BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2314–2321
doi:10.1093/bioinformatics/btm342

Systems biology

A statistical approach using network structure in the prediction of protein characteristics

Pao-Yang Chen*, Charlotte M. Deane and Gesine Reinert
Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK

Received on February 15, 2007; revised on June 8, 2007; accepted on June 22, 2007
Advance Access publication June 28, 2007
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The Majority Vote approach has demonstrated that protein–protein interactions can be used to predict the structure or function of a protein. In this article we propose a novel method for the prediction of such protein characteristics based on frequencies of pairwise interactions. In addition, we study a second new approach using the pattern frequencies of triplets of proteins, thus for the first time taking network structure explicitly into account. Both these methods are extended to jointly consider multiple organisms and multiple characteristics.

Results: Compared to the standard non-network-based method, namely the Majority Vote method, in large networks our predictions tend to be more accurate. For structure prediction, the Frequency-based method reaches up to 71% accuracy, and the Triplet-based method reaches up to 72% accuracy, whereas for function prediction, both the Triplet-based method and the Frequency-based method reach up to 90% accuracy. Function prediction on proteins without homologues showed slightly less but comparable accuracies. Including partially annotated proteins substantially increases the number of proteins for which our methods predict their characteristics with reasonable accuracy. We find that the enhanced Triplet-based method does not currently yield significantly better results than the enhanced Frequency-based method, suggesting that triplets of interactions do not contain substantially more information about protein characteristics than interaction pairs. Our methods offer two main improvements over current approaches—first, multiple protein characteristics are considered simultaneously, and second, data is integrated from multiple species. In addition, the Triplet-based method includes network structure more explicitly than the Majority Vote and the Frequency-based method.

Availability: The program is available upon request.

Contact: pchen@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The biological function of a protein within the cell is governed by its protein–protein interactions. While these interactions have recently become widely available for many organisms (e.g. Gavin *et al*., 2002; Uetz *et al*., 2000), they are not yet fully explored with regards to the insights into protein characteristics they might provide.

We now have (just about) enough information to see each protein not only in the context of its immediate neighbours, but also in the overall context of the whole protein–protein interaction network. Moreover, these data-sets allow the examination of multiple species data, its similarities and its differences (Sharan *et al*., 2005), so we can start tackling the question whether data from multiple species improves our ability to predict protein characteristics.

A protein interaction network (*interactome*) is conceptualized as a non-directional graph; proteins are nodes, and interactions between proteins are edges, see, e.g. (Barabasi and Oltvai, 2004; Liu *et al*., 2005). The distances in the network, therefore, refer to graph distance rather than to physical distance, thus focussing on topological properties. In this article, we predict protein function and structure by using not only pairwise protein–protein interactions, but also by explicitly including network structure. We shall see that, beyond pairwise interactions, additional network information does not significantly improve our ability to predict protein characteristics.

Biologically, such lack of improvement is a surprise, and may be due to poor data quality. Protein characteristics, such as function, structure and subcellular location, all affect and are affected by the protein interaction network (Aloy and Russell, 2003; Chou, 2000; Spirin and Mirny, 2003). For instance, functional proteins have been shown to group within the network (Spirin and Mirny, 2003). This effect is expected—proteins will act together to achieve a complex biochemical function, so often neighbours within the network will share common biochemical functions, although not identical chemical functions. In contrast, for protein 3D structures we do not expect clumping of identical structures within the network; instead we would expect patterns of preferred structural partners (Aloy *et al*., 2004). As protein interactions are specific, the 3D structure of the proteins involved should also be specific.

These patterns of interactions for both structure and function have led to the development of prediction algorithms based on the position of a protein within the interaction network (Nabieva *et al*., 2005; Schwikowski *et al*., 2000). In functional prediction, the most popular approach is to observe the functional characteristics which the nearest neighbours of the target protein possess, and to select the function which occurs

*To whom correspondence should be addressed.

most frequently. This simple method, called the *Majority Vote* approach (Schwikowski *et al*., 2000), is one of the most accurate ways of predicting protein function to date. It reached 72% accuracy in predicting 42 functional categories in the top three predictions for yeast. We demonstrate that taking pairwise interactions or triplets of interactions into account can improve on this popular method.

For both structure and function prediction, methods based on interaction networks have so far had limited success (Aloy *et al*., 2004); far more useful techniques use homology of a target sequence to an already solved protein (Zhang and Skolnick, 2005). However, a large number of proteins do not have known homologues (Burley *et al*., 1999; Iliopoulos *et al*., 2001), and these are the target proteins where our methods based on the interactome could provide considerable progress.

Both our methods rely on an upcast set of categories. A protein x is annotated with a set of categories $S(x)$; these categories could relate for example to structure, to function or to subcellular location. The protein–protein interaction network provides a set $B(x)$ of proteins interacting with protein x. The characteristics of these interacting partners, together with the characteristics of x, give an upcast set of triples.

For the *Frequency-based method*, we give a category score based on the counts of relative frequencies of pairwise category–category interactions, and we predict the category with the highest score, which is the most common category in interaction pairs that the protein x is involved in. The method differs from the Majority Vote in that relative frequencies of all category pairs are taken into account.

The *Triple method* and its variants use the lines and triangles of the category interactions in the prediction of protein characteristics. Heuristically, for a protein x we look at all the triples that x is involved in. We then translate these triples into category triples. In the network, the frequencies of different category triples differ considerably. We predict, for x, the category which is 'most common' in the type of triples that x is involved with.

In addition, protein characteristics, such as structure, function and subcellular location are far from independent. We make use of this dependence to improve our predictions by overlaying many characteristics onto the pairs and triples, then use this mixture of information to predict a single characteristic. For instance, the patterns of proteins with a particular structure and subcellular location can be employed to aid prediction of functional category.

Both the Frequency-based method and the Triple method are extended to include additional information on neighbouring protein characteristics, an approach which is not feasible for the Majority Vote method. This inclusion of multiple protein characteristics shows a marked improvement over simpler methods. In the case of function prediction, the use of additional information can improve the accuracy from 61 to 71%. When partially annotated proteins are included to provide more information, the number of prediction is increased. In the case of the Enhanced Frequency-based method by 143 and 63% for structure and function prediction, respectively.

The methods utility is shown in that for function prediction on proteins without homologues (that are therefore not predictable by sequence based approaches) accuracy reaches 89%.

Finally, the inclusion of multiple species data does not dramatically improve the results. It appears that while eukaryotic networks have some predictive power for other eukaryotes, and similarly prokaryotic networks have some predictive power for other prokaryotes, the inclusion of eukaryotic data does not improve prediction for prokaryotic proteins and vice versa. Therefore eukaryotic and prokaryotic protein interaction networks should be treated separately. In particular, this study leads us to propose that there may be some fundamental differences between networks from different kingdoms, whereas networks from the same kingdom display enough similarity to possess some predictive power.

Summarizing, our method offers two main improvements over current approaches—first, multiple protein characteristics are considered simultaneously, and second, data is integrated from multiple species. The results suggest three conclusions: first, that a model for protein–protein interaction networks which is based on pairwise interactions might be suitable. Second, that structure includes information on function, and vice versa; but location may be surplus to requirements when both function and structure information are included. Third, protein–protein interaction networks from different kingdoms are substantially different.

2 MATERIALS AND METHODS

A protein–protein interaction network can be represented as a graph in which proteins are nodes, and two nodes are linked by an (undirected) edge if the corresponding proteins interact. We assume that the network is known and that function and structure are known for all but one protein. The task is to predict function and/or structure for that unknown protein. Our approach is to take the local network structure around the target protein into account. For that purpose, the network structure is modelled as dependent subnets. We construct an auxiliary set of dependent subnets based on the categories which the proteins possess. Due to the lack of annotation for many of the proteins only simple network structures, namely pairs, triples, lines and triangles of three nodes are taken into account.

A triple is a subnet formed by a centre node and two of its neighbours. A triangle is a triple where all three nodes are connected to each other by an edge. A line, by contrast, is a triple in which the two flanking nodes are not connected by an edge.

Rather than only working with data from the organism under study, we enhance our models by using pattern frequencies obtained from pooled interactions in other organisms, an extension which is not feasible for the Majority Vote method. We establish three prior data bases pooling protein–protein interactions collected from, prokaryotes, eukaryotes and both kingdoms. The frequencies of pairs, triples, lines and triangles are counted in these three prior data bases and are employed to predict protein structure and function for the target protein.

In addition to the integration of protein interactions, multiple protein characteristics are considered simultaneously. For the prediction of one protein characteristic, we treat the three cases that one, two or three characteristics, namely structure, function and location, are available for the proteins in our model.

The details of the statistical approach are below. Performance in all cases is evaluated using a leave-one-out cross-validation.

2.1 Protein-protein interaction networks

Experimental protein–protein interactions, excluding self interactions, were obtained from DIP. Self interactions (< 3% of all interactions) are not included in this article so that all triples are constructed of three different proteins. Our method is applied to *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Mus musculus*, *Homo sapiens* and *Escherichia coli*. We have also used *Halicobacter pylori* data when exploring the use of expanding prior datasets, but not for method comparison due to the small sample size. In total, the dataset contains 18 772 proteins with 52 568 interactions.

2.2 Classifications of structure, function and subcellular location

We classify the proteins in our dataset into SCOP classes (Murzin *et al.*, 1995) using the SUPERFAMILY databases (Gough and Chothia, 2002). Between 61 and 89% of proteins are classified. Proteins are classified into 7 distinct classes at the top level of the SCOP hierarchy; see Supplementary Material Table A1. In our analysis, a protein is on average found to be assigned to 1.3 classes.

The protein function categorization we use is based on the 24 functional groups from the second level of molecular function in the Gene Ontology (Ashburner *et al.*, 2000) (see Supplementary Material Table A2). Molecular function ontology in GO has 188 secondary level categories, excluding the categories ‘obsolete’ and ‘unknown’. The 24 groups are those that are most frequently observed. An annotated protein is assigned several nodes in GO, which can be traced back to one or multiple nodes (groups). In our analysis, a protein is on average assigned to 1.2 functional groups. The annotation of 13 subcellular locations from MIPS (Mewes *et al.*, 2002) for yeast is used in our dataset (see Supplementary Material Table A3).

2.3 The upcast set of category–category interactions

From the protein–protein interaction network, we build an upcast set of category–category interactions. A category–category interaction is constructed by two characteristic categories from two interacting proteins.

Consider a protein x, within the set of all characteristic categories S, $S(x)$ includes the categories that protein x is classified into. If two proteins x and y interact, the category–category interaction is the edge between two characteristic categories, a and b ($a \in S(x)$, $b \in S(y)$), from each of two proteins (denoted by $a \sim b$). The upcast set of category–category interactions is a collection of all category–category interactions extracted from the protein–protein interaction network, which may be from one or multiple organisms.

2.4 The Frequency-based method

First, we provide a Frequency-based method, see also Chen (2005), to predict a protein characteristic using protein interactions.

The score for the query protein x with annotated neighbours $B(x)$, to be in a specific category a is proportional to the product $C(a, x)$. This is the product of the relative frequencies f of observing category a for all category–category interactions of x's neighbours in the prior data base;

$$C(a, x) = \prod_{\substack{b \in S(n) \\ n \in B(x)}} f(a \sim b), \tag{1}$$

where $f(a \sim b)$ is the relative frequency of category–category interaction $\{a \sim b\}$ among all category–category interactions.

We define our score $F(a, S(x))$ by

$$F(a, S(x)) := \frac{C(a, x)}{\sum_{k \in S} C(k, x)}.$$

This score is derived as an analogy of the likelihood of observing category a in $S(x)$ if all edges in the category interaction network occurred independently. Heuristically, this score serves as a measure for the chance of protein x having characteristic a.

The protein is then predicted to possess the characteristic category, or categories, with the highest score. This Frequency-based method takes account of both the categories observed in the neighbourhood and their global distribution, while it does not explicitly consider network structure beyond pairwise interactions.

2.5 The Enhanced Frequency-based method

The Frequency-based method can be extended to include two or more protein characteristics in the prediction of a specific protein characteristic. The Enhanced Frequency-based method is similar to the Frequency-based method, only the category in a category–category interaction is now a vector containing all characteristics of the protein. In the case of two protein characteristics, S_1 and S_2, a characteristic vector is a 2-vector with two characteristic categories from S_1 and S_2. While S is now the set of all characteristic vectors, $S(x)$ is the subset of S of the characteristic vectors of protein x,

$$S(x) = \left\{ [s_1, s_2] \middle| s_1 \in S_1(x), s_2 \in S_2(x) \right\}.$$

Given the characteristics of the neighbours, the product of the frequencies of category–category interactions $C(a, x; S_i)$ for a protein x to be in the characteristic category a of S_i is defined, in a similar way to (1), as

$$C(a, x; S_i) = \prod_{\substack{v_b \in S(n) \\ n \in B(x)}} f(v_a \sim v_b, v_{ai} = a),$$

where $v_a = [v_{a1}, v_{a2}]$, $v_{aj} \in S_j(x)\ (j \neq i)$. We add 1 to the relative frequency $f(v_a \sim v_b)$ to avoid the case when $v_a \sim v_b$ exists in unobserved interactions.

The enhanced method requires the target protein annotated with all characteristics except one unknown characteristic and the neighbouring proteins annotated with multiple characteristics. However, there are many partially annotated proteins in the neighbourhood that may provide useful information. These proteins are particularly important when only a few fully annotated ones are available. In Supplementary Material B1, an extended version of our enhanced method is provided which includes partially annotated proteins in the scores.

2.6 The upcast set of triples of characteristic categories

Similar to the upcast set of category–category interactions, we extend the concept of pairwise category–category interactions (as in Section 2.3) to triples of characteristic categories; see Figure 1 for example. A triple is a specific pattern constructed by three categories with two (a line) or three (a triangle) category interactions among them.

For an unannotated protein x and its interacting protein partners u and v, where u and v may or may not interact, the combination of their characteristic categories forms various patterns of triples, $\{b \sim a \sim c \mid b \in S(u), a \in S(x), c \in S(v)\}$. We call such a triple $\{b \sim a \sim c\}$ with $a \in S(x)$ a *triple around protein x*. In order to estimate how frequently a certain triple will occur for protein x and its neighbours, we simply use the frequency of this triple in the upcast set of protein–protein interactions.

2.7 The Triple method and the Line-Triangle method

We now take triples of proteins into account to derive a new score for a protein x to be in a specific characteristic category a; this score is proportional to the product $t(a, x)$ of the relative frequencies f of

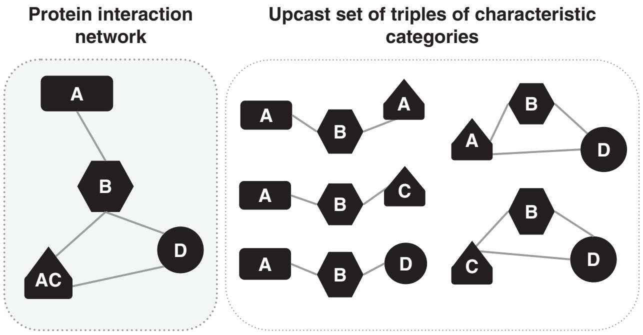
Protein interaction network
Upcast set of triples of characteristic categories
A
B
AC
D
A
B
A
A
B
D
A
B
C
B
A
B
D
C
D

Fig. 1. Upcast set of triples of characteristic categories.
In this example, three single-category proteins and one two-category protein in the protein interaction network result in an upcast set of five triples (three lines and two triangles).

observing category a throughout all triples of x's neighbours in the prior data base;

$$t(a,x) = \prod_{\substack{u \neq v \\ u,v \in B(x)}} \sum_{\substack{b \in S(u) \\ c \in S(v)}} f(b \sim a \sim c),$$

where $f(b \sim a \sim c)$ is the relative frequency of triple $\{b \sim a \sim c\}$ among all triples.

It is possible that some triples are not observed in the prior data base. This may be the case because either the triples do not occur in the real network, or because they have not yet been observed in a study. We therefore add 1 to all frequencies. The more different observed triples the target protein occurs in, the more confident we can be in predicting the characteristic. This confidence is reflected in the following weighting scheme. For each potential characteristic category a in the query protein x, the weight $w(a,x)$ is $(\frac{o}{h})^h$, where $o(a,x)$ is the number of triples around x, assuming x is in category a, observed in the prior data base. Here $h(x) = \sum_{u,v \in B(x)} |S(u)||S(v)|$ is the number of all potential triples around protein x. For example, in structure prediction, suppose that the query protein x has only two neighbours u and v, and that u is in category b and v in c and d. If x is in category a, then the two possible triples $\{b \sim a \sim c\}$ and $\{b \sim a \sim d\}$ result in $h=2$. With only the triple $\{b \sim a \sim c\}$ being observed in the prior data base, which gives $o=1$, the weight $w(a,x)$ is $(\frac{1}{2})^2$. A higher weight indicates more different triples, implying greater confidence in constructing the probability score. We define the weighted score $Q(a, S(x))$ by

$$Q(a, S(x)) := \frac{t(a,x) \cdot w(a,x)}{\sum_{k \in S}[t(k,x) \cdot w(k,x)]}. \tag{2}$$

We predict that protein x possesses characteristic a if a maximizes the weighted score $Q(a, S(x))$.

The Line-Triangle method is an easy extension of the triple method; we separate triples into lines and triangles and use the respective counts. Here, a triple $\{b \sim a \sim c\}$ around protein x is called a *line* if $b \not\sim c$, and it is called a *triangle* if $b \sim c$. If the query protein x is in a protein triangle so that $u \in B(x)$, $v \in B(x)$ and $u \in B(v)$ in the protein interaction network, then all corresponding category triangles are counted. If the query protein x is in a protein line so that $u \in B(x)$, $v \in B(x)$ and $u \notin B(v)$ in the protein interaction network, then all corresponding category lines in the upcast set are counted. The weights are adjusted relating to these frequencies.

We could think of our score as a model of the type

$$Pr(a \in S(x)|X_x^c) \propto \exp\{\log w(a,x) + \log t(a,x)\}, \tag{3}$$

where X_x^c is the network complimentary to node x and edges connecting with it. Related models, called $p^\star$ model, have been in use in social network analysis, see Wasserman and Pattison (1996).

This model (5) assumes that the probability of a characteristic category a is proportional to the log frequencies of the triples. Other factors such as network diameters are not included in the model.

2.8 The Enhanced Triple method and the Enhanced Line-Triangle method

Similar to the Enhanced Frequency-based method in previous Section 2.5, the Enhanced Triple method is an extension of the triple method to include multiple protein characteristics in the prediction of a specific protein characteristic.

Given the characteristics of the neighbours, the product of triple frequencies $t(a,x;S_i)$ for a protein x to be in the characteristic category a of S_i is defined, similarly to (2), as

$$t(a,x;S_i) = \prod_{\substack{u \neq v \\ u,v \in B(x)}} \sum_{\substack{v_b \in S(u) \\ v_c \in S(v)}} f(v_b \sim v_a \sim v_c, v_{ai} = a),$$

where $v_a = [v_{a1}, v_{a2}]$, $v_{aj} \in S_j(x)$ $(j \neq i)$. The weighted score $Q(a, S_i(x))$ in the Enhanced Triple method is given by applying $t(a,x;S_i)$ in (2).

Again, the Enhanced Line-Triangle method is an easy extension of the Enhanced Triple method; we separate triples into lines and triangles and use the respective counts. To include partially annotated proteins, an extended version of the Enhanced Triple method and the Enhanced Line-Triangle method is provided in Supplementary Material B2.

2.9 Combining the enhanced methods

Given the scores from the Enhanced Frequency-based method and from the Enhanced Line-Triangle method, an obvious way forward is to combine these scores. In particular the Line-Triangle score can be used to correct for over-counting in the frequency-based score, in the sense that in the frequency score a triangle would be translated into three counts of pairwise interactions. If there is a strong tendency for transitivity, i.e. for completing the triangle $\{a \sim b \sim c \sim a\}$ given that we see $\{a \sim b \sim c\}$, then the interaction $c \sim a$ is not very surprising, thus should be discounted for. As a guidance for a potential linear relationship between the scores, the coefficients in linear regression are estimated, see Supplementary Material Table H1.

In addition we used a rank-sum approach to reconcile differing predictions from the Enhanced Frequency-based method and from the Enhanced Line-Triangle method, as well as taking the average score, see Supplementary Material Table H2.

3 RESULTS

3.1 Comparison of methods

The DIP subsets from six organisms, *D.melanogaster*, *C.elegans*, *S.cerevisiae*, *M.musculus*, *H.sapiens* and *E.coli*, are analysed. A leave-one-out cross-validation is carried out as follows. Each time a single protein is left out of the prior data base and used as the test data, whereas the other proteins from the same organism are the training data (the prior data base). The frequencies of pairs, triples, lines and triangles are counted in the training data. We then apply the respective weighted probability score using the information from the protein interaction partners of our target protein. Only those proteins which interact with at least two proteins in the prior data base are selected as target proteins.

Table 1. Number of proteins with structure and function annotation

Organism (DIP)	Annotated proteins	Clustering coefficient[a]
D.melanogaster (D.M)	2195	0.03
C.elegans (C.E)	288	0.10
S.cerevisiae (S.C)	2160	0.19
M.musculus (M.M)	138	0.22
H.sapiens (H.S)	594	0.39
E.coli (E.C)	525	0.64

[a]The clustering coefficients are calculated from proteins with at least two annotated neighbours.

Table 2. The accuracies of structure and function prediction using different methods[a]

Organism (DIP)	Predicted proteins	M.V.	F.	E. F.	T.	L-T	E. T.	E. L-T
Structure[b]								
D.M	1262	0.35	0.17	**0.44**	0.15	0.15	0.41	0.41
C.E	78	0.36	0.37	0.49	0.38	0.40	0.45	0.46
S.C	1608	0.39	0.31	**0.54**	0.33	0.31	0.52	0.50
E.C	150	0.57	0.70	**0.71**	0.41	0.35	0.65	0.61
M.M	32	0.72	0.50	0.69	0.69	0.69	0.72	0.72
H.S	273	0.44	0.47	0.71	0.47	0.43	0.71	0.70
Function[c]								
D.M	1275	0.53	0.67	**0.69**	0.67	0.67	0.65	0.65
C.E	85	0.38	0.55	0.71	0.60	0.60	0.64	0.66
S.C	1618	0.67	0.61	0.67	0.67	0.67	0.68	*0.69*
E.C	154	0.69	0.69	0.70	0.69	0.69	0.68	0.66
M.M	32	0.59	0.88	0.81	0.91	0.91	0.88	0.88
H.S	274	0.79	0.90	0.89	0.90	0.89	0.90	0.89

[a]Predicting methods are Majority Vote (M.V.), the Frequency-based method (F.), the Enhanced Frequency-based method (E.F.), the Triple method (T.), the Line-Triangle method (L-T), the Enhanced Triple method (E.T.) and the Enhanced Line-Triangle method (E L-T).

[b]The protein structure is predicted the class with the highest probability.

[c]A function prediction is counted as correct if one of the best three predicted categories is correct.

Underline: where the result outperforms M.V. with statistical significance.

Bold: where E.F. outperforms E.L-T with statistical significance.

Italic: where E.L-T outperforms E.F. with statistical significance.

Our methods are compared to Majority Vote (Schwikowski *et al.*, 2000), which takes interactions, but not network structure, into account.

Tables 1 gives the number of proteins in interaction triples for which both neighbouring proteins have structure as well as function annotation, as required for the Enhanced Line-Triangle method. This is this smallest common datasets that we use for the comparison among different methods.

Tables 2 gives our results for both structure and function prediction. The accuracy is calculated as the ratio between the number of correctly predicted proteins and all predicted proteins. For structure prediction the predicted protein characteristic is the category with the highest probability score. Functional prediction (see Tables 2) is measured based on the highest three probability scores. To help judge statistical significance for accuracies between network and non-network-based methods, we use a normal approximation and perform paired z-tests (Supplementary Material Table D1 for *P*-values). We note that the paired z-test assumes that the predictions for different proteins are independent. While this assumption is most likely not satisfied, we postulate that the dependence is weak enough to still warrant a normal approximation; yet the *P*–values have to be viewed as approximate rather than as exact.

The results both for structure prediction and for function prediction (see Table 2) show that in many organisms the Enhanced Frequency-based method and the Enhanced Line-Triangle method outperform Majority Vote; when they do not outperform the Majority Vote, they gives comparable results. The Enhanced Triple method and the Enhanced Line-Triangle method do not, however, generally outperform the Enhanced Frequency-based method.

The accuracies of function prediction on proteins without homologues was also tested. Non-homologue proteins are selected from Table 1 that meet the criteria for the Enhanced Line-Triangle method (i.e. at least two neighbours and have structure and function annotation). The non-homologue proteins are those having no similar sequence (*E*-value < 0.001) within the same functional group among all DIP organisms. These proteins are not predictable by sequence-based approaches and are considered more difficult targets. Although only a limited number of proteins are tested, the accuracies from our methods are still comparable, see Supplementary Material E.

Table 1 contains the clustering coefficient for the various protein interaction networks. The clustering coefficient for a protein is the ratio between the number of interacting protein pairs in the neighbourhood and all protein pairs in the neighbourhood. The average clustering coefficient for an organism is defined as the average of all clustering coefficients from all proteins with at least two neighbours; it provides a measure of the density of interaction in a network, see, e.g. Dorogovtsev and Mendes (2003). A network clustering coefficient may be affected by the experimental methods used to identify the interactions. For instance small scale techniques concentrating on specific proteins such as those used in the DIP subsets of *M.musculus* and *H.sapiens* give rise to high clustering coefficients. In structure prediction, we observe a trend of increasing accuracy with organisms of higher clustering coefficients suggesting that the prediction improves with clustering, see Supplementary Material Figure D1.

When combining the enhanced methods we were surprised not to find any marked improvements over the single methods. Typically, when the Enhanced Frequency-based method predicted correctly, so does the Enhanced Line-Triangle method, and vice versa. The results are given in Supplementary Material Table H2. In contrast, there is a clear tendency for transitivity, as displayed in Supplementary Material Table I1. While there clearly is more information in the triples compared to the triangles, we conjecture that the noise in the data is to date too high to allow for making good use of the information in the triples.

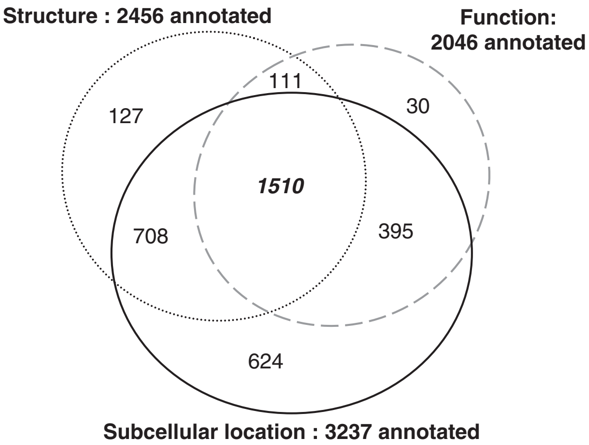
Structure : 2456 annotated
Function:
2046 annotated
127
111
30
1510
708
395
624
Subcellular location : 3237 annotated

Fig. 2. Venn diagram of annotated proteins. The Venn diagram shows the number of proteins having at least two interacting partners with one or multiple annotations for structure, function and subcellular location.

3.2 Integration of structure, function and subcellular location

To date, only for yeast is sufficiently reliable information on structure, function and subcellular location available to warrant including for prediction. These three protein characteristics are analysed separately, pairwise and all combined, to see how additional information aids prediction. In total, 7 structure classes (SCOP), 24 functional groups and 13 subcellular locations are used. Among 4554 proteins, the coverage of annotated proteins are 69, 57 and 86% in structure, function and locations, respectively. The number of proteins having at least two interacting partners with one, two or three annotations are shown in the Venn diagram (Fig. 2).

We start by predicting one characteristic without adding additional information using the Frequency-based method. Then the information from another characteristic is added and the Enhanced Frequency-based method is used. Finally, the information from all three characteristics is included in the model.

Figure 3 shows the result for function prediction. Including both structure and function information significantly improves the predictions compared to only including structure or function information, see Supplementary Material Table F1. Equally, adding location information compared to only including structure or function information significantly improves the prediction. However, once both structure and function are included in the model, it is only for function prediction that including location information still improves the prediction. It appears that only limited additional information can be extracted from the third characteristic. We, therefore, expect our enhanced methods to have similar performance on those organisms without location information.

3.3 Inclusion of partially annotated interacting proteins

The enhanced methods improve the accuracy by integrating multiple characteristics. They also require information from fully annotated neighbours, which may reduce the number of predictions. As described earlier in Section 2.8, our score can be extended to absorb information from partially annotated neighbours and to predict totally unknown proteins. Here, we compare the number of predictions (structure and function) given by different methods. The prediction on the six organisms in Table 1 are pooled and the accuracies are calculated for each method.

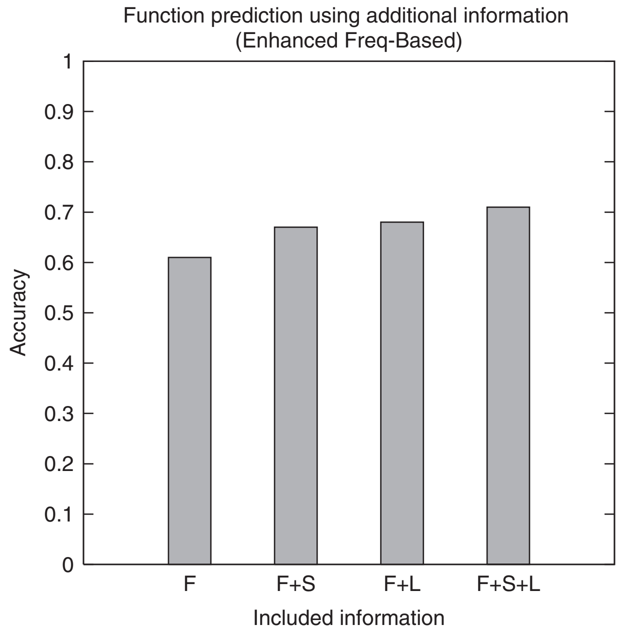

Fig. 3. Function prediction using additional information. The information included in the model function (F) with the additional information from structure (+S), location (+L) and both (+S+L). A prediction is counted as correct if one of the best three predicted functional categories is correct.

Figure 4 shows the results from structure and function prediction. The extended Enhanced Frequency-based method allows far greater coverage for both structure and function predictions with only a small decrease in accuracy. The inclusion of partially annotated proteins considerably improves the coverage of the model.

3.4 Prior data base from pooled protein–protein interactions

When predicting an unknown protein, the frequencies of pairs, lines and triangles suggest how often they are observed in a cell and provide biological information of which categories might interact. They are the basis of the probability score. These frequencies can be obtained by using pooled interactions from multiple organisms as a prior data base. Using a larger prior data base created from multiple species may help in the prediction of a less studied organism, both in specificity and in sensitivity. Here we group protein interactions into prokaryotes, including *E.coli* and *H.pylori*, and eukaryotes, including *C.elegans*, *S.cerevisiae*, *D.melanogaster*, *M.musculus* and

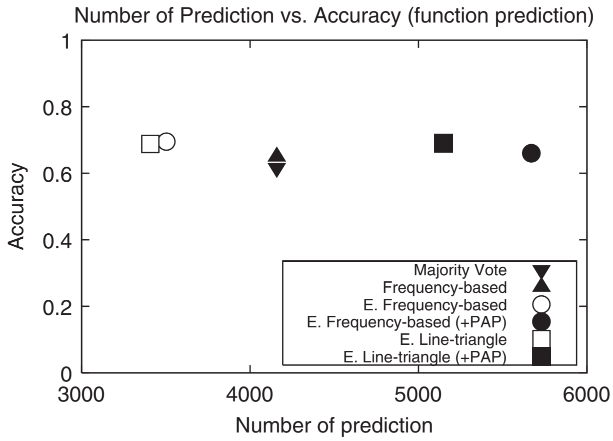

Fig. 4. Number of prediction by methods. The sizes of prediction and the accuracies by different methods in structure (squares) and function prediction (circles). The extended methods (including partially annotated proteins) is noted by +PAP.

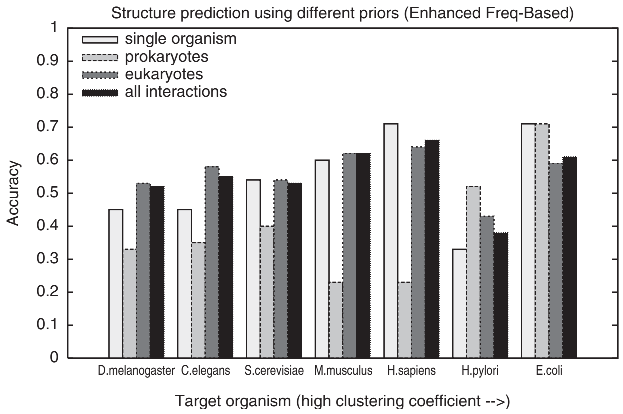

Fig. 5. Structure prediction using pooled interactions as prior data base.

H.sapiens, and a final global pooled dataset including all interactions. We once again predict the structure (7 SCOP classes) and function (24 functional groups) of DIP subsets using the Enhanced Frequency-based method.

In the prediction of structure as shown in Figure 5, when we predict eukaryotes with sparsely clustered protein interactions networks, such as *D.melanogaster*, *C.elegans*, and *S.cerevisiae*, using a prior data base from eukaryotes only, we gain a higher accuracy. On the other hand, the use of pooled interactions does not significantly improve predictions for *M.musculus*, *H.sapiens*, *H.pylori* and *E.coli* (the three highly clustered protein–protein interaction networks with many interactions). For function prediction, see Supplementary Material Figure G2., for *D.melanogaster*, *S.cerevisiae* and *H.sapiens*, the predictions significantly deteriorate when using prokaryotes as prior data. These results indicate that the quality of the prior data base may be more important than the quantity of data.

4 CONCLUSION

We begin our conclusion with two caveats. First, incomplete and biased data in protein interactions make their use in prediction challenging (Deane *et al.*, 2002). Second, our scoring method is based on the heuristic assumption that the likelihood for a specific category to be observed in the query protein is roughly proportional to the product of the relative frequencies of observing this category in all pairs or triples around the neighbours of a query protein, see Equation (2). Our multiplicative scheme has a tendency of to give a high score in the most likely category while the other categories share only a small proportion of the score. Two other scoring schemes, namely the summation and the maximum, were also tested; the multiplicative scheme gave the best results. Our probability score functions can be related to models proposed in social network analysis (Wasserman and Pattison. 1996) and can indeed be viewed as pseudo-likelihoods evaluated at their maximum-likelihood estimates, see e.g. (Cox, 2006); consistency is an issue if the dependence in the data is strong.

Our methods based on network structure show substantial improvement in the prediction of protein characteristics and offer an alternative to sequence-based approaches. Our Enhanced Frequency-based method is never outperformed by Majority Vote, but in contrast significantly improves over Majority Vote in a number of organisms.

It has been previously suggested that it is important to integrate biological information for the prediction of protein characteristics. Our Enhanced methods demonstrate the increased precision which using additional information in the enhanced model can give. The accuracies for function prediction range between 61 and 71% dependent on the amount of additional information. Moreover, the Enhanced methods can be extended to make use of the information from partially annotated proteins. The number of predictions is increased while the accuracy is still higher than Majority Vote.

The results from predicting proteins without homologues show that our methods are able to predict proteins that are not predictable by sequence-based procedures. Our methods can serve as an alternative approach for protein characteristic annotation.

A comparison shows that the Enhanced Triple methods show no marked improvement over the considerably simpler Enhanced Frequency-based method. This phenomenon is in contrast to the tendency towards transitivity in the networks, and warrants further observation. Our explanation is that to date the data in the triangle structures contains too much noise to be of much predictive power.

The results when using prior data bases of pooled interactions from other organisms show that the choice of prior data base is important. A prior data base pooling a large number of interactions can improve the prediction for a poorly studied organism, such as *D.melanogaster* (achieving a higher accuracy and a large number of predictions). But in predicting a eukaryotic organism, the prior data base built only from eukaryotes tends to give more accurate predictions than one from prokaryotes and vice versa. This suggests that there might be a network similarity within kingdoms and that the interaction networks in prokaryotes and eukaryotes may be different. Therefore, it may be more helpful to carefully construct prior data bases from a few well understood organisms from each kingdom rather than to accumulate far more data with low reliability.

Finally, we note that the accuracy of predictions tends to increase with the clustering coefficient. As more physical interactions are experimentally detected, it is anticipated that protein–protein interaction networks will become more compact and therefore our methods will become more accurate.

ACKNOWLEDGEMENTS

We would like to thank the associate editor and the referees for their helpful comments and criticisms. Funded in part by MMCOMNET Grant No. FP6-2003-BEST-Path-012999. Funding to pay the Open Access publication charges for this article was provided by Department of Statistics, Oxford University.

Conflict of Interest: none declared.

REFERENCES

Aloy,P. and Russell,R.B. (2003) Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162.

Aloy,P. *et al*. (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.

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

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet*., **5**, 101–113.

Burley,S.K. *et al*. (1999) Structural genomics: beyond the human genome project. *Nat. Genet*., **23**, 151–157.

Chen,P. (2005) A bayesian approach to predicting protein–protein interactions.. *Transfer report*. Oxford University.

Chou,K.C. (2000) Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci*., **1**, 171–208.

Cox,D. (2006) *Principles of Statistical Inference*, Cambridge University Press, Cambridge. Section 7.6.6.

Deane,C.M. *et al*. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell Proteomics*, **1**, 349–356,1535–9476.

Dorogovtsev,S.N. and Mendes,J.F.F. (2003) *Evolution of Networks : from Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gough,J. and Chothia,C. (2002) Superfamily: Hmms representing all proteins of known structure. scop sequence searches, alignments and genome assignments. *Nucleic Acids Res*., **30**, 268–272.

Iliopoulos,I. *et al*. (2001) Genome sequences and great expectations. *Genome Biol*. **2**, interactions0001.1-0001.3.

Liu,Y. *et al*. (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**, 3279–3285.

Mewes,H.W. *et al*. (2002) Mips: a database for genomes and protein sequences. *Nucleic Acids Res*., **30**, 31–34.

Murzin,A.G. *et al*. (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Nabieva,E. *et al*. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**, I302–I310.

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

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

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

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

Wasserman,S. and Pattison,P. (1996) Logit models and logistic regressions for social networks .1. an introduction to markov graphs and p. *Psychometrika*, **61**, 401–425.

Zhang,Y. and Skolnick,J. (2005) The protein structure prediction problem could be solved using the current pdb library. *Proc. Nat. Acad. Sci. USA*, **102**, 1029–1034.