Available Online through
www.ijpbs.com (or) www.ijpbsonline.com

IJPBS |Volume 4| Issue 2|APR-JUN|2014|55-65

Research Article
Biological Sciences

PROTEIN FUNCTION PREDICTION FROM PROTEIN INTERACTION NETWORK USING PHYSICO-CHEMICAL PROPERTIES OF AMINO ACIDS

Sovan Saha[1] & Piyali Chatterjee[2]

[1]Department of Computer Science and Engineering,
Dr.Sudhir Chandra Sur Degree Engineering College,Dumdum, Kolkata -700 074, India
[2]Department of Computer Science and Engineering,
Netaji Subhash Engineering College, Garia, Kolkata-700152, India,
*Corresponding Author Email: sovansaha12@gmail.com

ABSTRACT

When function of protein cannot be experimentally determined, it can often be inferred from sequence similarity. Analysis of the protein structure can provide functional clues or confirm tentative functional assignments inferred from the sequence. Many structure based approaches exist (e.g. fold similarity, three-dimensional templates), but as no single method can be expected to be successful in all cases, a more prudent approach involves combining multiple methods. In this work, we present a new approach to predict protein function that combines sequential, structural information into protein-protein interaction network. Our PPI network, derivable from protein sequence and structure only, is competitive with other function prediction methods that require additional protein information, such as the size of surface pockets. If we include this extra information about structure and sequence into our protein network, our method yields significantly higher accuracy levels than the others.

KEY WORDS

Protein interaction network, Protein function prediction, Functional groups, match Rate, Neighbourhood ratio, Physico-Chemical Properties (PCP), Amino acid sequence, Physico-Chemical Properties Score (PCPScore).

INTRODUCTION

Understanding the molecular mechanisms of life requires the decoding of the functions of proteins in an organism. Tens of thousands of proteins have been sequenced over recent years, and the structures of thousands of proteins have been resolved so far [1]. Still, the experimental determination of the function of a protein with known sequence and structure remains a difficult, time- and cost-intensive task. Computational approaches to correct protein function prediction would allow us to determine the function of whole proteomes faster and more cheaply. Simulating the molecular and atomic mechanisms that define the function of a protein is beyond the current knowledge of biochemistry and the capacity of available computational power. Similarity search among proteins with known function is consequently the basis of current function prediction [2]. A newly discovered protein is predicted to exert the same function as the most similar proteins in a database of known proteins. This similarity among proteins can be defined in a multitude of ways: two proteins can be regarded to be similar, if their sequences align well [3], if their structures match well [4], if both have common surface clefts or bindings sites [5], similar chemical features or common interaction partners [6], if both contain certain motifs of amino acids (AAs) [7] or if both appear in the same range of species [8]. A collection of protein function prediction systems that measure protein similarity by one of the conditions above has been developed. Each of these conditions is based on a

biological hypothesis; e.g. structural similarity implies that two proteins could share a common ancestor and that they both could perform the same function as this common ancestor [9]. These assumptions are not universally valid. Hegyi and Gerstein [10], showed that proteins with similar function may have dissimilar structures and proteins with similar structures may exert distinct functions. Furthermore, a single mutation can alter the function of a protein and make a pair of structurally closely related proteins functionally different [11]. Exceptions are also numerous if similarity is measured by means other than structure [2]. Due to these exceptions, none of the existing function prediction systems can guarantee generally good accuracy. The remedy is to integrate different protein data sources, i.e. to combine several similarity measures, based on several different data types. If two proteins are similar on more than one scale, then the prediction of their function will be more reliable. In this article, we present how to reach this data integration via two routes: We design a graph model for proteins that can represent several types of information and we define and employ sequence derived features for combining several sources of protein data, namely PPI network and amino acid sequences. Jensen et al. [12, 13] proposed the human protein function from post-translational modifications and localization features. The prediction method involved the use of sequence derived features for human protein function prediction. The Post translational modifications (PTMs) are the changes that occur to the protein after its production by the process of translation. They extracted the sequence derived features from the different servers like Expasy, PSORT. Al-Shahib et al. [14] calculated the frequency, total number of each amino acid and the set of amino acids for the input protein sequence. To encode distributional features, they also determined the number and size of continuous stretches of each amino acid or amino acid set. They subdivided every protein into four equally sized fragments and calculated the same feature values for each fragment and combination of fragments. In addition, the other features like the secondary structure were predicted using Prof [15], the position of putative transmembrane helices using TMHMM [16] and of disordered regions using DisEMBL [17]. The features were used for protein function prediction. Kanakubo et al. [18] stated that association rule mining was one of the most important issues in data mining. With apriori methods, the problem becomes incomputable when the total numbers of items are large. On the other hand, bottom-up approaches such as artificial life approaches were opposite of the top-down approaches of searches covering all transactions and may provide new methods of breaking away from the completeness of searches in conventional algorithms. Here, an artificial life data mining technique was proposed in which one transaction was considered as one individual and association rules were accumulated by the interaction of randomly selected individuals. The proposed algorithm was compared to other methods in application to a large scale actual dataset and it was verified that its performance was greatly superior to that of the method using transaction data virtually divided and that of apriori method by sampling approach, thus demonstrating its usefulness. Gupta et al. [19] proposed a novel feature vector based on physicochemical property of amino acids for prediction protein structural classes. They presented a wavelet-based time-series technique for extracting features from mapped amino acid sequence and a fixed length feature vector for classification is constructed. Wavelet transform is a technique that decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. The proposed feature vector contains information about the variability of ten physiochemical properties of protein sequences over different scales. The variability of physiochemical properties was represented in terms of wavelet variance. Jaiswal et al. [20] studied that the identification of specific target proteins for any diseased condition involves extensive characterization of the potentially involved proteins. Members of a protein family demonstrating comparable features may show certain unusual features when implicated in a pathological condition. They studied the Human matrix metalloproteinase (MMP) family of endopeptidases and discovered their role in various pathological conditions such as arthritis, atherosclerosis, cancer, liver fibrosis, cardio-vascular and neurodegenerative disorders, little is known about the specific involvement of members of the large MMP family in diseases. They hypothesized

that cysteine rich and highly thermo stable MMPs might be key players in diseased conditions and hence signify the importance of sequence derived features.

I. PRESENT WORK

- **Motivation:** Different works have been explored in different fields of function prediction which has been discussed in the earlier section. All these enlighten the fact that there is a scope for improvement and to apply domain specific knowledge.Motivated by that fact, we have computed neighborhood ratio of uncharacterized protein from a protein interaction network and assigned its function [22].Now after analysis it is observed that there is a possibility of highlighting a new way of predicting protein function by incorporating amino acid sequences with PPI network using physico-chemical properties A two pronged strategy i.e. utilizing physicochemical properties of amino acids in one way and neighborhood pattern of unannotated protein in its interaction network, on the other hand, can strengthen the prediction ability of this classifier.In this work, we have proposed two above mentioned approaches associated with various intelligent techniques.

- **Dataset:** We have used MIPS database here. The Munich Information Center for Protein Sequences (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 (ftp://ftpmips.gsf.de/yeast/PPI/) is incorporated with protein-protein interaction data of yeast (Saccharomyces Cerevisiae), is collected which contains 15613 genetic and physical interactions. Self-interactions are discarded. A set of 12487 unique binary interactions involving 4648 proteins are taken as data. In our proposed method 8 functional groups are considered. They are cycle control (O_1), cell polarity (O_2), cell wall organization and biogenesis (O_3), chromatin chromosome structure (O_4), Coimmunoprecipitation (O_5), copurification (O_6), DNA Repair(O_7), lipid metabolism (O_8), nuclear-cytoplasmic transport (O_9), pol II transcription (O_{10}), protein folding (O_{11}), protein modification (O_{12}), protein synthesis(O_{13}), small molecule transport (O_{14}) and vesicular transport (O_{15}). For each functional group, 90% protein pairs are taken as training samples and rest 20% among them are considered as test samples. Table-I shows the number of proteins involved in particular functionality.

Table-I: Proteins of 15 functional groups

Functional groups	Annotated proteins	Proteins considered as unannotated
cell cycle control	78	15
cell polarity	90	19
cell wall organization and biogenesis	85	17
chromatin chromosome structure	122	24
nuclear-cytoplasmic transport	18	4
pol II transcription	85	17
protein folding	29	6
protein modification	81	16
Coimmunoprecipitation	209	41
copurification	61	12
DNA Repair	79	15
lipid metabolism	84	17
protein synthesis	86	17
small molecule transport	53	11
vesicular transport	117	23

- **Basic terminologies:**

Protein interaction network: Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. Many of the most important molecular processes in the cell such as DNA replication are carried out by large molecular machines that are built from a large number of protein components organized by their protein–protein interactions. These protein interactions form a network like structure which is known as Protein interaction network. Here protein interaction network is represented as a graph G_P which consists of a set of vertex (nodes) V connected by edges (links) E. Thus G_P = (V, E).Here each protein is represented as a node and their interconnections are represented by edges.

Sub graph: A graph G'_P is a sub graph of a graph 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 as sub graph of G_P if $V' \subseteq V$ and $E' \subseteq E$. G'_P may be defined as a set of $\{K \cup U\}$ where K represents the set of un-annotated proteins while U represents the set of annotated protein.

Level-1 neighbors: In G'_P, the directly connected neighbors of a particular vertex are called level-1 neighbors.

Level-2 neighbors: In G'_P, level-2 neighbors are those who are directly connected neighbors of level-1 neighbors of that particular vertex.

Neighborhood ratio ($P^{l(=1,2)}_{O_{i(=1..15)}}$): The neighborhood ratio $P^{l}_{O_i}$ is defined as the ratio of no. of level-1(or level-2) neighbours (K) corresponding to a functional group O_i and total no. of level-1(or level-2) neighbors (P). Here, O_i represents any element of 15 functional groups and l denotes level-1 and level-2. It may be defined as.

$$P^{l(=1,2)}_{O_{i(=1..15)}}) = \frac{K}{P}$$

Amino Acids: Amino acids play central roles as building blocks of proteins.20 amino acids within proteins convey a vast array of chemical versatility. The precise amino acid content, and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein. The chemical properties of the amino acids of proteins determine the biological activity of the protein. Proteins not only catalyze all (or most) of the reactions in living cells, they control virtually all cellular process. In addition, proteins contain within their amino acid sequences the necessary information to determine how that protein will fold into a three dimensional structure, and the stability of the resulting structure. The field of protein folding and stability has been a critically important area of research for years, and remains today one of the great unsolved mysteries. It is, however, being actively investigated, and progress is being made every day.

Physico-Chemical Properties (PCP): Physico-Chemical Properties are the various features of amino acid which are used to predict protein class. These properties are very important in protein class prediction. Here we have considered some vital physico-Chemical properties [23] which are given below:

- **Extinction Coefficient ($E_{protein}$):** Extinction Coefficient is a protein parameter that is commonly used in the laboratory for determining the protein concentration in a solution by spectrophotometry. It describes to what extent light is absorbed by the protein and depends upon the protein size and composition as well as the wavelength of the light. For proteins measured in water at wavelength of 280nm, the value of the Extinction coefficient can be determined from the composition of Tyrosine, Tryptophan and Cystine.
  Mathematically:
  $E_{protein} = N_{tyr} * E_{tyr} + N_{trp} * E_{trp} + N_{cys} * E_{cys}$
  Where E_{tyr}=1490, E_{trp} =5500, E_{cys} =125 are the Extinction coefficients of the individual amino acid residues.
- **Absorbance (Optical Density):** For proteins measured in water at wavelength of 280nm the absorbance can be determined by the ratio of Extinction coefficient and the molecular weight of the protein. It is a representation of a material's light blocking ability.
  Mathematically:
  Absorbance = Eprotein / Molecular Weight
- **Number of Negatively Charged Residues (Nneg):** This can be calculated from the composition of Aspartic acid and Glutamic acid.
- **Number of Positively Charged Residues (Npos):** This can be calculated from the composition of Arginine and Lysine.

- **Aliphatic Index (AI):** The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins.
  Mathematically:
  $AI = X_{ala} + a * X_{val} + b * (X_{ile} + X_{leu})$
  Where X_{ala} , X_{val} , X_{ile} and X_{leu} are the mole percentages of alanine, valine, isoleucine and leucine respectively. Coefficients a and b are the relative volume of valine side chain and side chains to the side chain of alanine i.e. a = 2.9 and b = 3.9.
- **IP/mol weight:** It calculates the isoelectric point by molecular weight of the input amino acid sequence. IP stands for isoelectric point of the input amino acid sequence. Mol weight stands for molecular weight of the input amino acid sequence.
- **Hydrophobicity (Hphb) and hydrophilic (Hph):** The hydrophobic effect represents the tendency of water to exclude non-polar molecules. **Hydrophobicity scales** are values that define relative hydrophobicity of amino acid residues. The more positive the value, the more hydrophobic are the amino acids located in that region of the protein. These scales are commonly used to predict the transmembrane alpha-helices of membrane proteins. When consecutively measuring amino acids of a protein, changes in value indicate attraction of specific protein regions towards the hydrophobic region inside lipid bilayer. While hydrophilic property represents a molecule or portion of a molecule that has a tendency to interact with or be dissolved by water and other polar substances.
- **Physico-Chemical Properties Score (PCP Score):** PCPScore is defined as scaling of the mean value obtained from physico-chemical properties mention above.
  - **Proposed Network:** We have proposed the method to predict protein function from the protein interaction network in our earlier work [22] which is indeed very simpler where the selection of unannotated protein is done randomly i.e.10% of proteins are taken from functional group and prediction technique is based on the value of neighborhood ratio. Neighborhood ratio of unknown protein is computed as stated above. This method attempts to find the maximum neighborhood ratio and assign this protein to its corresponding functional group.

A. Method I

It has been observed that a new way can be established if we introduce physico-chemical properties after constructing ppi network from the above mentioned process and use it to predict protein function by using a scoring function (**PCP** Score). Steps associated with this method are described below: Given G'_P, a sub graph of protein interaction network, consisting of proteins as nodes associated with any element of set O= {O_1, O_2, O_3,….,O_{15}} where O_i represents a particular functional group, this method maps the elements of the set of un-annotated proteins U to any element of set O. Steps associated with this method is described as follows:

- **Step 1**: Take any protein as an element from set U.
- **Step 2**: Find Level-1 and level-2 neighbors of that protein in G'_P associated with set O. each level-1 and level-2 protein's respective amino acid sequence are considered.
- **Step3:** Compute $E_{protein}$, Absorbance (Optical Density), Nneg , Npos, AI,IP/mol weight, Hphb and Hph.
- **Step 4:** Compute mean (PCPScore) of the above features and scale the values within 0 and 1.
- **Step 5**: Compute $(\mathbf{PCPScore})^{\mathbf{l}(=1,2)}_{\mathbf{O_{i(=1,..,15)}}}$.

  Step 6: For each functional group O_i, find Level-1 and level-2 protein with Maximum PCPScore and also find the neighbor protein with Maximum PCPScore between these two levels and among all functional groups. Assign functional group of that neighbor protein with highest PCPScore to Unknown protein, i.e.

$$\mathbf{PCPScore^{l}_{O_k}} = \mathrm{Max}(\left(\max(\mathbf{PCPScore^{1}_{O_{i(=1,..,15)}}}\right), \left(\max(\mathbf{PCPScore^{2}_{O_{i(=1,..,15)}}}\right))$$

Assign un-annotated protein from the set U to functional group O_k.

Illustration of Method-I with an example:

An unannotated protein YER014c-a is taken from our dataset U. Here it has been explained how this protein is assigned to a particular functional group among 15 functional groups.

Level-1 proteins of YER014c-a are YKL190w and YHR181w.Level-2 proteins of YER014c-a are YAL026c, YGL070c, YLR399c and YKL194c.

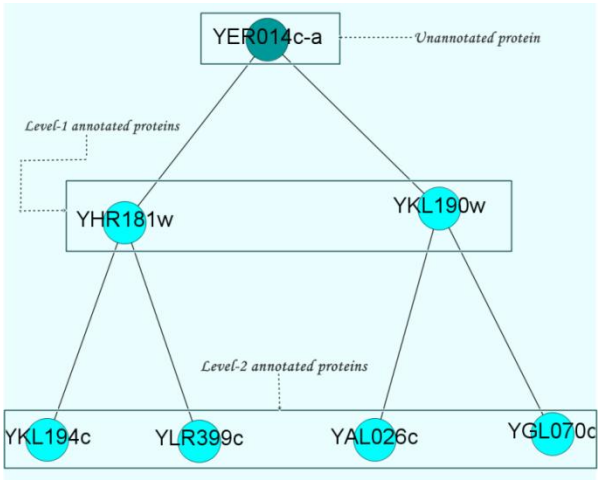


Fig.1 Prediction of functional group of Protein YER014c using the above method Consider the amino acid sequence of YHR181w:

MLLELISYAGTVSGFLFLTLSIASGLYYISELVEEHTEPTRRFLTRA
IYGIILILILLLLLDGFPFKLTLFSIACYIVYYQNLKSFPFISLTSPTFLL
SCVCVVLNHYFWFKYFNDTEVPPQFKFDPNYIPRRRASFAEVA
SFFGICVWFIPFALFVSLSAGDYVLPTTSEQHMAKKNDDITTN
NQPKFRKRAVGLARVVINSVRKYIYSLARVFGYEIEPDFDRLAV

Estimate physico-chemical properties values of YHR181w:

[1] $E_{protein}$: 26400
[2] Absorbance/Optical Density : 1.353
[3] Nneg : 13
[4] Npos : 18
[5] AI : 88.15
[6] IP/mol weight: 0.047
[7] Hphb :10.05

PCPScore of **YHR181w** obtained after scaling the mean value obtained from the above data between the ranges of 0-1: **0.884**.Similarly PCPScore of **YKL190w** obtained after scaling: **0.861**.After level-1 PCPScore calculation, PCPScore of level-2 is estimated shown below:

Table-II: PCPScore of proteins

Protein Name	PCPScore
YAL026c	0.90
YGL070c	0.89
YLR399c	0.88
YKL194c	0.89

Then steps of the Method-I is applied as discussed in the last algorithm and thus YER014c-a is assigned to one out of 15 functional groups.

B. Method-II

Further assessment makes us realize the fact that the result obtained in the above method can be further enhanced if we combine PCP Score and neighborhood ratio (P) in the process of predicting protein function using PPI network. So we modify the above algorithm which has been described below:

- **Step 1**: Take any protein as an element from set U.
- **Step 2**: Find Level-1 and level-2 neighbors of that protein in G'_Passociated with set O. each level-1 and level-2 protein's respective amino acid sequence are considered.
- **Step3:** Compute Eprotein, Absorbance (Optical Density), Nneg , Npos, AI,IP/mol weight, Hphb and Hph.
- **Step 4:** Compute mean (PCPScore) of the above features and scale the values within 0 and 1.
- **Step 5**: Compute $(\mathbf{PCPScore})^{\mathbf{l}(=1,2)}_{\mathbf{O_{i(=1,..,15)}}}$.
- **Step 6:** For each functional group O_i, find Level-1 and level-2 protein with Maximum (PCPScore+P) and also find the neighbor protein with Maximum (PCPScore+P) between these two levels and among all functional groups. Assign functional group of that neighbor protein with highest (PCPScore+P) to Unknown protein, i.e.

$$(\mathbf{PCPScore}+\mathbf{P})^{\mathbf{l}}_{\mathbf{O_k}} = \mathrm{Max}\Big(\max((\mathbf{PCPScore}+\mathbf{P})^{\mathbf{1}}_{\mathbf{O_{i(=1,..,15)}}}\Big), \Big(\max((\mathbf{PCPScore}+\mathbf{P})^{\mathbf{2}}_{\mathbf{O_{i(=1,..,15)}}}\Big)\Big)$$

Assign un-annotated protein from the set U to functional group O_k.

Hence, it can be concluded that method-II is basically a combination of method-I and neighborhood ratio method [22] where more emphasis is given on the method of unannotated protein function prediction procedure which in turn comparatively increases the results which will be discussed in the upcoming sections.

II. RESULTS AND DISCUSSION

We have taken 12487 interactions where 4648 proteins are involved. Eight functional groups cell cycle control (O_1), cell polarity (O_2), cell wall organization and biogenesis (O_3), chromatin chromosome structure (O_4), Coimmunoprecipitation (O_5), copurification (O_6), DNA Repair(O_7), lipid metabolism (O_8), nuclear-cytoplasmic transport (O_9), pol II transcription (O_{10}), protein folding (O_{11}), protein modification (O_{12}), protein synthesis(O_{13}), small molecule transport (O_{14}) and vesicular transport (O_{15}) are chosen from this network. In both the three methods 20% of proteins from each functional group are taken as unannotated proteins. A technique is followed in each of these methods in observing neighborhood pattern and associated functional group. We have used different performance measures to evaluate the prediction accuracy of our 2 methods. Accuracy1 is measured by dividing the number of predicted proteins with a functional group by the total number of observed proteins with the same functional group. Thus, we have 15 different accuracies for 15 functional groups.

$$\text{Accuracy}_1 = \frac{\sum_{i=1}^{15} \text{Number of predicted proteins having functional group Oi}}{\text{Total number of observed proteins having functional group Oi}}$$

In Fig. 2, we have depicted the performance analysis of the two methods for each functional group and finally accuracies of Method-I, Method-II are also given in Table-III.Method-II computes neighborhood ratio of unannotated protein for each functional group where level-1 and level-2 neighbors are considered. Then, it assigns the unannotated protein to a particular functional group having the maximum PCPScore considering level-1 and level-2 neighbors together. From the Table-II, using Method- II we have achieved good prediction accuracy for all functional groups except protein folding and cell polarity.

Table-III: Accuracy1 achieved in two methods

Methods	Total number of unannotated proteins	Total number of correctly predicted proteins	Overall Accuracy (%)
Method-I	194	167	86
Method-II	194	170	88

Evaluation metrics:

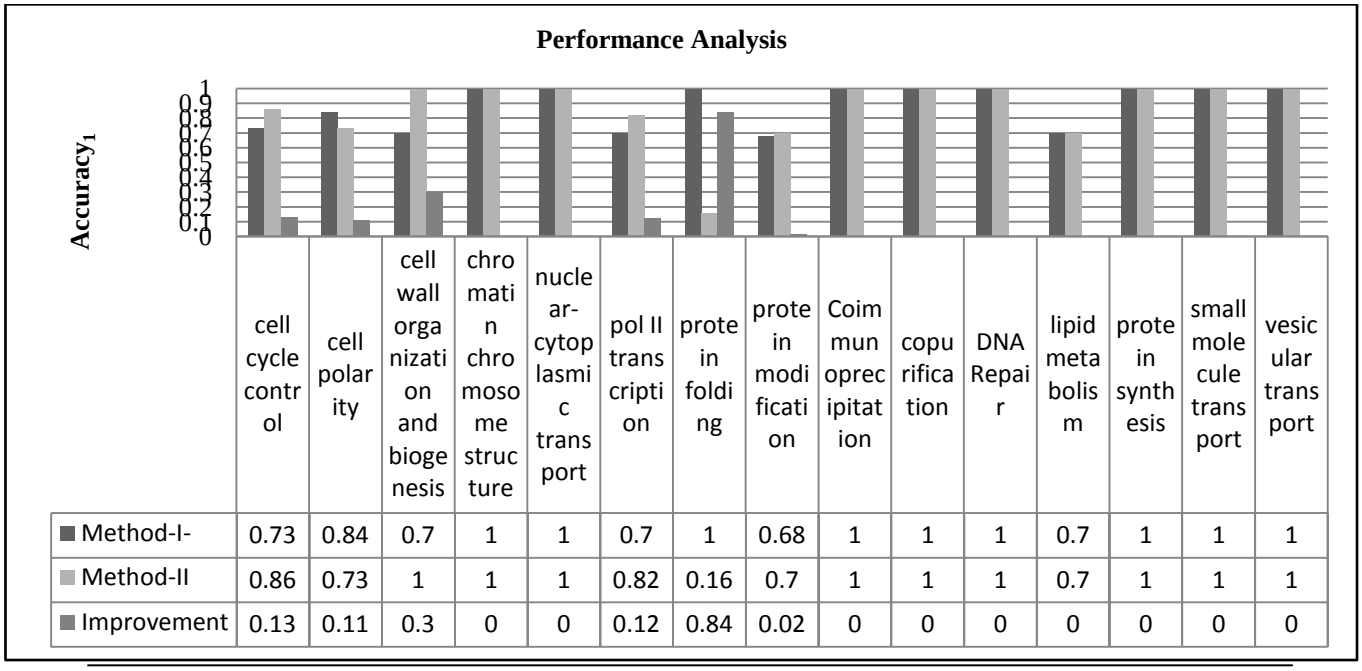
The training results are yet evaluated in another way using standard measures, such as the Accuracy2 (A), Recall, and Precision values, which are explained below:

$$\text{Accuracy}_2 \text{ (A)} = (1 - \text{Error}) = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{True positive rate / Recall / Sensitivity} = \frac{TP}{TP + FN}$$

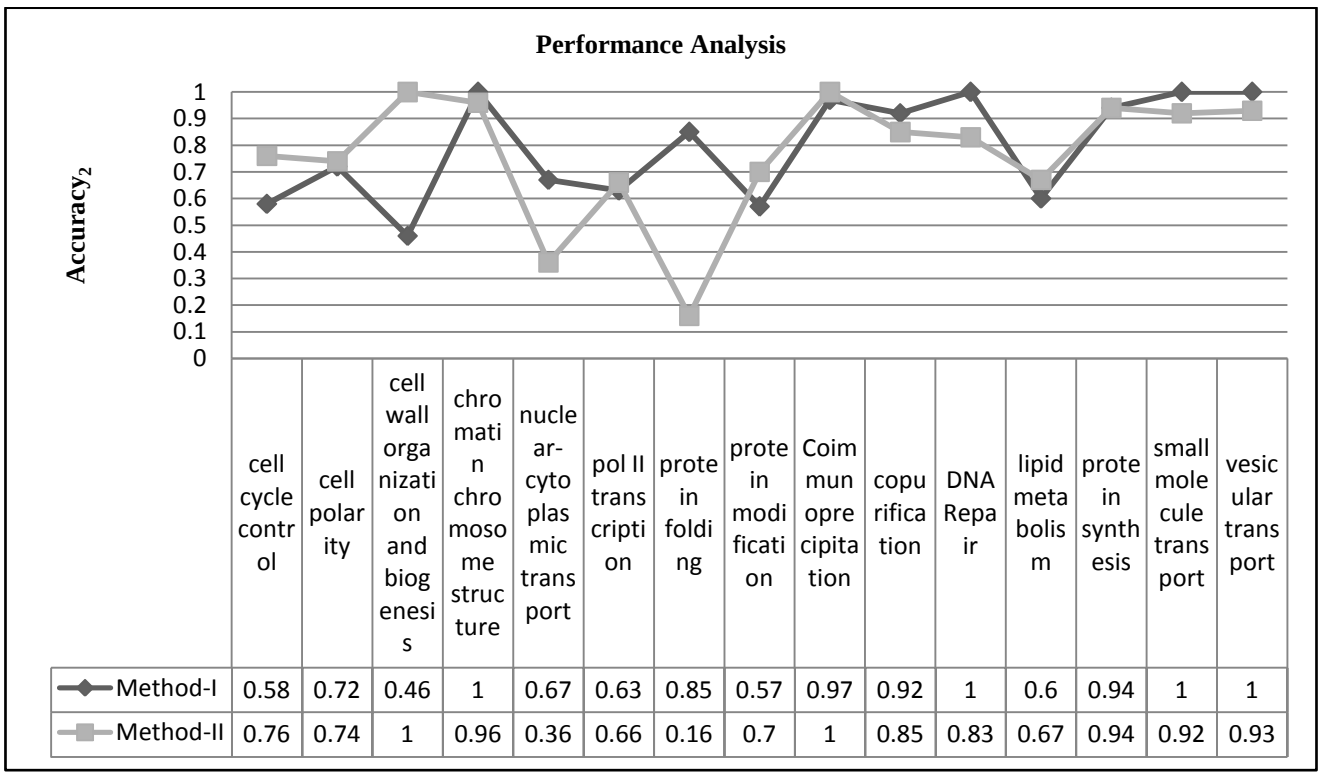
$$\text{Precision / Specificity} = \frac{TP}{TP + FP}$$

Where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives. The recall (R) corresponds to the percentage of correct positive predictions and the precision (P) measures the percentage of observed positives that are correctly predicted. The true positive rate (TPR) is described as either the recall or sensitivity measure, and the false positive rate (FPR) estimates the false alarm rate or fall-out values. The performance of the three methods for each type of functional groups is described by the recall R and the precision P.



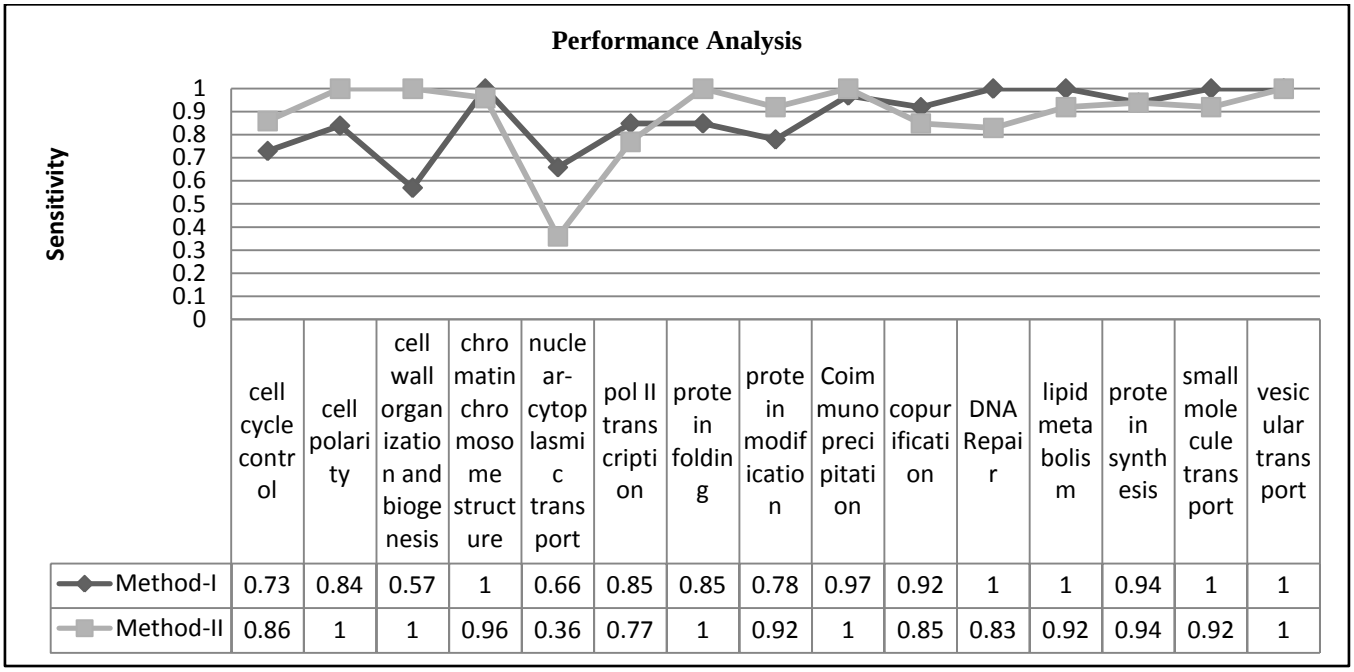
	cell cycle control	cell polarity	cell wall organization and biogenesis	chromatin chromosome structure	nuclear-cytoplasmic transport	pol II transcription	protein folding	protein modification	Coimmunoprecipitation	copurification	DNA Repair	lipid metabolism	protein synthesis	small molecule transport	vesicular transport
Method-I-	0.73	0.84	0.7	1	1	0.7	1	0.68	1	1	1	0.7	1	1	1
Method-II	0.86	0.73	1	1	1	0.82	0.16	0.7	1	1	1	0.7	1	1	1
Improvement	0.13	0.11	0.3	0	0	0.12	0.84	0.02	0	0	0	0	0	0	0

Fig. 2 Performance Analysis of the two methods over 15 functional groups



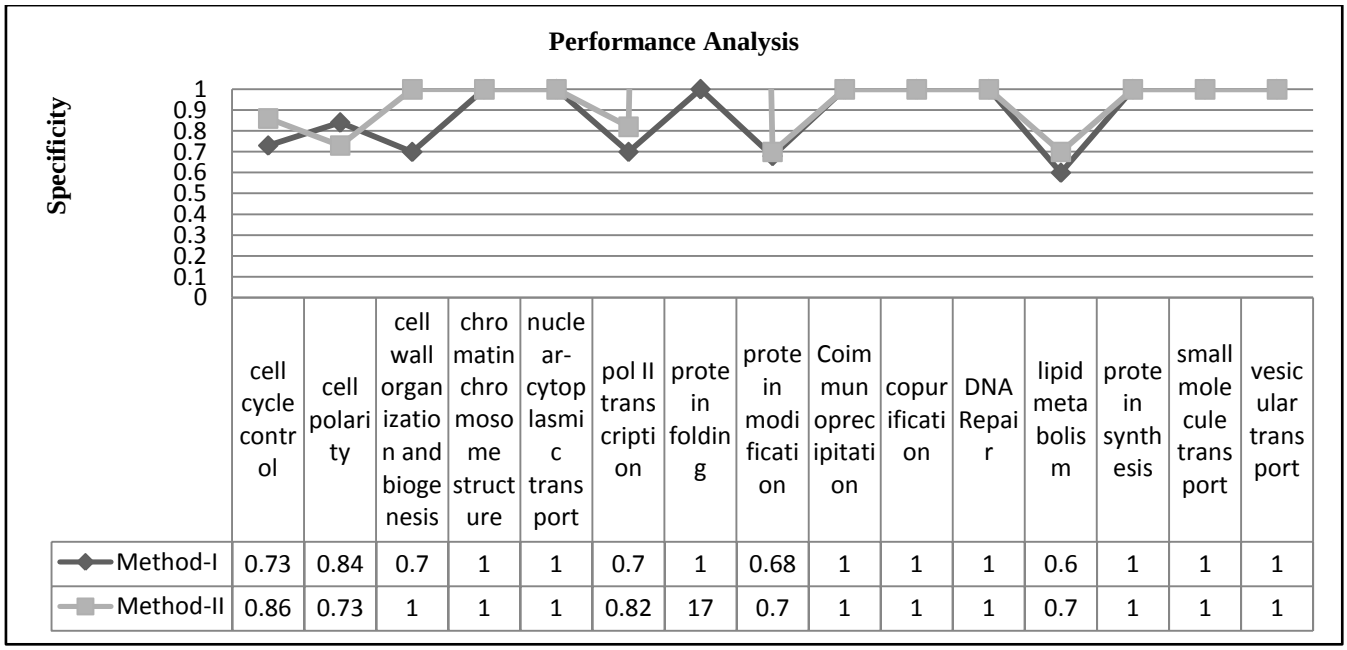
	cell cycle control	cell polarity	cell wall organization and biogenesis	chromatin chromosome structure	nuclear-cytoplasmic transport	pol II transcription	protein folding	protein modification	Coimmunoprecipitation	copurification	DNA Repair	lipid metabolism	protein synthesis	small molecule transport	vesicular transport
Method-I	0.58	0.72	0.46	1	0.67	0.63	0.85	0.57	0.97	0.92	1	0.6	0.94	1	1
Method-II	0.76	0.74	1	0.96	0.36	0.66	0.16	0.7	1	0.85	0.83	0.67	0.94	0.92	0.93

Fig. 3 Performance Analysis of Accuracy$_2$ (A) for two methods



	cell cycle control	cell polarity	cell wall organization and biogenesis	chromatin chromosome structure	nuclear-cytoplasmic transport	pol II transcription	protein folding	protein modification	Coimmunoprecipitation	copurification	DNA Repair	lipid metabolism	protein synthesis	small molecule transport	vesicular transport
Method-I	0.73	0.84	0.57	1	0.66	0.85	0.85	0.78	0.97	0.92	1	1	0.94	1	1
Method-II	0.86	1	1	0.96	0.36	0.77	1	0.92	1	0.85	0.83	0.92	0.94	0.92	1

Fig. 4 Performance Analysis of Sensitivity for two methods



	cell cycle control	cell polarity	cell wall organization and biogenesis	chromatin chromosome structure	nuclear-cytoplasmic transport	pol II transcription	protein folding	protein modification	Coimmunoprecipitation	copurification	DNA Repair	lipid metabolism	protein synthesis	small molecule transport	vesicular transport
Method-I	0.73	0.84	0.7	1	1	0.7	1	0.68	1	1	1	0.6	1	1	1
Method-II	0.86	0.73	1	1	1	0.82	17	0.7	1	1	1	0.7	1	1	1

Fig. 5 Performance Analysis of Specificity for two methods

In our previous work [22], we compute neighborhood ratio of unannotated protein for each functional group where level-1 and level-2 neighbors are considered. Then, it assigns the unannotated protein to a particular functional group having the maximum neighborhood ratio considering level-1 and level-2 neighbors together.

Then like method [22], method-I does not consider neighborhood ratio for predicting function from protein interaction network. It rather incorporates physico-chemical properties into PPI and use PCP Score for prediction. In one way, it is attractive than the previous one in the sense that it does not confine prediction procedure to PPI network only but it extends its limit to the amino acid sequence of each protein.

Method-II is a unique approach which is formed by combining both neighborhood method and method-I i.e. here the protein function is predicted by using both neighborhood ratio and PCP Score which leads to the enhancement of overall success rate as shown in Table-III.

REFERENCES

[1] Schwikowski, B., Uetz, P. and Fields, S.A network of protein-protein interactions in yeast. Nature Biotech.18, 1257-1261 (2000).

[2] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Tagaki, T. Assessment of prediction accuracy of protein functions from protein- protein interaction data. Yeast 18, 523-531 (2001).

[3] J. Chen, W. Hsu, M. L. Lee, and S. K. Ng. Labeling network motifs in protein interactomes for protein function prediction. Proc 23rd International Conference on Data Engineering (ICDE). 546- 555, 2007.

[4] Vazquez, "Global Protein Function Prediction from Protein-Protein Interaction Networks," Nature Biotechnology, vol. 21, pp. 0697- 700, June, 2003.

[5] U. Karaoz, T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor, and S. Kasif. Whole-genome annotation by using evidence Integration in functional-linkage.

[6] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, M. Singh. Whole Proteome prediction of protein functions via graph-theoretic analysis of interaction maps. Bioinformatics 21 (Suppl 1): i302– i310, 2005.

[7] M. Deng, Inferring domain-domain interactions from protein protein interactions. Genome Res. 12(10):1540-8, 2002.

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

[9] D. D. Wu, X. Hu, An efficient approach to detect a protein community from a seed. 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB2005).La Jolla CA, USA: IEEE pp. 135–141, 2005.

[10] Vazquez, "Global Protein Function Prediction from Protein-Protein Interaction Networks," Nature Biotechnology, vol. 21, pp. 697- 700, June 2003.

[11] M. P. Samanta,S. Liang, Predicting protein functions from redundancies in large scale protein interaction networks. ProcNatlAcadSci USA 100: 12579–12583, 2003.

[12] L. Jensen. "Prediction of Protein Function from Sequence Derived Protein Features",Ph.D. thesis, Technical University of Denmark, 2002

[13] L. Jensen, M. Skovgaard and S. Brunak. "Prediction of Novel Archaeal Enzymes from Sequence Derived Features", Protein Science, 11: 2894-2898, 2002

[14] Al-Shahib, R. Breitling, and D. R. Gilbert "Predicting protein function by machine learning on amino acid sequences – a critical evaluation" BMC Genomics, 8:1-10, 2007.

[15] M. Ouali, R.D. King "Cascaded multiple classifiers for secondary structure prediction" ProtSci., 9:1162–1176, 2000.

[16] http://www.cbs.dtu.dk/services/TMHMM/

[17] R. Linding, L. J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell "Protein disorder prediction: implications for structural proteomics" Structure, 11:1453-1459, 2003.

[18] M. Kanakubo and M. Hagiwara. "Speed up technique for Associative rule mining based on an Artificial Algorithm", GRC book on granular computing, 38(12):318-323, 2007.

[19] R. Gupta, A. Mittal, and K. Singh. "Time series based feature extraction approach for prediction of protein structural class", EURASIP Journal, 8(1): 1-7, 2008.

[20] Jaiswal, A. Chhabra, U. Malhotra, S. Kohli, V. Rani "Comparative analysis of human matrix metalloproteinases: Emerging therapeutic targets in diseases" Bioinformation 6(1): 23-30, 2011.

[21] http://www.cytoscape.org/documentation_users.html.

[22] Sovan Saha, Piyali Chatterjee, Subhadip basu, Mahantapas kundu,Mita Nasipuri,Improving

prediction of Protein Function from protein Interaction Network using Intelligent Neighborhood Approach, 978-1-4673-4698-6_©2012 IEEE.

[23] Predicting Protein Function using Decision Tree, Manpreet Singh, Parminder Kaur Wadhwa, and Surinder Kaur, World Academy of Science, Engineering and Technology 15 2008.

***Corresponding Author:**
Sovan Saha
Department of Computer Science and Engineering,
Dr.Sudhir Chandra Sur Degree Engineering College,
Dumdum, Kolkata -700 074, India,