BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 8 2006, pages 967–973
doi:10.1093/bioinformatics/btl042

Systems biology

Assessing semantic similarity measures for the characterization of human regulatory pathways

Xiang Guo[1,*], Rongxiang Liu[2], Craig D. Shriver[3], Hai Hu[1] and Michael N. Liebman[1]

[1]Windber Research Institute, Windber, PA 15963, USA, [2]GlaxoSmithKline Pharmaceutical R&D, King of Prussia, PA 19420, USA and [3]Walter Reed Army Medical Center, Washington, DC 20307, USA

Received on September 20, 2005; revised on January 16, 2006; accepted on February 3, 2006
Advance Access publication February 21, 2006
Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Pathway modeling requires the integration of multiple data including prior knowledge. In this study, we quantitatively assess the application of Gene Ontology (GO)-derived similarity measures for the characterization of direct and indirect interactions within human regulatory pathways. The characterization would help the integration of prior pathway knowledge for the modeling.

Results: Our analysis indicates information content-based measures outperform graph structure-based measures for stratifying protein interactions. Measures in terms of GO biological process and molecular function annotations can be used alone or together for the validation of protein interactions involved in the pathways. However, GO cellular component-derived measures may not have the ability to separate true positives from noise. Furthermore, we demonstrate that the functional similarity of proteins within known regulatory pathways decays rapidly as the path length between two proteins increases. Several logistic regression models are built to estimate the confidence of both direct and indirect interactions within a pathway, which may be used to score putative pathways inferred from a scaffold of molecular interactions.

Contact: s.guo@wriwindber.org

1 INTRODUCTION

The function of a biological system relies on a combinatory effect of many semantic elements, which interact non-linearly. We need to take a global view of the entire biological network, at many levels of abstraction, to manage complex biological states such as disease. Biological pathways and networks are built upon the identification of protein interactions. Traditionally, information about protein–protein interactions is collected from small-scale screening. The accuracy of each interaction is often validated with multiple experiments. With the development of high-throughput methods such as the two-hybrid assay and protein chip technology, the information within interaction databases has increased tremendously (Drewes and Bouwmeester, 2003). In addition, a number of computational methods have been developed for the prediction of protein–protein interactions based on protein structure and/or genomic information (Valencia and Pazos, 2002). The increased coverage of the protein–protein interaction map provides deeper insight into the global properties of the interaction networks. However, interaction data derived from large-scale assays and computational methods are often very noisy. Thus, it is essential to develop strategies to validate putative protein interactions such that pathways can be rebuilt from a scaffold of reliable molecular interactions (Chen and Xu, 2003).

Various genomic features exist in sequence, structure, functional annotation and expression-level databases which may be used for interaction prediction and validation (Valencia and Pazos, 2002). Recently, Lu *et al.* (2005) have evaluated the predictive power of 16 features, ranging from coexpression relationships to similar phylogenetic profiles. Among those features, semantic similarity between two proteins has the dominant performance in discriminating true interactions from noise. The maximum predictive power is approached by integrating only a few features including the functional similarity of protein pairs.

Semantic similarity is traditionally assessed as a function of the shared annotation of proteins in a controlled vocabulary system, such as Gene Ontology (GO) (Sprinzak *et al.*, 2003). GO terms and their relationships are represented in the form of directed acyclic graphs (DAGs). The ontology provides computationally accessible semantics about the gene functions they describe. GO comprises three categories: molecular function (MF), biological process (BP) and cellular component (CC). MF describes activities at the molecular level, and a BP is accomplished by one or more assemblies of MF (Ashburner *et al.*, 2000). Although interacting proteins often participate in the same BP, they are less likely to have the same MF. Jansen *et al.* calculate the similarity of a protein pair by identifying the set of GO terms shared by the two sets of protein annotations (2003). Their method can only use annotations derived from BP subontology, but not MF subontology. In addition, even though two annotations are different, they can be closely related via their common ancestors in DAG. Traditional methods also fail to take into account the specificity of GO terms. Although some proteins share the same GO terms, these terms may be too general to verify the functional association of the annotated proteins.

There are two strategies that can be used to overcome these limitations. The first strategy is based on the graph structure of GO. For each protein we may obtain an induced graph which includes the specific set of GO annotations for the protein and all parents of those GO terms. The similarity between two induced graphs can then be used to estimate the similarity between two proteins (Gentleman, 2005, http://www.bioconductor.org/repository/devel/vignette/GOvis.pdf). The second strategy is based on the assumption that the more information two terms

*To whom correspondence should be addressed.

share, the more similar they are. The shared information is indicated by the information content of the terms that subsume them in DAG. The information content is defined as the frequency of each term, or any of its children, occurring in an annotated dataset. Less frequently occurring terms are said to be 'more informative'. Given the information content of each term, several measures may be calculated to estimate the semantic similarity between annotated proteins (Lord *et al*., 2003b). Recently, both approaches have been applied in the analysis of protein interactome (Brown and Jurisica, 2005; Chen and Xu, 2004). However, a systematic evaluation of their performance remains to be done.

Given the large amount of protein interaction data, we can build a comprehensive scaffold of interactions. One popular paradigm for cellular modeling involves rebuilding pathways from this scaffold. The mining usually uses global data pertaining to molecular and cellular states such as gene expression profiles and protein post-translational modifications. The active subnetworks extracted from the large interaction scaffold may represent concrete hypotheses as to the underlying mechanisms governing the observed state change (Ideker and Lauffenburger, 2003). However, the noisy nature of both high-throughput interactions and state measurements makes pathway modeling extremely difficult. The integration of prior pathway knowledge would increase the reliability of newly inferred pathways. KEGG (Kyoto Encyclopedia of Genes and Genomes) includes current knowledge on molecular interaction networks such as pathways and complexes (Kanehisa *et al*., 2004). Characterization of KEGG pathways may help us to develop new methods for the pathway modeling.

In this study, we quantitatively assess the application of GO-based similarity methods in human protein–protein interaction and pathway analysis. First, receiver operating characteristic (ROC) analysis is used to assess the ability of GO graph structure and information content-based methods to stratify protein interactions. For each method, there are three measures in terms of BP, MF or CC annotations. We investigate the possibility to integrate the three measures by logistic regression for performance improvement. Based on the logistic regression model, we then estimate the reliability of several protein–protein interaction datasets. More importantly, we characterize semantic similarity of proteins within human regulatory pathways. Several logistic regression models are built to validate indirect protein interactions in a pathway. These models may be used to infer or rank putative pathways given the scaffold of protein interactions.

2 METHODS

2.1 Estimation of semantic similarity

Graph similarity-based measures are estimated using *GOstats* package of Bioconductor (Gentleman, 2005). Each protein is associated with an induced graph that is obtained by taking the most specific GO terms annotated with the protein and by finding all parents of those terms until the root node has been obtained. Two methods, union-intersection (UI) and longest shared path (LP), are used to calculate the between-graph similarity. The first method uses the number of nodes two induced graphs share divided by the total number of nodes in two graphs. The resulting similarity values are bounded between 0 and 1 with more similar proteins having values near 1. The second method, LP, adopts the depth of the longest path shared by two induced graphs as the similarity score. The larger the depth the more similar two proteins are. If two proteins are both quite specific and similar, they should have long shared path and thus high similarity score.

Information content-based measures are implemented using a locally installed GO database. We use the associations between GO terms and UniProt-Human (Bairoch *et al*., 2005) proteins to calculate the information content $p(t)$ which is the frequency of each GO term or any child term occurring within the corpus. Both 'is-a' and 'part-of' links are used to define the child term. Given the information content, we have applied the three measures to calculate the semantic similarity between terms. The first measure (Resnik) is solely based on the information content of shared parents of the two terms. If there is more than one shared parent, the minimum information content is taken. Then the similarity score is derived as shown in Equation (1).

$$\mathrm{sim}(t1,t2) = -\ln\left(\min_{t \in S(t1,t2)}\{p(t)\}\right), \quad (1)$$

where $S(t1, t2)$ is the set of parent terms shared by $t1$ and $t2$ (Resnik, 1999). Two other measures use not only the information content of the shared parents, but also that of the query terms. Given query terms $t1$ and $t2$, the Lin's similarity is defined as

$$\mathrm{sim}(t1,t2) = \frac{2 \times \ln\left(\min_{t \in S(t1,t2)}\{p(t)\}\right)}{\ln p(t1) + \ln p(t2)}, \quad (2)$$

where $p(t1)$, $p(t2)$ and $p(t)$ are information content values for $t1$, $t2$ and their parents, respectively (Lin, 1998). Lin's method generates normalized similarity values between 0 and 1. In contrast, Jiang's method uses the same components for the calculation, but generates semantic distance which can vary between infinity and 0 (Jiang and Conrath, 1997).

$$\mathrm{sim}(t1,t2) = 2 \times \ln\left(\min_{t \in S(t1,t2)}\{p(t)\}\right) - \ln p(t1) - \ln p(t2) \cdot \quad (3)$$

Given those measures, the semantic similarity between two proteins could be derived accordingly. If a protein is annotated with several GO terms, the maximum similarity between all terms is taken as the between protein similarity.

All five methods (UI, LP, Resnik, Lin and Jiang) are based on the April 2005 release of GO database. The mappings from Gene IDs to GO IDs can be restricted based on evidence codes. We drop those annotations inferred from physical interaction (IPI) to avoid circular reference. In addition, the annotations associated with 'BP unknown' (GO:0000004), 'MF unknown' (GO:0005554) and 'CC unknown' (GO:0008372) are eliminated from our analysis.

2.2 ROC curve analysis

These five methods are assessed for their ability to stratify human protein–protein interactions. Each method generates three sets of similarity values corresponding to BP, MF and CC categories of GO. The positive dataset is assembled from KEGG. It comprises pairwise interactions among proteins of the same complex and interactions of neighboring proteins within human regulatory pathways. After discarding proteins with indirect interaction effect, the interaction nature of neighboring proteins includes activation, inhibition, binding/association, dissociation, state change, phosphorylation, dephosphorylation, glycosylation, ubiquitination and methylation. As to the negative dataset, we randomly choose two distinct human proteins from Entrez Gene database as a non-interacting protein pair. This is valid since the chance of identifying protein–protein interactions at random is very small (0.024% based on the two-hybrid data by Utez *et al*., 2000).

An ROC curve depicts relative trade-offs between sensitivity and specificity of certain method for different values of the threshold. Sensitivity is defined as the ability to identify a true positive in a dataset. Specificity is defined as the ability to identify a true negative in a dataset. The area under an ROC curve (AUC) is generally used as a measure of the performance. It denotes the probability that the classification method will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Random guessing generates the diagonal line $y = x$, which has an AUC of

0.5. A realistic classification method must have an AUC larger than 0.5. Curves from different cross-validation runs are averaged by sampling at fixed thresholds, and standard deviations are used to visualize the variability across the runs (Fawcett, 2003). We use the ROC and ROCR libraries in R to draw the graph and calculate the AUCs (Sing *et al*., 2004).

2.3 Logistic regression

Multiple logistic regression is effective when the response variable is dichotomous and the input variables are continuous, categorical or dichotomous. It is a commonly used model for the prediction of true protein–protein interactions (Bader *et al*., 2004; Lin *et al*., 2005). The form of the model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k, \quad (4)$$

where p is the probability of a putative interaction to be true and X_1, $X_2, \ldots, X_k$ are independent variables such as semantic similarity measures. Logistic regression thus forms a predictor variable $\log[p/(1-p)]$ which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic function. We use the glm function in R to perform the logistic regression. Likelihood ratio test is applied to see if a model including a given independent variable provides more information than a model without this variable. The generalization error and performance of each logistic regression model is estimated by 10-fold cross-validation and ROC curve analysis.

2.4 Reliability estimation

Experimentally determined human protein–protein interactions have been collected in the Biomolecular Interaction Network Database (BIND) (Bader *et al*., 2003). Interaction data in BIND are organized into low-throughput (LTP) and high-throughput (HTP) sections based on the number of records in the same publication. HTP data are imported from papers that have more than 40 interaction results arising from the same experimental design and methodology. Examples include those derived from exhaustive 2-hybrid hybridizations, immunoprecipitations and microarray methods. LTP interactions are manually curated from papers with less than 40 interaction results identified by the same method. They include not only data identified by traditional small scale screening, but also two-hybrid assay and other newer approaches. Recently, an approach based on evolutionary cross-species comparisons has emerged for the completion of protein interaction maps (Matthews *et al*., 2001). Human protein–protein interactions may be predicted from lower eukaryotic protein interaction maps through the identification of orthologous genes between different species (Lehner and Fraser, 2004; Brown and Jurisica, 2005).

We compare the reliability of the three human protein interaction datasets using Resnik measures. Experimental datasets (LTP and HTP) are downloaded from BIND, and the orthology-inferred dataset (Ortho) is from the core dataset computed by Lehner and Fraser. The reliability of each dataset is estimated by the fraction of interactions with scores more than the defined threshold over all protein–protein interactions with corresponding measures available. For BP, MF and CC-derived measures, a different threshold is chosen to achieve maximum accuracy in discriminating true and false interactions for our training dataset described in Section 2.2. The accuracy is the weighted average of true positive and true negative rates. For the logistic regression model, 0.5 is used as the threshold.

2.5 Regulatory pathway analysis

KEGG Markup Language (KGML) facilitates computational analysis and modeling of protein pathways and networks (Kanehisa *et al*., 2004). Currently, there are approximately 30 human regulatory pathways with KGML files available. For each pathway, we calculate the semantic similarity values for proteins within the same complex, neighboring proteins and protein pairs with different distance in the pathway. Neighboring pairs represent proteins that directly interact with each other, while distant pairs represent proteins that interact indirectly through various numbers of bridge proteins. The distance of two proteins is defined as the length of their shortest path in the pathway. Mean similarity values are calculated for each category of protein pairs. Permutation test is used to see how often random chance would generate a mean similarity at least as high as the observed value. For each category, the same number of random pairs is picked from all proteins in the pathways, and the mean similarity value is calculated and compared with the original mean similarity. This process is repeated 1000 times, and the P-value is defined as the frequency that the random dataset generates mean similarity value equal or higher than the original value. In addition, the mean similarity (y) is fitted against the distance (x) with exponential distribution such that the rate of decay may be estimated by mean life of the distribution.

3 RESULTS

3.1 Performance of semantic similarity measures for stratifying protein–protein interactions

We assemble proteins within a complex or neighboring to each other in KEGG regulatory pathways as the positive protein–protein interaction dataset (total number 1649). Among them, there are 1500 protein pairs with BP annotations, 1425 pairs with MF annotations and 1255 pairs with CC annotations available for both proteins. The negative dataset with the same number of protein pairs is built by randomly choosing human proteins from Entrez Gene. As shown by the ROC curve analysis, similarity measures based on BP annotation have the highest ability to stratify protein–protein interactions (Figs 1 and 2). MF-derived measures follow, and CC-derived measures have the worst discriminating power. Since GO associations with evidence code TAS (Traceable Author Statement) are regarded as the most accurate, we investigate if the performance can be improved by restricting GO annotations to TAS only. Interestingly, no significant improvement is achieved while less protein pairs have similarity values available.

While the information on subcellular localizations can be used to define robust negative controls for protein interactions, our analysis indicates that localization-based similarity measures may not have the ability to separate true protein interactions from noise. The reason may be 2-fold. In contrast to the existence of over 9000 BP terms and over 7000 MF terms, the total number of CC terms is only around 1600. This subontology is much less complete and specific compared with the MF and BP subontologies, thus it may not be expressive enough to validate protein–protein interactions. The other possible reason is related to the bias in link type usage among the different subontologies. GO terms are placed within a structure of relationships with the link type of 'is-a' between parent and children as well as the type of 'part-of' between part and whole. Generally, only the 'is-a' links are considered for similarity measures (Resnik, 1999), but the omission of the 'part-of' links would result in orphan terms which make the semantic comparison impossible. Our similarity measures consider two links equally, which may not be optimal. The ratio of 'part-of' links versus 'is-a' links is 17% in BP category and there are only 2 'part-of' links in MF category, but the ratio increases to 70% in CC category. The high percentage of 'part-of' relationships may make the CC-derived measurement less accurate than the other measures.

In all three GO categories, the information theoretic methods consistently perform better than graph structure-based methods (Fig. 2). Among the five methods, UI has the worst performance

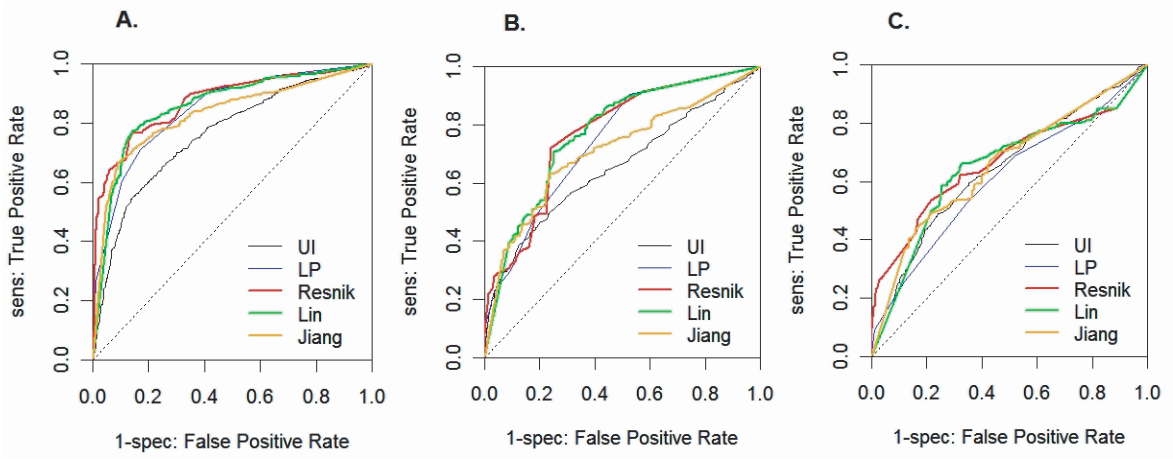

Fig. 1. ROC curves for the comparison of five semantic similarity estimation methods. They illustrate the trade-off between sensitivity and specificity for all possible thresholds of similarity measures in terms of (**A**) biological process, (**B**) molecular function and (**C**) cellular component annotations, respectively. The curve for a random classifier is shown as a line extending from the origin with a slope of 1.

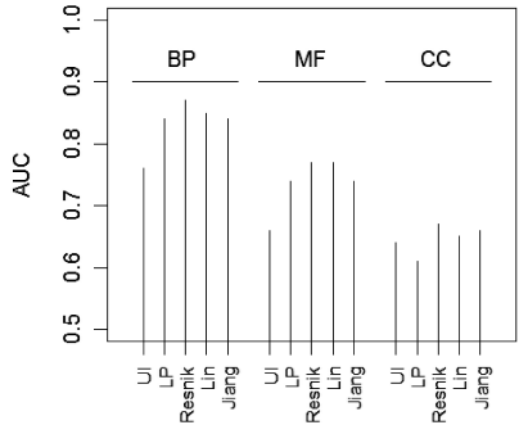

Fig. 2. AUC summary for 15 semantic similarity measures derived from 5 different methods in terms of 3 GO categories (BP, MF and CC).

in terms of BP and MF-derived similarity measures. This approach estimates the overlap of all GO terms and their parents associated with two proteins, but it does not discriminate general and specific terms. LP improves the performance by considering the specificity of shared annotations. Long shared path suggests that two proteins are both specific and similar. However, this method assumes that nodes and links in ontology are uniformly distributed. This assumption is not accurate in GO where link densities vary because of the vagaries of biological knowledge. Instead of solely using the structure of the ontology, information content-based methods explore the usage of GO terms within the corpus. They generate high AUC values indicating the better performance of those measures for the validation of protein–protein interactions. Resnik's measure seems to outperform Lin and Jiang's measures. This measurement has also been reported to be the most discriminatory in terms of the correlation between semantic similarity and sequence similarity, while Jiang's distance shows the weakest correlation (Lord *et al.*, 2003a). Therefore, we will use Resnik's approach for all other analyses in the article.

3.2 Integration of similarity measures by logistic regression

Based on Resnik's method, we explore the possibility of improved performance through integrating BP, MF and CC-derived measures. We select 974 positive protein pairs with annotations for all three GO categories, along with the same number of negative protein pairs. Using the three measures as explanatory variables, we have compared the performance of logistic regression models for the prediction of true protein–protein interactions. Likelihood ratio tests indicate that the model combining BP and MF-derived measures provides a better fit than the model using either measure alone ($p < 0.001$). However, the inclusion of CC-based measure does not improve the fit ($P > 0.05$), which is consistent with the poor performance of this measurement revealed by ROC analysis (Fig. 2). The results also suggest that the maximum predictive power of GO annotation is reached by integrating two features (BP and MF) only.

We then rebuild the logistic regression model with two measures using a larger dataset (2660 positive and negative protein pairs with both BP and MF annotations). The false positive rate is 20.6%, and the false negative rate is 17.4% if we use 0.5 as the threshold for discriminating positive and negative predictions. AUC for this set of data is 0.89. The generalization error is estimated by 10-fold cross-validation. The dataset is split in 10 parts, subsequently each part is used as a test set for the logistic regression model which is built from the remaining 9/10th of the data. ROC curves are generated from 10 sets of prediction obtained from the cross-validation, and these curves are combined by threshold averaging (Fig. 3). The total error rate is 18.8% and the cross-validated AUC estimate is 0.89 ± 0.04, indicating the model is not overfit.

3.3 Reliability of human protein–protein interaction datasets

Using Resnik measures, we have estimated the probability of experimental and computationally inferred protein–protein interactions to be involved in biological pathways (Table 1). As expected,

Table 1. Reliability of human protein–protein interaction datasets

Data Source	# (Total)	# (BP)	% (BP)	# (MF)	% (MF)	# (CC)	% (CC)	# (BP & MF)	% (BP & MF)
LTP	5783	3297	65	2980	60	2170	64	2515	63
HTP	12 747	6599	39	6441	45	5296	49	5206	39
Ortho	9283	5249	46	6080	41	3663	65	4461	43

LTP and HTP denote the low-throughput and high-throughput datasets retrieved from BIND, and Ortho denotes the orthology-inferred human protein–protein interactions. Columns (#) list the number of total protein pairs, and protein pairs with BP, MF, CC or both BP and MF annotations available. Columns (%) list the percentage of true interactions predicted by BP, MF, CC, or both BP and MF similarities integrated by logistic regression.

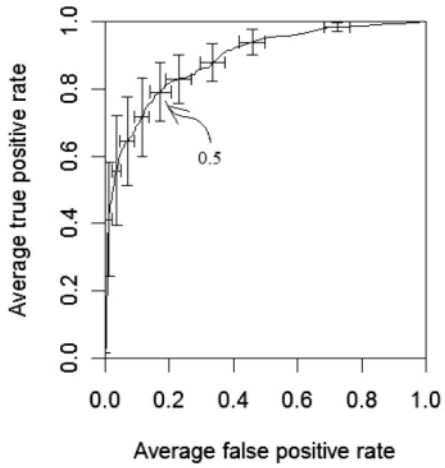

Fig. 3. ROC analysis of the logistic regression model for the discrimination of true protein–protein interactions from false positives. Threshold averaging is used to combine 10 ROC curves derived from 10-fold cross-validation into one curve with standard deviation bars. The curve position for the cutoff of 0.5 is specified in the graph.

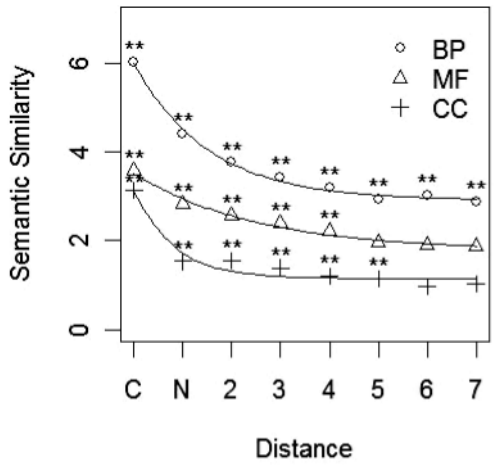

Fig. 4. Distance-dependent semantic similarity in human regulatory pathways. Mean similarity for protein pairs within the complex (C), neighboring proteins (N) and protein pairs with shortest path length of 2–7 in the pathway. The statistical significance is calculated based on permutation test (**$P < 0.01$, $n = 1000$). Curves denote the exponential distribution fit for the distance-dependent semantic similarity values.

LTP has the highest percentage of reliable interactions among three datasets based on BP-derived measure. The relatively low reliability (65%) may be caused by the definition of LTP and curation errors in BIND. LTP interactions are from publications with less than 40 results in one experiment, and some of them may be identified by two-hybrid assay and other less reliable experimental methods, which may increase the high false positive rate. HTP dataset has a reliability rate of 39%, which is consistent with the estimation by BIND's own support vector machine scoring system-Protein Interaction Confidence Kernel Scores (PICKS). Similar ratio has also been reported for high-throughput interactions in other species (Deane *et al*., 2002). Ortho dataset includes human protein interactions inferred from high-confidence 'core' protein interactions in worm, fly and yeast (Lehner and Fraser, 2004). Their reliability is comparable with that of high-throughput experimental data. Sharan *et al*. (2005) have shown that the orthology-based method has a success rate in the range of 40–52% based on two-hybrid tests of predicted yeast interactions. Our estimation is in line with their experimental verification.

Similar reliability estimations are seen when we calculate the percentage using either MF-derived measure or the logistic regression model integrating two measures. However, the CC-derived measure generates higher estimation for HTP and Ortho datasets, and the latter even has a higher reliability rate than LTP dataset. It verifies that the CC-based measure may not be applicable for the validation of protein interactions.

3.4 Semantic similarity of proteins within regulatory pathways

Biological networks have the properties of a small-world network. They have high clustering coefficient and short characteristic path length. These networks can be fragmented into clusters of proteins having similar characteristics. Our analysis shows that the semantic similarity between two proteins decreases as their distance within the pathway increases (Fig. 4). Proteins within the same complex and neighboring proteins in the pathway directly interact with each other, so they have the highest similarity in terms of all three GO categories. In addition, the complex proteins have higher similarity values than the neighboring proteins, which suggests a relationship between protein complex membership and GO-based semantic similarity measures. The higher the semantic similarity between two proteins, the more likely they are in the same complex.

KEGG pathways and GO BP are two annotation systems for the description of biological process in which each gene product participates. As expected, our analysis shows that all pairs of proteins within KEGG pathways have significantly higher similarity than

expected by chance in terms of BP. In contrast, similarity values of remote protein pairs are not different from those of random pairs in terms of MF and CC. As we know, a series of different functional steps comprise a pathway. Neighboring proteins perform one functional step, while distant proteins may play different functional roles in different cellular location. Our results are consistent with the pathway biology. In addition, CC-derived similarity values decrease in a stepwise pattern, since two or three sequential functional steps are likely to occur in the same cellular compartment.

The distance-dependent similarity fits an exponential decay model. The rate of decay is characterized by the mean life, which is the distance needed for the similarity to be reduced by a factor of e. BP, MF and CC-derived similarity values decay rapidly with mean lives of 1.51, 2.42 and 0.81, respectively.

Our study has shown that the logistic regression model can be used to separate direct interacting proteins from random protein pairs (Fig. 3). The reliability of a putative interaction may be estimated by this model. Similarly, indirect interacting proteins within a putative pathway may also be validated based on their semantic similarity. Following the same procedure, we have created three models using BP and MF-derived measures to assign confidence scores to protein pairs with distance of 2, 3 or 4 in a pathway. The 10-fold cross-validation shows that the prediction errors of these models are 26.9, 30.5 and 33.5%. Three models have AUC estimates of 0.82 ± 0.03, 0.79 ± 0.06 and 0.77 ± 0.06, respectively. These models may be used together to validate putative pathways by scoring both direct and indirect interactions in the pathway.

4 DISCUSSION

Although various functional similarity measures have been used in the interactome analysis, a systematic evaluation of their performance has not been reported. Our results demonstrate that information content-based measures have better performance than GO structure-based measures for the validation of protein interactions involved in human regulatory pathways. Among them, Resnik's approach seems to have the best performance. Measures in terms of either MF or BP can be used to stratify protein interactions. However, CC-derived measures may not be sensitive enough for this purpose.

The application of semantic similarity measures relies on the completeness and accuracy of GO annotation. Most of the proteins included in KEGG pathways have accurate and detailed annotation. However, there may be considerable amount of incorrect or under-annotated proteins in other databases. The performance of semantic similarity measures may be decreased when applied to a poorly annotated dataset. For example, if two proteins are annotated by a non-specific term 'signal transducer activity' (GO: 0004871) only, Lin similarity will be 1, Jiang distance will be 0, while UI, LP and Resnik measures generate low similarity scores. Therefore, in the case of under annotation, Lin and Jiang measures are more likely to generate false positives while more false negatives may be seen in other three measures. As the use of GO improves, the performance of those measures should improve when applied to experimental datasets.

Brown and Jurisica (2005) have recently adopted information content-based method to validate their protein interaction datasets. However, their method does not separate the three GO categories. The semantic similarity is determined by the maximum similarity from the set of all GO term pairs between interacting proteins. Our results show that BP-based measures produce higher similarity values than MF and CC-based measures (Fig. 4). If there are BP annotations available for a protein pair, then the similarity value derived from the method of Brown and Jurisica is most likely equal to our BP-based similarity value. Currently, BP annotation is the most comprehensive among the three GO categories. In our dataset, if an MF-based measure is defined for a protein pair, there is a 93% chance that a BP-based measure is also defined. Thus, information included in the MF annotation still remains largely unexplored by the method of Brown and Jurisica. Our results demonstrate that MF-derived measures can be used alone or integrated with BP-derived measures for the interactome analysis.

Our KEGG pathway analysis indicates that protein pairs with short path length have significantly higher semantic similarity values than expected by chance alone. These protein pairs can be separated from random protein pairs by logistic regression models. Current pathway modeling methods score candidate subnetworks based on various evidence including semantic similarity estimates for each protein interaction (Sharan *et al*., 2005). However, information about proteins, which interact indirectly through other bridge proteins, has not been utilized for pathway modeling. We propose to calculate confidence scores of not only direct interactions but also indirect interactions for the validation of putative pathways. The logistic regression model is our first step in this direction. Future work may include integration of more genomic features such as mRNA coexpression, and the development of a probabilistic model to score the candidate subnetworks based on the confidence values assigned to different protein pairs. We believe that new methods incorporating semantic similarity of proteins that interact directly and indirectly will greatly aid the extraction of active pathways and thus improve the interpretation of intriguing biological phenomenon.

ACKNOWLEDGEMENTS

We thank Dr Chen Yu of Monsanto Company for stimulating discussions and Nicholas Jacob, President of Windber Research Institute, for continuing support.

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Bader,G.D. *et al*. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.

Bader,J.S. *et al*. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol*., **22**, 78–85.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.

Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.

Chen,Y. and Xu,D. (2003) Computational analyses of high-throughput protein–protein interaction data. *Curr. Protein Pept. Sci*., **4**, 159–181.

Chen,Y. and Xu,D. (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids. Res*., **32**, 6414–6424.

Deane,C.M. *et al*. (2002) Protein interactions: two methods for the assessment of the reliability of high-throughput observations. *Mol. Cell Proteomics*, **1**, 349–356.

Drewes,G. and Bouwmeester,T. (2003) Global approaches to protein–protein interactions. *Curr. Opin. Cell Biol*., **15**, 199–205.

Fawcett,T. (2003) ROC graphs: notes and practical considerations for data mining researchers. *Technical report HPL-2003-4*.. HP Laboratories, Palo Alto, CA.

Gentleman,R. (2005) Visualizing and distances using GO.

Ideker,T. and Lauffenburger,D. (2003) Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol*., **21**, 255–262.

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of 10th International Conference on Research In Computational Linguistics,* Taiwan, 19–33.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids. Res*., **32**, D277–D280.

Lehner,B. and Fraser,A.G. (2004) A first-draft human protein-interaction map. *Genome Biol*., **5**, R63.

Lin,D. (1998) An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,* San Francisco, CA, 296–304.

Lin,N. *et al*. (2004) Information assessment on predicting protein–protein interactions. *BMC Bioinformatics*, **5**, 154.

Lord,P. *et al*. (2003a) Semantic similarity measures as tools for exploring the Gene Ontology. *Pac. Symp. Biocomput*., **8**, 601–612.

Lord,P.W. *et al*. (2003b) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Lu,L.J. *et al*. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*., **15**, 945–953.

Matthews,L.R. *et al*. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res*., **11**, 2120–2126.

Resnik,P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intel. Res*., **11**, 95–130.

Sharan,R. *et al*. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*, **102**, 1974–1979.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics,*, **21**, 3940–3941.

Sprinzak,E. *et al*. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol*., **327**, 919–923.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol*., **12**, 368–373.