VERSITA

CELLULAR & MOLECULAR BIOLOGY LETTERS
http://www.cmbl.org.pl

Received: 25 July 2014
Final form accepted: 20 November 2014
Published online: 25 November 2014

Volume 19 (2014) pp 675-691
DOI: 10.2478/s11658-014-0221-5

Short communication

FUNPRED-1: PROTEIN FUNCTION PREDICTION FROM A PROTEIN INTERACTION NETWORK USING NEIGHBORHOOD ANALYSIS

SOVAN SAHA[1], PIYALI CHATTERJEE[2], SUBHADIP BASU[3,] *, MAHANTAPAS KUNDU[3] and MITA NASIPURI[3]

[1]Department of Computer Science and Engineering, Dr. Sudhir Chandra Sur Degree Engineering College, Dumdum, Kolkata-700074, India, [2]Department of Computer Science and Engineering, Netaji Subhash Engineering College, Garia, Kolkata-700152, India, [3]Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India

Abstract: Proteins are responsible for all biological activities in living organisms. Thanks to genome sequencing projects, large amounts of DNA and protein sequence data are now available, but the biological functions of many proteins are still not annotated in most cases. The unknown function of such non-annotated proteins may be inferred or deduced from their neighbors in a protein interaction network. In this paper, we propose two new methods to predict protein functions based on network neighborhood properties. FunPred 1.1 uses a combination of three simple-yet-effective scoring techniques: the neighborhood ratio, the protein path connectivity and the relative functional similarity. FunPred 1.2 applies a heuristic approach using the edge clustering coefficient to reduce the search space by identifying densely connected neighborhood regions. The overall accuracy achieved in FunPred 1.2 over 8 functional groups involving hetero-interactions in 650 yeast proteins is around 87%, which is higher than the accuracy with FunPred 1.1. It is also higher than the accuracy of many of the state-of-the-art protein function prediction methods described in the literature. The test datasets and the complete source code of the developed software are now freely available at https://code.google.com/p/cmater-bioinfo/.

* Author for correspondence. Email: subhadip@cse.jdvu.ac.in

Abbreviations used: BIND – bimolecular interaction network database; DIP – Database of Interacting Proteins; ECC – edge clustering coefficient; HCS – highly connected sub-graphs; LNPC – Laplacian network partitioning correlations; MCODE – molecular complex detection; MIPS – Munich Information Center for Protein Sequences; NMF – non-negative matrix factorization; PPI – protein-protein interactions; RNCS – restricted neighborhood search clustering algorithm; SVM – support vector machine

Keywords: Protein interaction network, Protein function prediction, Functional groups, Neighborhood analysis, Relative functional similarity, Edge clustering coefficient

INTRODUCTION

Proteins are the most versatile macromolecules in living systems. They serve crucial functions in essentially all biological processes. Determining protein functions experimentally is a laborious and time-consuming task, so computational methods of predicting functions are the focus of extensive research. These methods are based on aspects of molecular biology such as the gene and protein sequence and structure, the gene neighborhood, gene fusions, cellular localization, and protein–protein interactions (PPI).

The prediction of protein functions based on protein interaction information is an emerging area of research. In this approach, the functions of non-annotated proteins are determined by looking at their neighborhood properties in the protein interaction network. It is reliant on the fact that the neighbors of a given protein have similar function.

In the work of Schwikowski [1], a neighborhood-counting method is proposed to assign k functions to a protein by identifying the k most frequent functional labels among its interacting partners. It is simple and effective, but the full topology is not considered and no confidence scores are assigned for the annotations.

In the chi-square method, Hishigaki et al. [2] assigns k functions to a protein with the k largest *chi-square* scores. For a protein P, each function f is assigned a score:

$$\frac{(n_f - e_f)^2}{e_f}$$

where n_f is the number of proteins in the n-neighborhood of P that have the function f and e_f is the expectation of this number based on the frequency of f among all proteins in the network.

Chen et al. [3] extends this neighborhood property to higher levels in the network. They developed an algorithm to assess the functional similarity between a protein and its neighbors from its Level –1 and Level –2.

Many graph algorithms have been applied for this type of functional analysis. Vazquez et al. [4] assign proteins to a function to maximize the connectivity of a protein assigned with the same function. They map this problem into an optimization problem using simulated annealing, where they maximize the number of edges that connect proteins (non-annotated or previously annotated) assigned with the same function. Karaoz et al. [5] apply a similar approach to a collection of PPI data and gene expression data. They construct a distinct network for each function in gene ontology. For a particular state of function of each annotated protein v equals +1 if v has the function f, and –1 if v has a different function.

Nabieva et al. [6] proposes a flow-based approach to predict protein function from the protein interaction network. Considering both the local and global properties of the graph, this approach assigns a function to a given non-annotated protein based on the amount of flow it receives during simulation, whereas each annotated protein is the source of functional flow.
Deng et al. [7] proposes an approach employing the theory of the Markov random field where they estimate the posterior probability of a protein of interest. Letvsky and Kasif [8] use loopy belief propagation with the assumption of a binomial model for local neighbors of protein annotated with a given time. Similarly, Wu et al. [9] propose a related probabilistic model to annotate functions of unknown proteins and PPI networks based on the structure of the PPI network.
In the work of Samanta et al. [10], a network-based statistical algorithm is proposed. It assumes that if two proteins share a significantly larger number of common interacting partners, they share a common functionality. Another application, proposed by Arnau et al., is UVCLUSTER. It is based on bi-clustering that iteratively explored distance datasets [11].
Bader and Hogue [12] proposed molecular complex detection (MCODE), where dense regions are detected according to some heuristic parameters. Altaf-ul-Amin et al. [13] also use a clustering approach. It starts from a single node in a graph and clusters are gradually grown until the similarity of every added node within a cluster and the density of clusters reaches a certain limit.
Spirin and Mirny [14] use a graph clustering approach where they detect densely connected modules within themselves and modules that are sparsely connected with the rest of the network based on super paramagnetic clustering and the Monte Carlo algorithm. Pruzli et al. [15] use a graph theoretic approach where clusters are identified using Leda's routine components and those clusters are analyzed by highly connected sub-graphs (HCS) algorithm. King et al. [16] proposed applying the restricted neighborhood search clustering algorithm (RNCS) to partition the interaction networks into clusters using a cost function. Clusters are then filtered according to their size, density and functional homogeneity. Krogan et al. [17] used the Markov clustering algorithm to predict protein function.
In the work of Wang and Ding [18], the problem of predicting protein interactions is formulated from a new mathematical perspective: sparse matrix completion. A novel non-negative matrix factorization-based matrix completion approach is proposed to predict new protein interactions from existing protein interaction networks. Via manifold regularization, this method has been developed to integrate different biological data sources, such as protein sequences, gene expressions, and protein structure information. Extensive experimental results on four species (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Homo sapiens* and *Caenorhabditis elegans*) have shown that these new methods outperform related state-of-the-art protein interaction prediction methods.

This survey reveals that there is scope for the application of domain-specific knowledge to improve the performance of protein function prediction from protein interaction network. Motivated by this, we propose a neighborhood-based method for predicting the function of an uncharacterized protein by computing the neighborhood scores on the basis of protein functions. The uncharacterized protein is assigned with the function corresponding to the highest neighborhood score. We have developed two neighborhood selection approaches where protein functions are predicted on the basis of the functions of direct and indirect neighbors.

DATASET

We used the Munich Information Center for Protein Sequences (MIPS; ftp://ftpmips.gsf.de/yeast/PPI/) database for this study. MIPS is located at the Institute for Bioinformatics (IBI), which is part of the GSF National Research Center for Environment and Health. The MIPS focuses on genome-oriented bioinformatics, in particular the systematic analysis of genome information including the development and application of bioinformatics methods in genome annotation, expression analysis and proteomics.

The database incorporates the protein–protein interaction data of yeast (*Saccharomyces cerevisiae*), which contains 15613 genetic and physical interactions. Discarding self-interactions leaves a set of 12487 unique binary interactions involving 4648 proteins. The complete dataset and the lists of all functional groups are given as supplementary files and available at https://code.google.com/p/cmater-bioinfo/.

It has been observed that there are over 2000 different interaction types. However, most of the proteins are found to be involved in the 8 functional groups considered in our work. These functional groups are cell cycle control (O_1), cell polarity (O_2), cell wall organization and biogenesis (O_3), chromatin chromosome structure (O_4), nuclear-cytoplasmic transport (O_5), pol II transcription (O_6), protein folding (O_7) and protein modification (O_8). For each functional group, 90% of the proteins are chosen as training samples using a random sub-sampling technique and the remaining 10% are considered as test samples.

Since we have considered both Level –1 and Level –2 neighbors, the protein interaction network formed for each protein in any functional group is large and complex. Therefore, in this study, we are focusing on only 10% of the available proteins in each functional group as the test set. Fig. 1 shows the complete protein interaction network considered in this study, where the training set proteins are marked with an elliptical shape and the non-annotated proteins are indicated with a triangle. Table 1 shows the detailed composition of the test dataset for the various functional groups.

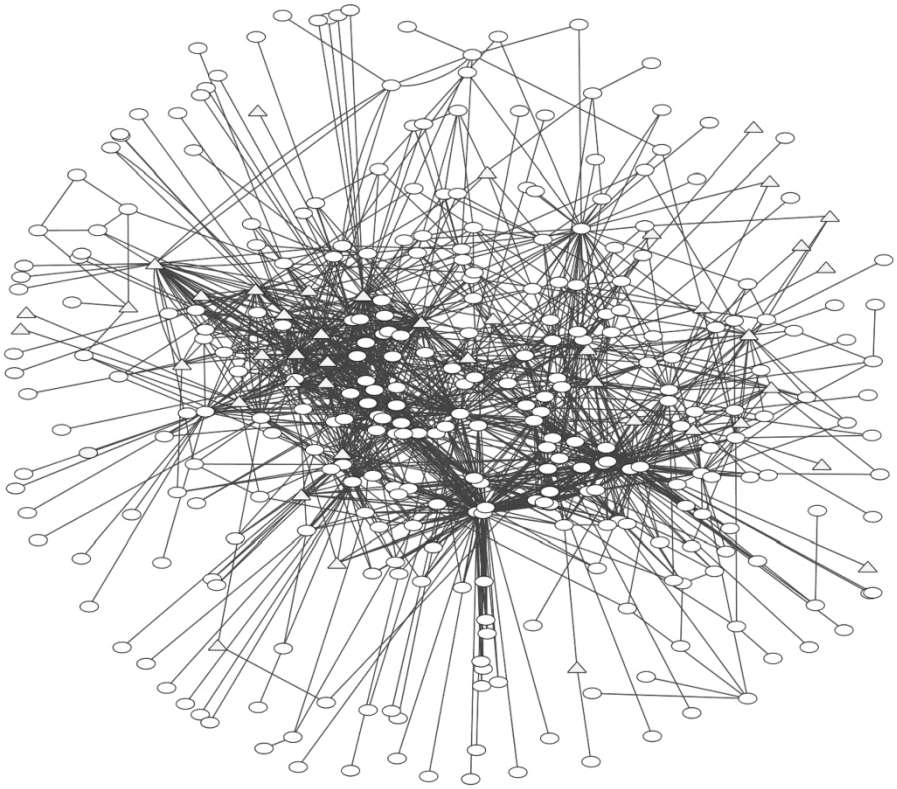

Fig. 1. The overall PPI network considered in this study is shown. The training set proteins are marked as elliptical shapes and the non-annotated proteins are highlighted with triangular shapes.

Table 1. Protein pairs distributed in 8 functional groups that are considered in this study.

Functional groups	Annotated proteins	Non-annotated proteins
Cell cycle control	78	8
Cell polarity	90	9
Cell wall organization and biogenesis	85	9
Chromatin chromosome structure	122	13
Nuclear-cytoplasmic transport	18	2
Pol II transcription	85	9
Protein folding	29	3
Protein modification	81	9
Total	588	62

DEFINITIONS AND NOTATIONS

Protein–protein interaction network

Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. These protein interactions form a network-like structure that is known as a protein interaction network. A protein interaction network is generally represented as a graph consisting of a set of nodes connected by edges or links. Proteins are represented as nodes in the graph and the edges signify interactions between two proteins. Here, the protein interaction network is represented as a graph G_p, which consists of a set of vertices V (nodes) connected by edges E (links). Thus, $G = (V, E)$.

Sub-graph

A graph G'_p is a sub-graph of G_p if the vertex set of G'_p is a subset of the vertex set of G_p and if the edge set of G'_p is a subset of the edge set of G_p. That is, if $G'_p = (V', E')$ and $G_p = (V, E)$ then G'_p is called a sub-graph of G_p if $V' \subseteq V$ and $E' \subseteq E$. G'_p can be defined as a set of $\{P_U | P_A\}$, where P_U represents the set of non-annotated proteins while P_A represents the set of annotated protein.

Level –1 Neighbors

For any vertex v in G'_p all those vertices in G'_p that are connected with v through an edge are deemed Level –1 neighbors of v.

Level –2 Neighbors

In G'_p, Level –2 neighbors are those that are directly connected neighbors of Level –1 neighbors of that particular vertex.

Neighborhood ratio

The neighborhood ratio $P^{l}_{O_i}$ for any protein is defined as the ratio of the number of Level $-$ 1 (or Level $-$ 2) neighbors (K) corresponding to a functional group O_i to the total number of Level $-$ 1 (or Level $-$ 2) neighbors (P). Here, O_i represents any element of 8 functional groups and l denotes Level –1 and Level –2 . It may be defined as $P^{l(=1,2)}_{O_{i(=1..8)}} = \frac{K}{P}$.

Protein neighborhood ratio score

The protein neighborhood ratio score $Pscore^{l(=1,2)}$ is defined as the neighborhood ratio $P^{l}_{O_i}$ of a particular functional group assigned to a unique protein belonging to that respective functional group. Here, O_i represents any element of 8 functional groups, and l denotes Level –1 and Level –2 .

Fig. 2 shows a detailed illustration of the neighborhood relationship between proteins with a non-annotated protein YAL003w from our test dataset. From G_p, $G'_{YAL003w}$ is taken as an example and its Level $-$ 1 neighbors are YAL023c, YAL028w and YAL013w. The Level $-$ 2 neighbors of YAL003w are YFR028c, YAR033w, YOR284w, YIL169c, YAL041w and YNL126w. Two functional

groups (protein folding and cell polarity) are involved in both Level − 1 and Level − 2 which is shown in Fig. 1. Of the Level − 1 neighbors, 2 are for protein folding (O_7) and 1 is for cell polarity (O_2). So, the neighborhood ratio of this protein in Level − 1 for functional group protein folding is computed as: $P_{O_7}^1 = \frac{2}{3} = 0.666$. Using the same procedure, $P_{O_2}^1$, $P_{O_7}^2$ and $P_{O_2}^2$ are computed and their respective values are 0.333, 0.50 and 0.50. We assigned these computed neighborhood ratios for a given functional group to the protein ($Pscore^{l(=1,2)}$) of Level –1 and Level –2 belonging to that group.

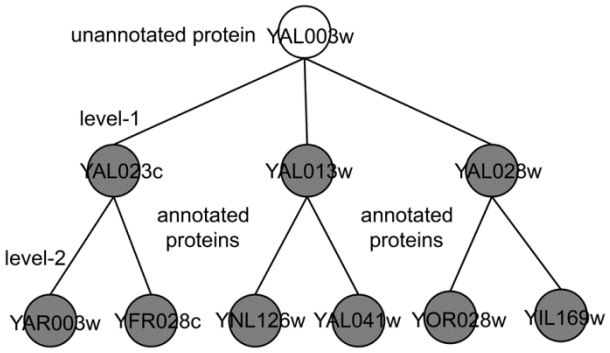

Fig. 2. The sub-graph G'_p of Protein YALOO3w and its Level –1 and Level –2 neighbors.

Relative functional similarity

To compute the functional centrality of the proteins, we take into consideration that the protein functional annotations are in ancestor–descendent relationships. Therefore, the higher the relative functional similarity score [20], the greater the functional similarity between the two proteins. Thus, the relative functional similarity method is a quantitative measure of the similarity of functions between two proteins taking into account its hierarchical structure. It is defined as:

$$W_{u,v}^{l(=1,2)} = ((maxdepth/maxdepth + k) * j) \,/\, (j + h);$$

where, $u \in P_A$, $v \in P_U$, *maxdepth* is the maximum depth of the PPI network, j measures maximum number of common ancestors shared between u and v in a single path, h is the value of the longer distance between u and v to their closest leaf node, k measures shortest distance between u and v, and l denotes Level –1 or Level –2.

Protein path connectivity score

Protein path connectivity score [21] is defined as a measure for network connectivity. It is based on paths between two proteins in an interaction network and is calculated as:

$$Q_{u,v}^{l(=1,2)} = \sum\nolimits_{p \in Ux} 1/L\,(P)$$

where $u \in P_A$, $v \in P_U$, U_x is the set of all paths between u and v with a maximum length of x, p is a path with the length $L(p)$ as a member of U_x, and l denotes the level number (Level –1 and Level –2). According to this definition, proteins with more paths and shorter path-lengths are more tightly connected in the interaction network. See Fig. 3 for an illustration of the sub-graph created for a pair of annotated and non-annotated proteins. The corresponding path connectivity score is shown in Table 2.

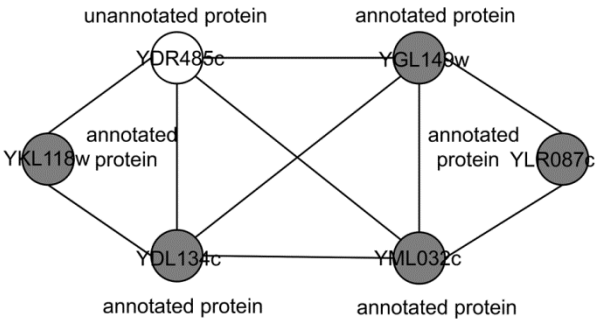

Fig. 3. The sub-graph G'_p for $x = 3$, considering two nodes YDR485c (non-annotated protein) and YGL149w (annotated protein).

Table 2. Estimation of the path connectivity score from Fig. 2.

Connectivity (x)	Path description	Score ($1/L(P)$)
$1-way$	YDR485c - YGL149w	1/1
$2-way$	YDR485c - YML032C - YGL149w	1/2
$2-way$	YDR485c - YDL134c - YGL149w	1/2
$3-way$	YDR485c - YML032c - YLR087c - YGL149w	1/3
$3-way$	YDR485c - YDL134c - YML032c -YGL149w	1/3
$3-way$	YDR485c - YKL118w - YDL134c - YGL149w	1/3
$3-way$	YDR485c - YML032c - YDL134c - YGL149w	1/3
$Q^1_{(YGL149w,YDR485c)} = 1/1 + 1/2 + 1/2 + 1/3 + 1/3 + 1/3 + 1/3 = 3.33$		

Edge clustering coefficient

The connections between nodes are denser in a PPI complex than with the rest in network. The edge clustering coefficient [22] describes how close two proteins are. It is widely used to identify the modularity of networks. The edges with a higher clustering coefficient are more likely to be involved in the community structure in a network. Therefore, a node has a high probability of being essential if it possesses more adjacent edges with higher edge clustering coefficients. Proteins with a high degree may become non-essential if the edge clustering coefficients of their adjacent edges are relatively low. By contrast, those proteins with low connectivity are essential because the edge clustering

coefficients of their adjacent edges are relatively high. The edge clustering coefficient is mathematically defined as follows:

$$ECC_{\mathrm{u,v}}^{\mathrm{l}(=1,2)} = Z_{u,v}/min\ (K_u - 1, K_v - 1)$$

where $u \in P_A$, $v \in P_U$, $Z_{u,v}$ is the number of triangles built on the edge (u, v), K_u and K_v are the degrees of nodes u and v respectively, $min(K_u - 1,\ K_v - 1)$ is the maximal possible number of triangles that might potentially include the edge (u, v), and l is the level-number.

Neighborhood score

The neighborhood score $N_{(\mathrm{O_k})}^{l}$ for any protein is defined initially as the summation of $Pscore^{l(=1,2)}$, $Q_{u,v}^{l(=1,2)}$ and $W_{\mathrm{u,v}}^{\mathrm{l}(=1,2)}$ in FunPred 1.1. In FunPred 1.2 it also incorporates $ECC_{\mathrm{u,v}}^{\mathrm{l}(=1,2)}$. Here, $\mathrm{O_K}$ represents any element of 8 functional groups and *l* denotes Level –1 and Level –2.

PROPOSED METHOD

We have proposed two methods for the prediction of protein function from the protein interaction network. These two methods differ in selection of the neighborhood of the non-annotated proteins over different aspects of neighborhood properties defined in the previous section.

FunPred 1.1

In FunPred 1.1, the prediction technique is based on the combined score of the neighborhood ratio, protein path connectivity and relative functional similarity (as discussed before). This method attempts to find the maximum of the summation of three scores thus obtained in each level and assign the non-annotated protein to the corresponding functional group of the protein having the maximum value. Given G'_P, a sub-graph of the protein interaction network, consisting of proteins as nodes associated with any protein of set $O = \{O_1, O_2, O_3, \ldots, O_8\}$; where O_i represents a particular functional group, this method maps the elements of the set of non-annotated proteins P_U to any element of set O. Steps associated with this method are described as Algorithm 1.

Algorithm 1. Basic methodology of FunPred 1.1

Input: Non-annotated protein set P_U

Output: The elements of the set of non-annotated proteins P_U are mapped to any element of set O.

Step 1: Take any protein as an element from set P_U.

Step 2: Count Level –1 and Level –2 neighbors of that protein in G'_P associated with set O.

Step 3: Compute $P_{O_{i(=1,\ldots,8)}}^{l(=1,2)}$ and assign this score to each protein $(Pscore^{l(=1,2)}) \in P_A$, belonging to the respective functional group.

Step 4: Compute $Q_{\mathrm{u,v}}^{\mathrm{l}(=1,2)}$, $W_{\mathrm{u,v}}^{\mathrm{l}(=1,2)}$ for each edge in Level –1 and Level –2.

Step 5: Obtain the neighborhood score, i.e.

$$N^{l}_{(\mathrm{O_k})} = Max((\max(Pscore^{1} + Q^{1}_{\mathrm{u,v}} + W^{1}_{\mathrm{u,v}})), (\max(Pscore^{2} + Q^{2}_{\mathrm{u,v}} + W^{2}_{\mathrm{u,v}})))$$

Step 6: Assign the non-annotated protein from the set P_U to the functional group O_k.

FunPred 1.2

In FunPred 1.1, for any non-annotated protein, we consider all Level –1 neighbors and Level –2 neighbors belonging to any of 8 functional groups. Prediction is done on the basis of neighborhood property where computation considers all Level –1 and Level –2 neighbors.

However, if the computation is done only on significant neighbors that have maximum neighborhood influence on the protein of interest, then exclusion of non-essential neighbors may reduce the time of computation. This is the basis of our heuristic adopted in FunPred 1.2. Here, we only consider a region or portion of a graph where neighbors are more connected; i.e., densely connected neighbors are considered to be more significant. Using the heuristic that a higher neighborhood ratio may exist in densely connected sub-graphs, the search space in FunPred 1.2 is reduced. It may not always happen that in a densely connected region, the neighbors belong to same functional group. A protein may have many neighbors from different functional groups. Without calculating neighborhood ratios for all of them, this method looks for the promising regions and only then is the calculation of $N^{l}_{(\mathrm{O_k})}$ done. Here, the edge clustering coefficient (ECC) of each edge in Level –1 and Level –2 (as mentioned in the earlier section) is calculated. The edges with relatively low ECC get eliminated. The original network is thus reduced. The original FunPred 1.1 algorithm is applied to this reduced PPI network with the incorporation of the $ECC^{\mathrm{l}(=1,2)}_{\mathrm{u,v}}$ value for each edge in each of the two levels while calculating the value of neighborhood score $N^{l}_{(\mathrm{O_k})}$. The computational steps associated with FunPred 1.2 are described as Algorithm 2.

Algorithm 2. Basic methodology of FunPred 1.2

Input: Non-annotated protein set P_U

Output: The elements of the set of non-annotated proteins P_U are mapped to any element of set O.

Step 1: Take any protein as an element from set P_U.

Step 2: The protein interaction network of the selected protein is constructed with identification of its Level –1 and Level –2 neighbors.

Step 3: Compute $ECC^{\mathrm{l}(=1,2)}_{\mathrm{u,v}}$ for each edge in Level –1 and Level –2.

Step 4: Eliminate non-essential annotated proteins (neighbors) associated with edges having lower values of $\mathrm{ECC}^{\mathrm{l}(=1,2)}_{\mathrm{u,v}}$ both in Level –1 and Level –2, thus generating a densely connected reduced protein interaction network.

Step 5: Count Level –1 and Level –2 neighbors of that protein in $\mathrm{G'_P}$ associated with set O.

Step 6: Compute $P_{O_{i(=1,..,8)}}^{l(=1,2)}$ and assign this score to each protein $(Pscore^{l(=1,2)}) \in P_A$ belonging to the respective functional group.
Step 7: Compute $Q_{u,v}^{l(=1,2)}$, $W_{u,v}^{l(=1,2)}$ for each edge in Level –1 and Level –2.
Step 8: Obtain the neighborhood score i.e.
$N_{(O_k)}^{l} = Max((\max(Pscore^1 + Q_{u,v}^1 + W_{u,v}^1 + ECC_{u,v}^1)), (\max(Pscore^2 + Q_{u,v}^2 + W_{u,v2} + ECC_{u,v2}))$
Step 9: Assign the non-annotated protein from the set P_U to the functional group O_k.

The FunPred 1.2 algorithm is illustrated with an example in Fig. 4. Let us consider the non-annotated protein YCL003w taken from our test dataset. From the sub-graph shown in Fig. 4, the Level –1 neighbors of YCL003w are YAL013w and YAL028w, and their respective next immediate neighbors are YFR028c, YAL041w and YNL126w and YAR033w and YIL169c. Now the edge clustering coefficient of each edge is calculated. The ECC value of the edge (YCL003w, YAL013w) is estimated as 0.581 while that of (YCL003w, YAL028w) is 0.081, which is below our heuristically estimated threshold, $\tau < 0.1$. Therefore, the edge (YCL003w, YAL028w) gets eliminated, resulting in the formation of a densely connected sub-graph. In this process, YAL028w and its next level neighbors are ignored. The computation of the neighborhood score $N_{O_k}^{l}$ is the same as for FunPred 1.1, except for the addition of an ECC score into $N_{O_k}^{l}$.

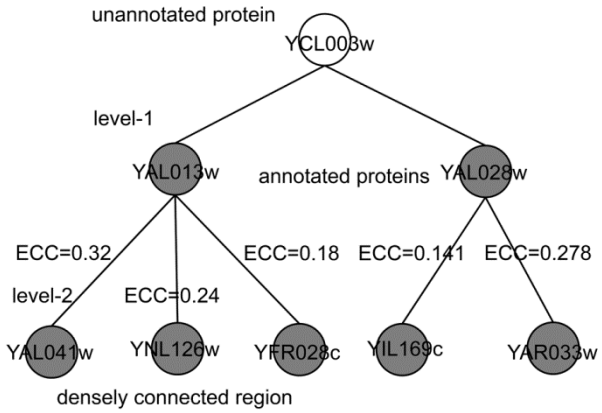

Fig. 4. The sub-graph of YCL003w, including its Level –1 and Level –2 neighbors. Here, only the densely connected region with protein YAL013w is retained and highlighted after ECC computation.

RESULTS AND DISCUSSION

As discussed in the database section, we analyzed 12487 hetero-interactions involving 4648 proteins. However, the 8 functional groups considered in this study cover 650 proteins (see Table 1 for details). In both FunPred 1.1 and FunPred 1.2, 10% of the proteins from each functional group are considered as

non-annotated proteins using random sub-sampling. The training results for the i^{th} functional group are evaluated using standard performance measures, such as Precision (*P*), Recall (*R*) and F-Score (*F*) values, which are calculated using the following equations:

$$P_i = \frac{\sum_{p \in V} k_p}{\sum_{p \in V} m_p}, \quad R_i = \frac{\sum_{p \in V} k_p}{\sum_{p \in V} n_p}, F_i = \frac{2(P_i \times R_i)}{(P_i + R_i)}$$

where for any protein $p \in V$, k_p is the number of correctly predicted proteins for the i^{th} functional group, $m_p$is the total number of proteins predicted in the i^{th} functional group and n_p is the number of all protein in the functional group i. Table 3 shows detailed performance analysis of the two methods for each functional group with respect to Precision, Recall and F-score values. The overall accuracy of any method is estimated using the following formulation:

$$\text{Accuracy} = \frac{\sum_{i=1}^{8} k_p}{Total\ number\ of\ unannotated\ proteins}.$$

The overall accuracies of FunPred 1.1 and FunPred 1.2 are 75.8% and 87%, respectively. Functional group-wise Precision, Recall and F-scores of the two methods are given in Table 3. The average Precision of FunPred 1.2 is estimated as 85.9 ± 11.9% (Table 4). Although we observe relatively low values of Recall and F-scores for the two methods, the high Precision scores indicate that our algorithm returns substantially more relevant results than irrelevant ones.

Table 3. Performance evaluation of FunPred 1.1 and FunPred 1.2 for eight functional groups.

Functional groups	Methods	Precision	Recall	F-Score
Cell cycle control	FunPred-1.1	0.62	0.35	0.46
	FunPred-1.2	0.87	0.50	0.64
Cell polarity	FunPred-1.1	0.88	0.50	0.64
	FunPred-1.2	1	0.56	0.72
Cell wall organization and biogenesis	FunPred-1.1	0.88	0.66	0.75
	FunPred-1.2	0.88	0.66	0.75
Chromatin chromosome structure	FunPred-1.1	0.84	0.78	0.81
	FunPred-1.2	0.92	0.85	0.88
Nuclear-cytoplasmic transport	FunPred-1.1	1	0.66	0.80
	FunPred-1.2	1	0.66	0.80
Pol-II transcription	FunPred-1.1	0.77	0.43	0.55
	FunPred-1.2	0.77	0.43	0.55
Protein folding	FunPred-1.1	0.66	0.40	0.50
	FunPred-1.2	0.66	0.40	0.50
Protein modification	FunPred-1.1	0.44	0.25	0.32
	FunPred-1.2	0.77	0.43	0.55

F-score is a balanced performance measure that estimates the harmonic mean of Precision and Recall. For several functional groups, such as chromatin chromosome structure and nuclear-cytoplasmic transport, we obtained high F-score values. The high standard deviation across the estimated performance measures across different functional groups indicates the inherent complex nature of the problem and wide variability in the data samples.

Table 4. Means and standard deviations of Recall, Precision and F-Score for FunPred 1.1 and FunPred 1.2.

Methods	Mean/SD	Precision	Recall	F-Score
FunPred-1.1	Mean	0.7613	0.5038	0.6019
	Standard deviation	0.1792	0.1810	0.1770
FunPred-1.2	Mean	0.8588	0.5613	0.6724
	Standard deviation	0.1192	0.1545	0.1357

As discussed in the previous section, FunPred 1.2 significantly reduces the neighborhood network. As a result, FunPred 1.2 performs better than FunPred 1.1. For example, in Table 3, we can see an improvement in Precision of 33% in the functional group protein modification and 25% in cell cycle control. In our experiment, the protein folding functional group performs worse than the other groups. In almost all other cases we registered good prediction performances with FunPred 1.1 or obtained significant performance improvement with FunPred 1.2.

A limitation of our current study is the overall low Recall scores, which indicates that we are unable to retrieve most of the relevant results. The best performance is achieved in the functional group chromatin chromosome structure, where the respective Recall scores for FunPred 1.1 and FunPred 1.2 were 78% and 85%, respectively. It may also be worth mentioning here that in our database, we have the highest number of test samples for this particular functional group, and we could annotate this important protein group successfully. Likewise, the low performance in protein folding may be attributed to the lack of availability of annotated proteins.

We identified four state-of-the-art neighborhood analysis methods and compared their performances for our *Saccharomyces cerevisiae* dataset with each other and with our methods. We chose the neighborhood counting method of Schwikowski et al. [1], the chi-square method of Hishigaki et al. [2], a recent version of the neighbor relativity coefficient (NRC) of Moosavi et al. [21] and the FS-weight based method of Chua et al. [23].

The work of Moosavi et al. [21], clearly the strongest of the four methods, focuses on the prediction of three functional groups. The average Precision, Recall and F-scores obtained using their NRC method are 0.374, 0.434 and 0.368, respectively. The performance of our method across 8 functional groups (Table 4) highlights the fact that in terms of average prediction scores, our method is better than the NRC method. This may be because we considered both Level –1 and

Level –2 neighbors and explored a variety of scoring techniques in the protein interaction network, such as neighborhood ratio, protein path connectivity and relative functional similarity. We not only incorporated successors of a specific non-annotated protein but also its ancestors (in relative functional similarity feature) while estimating neighborhood score for function prediction.

For chi-square methods (Chi-square #1 and Chi-square #1 and #2), the weak prediction outcomes may be due to the network sparseness. As claimed by Hishigaki et al. [2], the chi-square methods work better on dense parts of the interaction network.

Simultaneous use of both Level –1 and Level –2 neighbors increases the prediction performance for all other methods considered here except the chi-square method. The neighborhood counting method, despite its simplicity, has notable performance benefits when it uses both Level –1 and Level –2 neighbors. However, since it does not consider any difference between direct and indirect neighbors, it produces lower performance than NRC, FS-weight #1 (only direct neighbors are considered) and FS-weight #1 and #2 (both direct and indirect neighbors are considered) methods in most cases. None of these methods reduces network size by eliminating edges through some scoring techniques to improve prediction accuracy as implemented in FunPred 1.2. Fig. 5 shows a detailed performance comparison of among the four methods (and their variants) along with our proposed systems.

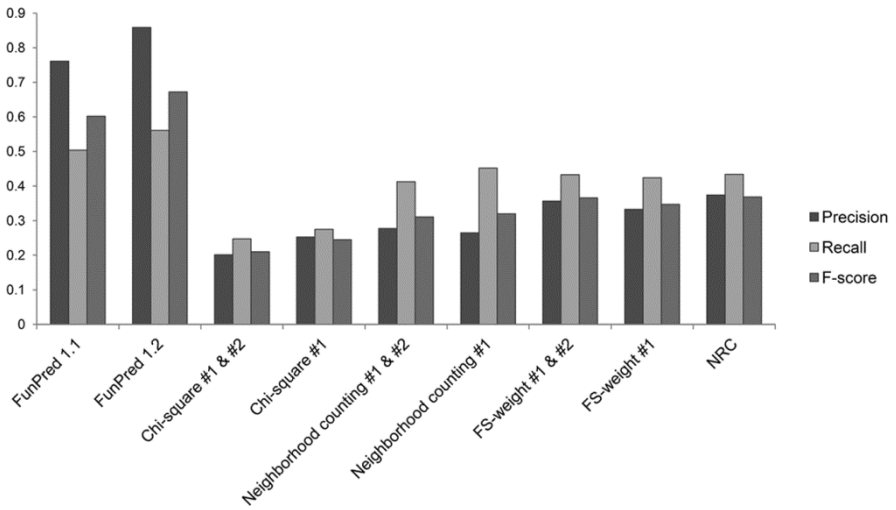

Fig. 5. Comparative analysis of the function prediction performances of related studies on the yeast interaction network is shown. The analysis involved various methods: NRC, FS-weight #1, FS-weight #1 & #2, neighborhood counting using direct neighbors (Neighborhood counting #1), neighborhood counting using both Level –1 and Level –2 neighbors (Neighborhood counting #1 & #2), Chi-square with n = 1 (Chi-square #1) and Chi-square with n = 2 (Chi-square #1 & #2), along with our methods FunPred 1.1 and FunPred 1.2.

Our data show that our proposed FunPred 1 software has better performance than existing function prediction methods. They also show that the network structure may be pruned based on the edge coefficients, leading to improved and faster functional prediction in complex and diverse protein–protein interaction networks. The dataset used in this study and the complete source codes of the FunPred 1 software package are available in the public domain for non-commercial research use at http://code.google.com/p/cmater-bioinfo/.

For performance improvements, domain–domain affinity information may be incorporated in prediction of the protein functions. The PPIs may be decomposed into physical interactions between constituent domains of proteins, using the method proposed in one of our earlier works [19]. The use of domain interaction information in the prediction of protein function may be considered as a future extension of this study.

Different physicochemical properties of amino acids have strong influence on protein structures and their interactions [24]. The use of the physicochemical properties of a protein sequence may give useful information in the prediction of protein function and can be considered as a future study.

This study currently considers 8 functional groups in the yeast PPI network. We would like to extend this to encompass other significant functional groups. Also, we will explore the effectiveness of this method in other organisms, such as in human protein–protein interactions with even more complex network architecture. In a nutshell, we are proposing two useful sets of features for the prediction of protein functions in the complex yeast PPI network with reasonable accuracy.

REFERENCES

1. Schwikowski, B., Uetz, P. and Fields, S. A network of protein-protein interactions in yeast. **Nat. Biotechnol.** 18 (2000) 1257–1261.
2. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. Assessment of prediction accuracy of protein function from protein–protein interaction data. **Yeast (Chichester, England)** 18 (2001) 523–531.
3. Chen, J., Hsu, W., Lee, M.L. and Ng. S.K. Labeling network motifs in protein interactomes for protein function prediction. **IEEE 23rd International Conference on Data Engineering** (2007) 546–555.
4. Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. Global protein function prediction from protein-protein interaction networks. **Nat. Biotechnol.** 21 (2003) 697–700.
5. Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R. and Kasif, S. Whole-genome annotation by using evidence integration in functional-linkage networks. **Proc. Natl. Acad. Sci. USA** 101 (2004) 2888–2893.
6. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. and Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. **Bioinformatics** 21 (2005) i302–i310.

7. Deng, M., Mehta, S., Sun, F. and Chen, T. Inferring domain–domain interactions from protein–protein interactions. **Genome Res.** (2002) 1540–1548.
8. Letovsky, S. and Kasif, S. Predicting protein function from protein/protein interaction data: a probabilistic approach. **Bioinformatics** 19 (2003) i197–i204.
9. Wu, D.D. An efficient approach to detect a protein community from a seed. **Proc. IEEE Symp. Comput. Intel. Bioinforma. Comput. Biol.** (2005) 1–7.
10. Samanta, M.P. and Liang, S. Predicting protein functions from redundancies in large-scale protein interaction networks. **Proc. Natl. Acad. Sci. USA** 100 (2003) 12579–12583.
11. Arnau, V., Mars, S. and Marín, I. Iterative cluster analysis of protein interaction data. **Bioinformatics** 21 (2005) 364–378.
12. Bader, G.D. and Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. **BMC Bioinformatics** 27 (2003) 1–27.
13. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. and Kanaya, S. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. **BMC Bioinformatics** 7 (2006) DOI: 10.1186/1471-2105-7-207.
14. Spirin, V. and Mirny, L.A. Protein complexes and functional modules in molecular networks. **Proc. Natl. Acad. Sci. USA** 100 (2003) 12123–12128.
15. King, A.D., Przulj, N. and Jurisica, I. Protein complex prediction via cost-based clustering. **Bioinformatics** 20 (2004) 3013–3020.
16. Asthana, S., King, O.D., Gibbons, F.D. and Roth, F.P. Predicting protein complex membership using probabilistic network reliability. **Genome Res.** 14 (2004) 1170–1175.
17. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A. , Canete, M.M., Vlasblom, J. Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. **Nature** 440 (2006) 637–643.
18. Wang, H., Huang, H., Ding, C. and Nie, F. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. **J. Comput. Biol.** 20 (2013) 344–358.
19. Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M. and Plewczynski, D. PPI_SVM: prediction of protein-protein interactions using machine learning,

domain-domain affinities and frequency tables. **Cell. Mol. Biol. Lett.** 16 (2011) 264–278.
20. Wu, X., Zhu, L., Guo, J., Zhang, D.Y. and Lin, K. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. **Nucleic Acids Res.** 34 (2006) 2137–2150.
21. Moosavi, S., Rahgozar, M. and Rahimi, A. Protein function prediction using neighbor relativity in protein-protein interaction network. **Comput. Biol. Chem.** 43 (2013) DOI: 10.1016/j.compbiolchem.2012.12.003.
22. Peng, W., Wang, J., Wang, W., Liu, Q., Wu, F.X. and Pan, Y. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. **BMC Syst. Biol.** 6 (2012) DOI: 10.1186/1752-0509-6-87.
23. Chua, H.N., Sung, W.K. and Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. **Bioinformatics** 22 (2006) 1623–1630.
24. Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., and Plewczynski, D. PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. **J. Mol. Model.** 17 (2011) 2191–2201.