the MIPS (Munich Information Center for Protein Sequence) complexes catalog[28] since the proteins in a complex are guaranteed to bind to each other. The number of gold-standard positive pairs used in our experiments was 7727. A gold-standard negative data set is harder to define. Jansen *et al.* created such a list by picking pairs of proteins known to be localized in separate subcellular compartments[15], resulting in a total of 1838501 pairs.

### 3.2. Results on Using the Mixture Model

The similarity between genes based on the *Biological Process* categorization of the GO hierarchy was calculated using Eq.(1) and Eq.(2). The method was separately applied to the *BioGRID* data set, in which PPI data have non-negative, integral confirmation numbers $k$. Interacting pairs of proteins from *BioGRID* data set were grouped based on their confirmation number. It is clear that the PPI data set may include a large number of false positives. Thus, the challenge is to differentiate the true interactions from the false ones. We hypothesize that each of these groups generated above contains two subgroups, one representing pairs of proteins that interact with high support, and the other representing pairs that interact with low support. Data sets with larger confirmation numbers are expected to have less noise.

As shown in Figure 3, a histogram of the (logarithm of) similarity measure (using the *Resnik* method for similarity of GO terms) was plotted for pairs of genes within each group (i.e., same degree of confirmation from the PPI data set). In order to visualize the whole histogram, we have arbitrarily chosen $\log(0) = \log(0.01) \approx -4.61$. Based on our earlier assumptions, we conjectured that each of these histograms can be modeled as a mixture of two normal distributions (since the original is a mixture two log-normal distribution), one for the Group 1, and the other for the Group 2.

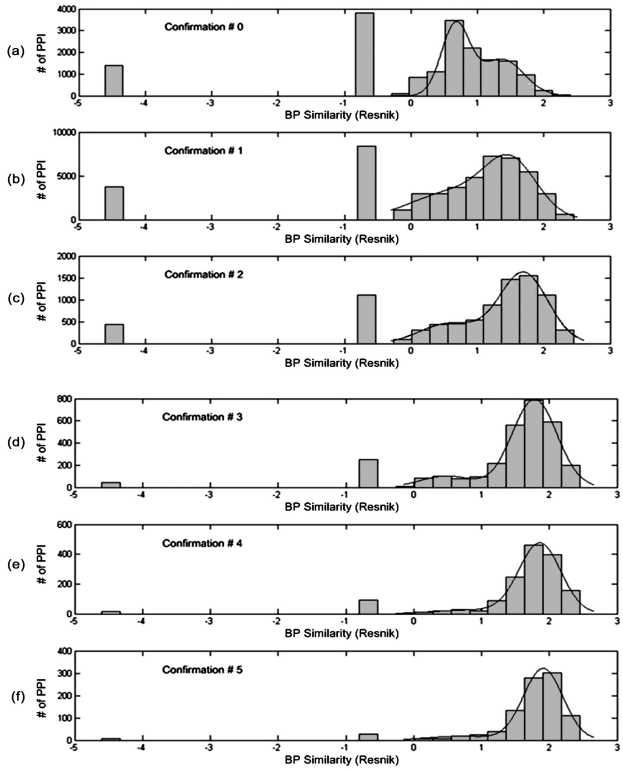


**Fig. 3.** Distribution of similarity of genes based on the Resnik method using: (a) the entire PPI data set, (b) 1 or more independent confirmations, (c) 2 or more independent confirmations, (d) 3 or more independent confirmations, (e) 4 or more independent confirmations, (f) 5 or more independent confirmations.

All the plots in Figure 3 have three well-separated subgroups. Note that the leftmost subgroup corresponds to those pairs of genes for which at least one has the GO terms associated with the root of the GO hierarchy; the subgroup in the middle cor-

responds to those pairs of genes at least one of which is associated with a node close to the root of the GO hierarchy. The reason for the existence of these two subgroups is that there are some PPI data sets containing genes with very non-specific functional annotations. As we can see from Figure 3, the larger the confirmation number, the less pronounced are these two groups. Thus, for the two leftmost subgroups, similarity of genes based on GO annotation cannot be used to differentiate signal from noise (Thus function prediction for these genes are necessary and is an important focus of this paper). However, for PPI data containing genes with specific functions (i.e., the rightmost group in the plots of Figure 3), similarity of genes in this group was fitted to a mixture model as described in section 2.2. In fact, a fit of the rightmost group with two normal distributions is also shown in the plots of Figure 3. The fit is excellent (with R-squared value more than 98 percent for the data set with confirmation number 1 or more). The details are shown in Figure 4. We are particularly interested in the fit of the data set with confirmation 1 and above. The estimated parameters are $\mu_1 = 0.3815$, $\sigma_1 = 0.4011$, $\mu_2 = 1.552$, $\sigma_2 = 0.4541$, $w_1 = 0.23$, and $w_2 = 0.77$. From the fit, we can calculate a value $s = 0.9498$ such that for any $x > s$, $p(x \in \text{Group 2}) > p(x \in \text{Group 1})$. This is the threshold meant to differentiate the two groups. The further away the point is from $s$, the greater the confidence. Furthermore, the confidence can be measured by computing the *p-value* since the parameters of the distribution are known. Further investigation of these two groups reveal that proteins pairs in Group 2 contain proteins that have been well annotated (associating with GO terms that have levels larger or equal to 3). The components of Group 1 are more complicated. It consists of the interactions between two poorly annotated genes, the interactions between a well annotated gene and a poorly annotated gene, and the interactions between two well annotated genes.

The results of further experiments performed on the PPI data sets from the human proteome[27] also displayed similar results (data not shown).

To test the power of our estimation, we applied it to the gold-standard data set. In particular, for each pair of genes in the gold-standard data set, we calculated the similarity between the genes in that pair and compared it to the threshold value $s = 0.9498$. If the similarity was larger than $s$, we labeled it as Group 2, otherwise, as Group 1. We then calculated the percentage of pairs of proteins in Group 2 and Group 1 in the gold-standard positive and negative data sets.

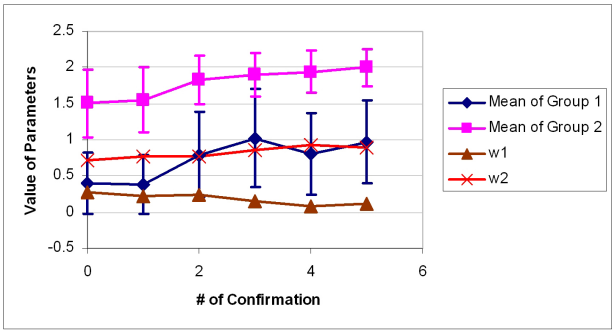


**Fig. 4.** Parameters for the density function, fitting $p(x) = w_1 p_1(x) + w_2 p_2(x)$ for the metric of support for PPI data with different confirmation numbers. Group 1 corresponds to noise, and Group 2 to signal.

As shown in Table 1, majority of the pairs in the gold-standard positive data (GSPD) set were labeled correctly as Group 2 (99.61%), and most of the pairs in the gold-standard negative data set (GSND) were correctly labeled as Group 1 (83.03%). These high percentage values provide further support for our mixture-model based technique. It is worth pointing out that the GSPD set is clearly more reliable than the GSND set as described in section 3.1.2.

**Table 1.** Mixture model on gold-standard data set.

| | total PPI pairs | subgroup PPI pairs | percentage |
|---|---|---|---|
| GSPD[a] | 7727 | 7696[1] | 99.61 |
| GSND[b] | 1838501 | 1526467[2] | 83.03 |

[a] Golden Standard Positive Data set.
[b] Golden Standard Negative Data set.
[1] Number of PPI pairs in Group 2.
[2] Number of PPI pairs in Group 1.

One possible objection to the application in this paper is that the results of the mixture model is an artifact of functional bias in the PPI data set. To address this objection, we applied the mixture model to PPI data after separating out the data from the three main different high-throughput methods, i.e.,

yeast two-hybrid systems, mass spectrometry, and synthetic lethality experiments. As reported by Mering *et al.*[1], the overlap of PPI data detected by the different methods is small, and each technique produces a unique distribution of interactions with respect to functional categories of interacting proteins. In other words, each method tends to discover different types of interactions. For example, the yeast two-hybrid system largely fails to discover interactions between proteins involved in translation; mass spectrometry method predicts relatively few interactions for proteins involved in transport and sensing.

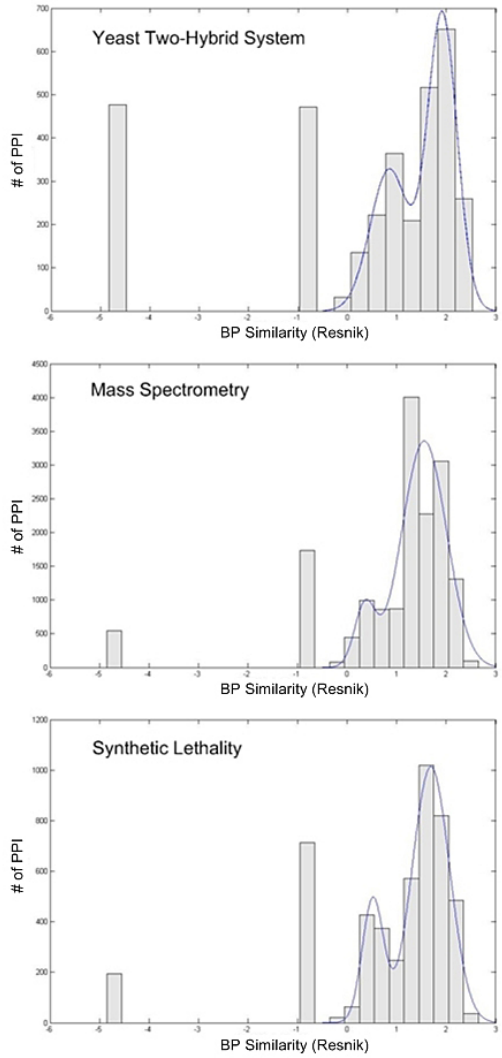


**Fig. 5.** Distribution of similarity of pairs of genes based on the method by Resnik for PPI data generated by high-throughput methods yeast two-hybrid systems (top), mass spectrometry (middle), and Synthetic Lethality (bottom).

In summary, if the PPI data set has a functional bias, then the PPI data produced by individual methods should have an even greater functional bias, with each one biased toward different functional categories.

Despite the unique functional bias of each method, the mixture model when applied to the PPI data from the individual methods showed the same bimodal mixture distribution (Figure 5), indicating that the mixture model is tolerant to severe functional bias and is therefore very likely to be a the reflection of inherent features of the PPI data.

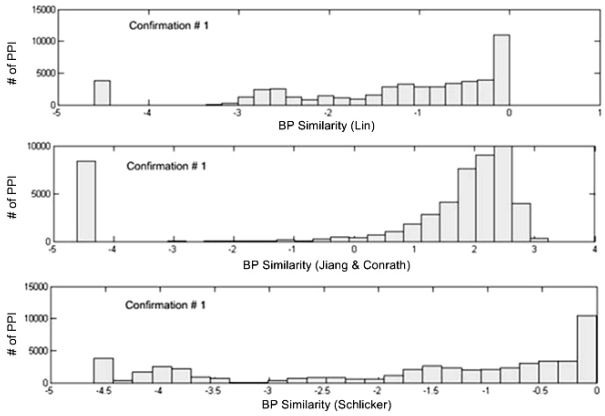


**Fig. 6.** Distribution of similarity of genes based on method Lin, Jiang-Conrath, and Schlicker for PPI data with confirmation number of 1 and more (Confirmation # 1).

In order to justify our choice of the Resnik similarity measure, we also applied the Lin (Eq.(4)), Jiang-Conrath (Eq.(3)), and Schlicker (Eq.(5)) methods to the PPI data set with confirmation number 1 or more. The results shown in Figure 6 confirms our analysis that the Lin and Jiang-Conrath methods are inappropriate for similarity computation. As shown in Figure 6, the histogram of similarity values between genes calculated by Lin's formula has a peak at the rightmost end. Additionally, the rest of the histogram fails to display a bimodal distribution, which is necessary to flush out the false positives. Furthermore, the peak increases with increasing confirmation number (data not shown). In contrast, the histograms of distance measures between genes calculated by the Jiang-Conrath's method (middle in Figures 6) produces a peak at its leftmost end with a unimodal distribution for the remaining, thus showing that the mix-

ture model is unlikely to produce meaningful results. Schlicker's method was devised to combine Lin's and Resnik's methods. However, its performance was similar to that of Lin's method (see in Figure 6). We also applied these methods to the same PPI data set, but with higher confirmation numbers (data not shown). Since those data sets are likely to have fewer false positives, it is no surprise that the histograms were even less useful for discriminatory purpose.

Finally, we tried our methods on the other two GO categorizations, i.e., *cellular component* and *molecular function*. Since those categorizations are less comprehensive with a large number of unannotated genes, similarity calculations based on the them did not adequately reflect the reliability of PPI data (results not shown).

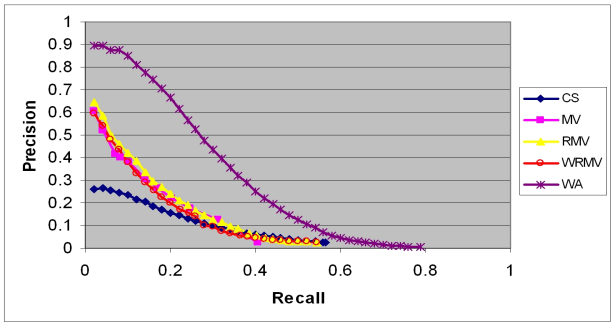


**Fig. 7.** Precision-recall analysis of five function prediction methods using the conventional evaluation metric as described in Eq.(13) for 1) Chi-Square method (CS), 2) Majority-Voting method (MV), 3) Reliable Majority-Voting method (RMV), 4) Weighted Reliable Majority-Voting (WRMV), and 5) FS-Weighted Averaging method (WA).

### 3.3. Function Prediction

Five different function prediction approaches based on neighborhood-counting – three introduced in section 2.3, one introduced in section 1, and the last one called FS-Weighted Averaging method (WA) developed by Hua *et al.*[8] – were compared. We note that in our implementation of the WA method, we use the similarity measure given in Eq.(2) from Hua *et al.*[8] to compute the reliability measure, $w_{v,u}$, in Eq.(11) of this paper. The precision and recall for each approach was calculated on the *BioGRID* PPI data set using 5-fold cross validation. First, a conventional evaluation method was employed, which consisted of computing precision and recall as a simple count of the predictions for the gold-standard positive and negative sets. As shown in Figure 7, when conventional evaluation methods were applied to calculate the precision and recall, the FS-Weighted Averaging (WA) method performed the best, and there was no significant difference among the other three methods (MV, RMV, and WRMV). However, when the new evaluation method (as described in Eq.(14) and Eq.(15)) was applied, both WA and WRMV performed well (see Figure 8). Among the three versions of Majority-Voting methods (MV, RMV, and WRMV), Weighted Reliable Majority-Voting method performed the best, and the conventional Majority-Voting method performed the worst.

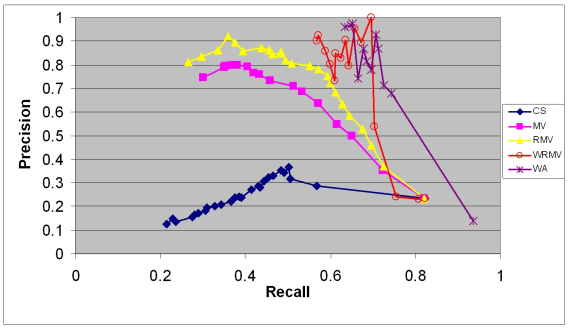


**Fig. 8.** Precision-recall analysis of five function prediction methods using new evaluation metric as described in Eq.(14) and Eq.(15) for 1) Chi-Square method (CS), 2) Majority-Voting method (MV), 3) Reliable Majority-Voting method (RMV), 4) Weighted Reliable Majority-Voting method (WRMV), and 5) FS-Weighted Averaging method (WA).

In order to see the effectiveness of the new evaluation metric, the precision-recall curves of the three function prediction methods (RMV, WRMV and WA) using new and conventional evaluation metrics are compared by combining the related curves in Figure 7 and Figure 8. As shown in Figure 9, the proposed new evaluation method has two advantages over the conventional one. First, the new evaluation method provides wider precision and recall coverage, that is, at the same precision (recall) value, the recall (precision) calculated by the new method is larger than that calculated by the old one. This is due to the strict definition of conventional precision and recall, while ignoring the fact that some pairs of true and predicted annotations are very similar to each other. Second, the new evaluation method has more power to measure the performance of function pre-

diction methods. For example, the precision-recall curves of the function prediction methods RMV and WRMV diverge based on the new evaluation metric, but are roughly indistinguishable based on the conventional metric (Figure 9).

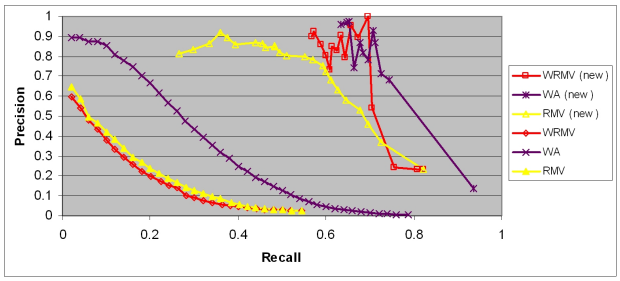


**Fig. 9.** Comparison of precision-recall analysis of three Majority-Voting function prediction methods using new evaluation metric as described in Eq.(14) and Eq.(15) for 1) Weighted Reliable Majority-Voting method (WRMV new), 2) FS-Weighted Averaging method, (WA new), and 3) Reliable Majority-Voting method (RMV new), and conventional metric as described in Eq.(13) for 4) Weighted Reliable Majority-Voting method (WRMV), 5) FS-Weighted Averaging method, (WA), and 6) Reliable Majority-Voting method (RMV).

## 4. DISCUSSION AND CONCLUSIONS

Function predictions based on PPI data were performed using two sources of data: GO annotation data and BioGRID PPI data. Previous research on this topic focused on the interaction network inferred from PPI data, while ignoring the topology of the hierarchy representing the annotations. In some cases, only a fraction of the terms were used. Thus the resulting predictions were not comparable. For PPI data, quantitatively assessment of confidence for pairs of proteins becomes a pressing need.

The research described in this paper addresses the above shortcomings. Our significant contributions are:

(1) A mixture model was introduced to model PPI data. The parameters of the model were estimated from the similarity of genes in the PPI data set. This mixture model was used to devise a metric of support for protein-protein interaction data. It is based on the assumption that proteins having similar functional annotations are more likely to interact.
(2) New function prediction methods were proposed to predict the function of proteins in a consistent way based on the entire hierarchical annotation system. Results show that the performance of the predictions was improved significantly by integrating the mixture model described above into the function prediction methods.
(3) A newly proposed evaluation method provides the means by which systematic, consistent, and comprehensive comparison of different function prediction methods is possible.

In this paper, we have mainly focused on introducing a metric of support for the PPI data using GO information, and the application of such a metric in function prediction for uncharacterized proteins. Although the fact that proteins having similar function annotations are more likely to have interactions has been confirmed in the literature, we provide a quantitative measure to estimate the similarity, and to uncover the relationship between the metric and the support of PPI data. GO annotations are generated by integrating information from multiple data sources, many of which have been manually curated by human experts. Thus assessing PPI data using the GO hierarchy is a way in which multiple data sources are integrated. The comprehensive comparison of the method to assess PPI data using GO information and other counterparts as described in section 1 is necessary and will be addressed elsewhere.

## Acknowledgments

This research is supported by the program Molecular Assemblies, Genes, and Genomics Integrated Efficiently (MAGGIE) funded by the Office of Science, Office of Biological and Environmental Research, U.S. Department of Energy, under contract number DE-AC02-05CH11231. GN was supported by NIH Grant P01 DA15027-01 and NIH/NIGMS S06 GM008205, and EZ was supported by FIU Dissertation Year Fellowship.

## References

1. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**(6887) (May 2002) 399–403
2. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular Systems Biology **3**(88) (2007) 1–13

3. Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V.: The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28**(1) (January 2000) 33–36
4. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., Mewes, H.W.: The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res **32**(18) (2004) 5539–5545
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet **25**(1) (May 2000) 25–29
6. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nat Biotechnol **18**(12) (December 2000) 1257–1261
7. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast **18**(6) (April 2001) 523–531
8. Chua, H.N., Sung, W.K., Wong, L.: Exploiting indirect neighbours and topological weight to predict protein function from proteinprotein interactions. Bioinformatics **22**(13) (July 2006) 1623–1630
9. Vzquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nat Biotechnol **21**(6) (June 2003) 697–700
10. Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., R.Cantor, C., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci U S A **101**(9) (March 2004) 2888–2893
11. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics **19 Suppl 1** (2003) i197–i204
12. Deng, M., Tu, Z., Sun, F., Chen, T.: Mapping gene ontology to proteins based on protein–protein interaction data. Bioinformatics **20**(6) (2004) 895–902
13. Wu, Y., Lonardi, S.: A linear-time algorithm for predicting functional annotations from protein-protein interaction networks. In: Proceedings of the Workshop on Data Mining in Bioinformatics (BIOKDD'07). (2007) 35–41
14. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics **21 Suppl 1** (June 2005) i302–i310
15. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science **302**(5644) (October 2003) 449–453
16. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol **22**(1) (January 2004) 78–85
17. Yu, J., Fotouhi, F.: Computational approaches for predicting protein—protein interactions: A survey. J. Med. Syst. **30**(1) (2006) 39–44
18. Ben-Hur, A., Noble, W.S.: Kernel methods for predicting protein-protein interactions. Bioinformatics **21 Suppl 1** (June 2005)
19. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M.: A probabilistic functional network of yeast genes. Science **306**(5701) (November 2004) 1555–1558
20. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. PROTEINS: Structure, Function, and Bioinformatics **3**(63) (May 2006) 490–500
21. Resnik, P.: Using information content to evaluate semantic similarity. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. (1995) 448–453
22. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics. (1997)
23. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. (1998)
24. Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., Lengauer, T.: A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics **7** (June 2006) 302–317
25. Lord, P.W., Stevens, R.D., Brass, A., Goble, C.A.: Semantic similarity measures as tools for exploring the gene ontology. Pac Symp Biocomput (2003) 601–612
26. Murali, T., Wu, C., Kasif, S.: The art of gene function prediction. Nat Biotechnol **24**(12) (2006) 1474–1475
27. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res **34**(Database issue) (January 2006)
28. Mewes, H., Gruber, F., Geier, C., Haase, B., Kaps, D., Lemcke, A., Mannhaupt, K., Pfeiffer, G., Schuller, F.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res **30**(1) (2002) 31–34