International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.31, July 2024

A Review on Protein Function Prediction Methods using Protein-Protein Interaction Networks

Saima Khan
Faculty, Computer Science and Engineering
University of Development Alternative
Dhaka, Bangladesh

Md. Abidur Rahman Khan
Computer Science and Engineering
University of Asia Pacific
Dhaka, Bangladesh

ABSTRACT

Proteins are essential components of all living organisms, performing a myriad of biological functions within living bodies and systems. Understanding protein functions is crucial for researchers, as it enables the development of various evolutionary medicines, treatments, and other beneficial products. However, many protein functions remain unknown. Computational methods have gained popularity over traditional physical experiments for predicting protein functions. These methods include approaches based on sequence and structure knowledge, gene expression data, and protein-protein interaction data. Notably, protein function prediction methods utilizing protein-protein interaction networks have yielded more satisfactory results compared to those using other attributes. Proteins rarely function in isolation; they typically operate in conjunction with their interacting partners. Numerous researchers have proposed and implemented various techniques for accurately predicting the functions of unknown proteins. This paper presents a comprehensive review of various methods proposed and utilized by researchers for predicting protein functions using protein-protein interaction networks. The descriptions include essential tables and figures, accompanied by appropriate citations and references. The aim of this paper is to assist other researchers in understanding these techniques and to encourage the development of enhanced approaches for predicting protein functions through protein-protein interaction networks.

General Terms

Protein, prediction, support, wet lab, dry lab, accuracy

Keywords

Protein function prediction, protein-protein interaction network, common neighbor, majority, neighbor protein

1. INTRODUCTION

Protein function prediction represents a pivotal area of inquiry within the realm of bioinformatics. Despite significant strides, many protein functions remain elusive to researchers. A comprehensive understanding of protein functions holds promise for addressing a myriad of biological and medical challenges. Traditional wet lab methodologies have historically served as the cornerstone for elucidating protein functions. However, such approaches are characterized by inherent limitations, including time intensiveness, financial burden, and technical complexity. Consequently, computational methods have emerged as an increasingly favored alternative for predicting protein functions. These computational techniques offer expedited and streamlined predictions, entail reduced financial expenditure, and necessitate diminished human intervention compared to conventional wet lab methodologies. Wet lab-based experiments involve direct manipulation and testing of biological or chemical materials in a laboratory environment. These experiments necessitate the use of physical samples and reagents, often employing techniques like mixing, culturing, and measuring under controlled conditions. Dry lab-based experiment refers a form of research or study carried out without the conventional wet lab methods, which usually entail working with chemicals, biological specimens, or other physically manipulable materials. Rather, dry lab experiments focus on computational techniques, simulations, data analysis, and theoretical modeling.

Several computational methods exist for forecasting protein functions, including sequence-based, structure-based, protein-protein interaction network-based, gene expression data-based, and pathway analysis from gene expression data-based methods [1]. However, the limited availability of protein structure data restricts the effective and widespread use of homology-based approaches in protein function prediction [1]. Many databases like SWISSPROT [25], DIP [26], NCBI [27], STRING [28], and PDB [29] are available for using protein function prediction experiments.

Among various computational methods for protein function prediction, leveraging protein-protein interaction networks emerges as a potent strategy for efficiently and swiftly predicting precise protein functions. Since proteins generally function through interactions with other proteins, protein-protein interaction networks provide crucial insights for predicting their functions. PPI networks are particularly valuable for predicting protein functions because they provide a comprehensive view of how proteins interact within a cell. By mapping these interactions, researchers can infer the roles of unknown proteins based on their interaction partners and network positions. Various computational techniques are applied to analyze PPI networks, including network clustering, machine learning, and data integration methods.

2. THE CONCEPT OF PROTEIN FUNCTION PREDICTION USING PROTEIN-PROTEIN INTERACTION NETWORKS

To carry out a specific function, a protein necessitates interaction with another protein. This protein interaction is depicted in the form of a network termed the protein-protein interaction network. Leveraging the understanding of this interaction network, various computational techniques have been proposed for protein function prediction, utilizing one or more interaction networks. These approaches are classified into four categories [1]: neighbor-based techniques, clustering-based approaches, optimization-based techniques, and association analysis-based techniques. The neighbor-based techniques assign a level to an annotated protein by transferring labels within its neighborhood. Clustering-based approaches identify densely connected regions in the interaction network, termed clusters, and assign a label to an annotated protein based on the most dominant label in the corresponding cluster. Optimization-based techniques utilize the entire connectivity structure of the network. Association analysis-based techniques utilize association analysis algorithms to detect frequently occurring sets in the interaction network for protein function prediction. The term ”support” is commonly used in the context of predicting protein function. This term refers to the frequencies of the presence of specific functions within a protein-protein interaction network.

3. PROTEIN FUNCTION PREDICTION USING MAJORITY RULE

Majority rule [2, 5] states that the functions of unknown characteristics can be inferred based on the functions of their immediate neighbors. However, an unknown protein might have several neighbors with varying characteristics. This paper has tried to reduce the diversity of annotations linked to neighboring proteins.

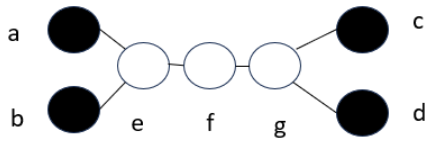

Fig. 1. In this figure, there is a total of 7 proteins (each node represents protein and each edge represents the link between proteins). The black nodes (a, b, c, d) are the proteins with known functions and and rest nodes (e, f, g) are the proteins with unknown functions. According to Majority rule [2, 5] protein e and g can be annotated with the functions of their direct neighbors but protein f cannot be annotate as it does not have immediate neighbors with known functions.

They have suggested categorizing proteins into functional classes based on their physical interaction networks, aiming to minimize the interactions between different functional categories. This approach often leads to multiple functional assignments due to the presence of several equivalent solutions. This method has been applied to analyze the protein-protein interaction network in yeast Saccharomyces cerevisiae. It has also been explored that interacting proteins could belong to at least one common functional class. This insight helps in understanding the relationships among proteins and it provides a basis for categorizing proteins with similar functions. Consequently, it aids in the functional classification of the remaining subset of uncharacterized proteins [2, 5, 6].

In their study [2], 2,709 published interactions have been analyzed that involves 2,039 proteins available from public databases [21, 22] and two large scale studies [23, 24]. Their analysis encompasses only those direct interactions that have been identified through biochemical experiments or two-hybrid studies. It does not take into account protein complexes for which the specific protein contacts remain unidentified. To visualize the interactions, they created a software application utilizing the graph-drawing library ”AGD” (http://www.mpi-sb.mpg.de/AGD). Interestingly, this approach yielded a single extensive network of protein interactions. This network consisted of 2,358 connections linking 1,548 distinct proteins. The subsequent largest network comprised merely 19 proteins. Additionally, there were nine networks comprising between 5 and 11 proteins each. Moreover, the remaining 193 networks comprised 4 or fewer proteins individually. Thus, there were a total of 204 autonomous networks, each interaction being unique to a single network.
Their examination of protein-protein interactions [2] in S. cerevisiae enables to integrate a full one-quarter of the proteins predicted from the genome sequence into a single extensive network. This network unveils overarching patterns of interactions among proteins within functional classes or localization assignments, as well as numerous potential interconnections. Utilizing the interaction data facilitates making functional predictions for uncharacterized proteins. The efficacy of this approach is underscored by an analysis demonstrating that 72% of characterized proteins with known partners could be appropriately assigned a functional category. While acknowledging that the large network does not present a completely accurate portrayal of cellular connections, it remains useful for scrutinizing protein function even when focusing on specific regions of the network.

4. PROTEIN FUNCTION PREDICTION USING NEIGHBORHOOD METHOD

Neighborhood [3, 5] expands upon Majority, predicting protein functions by examining all proteins within a specific radius and identifying overrepresented functional annotations. Unlike Majority, this method doesn’t take into account any aspects of network topology within the local neighborhood. For instance, if a radius of 2 is considered, it encompasses all proteins within that range, regardless of whether they are direct neighbors or not.

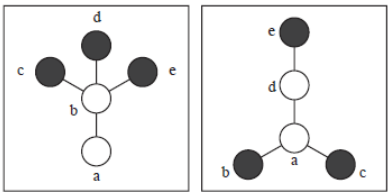

Fig. 2. Two protein interaction graphs [5] are treated in the same way with Neighborhood within a radius of 2 while annotating protein a. The black nodes are protein with known functions, and rest nodes are protein with unknown functions.

In Figure 2, two interaction networks are handled equally when assessing a radius of 2 and annotating protein a. However, in the initial scenario, there’s only one link linking protein a to the annotated ones and here a is not direct neighbor of the annotated proteins. In the second case, there are numerous separate paths connecting a to

the annotated proteins. Additionally, two of the annotated proteins are directly neighboring a.
In this approach [3], during the preprocessing phase, the physical interaction data were consolidated into a protein interaction map. Subsequently, the function of each protein in the map is forecasted based on the functions of its 'n-neighboring proteins,' referring to a group of proteins reachable through n physical interactions at most (where n is an integer parameter). The protein of interest is then attributed the function with the highest x^2 value among the functions of all n-neighboring proteins. For each function category member, the x^2 value is computed using the subsequent formula:

$$x^2 = \frac{(n_i - e_i)^2}{e_i}$$

Where,
i= a protein function
e_i= expected number of i n n-neighoring proteins
n_i= observed number of i in n-neighboring proteins
Afterwards, the function of a queried protein is predicted to be function i with the highest x^2 value. If there are several functions with the same maximum x^2 value, both functions are designated. The ideal n value is ascertained through a self-consistency examination, where the projected functions of all proteins in the map are juxtaposed with their annotated functions for each n.
In this study [3], several experimentally-determined protein-protein interaction datasets have been compiled to assess the accuracy of predicting protein function based on them. To mitigate bias arising from unequal distribution of proteins across functions, the x^2 value has been calculated for each function. Additionally, systematic efforts have been made to explore the potential inclusion of indirectly interacting proteins in prediction.
Three definitions of 'protein function' were examined here [3]. The simplest one, the subcellular localization site, could be predicted with the highest reliability at 72.7%. This rate appears reasonable since proteins from various localization sites may occasionally interact, and experimental errors cannot be overlooked. Predicting the cellular role of proteins achieved an accuracy of 63.6%. This definition of protein function might be most beneficial for subsequent experimental investigations, and it's understandable that the 'guilt-by-association' principle is effective in inferring the cellular role.

5. PROTEIN FUNCTION PREDICTION USING NETWORK FLOW BASED METHOD

Nabieva et al. (2005) [5] proposed a network flow based algorithm, FunctionalFlow that exploits the underlying structure of protein interaction maps in order to predict protein function. In cross validation testing on the yeast proteome, it has been shown that FunctionalFlow has improved performance over previous methods in predicting the function of proteins with few (or no) annotated protein neighbors. By comparing several methods that use protein interaction maps to predict protein function, it has been demonstrated that FunctionalFlow performs well because it takes advantage of both network topology and some measure of locality. Finally, it is shown that performance can be improved substantially as multiple data sources have been considered and used them to create weighted interaction networks. Here, protein–protein physical interaction network has been constructed by using the protein interaction dataset compiled by GRID. The resulting network is a simple undirected graph $G = (V, E)$, where there is a vertex or node $v \in V$ for each protein, and an edge between nodes u and v if the corresponding proteins are known to interact physically (as determined by one or more experiments). Initially, a graph with unit-weighted edges has been considered, and then considers weighting the edges by the 'confidence' in the edge. The weight of the edge between u and v is denoted by $w_{u,v}$ For all reported results, it is considered that only the proteins making up the largest connected component of the physical interaction map (4495 proteins and 12 531 physical interaction links). We have learnt about many methods for predicting protein function from this paper. We have seen a detailed approach to work with it. We have worked with an undirected graph as the same way it has been used here described in this paper.

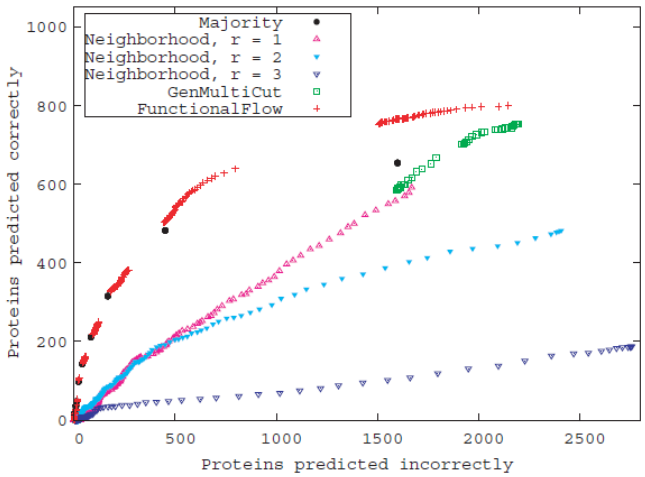

Fig. 3. ROC analysis was conducted by Nabieva et al. (2005) on Majority, Neighborhood, GenMultiCut, and FunctionalFlow using the unweighted physical interaction map of yeast [5]

The performance of four methods, Majority [2], Neighborhood [3], GenMultiCut [4], and FunctionalFlow [5], was evaluated using a 2-fold cross-validation on the unweighted yeast physical interaction map. Figure 3 illustrates, as a function of false positives (FP), the number of true positives (TP) predicted by each method, achieved by adjusting the scoring threshold. FunctionalFlow consistently identifies more TPs across all FPs compared to GenMultiCut or Neighborhood using radius 1, 2, or 3. Notably, FunctionalFlow outperforms Majority when proteins do not directly interact with at least three proteins of the same function.
From various techniques of protein function prediction using protein-protein interaction network, functional flow [5] is comparatively effective than other methods. Without parallel version of functional flow, it can't be applied practically for large scale of network. So, parallel version of functional flow algorithm has been proposed by Akkoyun et al. [30]. The major steps of this algorithms are described below:
First step is Transformation of input network and function annotations to a text based format that can be processed in a parallel way.
Second step is generation of a hash table that represents all interactions in the protein-protein interaction network and distribution of it to all computing nodes.
Third step is starting a number of processes concurrently to perform their own operations and generation of key value pairs where each pair shows an individual flow for a function. Only one biological function is considered by each process and it propagates a variety of flows assigned for that function.

Fourth step is accumulation of all the propagated flows and to combine them for calculation of the total amount of flows that enter and individual protein for each biological function of the network.
Fifth step is comparison of the total amount of flows coming from each biological function and then annotation of proteins with functions which has the highest value.

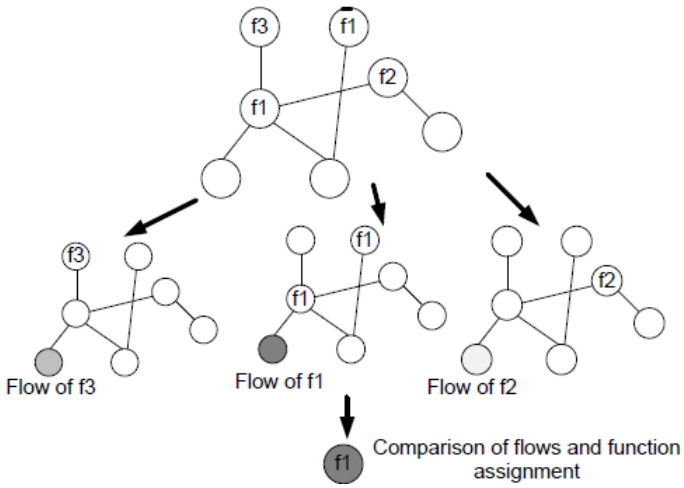

Fig. 4. An overview of the parallel functional flow [30]

6. GLOBAL OPTIMIZATION TECHNIQUE

Vazquez et al. (2003) [6] have suggested assigning functional classes to proteins based on their network of physical interactions, aiming to minimize the number of interactions between proteins from different categories. This functional assignment is performed on a global scale, relying on the overall connectivity pattern of the protein network. Due to the presence of multiple equivalent solutions, the method allows for multiple functional assignments.
The functional prediction strategy is based on a global optimization principle, where a score or energy is assigned to any given set of functional assignments for all unclassified proteins. This score is lower when interacting proteins share the same functional annotation. The novel aspect of this method is that the contribution to the total score of a functional assignment for an unclassified protein is calculated based on the number of neighboring classified and unclassified proteins with that function. Consequently, determining the functions of all unclassified proteins in the network is a global optimization problem and cannot be solved by considering only the local environment. The optimal function assignment corresponds to finding the minimal score for the entire network.
They applied their functional prediction method to the yeast Saccharomyces cerevisiae protein-protein interaction network. The interaction data, sourced from Reference [2], includes 1,826 proteins and 2,238 identified interactions. The functional classifications were obtained from the MIPS database [31], which features a detailed scheme with 424 functional categories, along with two categories for proteins without an assigned function: 'CLASSIFICATION NOT YET CLEAR-CUT' and 'UNCLASSIFIED PROTEINS'. This dataset includes 441 proteins in these two categories. Using their global optimization method, they assigned functions to all the proteins within these categories.
For each unclassified protein, they reported its degree (the number of proteins it is directly connected to) and listed up to three of the most probable predicted functions identified by their method. They attributed a higher level of certainty to the functions that appeared with a higher percentage of occurrence.

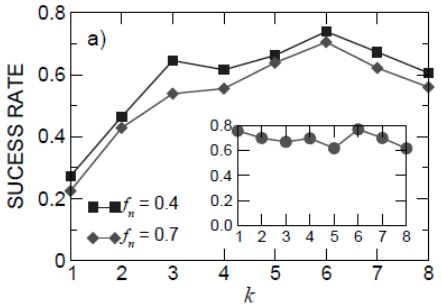

Fig. 5. The success rate of our method after setting a fraction fn of classified proteins as unclassified [6].

A key challenge in protein function predictions is assessing the reliability of the method given the incomplete knowledge of the interaction network. To establish an upper bound on the predictive power of their method, they ignored the functionality of a finite fraction f_n of the classified proteins and then measured the success rate of their predictions by comparing them with the actual classifications. This approach provides a quantitative estimate of the reliability of predictions based on the available network information. Figure 5 illustrates the percentage of successful predictions as a function of protein degree for different f_n values, using the most detailed functional classification scheme (424 classes). For unclassified proteins with a degree greater than 2, correct predictions can be made in 60%-70% of cases, even when up to 40% of the information is missing (f_n=0.4), and this accuracy is fairly independent of the protein's degree.

7. PROTEIN FUNCTION PREDICTION USING MINIMUM DISTANCE CLASSIFIER

Tania et al. [21] proposed a method that uses a minimum distance classifier to predict the function of unannotated protein. From the protein interaction network hyper geometric distribution value and correlation coefficient of every protein have been calculated and used as features for this method. Though proteins are involved in so many functions, for their work only five functional groups (cell polarity, DNA repair, lipid metabolism, protein modification and protein synthesis) have been considered. Two different methods have been used for this study. One is PFP_MINDSET1 and another method is PFP_MINDSET2.

PFP_MINDSET1

Hypergeometric p-value has been used as features for PFP_MINDSET1. This p-value describes the distance between a pair of proteins u and v based on every protein's interaction neighbors. The hypergeometric p-value is defined as:

$$P(N, n_1, n_2, m) = \frac{\binom{N}{m}\binom{N-m}{n_1-m}\binom{N-n_1}{n_2-m}}{\binom{N}{n_1}\binom{N}{n_2}}$$

Here,
N= all proteins in the interaction network

$m = |N_u \cap N_v|$
$n_1 = |N_u|$
$n_2 = |N_v|$
p-value shows the chances that proteins u and v share m neighbors in a network of N proteins where uhas n_1 neighbors and v has n_2 neighbors. The numbers of distinct ways are counted in which two proteins with n_1 and n_2 interaction partners have m in common. And this is counted to compute p-value.
For this process, protein interaction network G is given. Where each node represents a protein P_i and the edge between nodes are represented as (P_i, P_j) which stands for interaction sets f five output classes $(O_1, O_2, \ldots, O_5)$. Protein pair of unknown function (P_i, P_j) is mapped to a particular functional group such as $(P_i, P_j) \rightarrow O_k$ where $O_k \varepsilon O$. The steps of PFP_MINDSET1 is given below:
Step 1:
In this step, computation of hypergeometric values are performed for each protein pair in G.
Step 2:
In this step, calculation of mean p-value is performed for each class which is done by averaging the p-value of known pairs of protein belonging to that class.
Step 3:
This step if for computation of the distances between (P_i, P_j) and O_k by taking the differences between the p-value of (P_i, P_j) and mean p-value of each $O_{k(k=1,2,..,5)}$
Step 4:
This step is final step for assigning function to proteins that means assignment of (P_i, P_j) to $O_k(k = 1, 2, \ldots, 5)$

PFP_MINDSET2
PFP_MINDSET2 uses the correlation coefficient as features. Correlation coefficient is described by the distance a pair of proteins which is based on the adjacency matrix of the protein interaction network.
In protein interaction network, the binary vectors, the set of N proteins, are represented by $X_i = (X_{i1}, X_{i2}, \ldots, X_{iN})$. Here X_{ik} is 1 if the i^{th} protein interacts with k^{th} protein and 0 otherwise. Based on the adjacency matrix, a correlation coefficient S_{ij} is calculated, which is defined by the equation given below:

$$s_{mn} = |\frac{X_{mn} - n\overline{X_m X_n}}{\sqrt{X_{mm} - n\overline{(X_m{}^2)}(X_{nn} - n\overline{(X_n{}^2)})}}|$$

$X_{ij} = X_i.X_j$
X_{ij} is equal to the number of bits "on" in both vectors.
X_{ii} is equal to the number of bits "on" in one vector. It derives similarity distance between interacting protein pairs by using correlation coefficient from the equation below:

$$d_{ij} = |1 - S_{ij}|$$

The steps of PFP_MINDSET2 is given below:
Step 1
This step computes the correlation coefficient of each protein pair in protein interaction network G. It also calculates the similarity distance from the computed correlation coefficient.
Step 2
From this step, mean similarity of each class is got by averaging the similarity distance of known protein that belongs to that class.
Step 3
This step computes the similarity distance between protein pairs (P_i, P_j) and O_k (O_k is one of the functional groups from the considered 5 functional groups). Similarity distance is calculated by taking difference between similarity distance of (P_i, P_j) and mean similarity distance of each $O_{k(k=1,2,\ldots,5)}$
Step 4
This step assigns (P_i, P_j) to $O_{k(k=1,2,\ldots,5)}$ whose distance from (P_i, P_j) is minimum.
They have achieved satisfactory accuracy in predicting protein functions by using minimum the distance classifiers.

8. PROTEIN FUNCTION PREDICTION USING TWO-NODE FREQUENT PATTERN

Li et al. (2011) [20] proposed a two node frequent pattern based method to predict function of unannotated protein on the basis of frequent pattern mining in graph data. This method is processed in 3 steps:
i. The first step is neighbor finding steps
ii. Second step is pattern finding
iii. The third step is function annotation
This approach considers an unweighted undirected graph for protein-protein interaction network. Then they use three main concepts for predicting function for a protein. These three concepts are given below:
i. Two-Node Functional Pattern: A pattern with two function item sets where each function item set corresponds to a graph node in the protein-protein interaction network.
ii. Support of Two-Node Functional Pattern: Number of all patterns with the same function sets in the graph.
iii. Most Frequent Two-Node Functional Pattern: This denotes the two-node functional pattern whose support is the largest among all two node functional patterns. When there is two or more two-node functional pattern, then they are sorted arbitrarily. In figure 6, there are total six protein nodes. Among them n_1, n_2, n_3, n_4 and n_5 these 5 nodes are annotated. Node n_6 is not annotated. To annotate n_6, first task is to find the neighbors of n_6. Node n_6 is connected with node n_1 and n_5. Neighbor n_1 has function f_1 and neighbor n_5 has function $\{f_1, f_3\}$. Now, the next step is to find all two-node functional patterns.

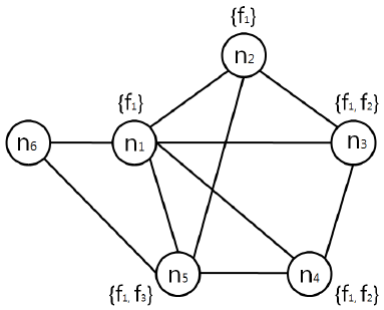

Fig. 6. A protein-protein interaction graph where n_6 is unannotated and n_1, n_2, n_3, n_4 and n_5 are annotated [20]

Table 1 lists all the two-node functional patterns including $\{f_1\}$, and the supports are 1, 3, and 2.

Table 2 lists all the two-node functional patterns including $\{f_1, f_3\}$, and the supports are 2, 1, and 0.

Table 1. Two-Node Functional Patterns Including $\{f_1\}$ and Corresponding Supports [20]

Two-Node Functional Pattern Including $\{f_1\}$	Support
$\{f_1\} - \{f_1, f_2\}$	3
$\{f_1\} - \{f_1, f_3\}$	2
$\{f_1\} - \{f_1\}$	1

Table 2. Two-Node Functional Patterns Including $\{f_1, f_3\}$ and Corresponding Supports [20]

Two-Node Functional Pattern Including $\{f_1\}$	Support
$\{f_1, f_3\} - \{f_1\}$	3
$\{f_1, f_3\} - \{f_1, f_2\}$	2
$\{f_1, f_3\} - \{f_1, f_3\}$	1

Table 3. Candidate Functions and Their Corresponding Supports [20]

Candidate Functions	Support
f_1	5
f_2	3
f_3	2

The most frequent two-node functional patterns corresponding to the two function categories are $\{f_1\} - \{f_1, f_2\}$ for $\{f_1\}$, which has the highest support of 3, and $\{f_1, f_3\} - \{f_1\}$ for $\{f_1, f_3\}$, which has a largest support of 2 respectively. Excluding the original function sets $\{f_1\}$ from $\{f_1\} - \{f_1, f_2\}$ and $\{f_1, f_3\}$ from $\{f_1\} - \{f_1, f_3\}$, each candidate function from $\{f_1, f_2\}$ and $\{f_1\}$ is sorted in table 3.

From the table 3, it is seen that f_1 has support 5, f_2 has support 3 and f_3 has support 2. Among all the two-node functional patterns, f_1 has the highest support of 5. So, node n_6 is annotated with function f_1.

The primary protein-protein interaction data used in their experiment for the baker's yeast Saccharomyces cerevisiae includes 1,274 protein nodes and 3,222 interactions among them. This dataset was obtained from the DIP [26] website. Functional annotation was conducted utilizing the Functional Catalogue (FunCat) [32]. As a preprocessing step, they removed all protein nodes that lacked interactions with other proteins from the dataset. Consequently, the experiment utilized 1,249 protein nodes and 2,985 interactions. After annotating each protein node, 16 functional categories were employed in the experiment. They depict a protein-protein interaction network as an unweighted, undirected graph, where the nodes represent proteins and the edges signify interactions between them.

Using the 10-fold cross-validation method, the proposed approach achieved an average partly prediction accuracy of 0.600. Table 4 presents the partly accuracy obtained by Two Node Frequent Pattern method [20].

9. PROTEIN FUNCTION PREDICTION USING NEARER NEIGHBOR PROTEINS INTERACTION

Khan and Tareeq (2024) [19] proposed a method to predict protein functions using the knowledge of functions of nearer neighbor proteins upto 2nd degree neighbor proteins. In their method, first they have clustered protein-protein interaction network with k mean clustering algorithm and secondly they have annotated functions of the proteins. The function annotaion process of their method contains 3 steps. In the first step, they have viewed the functions of the direct neighbor proteins, If any of the function is present with higher frequency, then that function is assigned to the input protein. Otherwise, they equally weight the support of the direct neighbor proteins and then go to 2nd and 3rd steps to annotate the functions of the input proteins. In the second step, this method examines the frequency of neighboring proteins that are located at a radius of 2. After finding the 2nd degree neighbors' functions, the method goes to 3rd step. In the third step, they find the number of common neighbors shared by the "input protein" and the "input proteins' direct neighbor proteins". The higher the common neighbors between the pairs, the higher chance is that they can share the same functions. So, based on the number of common neighbors, protein functions' presence are weighted in this step.

After completing these three steps, the final result is obtained by integrating their outcomes. Figure 7 shows the correctly predicted molecular and biological functions by this proposed method.

This method's performance has been evaluated against two established protein function prediction methods: Two-Node Frequent patterns and PFP_MINDSET1. When comparing with Two-Node Frequent patterns, we assessed the overall accuracy of both molecular function and biological process prediction. In contrast, when comparing with PFP_MINDSET1, they focused solely on the five most common functional categories to gauge and contrast sensitivity, specificity, and accuracy. Comparison 1 demonstrates that their proposed method outperforms Two-Node Frequent patterns [20], while Comparison 2 indicates that the performance of PFP_MINDSET1 [21] and their method is nearly equivalent. These comparisons collectively suggest that their proposed method [19] exhibits satisfactory performance.

Table 4. Partly Accuracy For Two Node Frequent Pattern [20]

Run	Partly Accuracy of Two Node Frequent Pattern
1	0.685
2	0.540
3	0.574
4	0.533
5	0.518
6	0.711
7	0.631
8	0.615
9	0.514
10	0.640
Average	0.600

10. CONCLUSION

Proteins are essential components of living organisms, playing critical roles in various biological processes. Understanding protein functions can significantly contribute to the development of systems, treatment procedures, medicines, and more. However, many protein functions remain unknown. To address this, researchers employ both wet lab and dry lab experiments to uncover these functions. Recently, dry lab methods, particularly computational approaches, have gained popularity due to their efficiency, lower cost, and reduced resource requirements compared to traditional wet lab techniques. Computational intelligence techniques, such as sequence and structural analysis, gene expression data analysis, pathway analysis, and protein-protein interaction (PPI) network analysis, are particularly effective for predicting protein functions. Since proteins often operate by interacting with other proteins, PPI networks are especially valuable for function prediction. This review

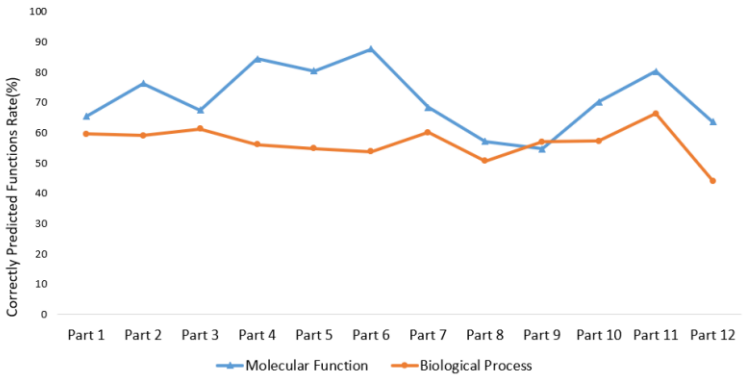

Fig. 7. Correctly Predicted Molecular and Biological Functions Rate (%) [19]

paper explores various protein function prediction methods using PPI networks, as proposed and utilized by researchers. It aims to facilitate a comprehensive understanding of these methods, enabling researchers to apply or develop new techniques for predicting protein functions.

11. REFERENCES

[1] Tiwari, A. K., Srivastava, R. (2014). A survey of computational intelligence techniques in protein function prediction. International journal of proteomics, 2014.

[2] Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein–protein interactions in yeast. Nature biotechnology, 18(12), 1257-1261.

[3] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast, 18(6), 523-531.

[4] Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., Kasif, S. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. Proceedings of the National Academy of Sciences, 101(9), 2888-2893.

[5] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 21, i302-i310.

[6] Vazquez, A., Flammini, A., Maritan, A., Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. Nature biotechnology, 21(6), 697-700.

[7] Pandey, G., Steinbach, M., Gupta, R., Garg, T., Kumar, V. (2007, August). Association analysis-based transformations for protein interaction networks: a function prediction case study. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 540-549).

[8] Sun, P., Tan, X., Guo, S., Zhang, J., Sun, B., Du, N., ... Sun, H. (2018). Protein function prediction using function associations in protein–protein interaction network. IEEE Access, 6, 30892-30902.

[9] Nguyen, C. D., Gardiner, K. J., Nguyen, D., Cios, K. J. (2008). Prediction of protein functions from protein interaction networks: a Naïve Bayes approach. In PRICAI 2008: Trends in Artificial Intelligence: 10th Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, December 15-19, 2008. Proceedings 10 (pp. 788-798). Springer Berlin Heidelberg.

[10] Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F. (2002, August). Prediction of protein function using protein-protein interaction data. In Proceedings. IEEE Computer Society Bioinformatics Conference (pp. 197-206). IEEE.

[11] Bogdanov, P., Singh, A. K. (2009). Molecular function prediction using neighborhood features. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(2), 208-217.

[12] Moosavi, S., Rahgozar, M., Rahimi, A. (2013). Protein function prediction using neighbor relativity in protein–protein interaction network. Computational Biology and Chemistry, 43, 11-16.

[13] Li, M., Wu, X., Wang, J., Pan, Y. (2012). Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. BMC bioinformatics, 13, 1-15.

[14] Letovsky, S., Kasif, S. (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics, 19(suppl-1), i197-i204.

[15] Xiong, W., Liu, H., Guan, J., Zhou, S. (2013). Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks. BMC bioinformatics, 14, 1-13.

[16] Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., ... Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415(6868), 141-147.

[17] H. Wang, H. Huang, and C. Ding, "Function-function correlated multi-label protein function prediction over interaction networks," Journal of Computational Biology, vol. 20, no. 4, pp. 322–343, 2013.

[18] Yanai, I., Mellor, J. C., DeLisi, C. (2002). Identifying functional links between genes using conserved chromosomal proximity. Trends in genetics, 18(4), 176-179.

[19] Khan, S., Tareeq, S. M. Protein Function Prediction Using Nearer Neighbor Proteins Interactions. International Journal of Computer Applications, 975, 8887. Volume-186, number-17, pp(15-22), 2024.

[20] Li, P., Heo, L., Li, M., Ryu, K. H., Pok, G. (2011, July). Protein function prediction using frequent patterns in protein-protein interaction networks. In 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (Vol. 3, pp. 1616-1620). IEEE.

[21] Tania Chatterjee Li and Piyali Chatterjee. "Protien Function Prediction by Minimum Dis- tance Classifier from Protien Interaction Network". In: (2012). doi: 978 - 1 - 4673 - 4700 - 6/12/$31.00.

[22] Costanzo, M.C. et al. The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res. 28, 73–76 (2000).

[23] Mewes, H.W. et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 28, 37–40 (2000).
Uetz, P. et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000).

[24] Ito, T. et al. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA 97, 1143–1147 (2000).

[25] B. Boeckmann, A. Bairoch, R. Apweiler et al., "The SWISS-SPROT protein knowledgebase and its supplement TrEMBL in 2003," Nucleic Acids Research, vol. 31, no. 1, pp. 365–370, 2003.

[26] I. Xenarios, L. Salw´ınski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," Nucleic Acids Research, vol. 30, no. 1, pp. 303–305, 2002.

[27] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Research, vol. 30, no. 1, pp. 207–210, 2002.

[28] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Research, vol. 39, supplement 1, pp. D561–D568, 2011.

[29] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," Nucleic Acids Research, vol. 28, no. 1, pp. 235–242, 2000.

[30] Akkoyun, E., Can, T. (2011, July). Parallelization of the functional flow algorithm for prediction of protein function using protein-protein interaction networks. In 2011 International Conference on High Performance Computing Simulation (pp. 56-62). IEEE.

[31] The MIPS Comprehensive Yeast Genome Database (CYGD), http://mips.gsf.de/proj/yeast/CYGD/db/.

[32] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., ... Mewes, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic acids research, 32(18), 5539-5545.