BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1623–1630
doi:10.1093/bioinformatics/btl145

Systems biology

Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions

Hon Nian Chua[1,*], Wing-Kin Sung[2] and Limsoon Wong[2]

[1]Graduate School for Integrated Sciences and Engineering and [2]School of Computing, National University of Singapore, Singapore

Received on October 15, 2005; revised on February 14, 2006; accepted on April 11, 2006
Advance Access publication April 21, 2006
Associate Editor: Alvis Brazma

ABSTRACT

Motivation: Most approaches in predicting protein function from protein–protein interaction data utilize the observation that a protein often share functions with proteins that interacts with it (its level-1 neighbours). However, proteins that interact with the same proteins (i.e. level-2 neighbours) may also have a greater likelihood of sharing similar physical or biochemical characteristics. We speculate that functional similarity between a protein and its neighbours from the two different levels arise from two distinct forms of functional association, and a protein is likely to share functions with its level-1 and/or level-2 neighbours. We are interested in finding out how significant is functional association between level-2 neighbours and how they can be exploited for protein function prediction.

Results: We made a statistical study on recent interaction data and observed that functional association between level-2 neighbours is clearly observable. A substantial number of proteins are observed to share functions with level-2 neighbours but not with level-1 neighbours. We develop an algorithm that predicts the functions of a protein in two steps: (1) assign a weight to each of its level-1 and level-2 neighbours by estimating its functional similarity with the protein using the local topology of the interaction network as well as the reliability of experimental sources and (2) scoring each function based on its weighted frequency in these neighbours. Using leave-one-out cross validation, we compare the performance of our method against that of several other existing approaches and show that our method performs relatively well.

Contact: g0306417@nus.edu.sg

1 INTRODUCTION

Conventional methods in predicting protein function from protein interaction data make use of the observation that the direct interaction partners of a protein are likely to share similar functions with it (Schwikowski *et al*., 2000; Hishigaki *et al*., 2001). This is reasonable as it is observed that 70–80% of proteins share at least one function with its interacting partner (Titz *et al*., 2004). Schwikowski *et al*. (2000) adopt a Neighbour Counting approach that labels a protein with the functions that occur most frequently in its interaction partners. Hishigaki *et al*. (2001) use chi-square statistics to identify protein functions that are over-represented in the interaction partners of a protein. However, using only interaction partners limits predictions to proteins that have at least one interaction partner with known annotation. Moreover, the possible annotations for an unknown protein are limited by the annotations of its interacting partners.

Many works have considered using other proteins in the interaction network for protein function prediction. Hishigaki *et al*. (2001) proposed extending the χ^2 statistics approach to include proteins in the interaction neighbourhood of a protein within an arbitrary radius. However, the performance declines when proteins beyond the interaction partners of a protein are considered. Brun *et al*. (2003) and Samanta and Liang (2003) applied clustering techniques to partition the proteome into functional classes (FCs) based on functional distance derived from protein–protein interactions. Others adapted global optimization techniques such as Markov random fields (Letovsky and Kasif, 2003; Deng *et al*., 2003) and simulated annealing (Vazquez *et al*., 2003) to predict protein function and have shown promising results. Lanckriet *et al*. (2004) introduced an integrated support vector machines classifier for function prediction, in which protein–protein interaction data were used to derive one of the kernels using pairwise interaction similarity between proteins based on interaction data. Nabieva *et al*. (2005) introduced a network-based algorithm that simulates functional flow between proteins. While these approaches demonstrated that the use of a variety of machine learning and statistical techniques can improve prediction performance, they bank on the same fundamental concept that the interaction partners of a protein are likely to share similar functions with it.

Conventional approaches associate protein interaction with the sharing of functions: if proteins A and B belong to the same functional pathway, A is likely to interact with B; therefore when A and B are observed to interact, they are likely to share functions. We refer to this as direct functional association. We observed that, in many cases, while a protein shares no function with its level-1 neighbours, it displays substantial function similarity with some of its level-2 neighbours. We refer to this as indirect functional association. The two forms of association are independent and either or both may be observed in the interaction neighbourhood of a protein. Although similar concepts have been utilized in deriving functional distances for some clustering techniques (Brun *et al*., 2003; Samanta and Liang, 2003), the concepts are used in the implicit forms of graphical distances and probabilistic functions.

In this paper we study the significance of indirect functional association in existing protein–protein interaction data from the *Saccharomyces cerevisiae* genome, and propose a new method of protein function prediction that takes into account indirect

*To whom correspondence should be addressed.

functional association. We compare our method with several existing approaches and show that it outperforms them.

2 MATERIALS AND METHODS

In this study, functional annotation scheme of proteins are taken from the most recent FunCat 2.0 functional classification scheme (Ruepp *et al*., 2004). FunCat annotations for *S.cerevisiae* are downloaded from the Comprehensive Yeast Genome Database of the Munich Information Center for Protein Sequences (MIPS) at the time of this work (May 2005). This version of the FunCat scheme consists of 473 FCs arranged in hierarchical order. A protein annotated with an FC is also annotated with all superclasses of that FC. To avoid arriving at misleading conclusions caused by biases in the annotations, we apply the concept of informative FCs from (Zhou *et al*., 2002) on the annotations. We define an informative FC as the one having (1) at least 30 proteins annotated with it and (2) no subclass meeting the requirement (1). A total of 117 informative FCs are derived in this way using the MIPS functional annotations for Yeast. Protein–protein interaction data are downloaded from the GRID database (Breitkreutz *et al*., 2003). The April 18, 2005 release of the YEAST GRID is used in this work. This release reports 19 452 pairs of interactions between yeast proteins, of which 17 811 are unique. The dataset comprises a total of 6701 proteins, of which 4162 are annotated.

2.1 Indirect functional association

A protein interaction network can be represented as an undirected graph $G = (V, E)$ that consist of a set of vertices V and a set of edges E. Each vertex $u \in V$ represents a unique protein, while each edge $(u, v) \in E$ represents an observed interaction between proteins u and v. We define a pair of protein u and v as level-k neighbours if there exists a path $\phi = (u, \ldots, v)$ of length k in G. To make subsequent discussion clearer, we define the set of all pairs of level-k neighbours as S_k. Note that any pair of proteins can be both level-k and level-k' neighbours, where $k \neq k'$. Hence any two sets S_k and $S_{k'}$, $k \neq k'$, may intersect.

Level-1 neighbours interact with each other and are likely to participate in some common pathways. Hence they have an increased likelihood of sharing some functions. This is the underlying biological relevance of direct functional association. The concept of indirect functional association is different, but no less intuitive. Level-2 neighbours interact with some common proteins. Hence they may share some physical or biochemistry characteristics that allow them to bind to these proteins. The more common proteins they interact with, the higher is the chance that they would share some functions.

Figure 1 shows two examples of indirect functional association that we found in existing biological data. In both examples, the level-1 neighbours of the target protein (underlined) did not share any function with it. However, they share functions with a number of their level-2 neighbours. Although we are able to find some specific examples of indirect functional association, these may be purely coincidental. In order to establish support for this form of hypothetical functional association, we try to search for evidence in existing biological data. Using the GRID protein interaction data and MIPS FunCat annotations described earlier, we perform some statistical analysis.

2.2 Significance

We are interested in finding out how often we would observe that a protein shares function with its level-2 neighbours instead of its level-1 neighbours.

From our datasets, we find that out of the 4162 annotated proteins, only 1999 or 48.0% share some function with its level-1 neighbours. Of the remaining proteins, 943 share some similarity with at least one of its level-2 neighbours, making up ∼22.7% of the ORFs. Less than 2% of the annotated proteins share functions exclusively with level-1 neighbours. The statistics are summarized in Table 1. Assuming that there is no unobserved interaction or annotation, indirect functional association would be a reasonable explanation for this observation.

To study the degree of functional similarity exhibited by various set of neighbour pairs, we consider five sets of protein pairs:

(1) Level-1 neighbours that are not Level-2 neighbours (i.e. $S_1 - S_2$);

(2) Level-2 neighbours that are not Level-1 neighbours (i.e. $S_2 - S_1$);

(3) Level-3 neighbours that are not Level-1 or Level-2 neighbours [i.e. $(S_3 - (S_2 \cup S_1)]$;

(4) Level-1 neighbours that are also Level-2 neighbours (i.e. $S_1 \cap S_2$);

(5) All protein pairs in the dataset

Figure 2 illustrates these five sets of protein pairs.

For each of the five sets of protein pairs, we compute the fraction of each set that share some functional similarity based on different levels of the MIPS annotation scheme Table 2 shows the number of protein pairs from each of these sets with known annotations at different levels of the MIPS annotation scheme. Higher levels depict more specific functional annotations. The results are presented in Figure 3.

We can see that protein pairs that are both level-1 and level-2 neighbours ($S_1 \cap S_2$) have the highest likelihood of sharing functions. This is expected since these neighbours display both direct and indirect functional associations with each other. The set of all protein pairs is used as a baseline to indicate the likelihood that any pair of proteins taken randomly from the dataset would show functional similarity with each other. We observe that the set of strict level-2 neighbours ($S_2 - S_1$) displays a higher likelihood of sharing functions than by chance. The set of strict level-3 neighbours $[S_3 - (S_2 \cup S_1)]$ are less likely to share functions but the likelihood is still higher than random. From these observations, we can see that the level-2 and level-3 neighbours of a protein can potentially contribute in inferring its functions. However, as higher level neighbours are defined over more interaction links, functional association between them is inevitably more sensitive to noise in the interaction data. Protein interaction data, as with other high-throughput biological data, contain much noise. In fact, it has been shown that the reliability of high-throughput yeast two-hybrid assays is only ∼50% (Sprinzak *et al*., 2003). Using higher level neighbours in function prediction therefore also increases the chance of including erroneous interaction information. Table 3 shows the number of pairs in each set of defined sets of protein pairs. With each increasing level k, the number of level-k neighbours substantially overwhelms those from the previous levels $(1, \ldots, k-1)$. Hence to improve function prediction by including higher level neighbours, we must first be able to reduce false positives effectively. For this study, we are only able to do this for level-2 neighbours and will hence focus on using these neighbours to improve function prediction.

2.3 Impact on function prediction

To study how well the different sets of neighbours of a protein can be used to infer its function, we use the neighbours in the sets $(S_1 - S_2)$, $(S_2 - S_1)$ and $(S_1 \cap S_2)$ of each protein to predict its functions using the Neighbour Counting method (Schwikowski *et al*., 2000). The Neighbour Counting method predicts the functions of each protein by counting the frequency in which its neighbour has each function. The function that is the n-th most frequent in a protein's level-1 neighbours will be predicted as the n-th most probable function of the protein. The rank of each predicted function is taken as its score. The performance of the predictions is evaluated by plotting precision against recall over varying thresholds as adopted in (Deng *et al*., 2003). For a given threshold β, Precision and Recall are defined as:

$$\text{Precision} = \frac{\sum_{p \in V} k_{p,\beta}}{\sum_{p \in V} m_{p,\beta}} \quad \text{Recall} = \frac{\sum_{p \in V} k_{p,\beta}}{\sum_{p \in V} n_p},$$

where n_p is the number of known functions of protein p; $m_{p,\beta}$ is the number of functions predicted for protein p at threshold β and $k_{p,\beta}$ is the number of functions predicted correctly for protein p at threshold β.

Precision is plotted against Recall for the predictions made using each set of neighbours over varying thresholds in Figure 4. We can see that over the same recall range, the predictions using only proteins from the set $(S_2 - S_1)$

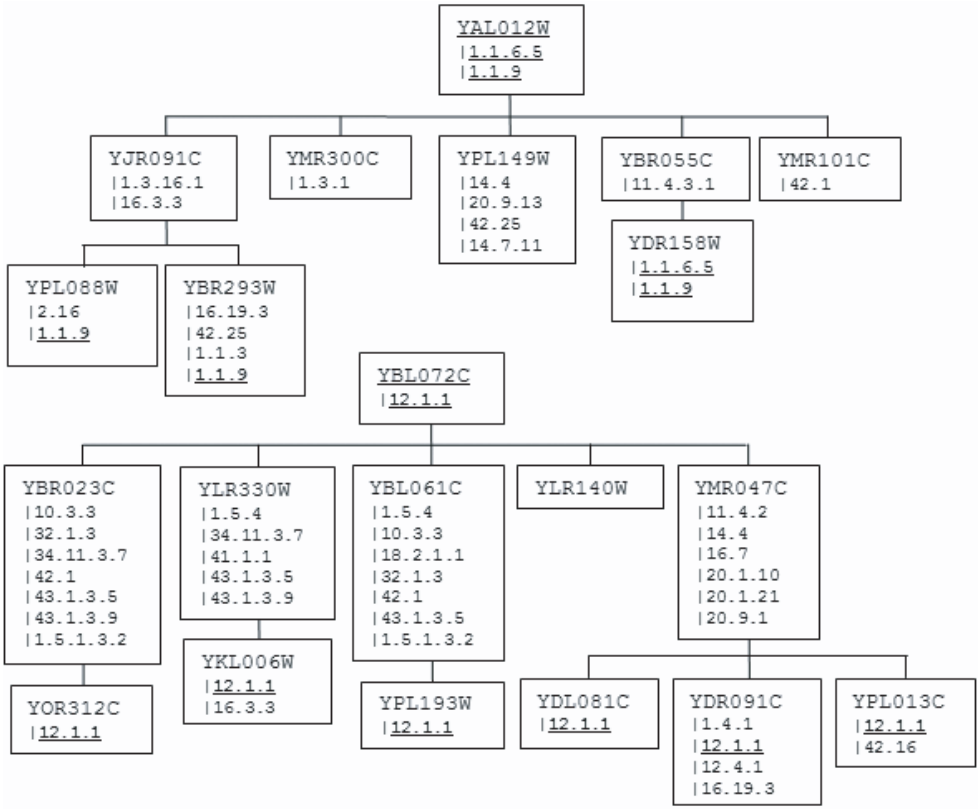


Fig. 1. Examples of indirect functional association in Yeast proteins. YAL012W and YBL072C are presented as the roots of trees in which their level-1 and level-2 neighbours correspond to the level-1 and level-2 child nodes. The level-2 neighbours share some functions (underlined) with the root protein while the level-1 neighbours do not share any functions with the root protein in both cases.

Table 1. Fraction of annotated yeast proteins that share function with (1) level-1 neighbours exclusively; (2) level-2 neighbours exclusively; (3) level-1 and level-2 neighbours; and (4) level-1 or level-2 neighbours

Shared functions with	Fraction
Level-1 neighbours exclusively	0.016338
Level-2 neighbours exclusively	0.226574
Level-1 and Level-2 neighbours	0.463960
Level-1 or Level-2 neighbours	0.706872

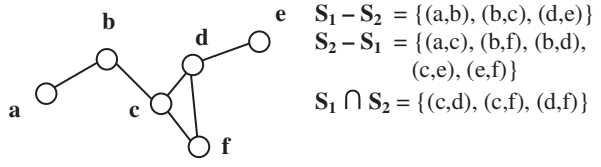


Fig. 2. Example to illustrate the neighbour pairs $(S_1 - S_2)$, $(S_2 - S_1)$ and $(S_1 \cap S_2)$.

have better precision than those from set $(S_1 - S_2)$. Since there is much more protein in the level-2 neighbours, a much broader recall range can also be achieved. This shows that while the level-2 neighbours of a protein may not share as much similarity with it relative to the level-1 neighbours, the functions shared by the level-2 neighbours may be more consistent, and therefore achieve better performance when used to infer the functions of the protein. We also observe that the proteins from the set $(S_1 \cap S_2)$ achieve the best results in inferring function.

Table 2. Number of protein pairs from different sets with known annotations at different levels of MIPS annotations

Annotation level	$S_1 - S_2$	$S_2 - S_1$	$S_3 - (S_2 \cup S_1)$	$S_1 \cap S_2$
0	6979	269 398	1 725 704	8169
1	6895	266 953	1 703 907	8150
2	6250	237 835	1 521 682	7400
3	3136	121 867	728 976	4718
4	497	185 79	94 592	1014
5	1014	80	250	11

3 ALGORITHM

As dicussed in the previous section, the functions of a protein are not only over-represented in its level-1 neighbours, but also in its level-2 and level-3 neighbours. However, we also know that higher level neighbours will inevitably contain more false positives. If we simply extend the Neighbour Counting technique to include level-2 neighbours, any increase in recall is more than offset by

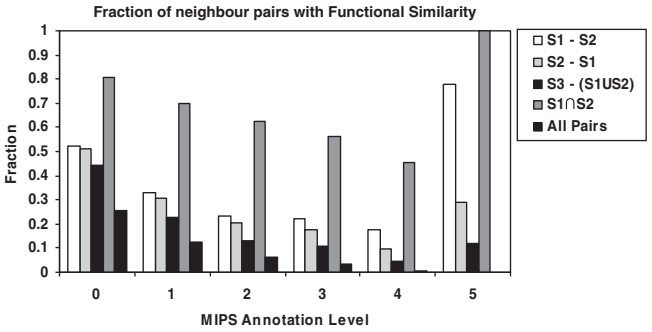


Fig. 3. Fraction of different sets of protein pairs with functional similarity over different levels of MIPS annotations. Higher annotation levels translate to more specific annotations.

Table 3. Pearson correlation values between different metrics and functional similarity for different sets of interaction neighbours

Neighbours	CD-distance	FS-Weight	FS-Weight R	Transitive FS-weight R
S_1	0.471810	0.498745	0.532596	0.532626
S_2	0.224705	0.298843	0.375317	0.381966
$S_1 \cup S_2$	0.224581	0.29629	0.363025	0.369378

the accompanying increase in false positives. Given the noisy nature of high-throughput protein interaction data, some form of filtering or weighting should be employed in order to reduce the effects of including erroneous interactions. We consider two forms of weighting based on local topology and the reliability of different interaction data experimental sources.

3.1 Functional similarity weight

Some existing approaches have suggested using the common interacting partners between two proteins as an estimate of their functional similarity. PRODISTIN (Brun *et al*., 2003) uses the Czekanowski-Dice distance (CD-Distance) as a metric for functional linkage. The CD-distance between two proteins u and v is given by

$$D(u,v) = \frac{|N_u \Delta N_v|}{|N_u \cup N_v| + |N_u \cap N_v|}, \quad (1)$$

where N_p refers to the set that contains p and its level-1 neighbours and $X \Delta Y$ refers to the symmetric difference between two sets X and Y. Note that $D(u,v) < 1$ if u and v are level-1 neighbours. If $N_u = N_v$, $D(u,v)$ will be evaluated to 0. On the other extreme, if $N_u \cap N_v = \emptyset$, $D(u,v)$ will be evaluated to 1. Figure 5 illustrates the computation of the CD-Distance. While the metric is adapted from a statistical measure for categorical data, its basis coincides with the concepts of direct and indirect functional association. When two proteins share many interactors, they are likely to share common functions that allow them to bind to the same proteins. The level-1 and level-2 neighbours of a protein have a CD-Distance of <1 from it while other proteins will have a CD-Distance of 1 from it.

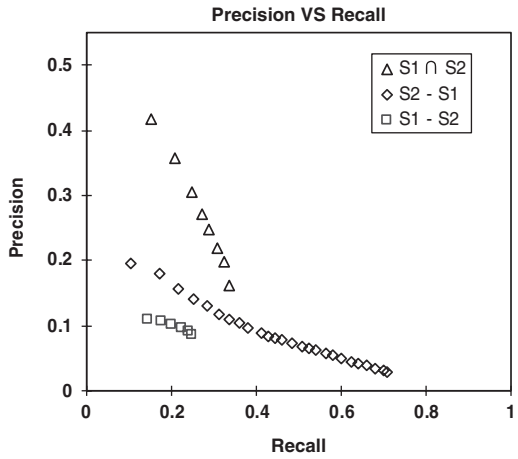


Fig. 4. Precision versus Recall for prediction of protein function using Neighbour Counting with different subsets of interaction neighbours.

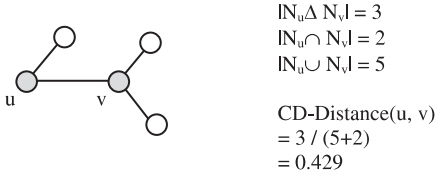


Fig. 5. CD-Distance computation for a pair of proteins u and v.

When proteins u and v interact with some common proteins, CD-Distance estimates the degree of functional similarity between them from the fraction of the level-1 neighbours of both proteins that are common. However, when two proteins interact with a common protein, they may not necessary bind to it at the same binding site. We feel that it would be more appropriate to suggest that when a fraction x of protein u's neighbours is common to protein v's neighbours, x is proportional to the probability that u's functions are shared with v through the common neighbours. Vice versa, if a fraction y of protein v's neighbours is common to protein u's neighbours, y is proportional to the probability that v's functions are shared with u through the common neighbours. Taking the two probabilities to be independent, we estimate the probability that u shares function with v as the product of x and y.

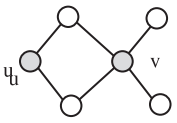
We devise a new measure, functional similarity weight (FS-Weight):

$$S_{\mathrm{FS}}(u,v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v| + \lambda_{u,v}} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v| + \lambda_{v,u}} \quad (2)$$

$\lambda_{u,v}$ is defined as

$$\lambda_{u,v} = \max(0, n_{\mathrm{avg}} - (|N_u - N_v| + |N_u \cap N_v|))$$

$\lambda_{u,v}$ is included in the computation to penalize similarity weights between protein pairs when any of the proteins has very few level-1 neighbours. n_{avg} is the average number of level-1 neighbours that each protein has in the dataset. Like the CD-distance measure,



CD-Distance(u, v)
= 4 / (6+2)
= 0.5(Similarity = 0.5)

FS-Weight(u, v)
= 4/(1+2(2)) x 4/(3+2(2))
= 0.457

Fig. 6. CD-Distance and FS-Weight computation.

FS-Weight gives greater weight to common neighbours than non-common ones.

Figure 6 illustrates the computation of FS-Weight for proteins A and B. For simplicity λ is not included in the computation.

To evaluate the effectiveness of the two measures as an estimator for functional similarity between protein pairs, we compute the Pearson's correlation between CD-Distance and functional similarity for all level-1 and level-2 neighbour pairs from our dataset. We define functional similarity between two proteins u and v, $S(u, v)$, as

$$S(u,v) = \frac{|F_u \cap F_v|}{|F_u \cup F_v|}, \tag{3}$$

where F_p is the set of functions that protein p has.

We categorize the protein pairs into three sets: S_1, S_2 and $S_1 \cup S_2$. Table 3 shows the respective correlation values. We can see that FS-Weight has greater correlation with functional similarity than CD-Distance for all cases.

3.2 Integrating reliability of experimental sources

As shown in Nabieva *et al*. (2005), different experimental sources of deriving protein–protein interaction may have different reliability. Nabieva *et al*. (2005) show that prediction result can be improved substantially when these differences in reliability are taken into account. To estimate the reliability of each experimental source, we follow the approach used in (Nabieva *et al*., 2005), which simply find the fraction of interaction pairs from each source that shares at least one function. The reliability values of the experimental sources derived from our dataset in this manner are presented in Table 4. For each interaction between a pair of proteins u and v, we estimate the reliability of that interaction using

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)^{n_{i,u,v}}, \tag{4}$$

where r_i is the reliability of experimental source i, $E_{u,v}$ is the set of experimental sources in which interaction between u and v is observed, and $n_{i,u,v}$ is the number of times which interaction between u and v is observed from experimental source i.

The reliability of an interaction increases with the number of times it is observed. Observations from different experimental sources contribute to the overall reliability in different degrees. We can now modify the FS-Weight measure defined earlier in (2) to take into account the reliability of each interaction:

$$S_R(u,v) = \frac{2\sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}}{(\sum_{w \in N_u} r_{u,w} + \sum_{w \in (N_u \cap N_v)} r_{u,w}(1 - r_{v,w})) + 2\sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w} + \lambda_{u,v}} \times \frac{2\sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w}}{(\sum_{w \in N_v} r_{v,w} + \sum_{w \in (N_u \cap N_v)} r_{v,w}(1 - r_{u,w})) + 2\sum_{w \in (N_u \cap N_v)} r_{u,w} r_{v,w} + \lambda_{v,u}} \tag{5}$$

Table 4. Estimated reliability for each experimental source in the GRID protein–protein interactions computed using Equation (4)

Source	Reliability
Affinity chromatography	0.823077
Affinity precipitation	0.455904
Biochemical assay	0.666667
Dosage lethality	0.5
Purified complex	0.891473
Reconstituted complex	0.5
Synthetic lethality	0.37386
Synthetic rescue	1
Two hybrid	0.265407

$\lambda_{u,v}$ is modified to take into account only reliable links:

$$\lambda_{u,v} = \max(0, n_{\text{avg}} r_{\text{int}} - (|N_u - N_v| + |N_u \cap N_v|)),$$

where r_{int} is the fraction of all interaction pairs that share some function. The modified FS-Weight measure is evaluated as described earlier and the results are tabulated in Table 3 under the name FS-Weight *R*. The modified measure displays markedly greater correlation with functional similarity for all the sets of neighbours.

3.3 Transitive functional association

If protein u is similar to protein w, and protein w is similar to protein v, proteins u and v may show some degree of similarity. We refer to this as transitive functional association. Independent of other evidence, we estimate the functional similarity between u and v by the product of the functional similarity between u and w, $S(u,w)$, and that between w and v. Taking transitive functional association into account, we modify the FS-Weight measure:

$$S_{TR}(u,v) = \max\left(S_R(u,v), \max_{w \in N_u} S_R(u,w) S_R(w,v)\right), \tag{6}$$

where $S_R(u,v)$ is the FS-Weight score between u and v defined in (5). We refer to this new measure as transitive FS-Weight *R*. The new measure is again evaluated and tabulated in Table 3. We can see that the new measure shows improved correlation with functional similarity for the protein pairs over the earlier measures.

3.4 Functional similarity weighted averaging

Using the FS-Weight measure, we propose a weighted averaging method, FS Weighted Averaging, to predict the function of a protein based on the functions of the level-1 and level-2 neighbours. The

likelihood that a protein p has a function x is estimated by

$$f_x(u) = \frac{1}{Z}\left[\lambda r_{\text{int}}\pi_x + \sum_{v\in N_u}\Big(S_{\text{TR}}(u,v)\delta(v,x) + \sum_{w\in N_v} S_{\text{TR}}(u,w)\delta(w,x)\Big)\right], \quad (7)$$

where $S_{\text{TR}}(u,v)$ is the Transitive FS-Weight R score for u and v defined in (6); r_{int} is the fraction of all interaction pairs that share some function as defined in (5); $\delta(p, x) = 1$ if p has function x, 0 otherwise; π_x is the frequency of function x in annotated proteins; $0 \le \lambda \le 1$ is the weight representing the contribution of background frequency to the score and Z is the sum of all weights, given by

$$Z = 1 + \sum_{v\in N_u}\Big(S_{\text{TR}}(u,v) + \sum_{w\in N_v} S_{\text{TR}}(u,w)\Big). \quad (8)$$

The function $f_x(u)$ is similar to the Neighbour Counting method, using the frequency of occurrence of a function in the neighbours of a protein to estimate the likelihood of the protein having that function. However, there are several key differences:

(1) Level-2 neighbours are included in the counting of function frequency;

(2) The instance of each protein is counted, i.e. if a level-2 neighbour interacts with two different level-1 neighbours, it will be counted twice; level-1 neighbours that are also level-2 neighbours will also contribute more to the score.

(3) A weight is assigned to each neighbour using the FS-Weight measure.

(4) The background frequency of function x, π_x, contributes to the score with a weight λ. When a protein has very few known neighbours or if the neighbours have very small weights, the background frequency will contribute more to the score. We set $\lambda = 1$. λ is a heuristic value and may be empirically determined based on classification performance.

(5) When the reliability is low, FS-Weight will compute lower scores for each neighbour pair. Since the estimation of background frequency will also be inaccurate, λ is multiplied with r_{int}.

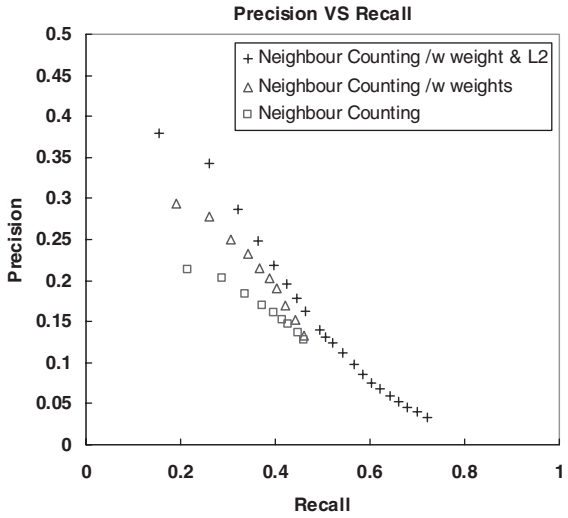


Fig. 7. Precision versus Recall curves for (1) Neighbour Counting; (2) Neighbour Counting with FS-Weight and (3) Neighbour Counting with FS-Weight and level-2 neighbours.

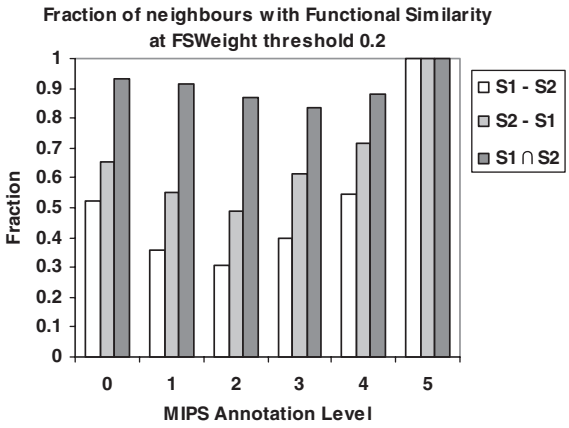


Fig. 8. Fraction of different set of protein neighbour pairs with functional similarity over different levels of MIPS annotations. The protein pairs are filtered with a FS-Weight threshold of 0.2.

4 RESULTS

4.1 Level-2 neighbours and FS-Weight

As discussed earlier, the true potential of using level-2 neighbours for functional prediction can only be unveiled when appropriate filtering is applied to reduce noise. Here we will repeat the statistical computations done for Figure 2 with only neighbours above an FS-Weight threshold of 0.2. The same sets of protein neighbour pairs are studied, without the set of all protein pairs since it is not relevant. The results are displayed in Figure 7. We can see that the fraction of the set $S_2 - S_1$ (exclusively level-2 neighbours) with similar functions has exceeded that of the set $S_1 - S_2$ (exclusively level-1 neighbours) substantially with the application of FS-Weight.

To provide a clear picture of how level-2 neighbours can provide real improvement to the prediction of protein functions, we modify the widely used Neighbour Counting method to include level-2 neighbours weighted with FS-Weight. Three approaches are followed: (1) the original Neighbour Counting; (2) Neighbour Counting with neighbours weighted with FS-Weight and (3) Neighbour Counting with neighbours weighted with FS-Weight and including level-2 neighbours. The Precision versus Recall graph is plotted for the three approaches and is shown in Figure 8. We can see that both the application of FS-Weight and the inclusion of level-2 neighbours substantially improve the performance of this simple prediction method.

4.2 Functional similarity weighted averaging

To evaluate how functional similarity weighted averaging fares, we compare against some of the well-known existing approaches. However, owing to the lack of details in some algorithms, we will

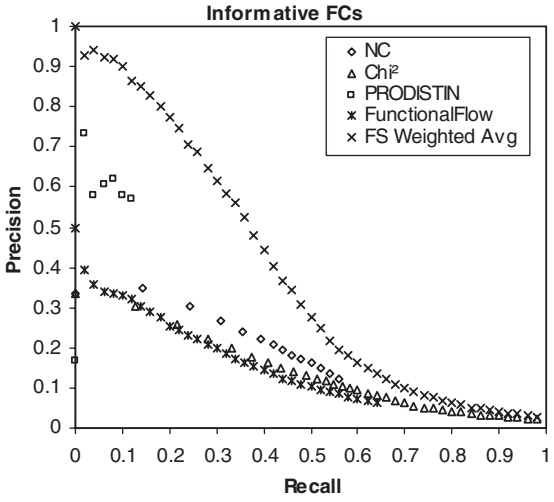


Fig. 9. Precision versus Recall curves for Neighbour Counting (NC), χ^2, PRODISTIN and FS Weighted Averaging in predicting the MIPS Functional Categories for proteins from the GRID interaction dataset.

compare with some approaches based on their datasets. We consider the following approaches:

Neighbour Counting approach. The Neighbour Counting approach labels a protein with the function that is most abundant in its level-1 neighbours. The k most frequent functions are assigned as the k most likely functions for that protein. The rank of the frequency for each function is used instead of the actual frequency count.

χ^2 approach. This is a statistical approach proposed by Hishigaki *et al*. (2001) that make use of χ^2 statistics to take into account the frequency of each function in the entire dataset. The χ^2 statistics of function j for protein i is computed by

$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)},$$

where $n_i(j)$ is the number of level-1 neighbours of i with function j and $e_i(j)$ is the expected number of level-1 neighbours of i with function j, which is derived from $n_i \times \pi_j$ where n_i is the size of i's level-1 neighbours and π_j is the frequency of function j in annotated proteins. The functions with the k largest χ^2 statistics are assigned as the k most likely functions for that protein.

PRODISTIN. It uses the CD-distance between each pair of proteins as a distance metric and clusters the proteins using the BIONJ algorithm. Only the largest connected component in a protein interaction network is used. The BIONJ algorithm produces a hierarchical classification tree. A PRODISTIN FC for a function is defined to be the largest possible subtree in the classification tree that (1) contains at least three proteins having the function and (2) has at least 50% of its annotated members having the function. Un-annotated proteins in the FC are then predicted with the function. We obtain different number of FC (and predictions) by varying the criteria in (2) between 50 and 100%.

Markov random fields. Deng *et al*. (2003) proposed a global optimization method based on random Markov fields and belief propagation to compute a probability that a protein has a function given the functions of all other proteins in the interaction dataset.

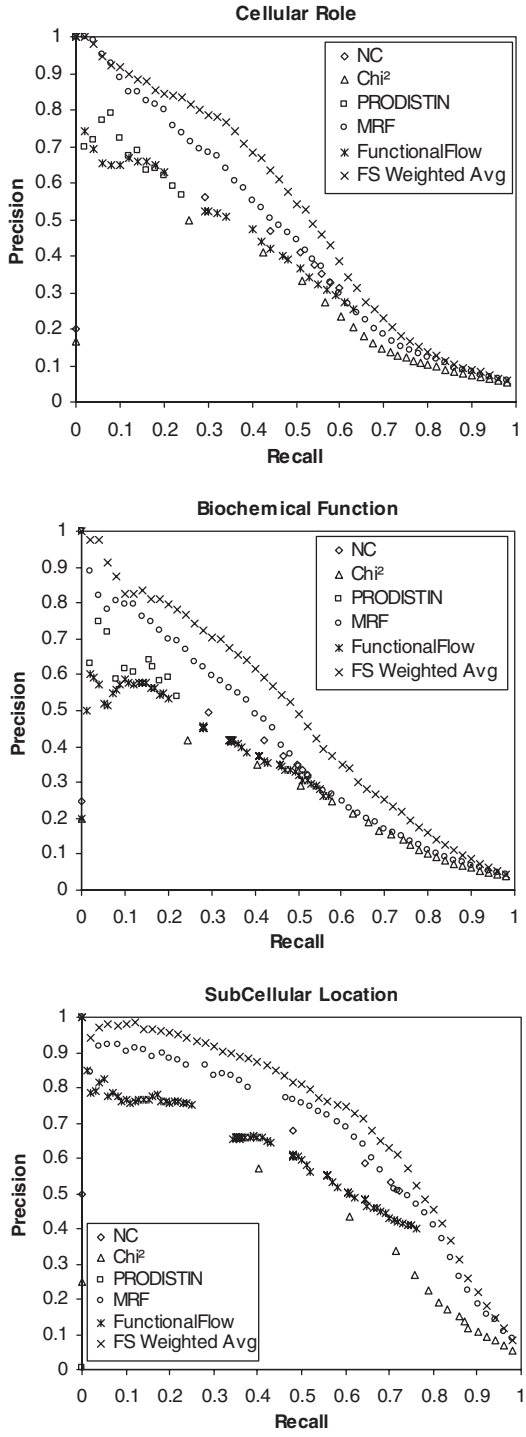


Fig. 10. Precision versus Recall curves for Neighbour Counting (NC), χ^2, Markov Random Fields (MRF), PRODISTIN and FS Weighted Averaging in predicting the Biochemical, Subcellular Locations and Cellular Role of proteins from protein interaction data.

It was shown in Deng *et al*. (2004) that the approach of Vazquez *et al*. (2004) models a special case of Deng *et al*. (2003) while the approach taken by Letovsky and Kasif (2003) is essentially similar to Deng *et al*. (2003).

Functional flow. Nabieva *et al*. (2005) proposes a network-based algorithm that simulates functional flow between proteins. Proteins are initially assigned infinite potential for a function if a protein is annotated with that function and 0 potential otherwise. Functions are then simulated to flow from proteins with higher potential to their level-1 neighbours that have lower potential. The amount of flow is influenced by the reliability of the interactions between interaction partners, which is derived similarly as in our approach. We implemented the Functional Flow algorithm according to the detailed description of the authors in Nabieva *et al*. (2005).

Proteins without any known interaction partners are removed from the dataset following the methodology of Deng *et al*. (2003) to provide a fairer comparison with methods that can only give a prediction to a protein when it has at least one annotated neighbour. This reduces the number of proteins to 4062, with 3326 annotated. Figure 9 shows the Precision versus Recall graph for the different methods based on varying thresholds using the GRID interaction dataset and MIPS FunCat annotations. We did not compare against MRF in this case as we did not implement the approach.

FS Weighted Average significantly outperformed other approaches. The second best approach in the comparison is PRODISTIN. PRODISTIN can only give a prediction for a smaller number of proteins but within its recall range, it achieves much better sensitivity than Neighbour Counting and χ^2.

To compare against the Markov random fields approach, we used the datasets and results provided by the authors, which consisted of protein–protein interaction data from MIPS and functional annotations from Gene Ontology. The functional annotation comprises of three broad categories: biochemical function, subcellular localization and cellular Role. These are predicted separately and the respective Precision versus Recall graphs for the various methods are presented in Figure 10.

As the interaction data for this dataset do not provide well-defined experimental sources, we manually categorized the interactions into several general types so that we can estimate their reliability. FS Weighted Average outperforms MRF as well as the rest of the methods in all the three categories of protein characterization. The relative performances of the different methods are consistent over the two datasets which used different protein interaction data, functional annotations and functional categorization schemes.

5 CONCLUSIONS

We have shown that level-2 and level-3 neighbours show increased likelihood of displaying functional similarity with each other than by random. We also devised a weighting function, FS-Weight, that leverage on both topology and reliability of interactions estimated from the frequency and sources of physical evidence to estimate functional similarity between level-1 and level-2 neighbors. Using FS-Weight, we developed a weighted averaging technique that combines weighted evidence from level-1 and level-2 neighbors with background frequencies to predict protein function and demonstrated that it substantially outperforms many existing methods. Although we found that level-3 neighbours also have higher-than-random likelihood of sharing functions, we are unable to reduce false positives effectively enough to make them useful for function prediction. We plan to study the feasibility of including evidence from level-3 neighbours through devising better ways to reduce false positives from these distant neighbours.

Conflict of Interest: none declared.

REFERENCES

Breitkreutz,B.J. *et al*. (2003) The GRID: the general repository for interaction datasets. *Genome Biol*., **4**, R23.

Brun,C. *et al*. (2003) Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. *Genome Biol*., **5**, R6.

Deng,M. *et al*. (2003) Prediction of protein function using protein–protein interaction data. *J. Comp. Biol*., **10**, 947–960.

Deng,M. *et al*. (2004) Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics*, **20**, 895–902.

Hishigaki,H. *et al*. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.

Lanckriet,G.R. *et al*. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput*., 300–311.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*., **19** (Suppl. 1), i197–i204.

Nabieva,E. *et al*. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21** (Suppl. 1), i302–i310.

Ruepp,A. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–5545.

Samanta,M.P. and Liang,S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl Acad. Sci. USA*, **100**, 12579–12583.

Schwikowski,B. *et al*. (2000) A network of interacting proteins in yeast. *Nat. Biotechnol*., **18**, 1257–1261.

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

Titz,B. *et al*. (2004) What do we learn from high-throughput protein interaction data? *Expert Rev. Proteomics*, **1**, 111–121.

Vazquez,A. *et al*. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol*., **21**, 697–670.

Zhou,X. *et al*. (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*, **99**, 12783–12788.