IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 3 Ver. V (May -Jun. 2014), PP 30-36
www.iosrjournals.org

PFP_Min: A graph Theoretic Approach for Prediction of Protein Function from Protein-Interaction Network

Chandrima Sinha Roy[1], Piyali Chatterjee[1]
[1] *Netaji Subhash Engineering College, Garia, Kolkata-700152, West Bengal, India*

***Abstract*:** *Proteins are responsible for all biological activities in a living object. With the advent of genome sequencing projects for different organisms, large amounts of DNA and protein sequence data is available, whereas their biological function is still un-annotated in most of the cases. Predicting protein function is the most challenging problem in post-genomic era. Using sequence homology, phylogenetic profiles, gene expression data function of un-annotated protein can be predicted. Recently, the large interaction networks constructed from high-throughput techniques like Yeast2Hybrid experiments are also used in protein function. In this paper, a graph-theoretic approach PFP_Min is proposed for prediction of protein interaction network. This approach considers protein Interaction network as a graph with every protein being an individual node where some of them are assumed to be of unknown function. The objective is to assign function to un-annotated protein based on the minimum cut set. While assigning function to unknown protein, a neighborhood heuristic is also taken to achieve better prediction accuracy.*

***Keywords*:** *Protein Interaction Network, Functional Group, Minimum Cut Set, Connectivity Index, and Neighborhood Heuristic.*

I. Introduction

Proteins serve the most crucial roles in essentially all biological process performed in a cell. With successful sequencing of several genomes, the challenging problem now is to determine functional group of protein in post genomic era. Determining protein functions experimentally is a laborious and time-consuming task involving many resources. Therefore, research is going on to predict protein functions using various computational methods. Protein function can be predicted from different sources such as Protein Sequence, Protein Structure, Gene Sequence, Gene Expression Data, Protein Interaction Network, Bio-Molecular Pathway etc. As experimental techniques (e.g. Isolation, purification, crystallization, and X-ray diffraction) for detection and validation of protein interactions are time consuming and costly e.g., *coimmunoprecipitation* [1] is the biochemical technique of precipitating a protein antigen out of a solution using an antibody that specifically binds to that particular protein. *Biomolecular fluorescence complementation* (BiFC) [2] is another method for actually observing the interactions of proteins based on the association of the fluorescent protein fragments that are attached to the components of the same macromolecular complex. Hence, use of computational methods to predict protein functions is becoming an important and popular area in Bioinformatics. Based on the concept that a protein performs similar function like its neighbor in Protein Interaction Network, a method *PFP_Min* is proposed to predict protein function using protein-protein interaction data. With sequencing projects, billion sequences of proteins have been submitted whose structure and function are not discovered yet, as laboratory based method (e.g. Isolation, purification, crystallization, and X-ray diffraction) is time consuming and costly. So, use of computational methods to predict protein functions is becoming an important area in Bioinformatics. A number of approaches to PPI prediction are based on the use of genome data. Pellegrini *et al*. [3] introduced such method at first which predicts an interaction between two proteins in a given organism if these two proteins have homologs in another organism. A subsequent expansion proposed by Marcotte *et al*. [4, 5] detects co-localization of two genes in different genomes. Two proteins in different organisms are predicted to interact if they have consecutive homologs in a single organism. Dandekar *et al*. [6] used the adjacency of genes in various bacterial genomes to predict functional relationships between the corresponding proteins. Proteins whose genes are physically close in the genomes of various organisms are predicted to interact. Jasen *et al*. [7] investigated the relationship between protein-protein interaction and mRNA expression levels by analyzing existing yeast data from a variety of sources and identifying general trends. Two different approaches were used to analyze the two types of available expression data; normalized differences were computed for absolute expression levels, while a more standard analysis of profile correlations was applied to relative expression levels. This investigation indicated that a strong relationship exists between expression data and most permanent protein complexes. By clustering based approaches, protein function can be predicted and the methods are Distance based clustering and Graph based clustering. In PPI network, the binary vectors $X_i = (x_{i1}, x_{i2}, \dots\dots, x_{iN})$ represent the set of protein purifications for *N* proteins, where x_{ik} is 1 if the i^{th} protein interacts with k^{th} protein (the k^{th} protein is presented in the i^{th} purification) and 0 otherwise. If a distance can be determined which fully

accounts for known protein complexes, unsupervised hierarchical clustering methods can be used to accurately assemble protein complexes from the data. A protein-protein interaction network is an unweighted graph in which the weight of each edge between any two proteins is either 1 or 0. Graph-based clustering techniques are explicitly presented in terms of a graph, thus converting the process of clustering a dataset into such graph-theoretical problems as finding a minimum cut or maximal subgraphs in the graph G. In Graph based clustering, several methods are there and the methods are Monte Carlo optimization, Molecular complex detection and Min-Cut. Among min-cut graph-theoretic approaches, Highly Connected Subgraph algorithm and Restricted Neighborhood Search Clustering algorithm are there. The Highly-connected subgraph or HCS method [8] is a graph-theoretic algorithm which separates a graph into several subgraphs using minimum cuts. The resulting subgraphs satisfy a specified density threshold. Despite its interest in density, this method differs from approaches discussed earlier which seek to identify the densest subgraphs. Rather, it exploits the inherent connectivity of the graph and cuts the most unimportant edges to find highly-connected subgraphs. Some graph-theoretic concepts should first be defined at this point. King *et al*. [9] proposed a cost-based local search algorithm. In the algorithm, a clustering of a graph $G = (V, E)$ is defined as a partitioning of the node set V. The process begins with an initial random or user-input clustering and defines a cost function. Nodes are then randomly added to or removed from clusters to find a partition with minimum cost. The cost function is based on the number of invalid connections. An invalid connection incident with v is a connection that exists between v and a node in a different cluster, or, alternatively, a connection that does not exist between v and a node u in the same cluster as v. The process begins with an initial random or user-input clustering and defines a cost function. Nodes are then randomly added to or removed from clusters to find a partition with minimum cost. The cost function is based on the number of invalid connections. In this algorithm they have discussed about two cost functions to find the minimum cut and one of them is naïve cost function and the other one is scaled cost function. Md. Altaf-Ul-Amin and co-authors [10] proposed a method of protein function prediction based on protein interaction network on min-cut considering degree of protein. From the above mentioned work, it is clear that there remains scope for work. The application of graph theory and the use of neighborhood property of an uncharacterized protein may be useful for prediction of its function. PFP_Min is based on this concept.

II. Materials and Methods

2.1 Basic Terminologies

Protein Interaction Network

A protein almost never performs its functions in isolation. Rather, it usually interacts with other biological entities such as DNA, RNA, as well as other proteins to accomplish a certain function. Hence the function of a protein may be inferred by looking at its interaction neighborhood. While interaction takes place between a known protein and an unknown protein, the behavior of unknown protein can be predicted from feature of the known protein. The interaction between proteins is the most significant characteristics of protein functions. Most proteins perform their functions by interacting with other proteins. Since a protein generally interacts with more than one protein, these interactions can be structured to form a network, and hence the name *protein interaction network.* A very common way of visualizing these networks is as undirected graphs, with the proteins acting as the nodes and the pair wise interactions acting as the edges of the graph. Such a representation can enable researchers to infer characteristics of proteins from those proteins not even directly interacting with it.

Cut-set

In a connected graph, a *cut–set* [11] is a set of edges whose removal from G leaves G disconnected, provided a removal of no proper subset of these edges disconnects G. A cut-set can also be defined as a minimal set of edges in a connected graph whose removal leaves G disconnected. For instance, in figure 1 the set of edges $\{a, c, d, f\}$ is a cut-set. There are many other cut-sets, such as edges $\{a, b, g\}, \{a, b, e, f\}$ and $\{d, h, f\}$. Edge $\{k\}$ alone is also a cut-set. The set of edges $\{a, c, h, d\}$ on the other hand, is not a cut-set, because one of its proper subsets, $\{a, c, h\}$, is a cut-set.

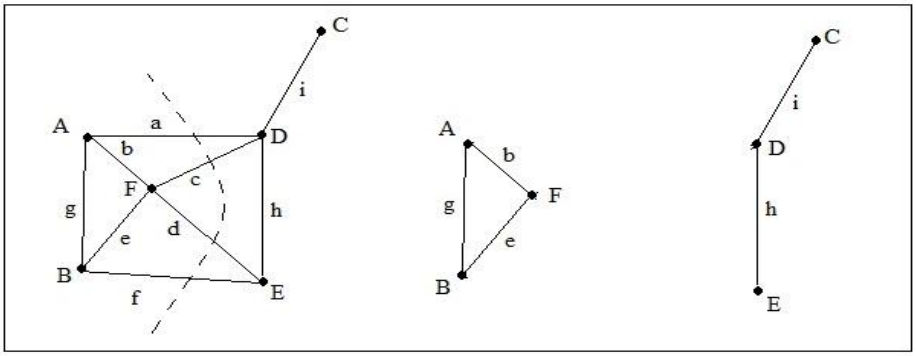

Figure 1. *Removal of a cut-set* $\{a, c, d, f\}$ *from a graph "cuts" it into two.*

Connectivity Index (CI_G)

Here, another graph parameter is proposed which is *Connectivity Index* (CI_G). It is computed by calculating interconnections among the vertices within a module of a graph G. Here, a module is denoted by a subgraph of proteins of the same functional group .It is discussed in detail in the next section.

2.2 PFP_Min : The Approach

Here a graph-theoretic approach, namely PFP_Min, is proposed which is for prediction of protein functions. It considers protein interaction network as a graph, where every protein being an individual node and some of them are assumed to be of unknown function. The objective is to assign function to un-annotated protein based on the minimum cut set. While assigning function to unknown protein, a neighborhood heuristic is also taken to achieve better prediction accuracy. In this work, different functionalities of proteins are considered. Cluster is formed with the proteins of similar functionality. Thus different clusters are created for different functionalities with the assumption that each protein is involved in a particular task. The graph contains different clusters of known protein and unknown proteins whose interaction information is known but not their functions. Considering the fact that proteins of similar functions are likely to interact, initially, unknown protein which has direct interaction with known protein, is assigned to its corresponding cluster. While assigning it, PFP_Min also inspects its immediate neighbor which is also unknown protein. If exists, then PFP_Min employs a neighborhood heuristic i.e. neighbor may be also a member of the same cluster. Using this Heuristic, direct unknown neighbor of the first unknown protein is also assigned to the same cluster. Neighborhood heuristic does not guarantee the actual assignment for a particular level-1 unknown protein in all the cases but it works accurately in most of the cases and ensures the better time complexity thereby executing simultaneous assignment. Thus, from various interaction information of unknown protein, several assignments can be done and a number of topology is obtained. In different topologies, assignment of unknown protein to different clusters is different. PFP_Min computes cut-set for every topology and saves one which has minimum cut-set. It also computes the interactions within clusters for every graph topology. PFP_Min considers a topology which has minimum cut-set and maximum connectivity index.

2.3 Illustration with an example

In this method, initially, uncharacterized protein (candidate) is assigned to a functional group (or cluster) depending upon its maximum interactions to that functional group and at the same time its level-1 uncharacterized neighbor (if exists) is also assigned to the same group. Thus every uncharacterized protein is assigned to particular functional group. But, at a particular instant of time, any unknown protein is considered to be either candidate protein or level-1 neighbor of any candidate protein. Thus, by exchanging the status of uncharacterized protein (either candidate or level-1 neighbor of candidate protein) a number of topologies can be found. The topology for which minimum cut-set is achieved is saved and assignment of uncharacterized protein in that topology is taken up as final assignment. This minimized cut set also leads every cluster having maximum intra connections. For example, if we look at figure 2, we can find interaction between known protein K_3 with unknown protein U_4, hence, U_4 is added to G_1, also U_1, U_2, U_3 proteins (which are level-1 proteins of U_4) are included into G_1. In the same manner U_5 and U_6 are included into G_1 and U_7 is included into G_3. Now, all unknown proteins are added into groups (U_8 not included as it is a level-2 protein and here we are not considering level-2 proteins) and CUT has been found by calculating inter group interactions between known–unknown proteins and unknown–unknown proteins. In this way we found cut-Set = 7 for figure 2.

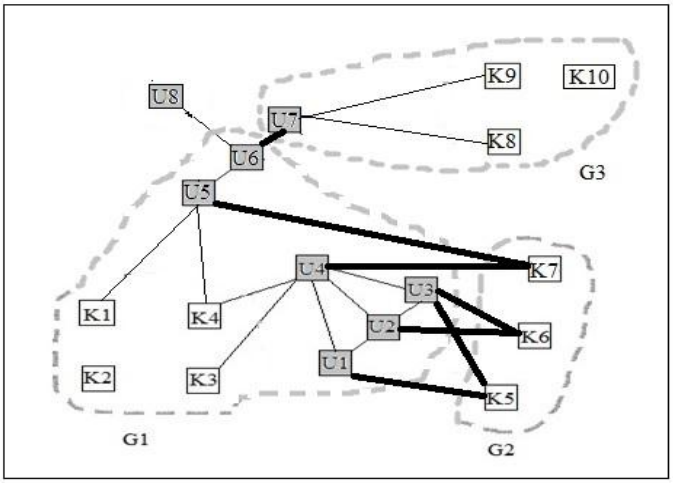

Figure 2. *Topology-A after assignment of unknown proteins*

In the next pass, all assignment of unknown proteins is withdrawn and new assignment is done from different interaction information. Now, again the same procedure is followed which is mentioned above in figure 2, only difference lies in the case of U_4 and its level-1 proteins which are considered to be the members of G_2, as interaction between U_4 with K_7 is found. In the same manner we have put U_5 and its level-1 (U_6 protein) into G_2 and U_7 in G_3. After inclusion of all unknown proteins into groups, we have calculated Cut-Set as 5 which is shown in figure 3.

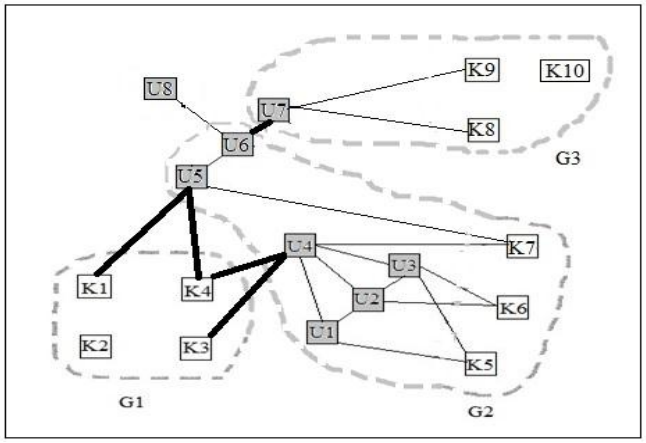

Figure 3. *Topology-B after different assignment of unknown proteins*

Between these two scenarios, we can see min-cut has been found for second scenario. Hence we are considering the second case as the best case between these two cases and predict function of unknown proteins in this way. Next section describes the computational steps of our approach. In case of figure 2, we found 7 edges as cut-set and for G_1 we can see connectivity index is 10 and for G_2 connectivity index is 0 and for G_3 connectivity index is 2, so total connectivity index has become 12, but in figure 3 we found cut-set as 5 and for $G_{1,}$ connectivity index is 0, for G_2 connectivity index is 12 and for G_3 we have 2 edges as connectivity index, so total connectivity index for figure 3 is 14. Hence, we can conclude that for minimized cut-set we can get maximized connectivity index.

2.4 Algorithm: PFP_Min

begin
for each vertex in *G do*
form a set *K* consisting of vertex with known Functional Group.
form a set *U* consisting of vertex without having any Functional Group.
for each vertex in *G* with known Functional Group,

form $G_1, G_2, \ldots\ldots, G_n$ set such that any G_i contains vertices with having same Functional Group and $G_1 \cap G_2 \cap \ldots\ldots \cap G_n = \emptyset$ and $K = G_1 \cup G_2 \cup \ldots\ldots \cup G_n$. Here, n represents the number of unique Functional Group.

repeat

Step 1: Select an element u_1 from *U*. Make u_1 an element of any G_1 or G_2 or …… G_n having edges with u_1.
find level-1 neighbor u_1', of u_1 which belongs to the set *U*.
remove u_1' from u_1 and make it an element of previous set of u_1 until *U* becomes empty.
save $T_j = G_1 \cup G_2 \cup \ldots\ldots \cup G_n$.

//min-cut calculation
Step 2: in T_j, count the number of edges among $G_1, G_2, \ldots\ldots, G_n$
$T_{j\ cut-set}$ = number of edges between $G_1, G_2, \ldots\ldots, G_n$
goto Step 1

Step 3: min-cut = minimum $(T_{j\ cut-set})$j>0 // j denotes the number of topologies
end

III. Result and Discussion

3.1 Data Set

The performance of PFP_Min is observed on some selected proteins of different functional groups. Here we have applied it to yeast *Saccharomyces Cerevisiae* protein-protein interaction network. We obtain the protein-protein interaction data from ftp://ftpmips.gsf.de/yeast/PPI/, which contains 15613 genetic and physical interactions. We have discarded self-interactions and extract a set of unique binary interactions. In this work, only 12 functional groups are considered, namely, Amino-acid Metabolism, Cell Cycle Control, Cell Polarity, Cell Structure, DNA Repair, DNA Replication, Lipid Metabolism, Meiosis, Mitosis, Protein Modification, Protein Synthesis and Vesicular Transport. Thus, a Protein interaction network is created from the 455 protein sequences where 368 protein sequences are of 12 functional groups and the remaining 87 protein sequences serve as test data as their functional groups are assumed to be unknown. Test protein sequences are selected randomly from the protein of known functional groups. Generally, 10% proteins from each functional group are chosen as test sequences. The network, thus formed, is on the basis of their interaction information.

Table 1. *Number of Proteins in different functional groups*

Name of Functional Groups	Number of Known Proteins
Amino-acid Metabolism	10
Cell Cycle Control	21
Cell Polarity	39
Cell Structure	22
DNA Repair	39
DNA Replication	6
Lipid Metabolism	38
Meiosis	21
Mitosis	22
Protein Modification	35
Protein Synthesis	46
Vesicular Transport	69

3.2 Performance Evaluation Metric

Overall Accuracy is a ratio of number of proteins whose functional groups are predicted to be correct and the total number of candidate proteins. Mathematically, it can be represented as,

Overall Accuracy = $\frac{Number\ of\ proteins\ whose\ predicted\ functional\ group\ is\ correct}{Total\ Number\ of\ candidate\ proteins} \times 100\%$

3.3 Performance Evaluation of PFP_Min

On execution of PFP_Min for twelve functional groups, 12 topologies have been found. It is clearly understood that Topology-3 is the optimal configuration though Topology-12 is giving the min-cut with lowest intergroup interactions (1115) but we are considering Topology-3 as the best min-cut because if we look at accuracy then Topology-3 offers the maximum prediction accuracy which is 35.14%, also intergroup

interactions and connectivity index of Topology-3 is very close to Topology-12, hence as per our heuristics we have chosen Topology-3 as the optimal configuration. Table 2 shows number of intergroup interactions, connectivity indices and accuracy measures achieved by PFP_Min for different topologies.

Table 2. *Accuracy Measures achieved by PFP_Min for twelve functional groups*

Group_No	Topology#x	Total Interactions	Accuracy (%)	Connectivity Index
1	Topology-1	1364	8.43	73
2	Topology-2	1346	7.79	158
3	Topology-3	1178	35.14	440
4	Topology-4	1322	7.45	176
5	Topology-5	1266	24	357
6	Topology-6	1327	4.28	84
7	Topology-7	1195	11.43	274
8	Topology-8	1266	10.29	219
9	Topology-9	1283	18.75	180
10	Topology-10	1215	9.84	165
11	Topology-11	1164	15.38	209
12	Topology-12	1115	26.98	457

In Figure 4, the nature of intergroup interconnections i.e. edges in Cut-Set vs. Connectivity Index is shown. It is understood from the graph that the less the no. of edges in the cut-set, more the connectivity index is in number. The topologies which have minimum cut set and maximum connectivity index offer better prediction accuracy compared to other topologies. From figure 5 we can see that Topology-12, 3, 5, 9 give better accuracies compared to other topologies.

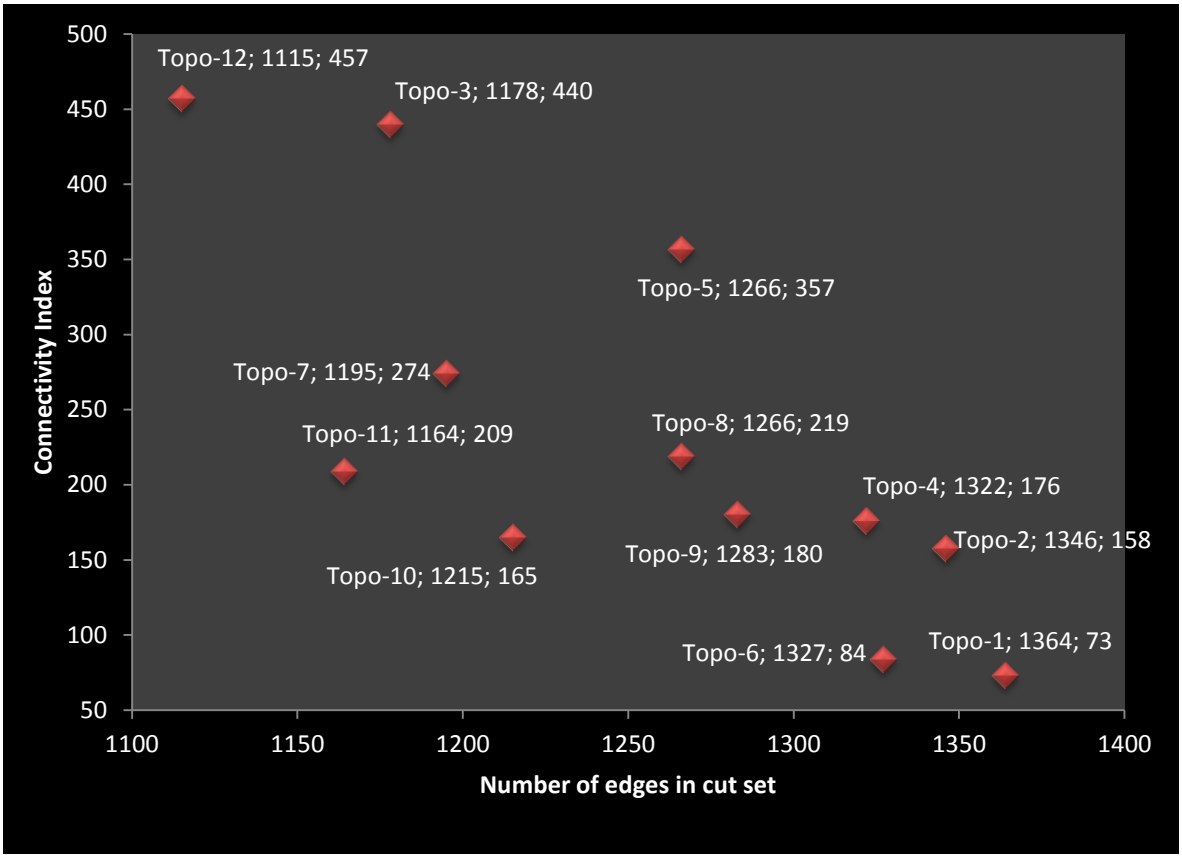

Figure 4. *Intergroup interactions Vs.Connectivity Index*

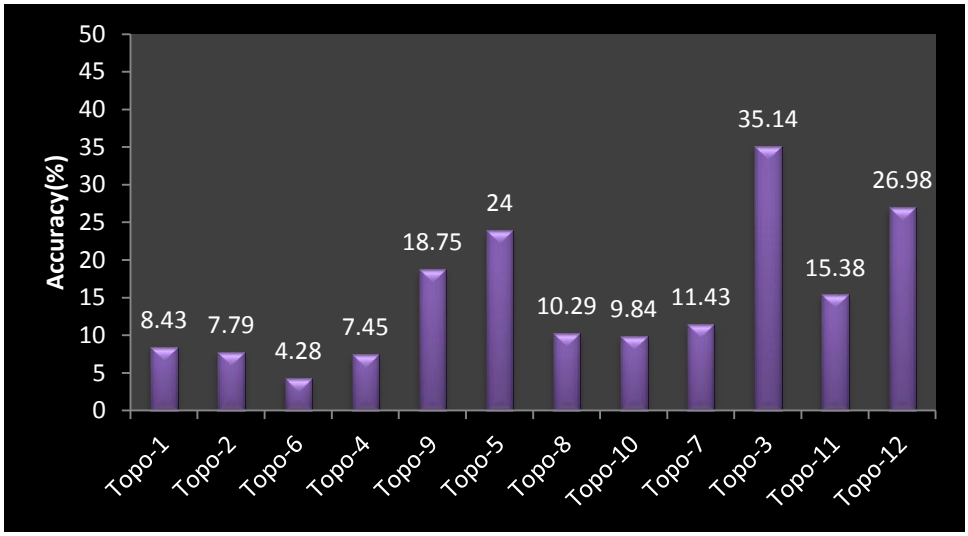

Figure 5. *Performance of PFP_Min on varying different topology*

IV. Conclusion

Candidate unknown protein selection and assigning it to appropriate functional group to minimize the cut-set and assigning its level-1 neighbors (unknown) to the same group thereby increasing connectivity of that group (heuristic applied) is the basis of this work. Application of two-pronged strategy (minimizing cut-set and maximizing connectivity-index) is novelty of this work. It is not possible for any topology to give better prediction with having minimal cut-set and maximal connectivity indices. So, a strategy is taken for selecting one configuration which balances these two parameters optimally. The present approach is designed to predict single function for each candidate unknown protein by considering the predictions in the case of the lowest Cut-Set as well as highest Connectivity Index. The proposed approach is a general approach and can be applied to any organism. In this paper, we show some evaluation results by applying the proposed method to yeast *Saccharomyces Cerevisiae* protein-protein interaction network. We have found the overall success rate to be 81.81% for three functional groups and 35.14% for twelve functional groups. This implies that if reasonable amount of interaction information is provided then reliable predictions can be made. The Experimental results also show that the number of cut-set has clear relation with the success rate. In PFP_Min method, we have not considered unknown proteins that are more distant from known proteins (path length is 2 or more has not been considered), we have only considered candidate unknown protein and its level-1 protein. Our future work is to find a better heuristic for determining the minimum cut-set and to increase path length to evaluate the proposed approach for other organisms and other types of functional classification.

Acknowledgement

The authors express their deep sense of gratitude to the Department of Computer Science and Engineering, Netaji Subhash Engineering College, West Bengal University of Technology.

References

Journal Papers:

[1]. Berthe Katrine, Fiil, Jin-Long Qiu1, Klaus Petersen, Morten Petersen and John Mundy. Coimmunoprecipitation (co-IP) of Nuclear Proteins and Chromatin Immunoprecipitation (ChIP) from *Arabidopsis*.

[2]. Kerppola TK. 2008. Bimolecular fluorescence complementation: visualization of molecular interactions in living cells.

[3]. Pellegrini, M. et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. PNAS, 96:4285–4288, 1999.

[4]. Marcotte, E.M. et al. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science, 285:751–753, 1999.

[5]. Marcotte, E.M. et al. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Nature, 402:83–86, 1999.

[6]. Dandekar, T. et al. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci., 23:324–328, 1998.

[7]. Jansen, R. et al. Relating Whole-Genome Expression Data with Protein-Protein Interactions. Genome Research, 12:37–46, 2002.

[8]. Hartuv, E., Shamir, R. A Clustering Algorithm based Graph Connectivity. Information Processing Letters, 76:175–181, 2000.

[9]. King, A. D., Przulj, N., Jurisica, I. Protein complex prediction via cost-based clustering. Bioinformatics, 20:3013–3020, 2004.

[10]. Prediction of Protein Functions Based on Protein-Protein Interaction Networks: A Min-Cut Approach: Md. Altaf-Ul-Amin, Toshihiro Koma, Ken Kurokawa, Shigehiko Kanaya

Book:

[11]. Narsingh Deo, Graph Theory with Applications to Engineering and Computer Science, Prentice-hall of India Private Limited.