International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.27, June 2024

Protein Function Prediction using Protein-Protein Interaction Networks Involving MCL and Majority Rule

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

Fatema Tuj Jahura
Assistant Programmer
Janata Bank PLC
Dhaka, Bangladesh

Shiplu Hawladar
Software Engineer
Google
Sunnyvale, California, United States

ABSTRACT

Protein is essential for all life processes, playing crucial roles such as providing structural integrity to the body and facilitating the transport of various substances within it. Understanding protein functions is critical for advancing biological science, as it aids in the improvement, regulation, and maintenance of numerous biological systems. Various methods exist to predict the functions of proteins with unknown roles, but many are time-consuming, complex, and costly. This study introduces a novel method that offers higher accuracy in predicting protein functions. It is easier, faster, and less expensive compared to many existing techniques. This new approach employs the Markov Clustering (MCL) Algorithm to cluster protein networks, followed by the application of the majority rule [3, 36] to predict protein functions.

General Terms

Protein, Prediction, Function

Keywords

Protein-protein interaction (PPI) network, Markov clustering (MCL) algorithm, protein function prediction, majority

1. INTRODUCTION

Protein is a fundamental component of the body, often referred to as its building block. It is crucial for constructing and repairing tissues, as well as regulating vital processes such as metabolism, digestion, and growth. Understanding protein functions is essential for developing effective medicines for living organisms. Leveraging knowledge about proteins with known characteristics can lead to the creation of important vaccines, medicines, and herbal products. Many protein functions remain unidentified. Discovering and predicting these functions will aid in the development of effective systems and treatments, ultimately benefiting many lives through improved therapies and medical interventions.

Wet lab based experiments involve things like test tubes, beakars, suitable labs. Wet laboratory-based experiments necessitate substantial investment in terms of personnel, financial resources, and extensive time commitments. Conversely, dry laboratory-based experimentation relies on computational methodologies, requiring computational resources, machinery, and necessitating a reduced personnel requirement, shorter timeframes, and notably lower costs compared to wet laboratory-based research endeavors. Consequently, computational-based approaches have gained prominence in recent years across numerous biological research domains, supplanting traditional wet laboratory methodologies with their attendant physical constraints.

Computational methods have become popular also in predicting protein functions. Protein functions can be predicted by following various ways like- based on the sequences and structures of the proteins, based on their interactions with other proteins, using the knowledge of gene expression data, using the knowledge of pathway analysis from gene expression data etc. [17].

A protein almost never performs its function in isolation. Rather, it usually interacts with other proteins in order to accomplish a certain function. However, in keeping with the complexity of the biological machinery, these interactions are of various kinds. At the highest level, they can be categorized into genetic and physical interactions. Proteins are more directly related to the process through which a protein accomplishes its functions. These interactions are of various kinds, such as the simultaneous membership of two proteins in the following biological systems [21]:

—A metabolic and/or signaling pathway

—A morphogenic pathway in order to perform a developmental function

—A protein complex and other such molecular machines

Since a protein generally interacts with more than one proteins, these interactions can be structured to form a network, and hence the name protein interaction networks. 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 of proteins not even directly interacting with it.

Due to the importance of the knowledge of these interactions, several high-throughput methods have been proposed for discovering those [22]. Again, depending on the final output, these methods can be categorized into two types [23], namely the discovery of pair wise interactions and extraction of protein complexes. While two-hybrid systems, protein chips and phage display are the most commonly known methods in the former category, the Tandem Affinity

Purification (TAP) approach is commonly used for extracting complexes.

2. LITERATURE REVIEW

Tiwari and Srivastava (2014) presents the computational intelligence techniques in protein function predictions [17]. This introduces with the existing computational techniques in protein function prediction. Four computational techniques are described here for predicting protein function. The techniques are

a. Using Sequences and Structures: There is a state of art comprehensive review of various computational intelligence techniques used in wide areas of applications like prediction of DNA and RNA binding sites, subcellular localization, enzyme functions, signal peptides, catalytic residue, nuclear/G protein coupled receptors, membrane protein using sequence and structures [24]. homology based method used this structure of a protein to identify protein using structure alignment technique [25]. The summary of the result obtained by various researchers using these techniques are also presented in this paper.

b. Using Protein Interaction Networks: Performing a specific function a protein must interact with another protein. The interaction of the protein is represented in the form of network called protein-protein interaction network. So by using the knowledge of this interaction network various computational techniques based approaches have been proposed for protein function prediction by using one or more interaction networks. Gaurav et al. (2006) [26] proposed an association analysis method based on h confidence.
Four categories of computational techniques for protein interaction network are

—Neighbor Based Techniques

—Clustering Based approaches

—Optimization Based Techniques

—Association Analysis Based Techniques

c. By Gene Expression Data: Gene expression is the process by which information from a gene is transformed into functional product such as protein or RNA by transcription and translation process. DNA micro arrays are used to analyze the gene expression level. Gene expression data are analyzed in the form of a matrix where each row represents a gene and each column represents a sample. Hon Nian et al. [9] developed the two step algorithm to predict the protein function.

d. In Pathway Analysis From Gene Expression Data Pathway: The pathway is a series of interconnected enzymatic steps linked with the production of intermediates that are used in the next enzymatic step so we can say that it is a series of consecutive enzymatic reactions that produce specific products. Pathway consists of genes that chemically act together for specific cellular or physiologic function so pathway analysis is useful for gene function prediction. Mikhail et al. [28] proposed a two phase approach to predict molecular function of not characterized gene by comparing their functional neighborhood to gene of known. Protein with similar functions have the similar type of patterned protein. Metabolic Pathway and Signaling Pathway are two types of pathway. In pathway analysis, each pathway will be ranked based on the score obtained either by the Enrichment analysis or by machine learning approaches. The highest score will be given to the pathway which has most relevant gene to related phenotype. To solve the issue, Mengfei et al. [29] proposed a diffusion state distance to capture a fined grained distance in proximity for function prediction in protein-protein interaction network. Xing-Ming et al. (2008) used SVM and genetic algorithm for the detection of protein interaction [30]. A case study for protein function prediction by using sequence derived properties is also provided in this paper. Various techniques have been described in this paper for predicting protein function. It is easier to predict a function of protein if its neighbor proteins function is known. Wei et al. (2013) have used an graph based centrality matrix to select proper candidate for labeling [31].
Hishigaki [4] proposed an objective prediction method has been developed that can systematically include the information of indirect interaction. This method can predict the subcellular localization, the cellular role and the biochemical function of yeast proteins with accuracies of 72.7%, 63.6% and 52.7%, respectively. The prediction accuracy rises for proteins with more than three binding partners and thus the open prediction results for 16 such proteins have been presented in this paper.
A protein interaction map has been considered here, where each node represents a protein and each edge represents the interaction between proteins. The function of each protein in the map is predicted, based on the functions of 'n-neighboring proteins', which are defined as a set of proteins reached via n physical interactions at most (n is an integer parameter). The protein of interest is assigned the function with the highest x2 value among functions of all n-neighboring proteins. For each member of the function category, the x2 value is calculated using the following formula:
$x^2 = \frac{(n_i - e_i)^2}{e_i}$
Where i denotes a protein function, e.g. 'Golgi', 'DNA repair' and 'transcription factor', e_i denotes an expectation number of i in n-neighboring proteins expected from the distribution on the total map, and n_i denotes an observed number of i in n-neighboring proteins. Then, the function of a query protein is predicted to be the function i with the maximum x^2 value. When there are multiple functions with the largest x^2 value, both functions are assigned. The optimal n value is determined by a so-called self-consistency test, where the predicted functions of all proteins in the map are compared with their annotated functions for each n.
Karaoz (2004) [8] proposed an effective methodology for combining biological evidence obtained in several high-throughput experimental screens and integrating this evidence in a way that provides consistent functional assignments to hypothetical genes. The visualization method of propagation diagrams has been used to illustrate the flow of functional evidence that supports the functional assignments produced by the algorithm. The results contain a number of predictions and furnish strong evidence that integration of functional information is indeed a promising direction for improving the accuracy and robustness of functional genomics. Mainly an effective method has been demonstrated here to interpret functional linkage networks as a medium for inferring gene function by integrating the evidence captured by protein-protein interaction and gene expression data. This framework provides two important capabilities. First of all, it provides a promising methodology for propagating functional information across functional-linkage graphs to genes that cannot be annotated with certainty solely by examining their neighbors in the graph and secondly it provides the integration of diverse types of experimental evidence about functional similarity with the propagation procedures. Some ideas have been got from this paper to infer gene function by protein-protein interaction and gene expression data though we have the goal to predict only protein function using protein-protein interaction data.
Majority rule [3, 36] says that it is possible to predict the functions of the proteins of unknown characteristics by the functions of their

neighbor proteins characteristics. But a protein of unknown characteristic may have many neighbors of different characteristics. The target of their research [3] was to minimize the number of different annotations that are associated with the neighboring proteins. They have proposed the assignment of proteins to functional classes on the basis of their network of physical interactions as determined by minimizing the number of protein interactions among different functional categories. This approach results in multiple functional assignments, a consequence of the existence of multiple equivalent solutions. A method has been applied to analyze the yeast Saccharomyces cerevisiae protein protein interaction network.
It has been also explored that the concept of interacting protein may belong to at least one common functional class and thus the knowledge of the functional classification of the remaining subset of not characterized proteins.
Use of traditional k-mean type algorithm is limited to numeric data. Qiao et al. (2020) [33] presents a clustering algorithm based on k-mean algorithm that works well for data with mixed numeric and categorical features. A new cost function and distance measure have been proposed here based on co-occurrence of values. The measures also take into account the significance of an attribute towards the clustering process. A modified description of cluster center has been proposed to overcome the numeric data only limitation of k-mean algorithm and provide a better characterization of clusters. The performance of this algorithm has been studied on real world data sets. Comparisons with other clustering algorithms illustrate the effectiveness of this approach. The proposed distance measure can work well for mixed as well as pure numeric and categorical data sets.
Nabieava (2015) [18] proposed a network flow based algorithm has been developed, 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).
GenMulticut is a generalization of well studied multi way k cut problem in computer science. This method is described elaborately by Nabieva et al. (2005) [18] and Vazquez et al. (2003) [36]. GenMulticut takes into account more global properties of interaction maps. It does not reward local proximity. For example in a particular network if there are two protein of known functions, the other protein of unknown functionalities would be assigned with any of these functions regardless of the network size.

Khan et al. (2024) [39] utilized protein-protein interaction networks, information on nearby neighbor proteins, and the presence of protein functions to predict the functions of unknown proteins.

3. PROTEINS: STRUCTURES AND THEIR RELATIONSHIP TO FUNCTIONS

3.1 Protein

proteins are the polymers consisted of two or more amino acid monomers. Amino acids are the building blocks of protein. amino acid consists of an amino group, a carboxyl group an alpha carbon. At one end of the alpha carbon, there is hydrogen and at another end of the alpha carbon, there is a variable group. This variable group determines the structures and functions of a particular protein.
Multiple amino acids are linked together to create polypeptide or protein. They can be linked together through a reaction called condensation. All amino acids are linked together through a peptide bond between carboxyl group and amino group.

3.2 Protein Structures

There are mainly four different structures of protein.

i. Primary Structure of Protein: Primary Structure of protein refers the orders of amino acids. It presents a sequence how the amino acids are linked together.

ii. Secondary Structure of Protein: Secondary protein structure is the particular shape that just a segment of the polypeptide chains take on. Secondary structure is formed when the sequence of amino acids are linked together through hydrogen bonds.
There are two types of secondary protein structures:
a. Alpha Helix: Alpha helix is a segment of a chain just forming "a helical" structure.
b. Beta Pleated Sheet: It is when the chain of the amino acid just fold over on itself to become this folded sheet.

iii. Tertiary Structure of Protein: Tertiary protein structure is the 3 Dimensional shape of the entire polypeptide. Different secondary structures fold themselves to become this overall globular 3 dimensional shape of protein.

iv. Quaternary Structure of Protein: When there is more than one polypeptide chain making a particular protein, then its structure is called quaternary protein structure. Not all proteins have quaternary structure.

3.3 Relation between Protein Structures and Functions

Each protein has its specific structure and performs specific tasks. When its structure is changed, the function also changes.
Example: Hemoglobin protein is responsible for transporting oxygen in human body. It is consisted of many amino acids. Among them if one is changed, then the entire structure and shape of the hemoglobin is changed. And for this reason the shape of red blood cells are also changed. And the changed shape of red blood cell which contains abnormal hemoglobin configuration, that is unable to transport oxygen which is harmful to human body and for this reason a person can die. Figure 1 presents normal and abnormal shape of hemoglobin protein.

4. CLUSTER ANALYSIS

4.1 Introduction with Clustering

Cluster analysis, also known as clustering, involves organizing a collection of items in a manner where items within the same group,

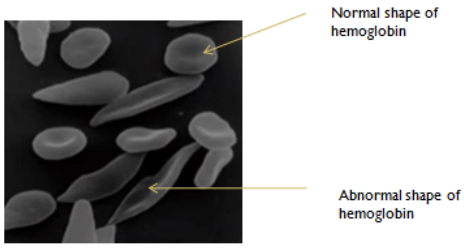

Fig. 1. Red blood cells (Normal and abnormal shape of hemoglobin) [38]

termed clusters, exhibit greater similarity to one another compared to items in different groups.
Cluster analysis doesn't represent a singular algorithm. Clustering algorithms vary greatly in their definitions of clusters and methods for identifying them efficiently. Common cluster definitions include groups with close member distances, dense regions within data space, intervals, or specific statistical distributions.
Genome scale data on protein interactions are generally represented as large networks, or graphs, where hundreds or thousands of proteins are linked to one another. Since proteins tend to function in groups, or complexes, an important goal has been to reliably identify protein complexes from these graphs. This task is commonly executed using clustering procedures, which aim at detecting densely connected regions within the interaction graphs.
Clustering can be performed in two ways- One is vector clustering and another is graph clustering. There are various algorithms for clustering such as K-mean algorithm, Markov clustering algorithm etc. Among them Markov clustering algorithm is one of the most efficient algorithms.

4.2 Markov Clustering (MCL) Algorithm

Markov Cluster algorithm is a fast and scalable algorithm for graphs. This is based on simulation of flow in graph. The main topic of this algorithm is mathematical theory. This algorithm concerns with issues of scalability, position in cluster analysis and graph clustering, performance criteria for graph clustering.
The MCL algorithm is straightforward, simulating flow through the alternation of two basic algebraic operations on a matrix. It is adaptable, emergent, scalable, intrinsic, and fast. There is a fundamental relationship between the MCL process and the cluster structure in graphs, which is particularly valuable given the numerous heuristic methods used in cluster analysis.
In the Markov clustering algorithm, highly connected nodes are likely to be grouped in the same cluster, while sparsely connected nodes may end up in different clusters. A random walk can start from any node; if it begins at node r and has a high probability of reaching node t, then r and t will be clustered together. The probability of a random walk taking an edge at node u depends solely on u and the edge, not on the previous path taken, simplifying the computation. A flow network is used to approximate the partitioning, with an initial flow introduced into each node. At each step, a portion of the flow moves from a node to its neighbors through the outgoing edges.
Edge weight in MCL is determined based on similarity between two nodes. The edge weight is considered as the bandwidth or connectivity. If an edge has higher weight than the other, then more flow will be flown over the edge. The amount of flow is proportional to the edge weight. If there is no edge weight, then the same weight can be assigned to all edges.
In MCL, a graph is partitioned in such a way that inter partition similarity is the highest and the intra partition similarity is the lowest. The number of Higher-Length paths in graph is large for pairs of vertices lying in the same dense cluster and Small for pairs of vertices belonging to different clusters. A Random Walk in G that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited.

5. DATA

Weighted graph data sets have been used in this research. The experiment has been done on Saccharomices cerevisiae protein. The protein protein interaction data sets have been collected from string-db.org [34].
Each node has been considered as a protein, and edges have been considered as a link between two proteins. Weights have been given based on the distance between proteins. Nearest proteins get the highest weights. Weights decrease when the distances increase.
There are two main parts of this research. First task of this experiment was to cluster proteins. The protein–protein physical interaction network has been constructed using the protein interaction dataset compiled by dataset (string-db.org). For clustering proteins, a dataset has been used that was consisted with a set of proteins and a score (score based on weight). The resulting network is a directed 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. The output provides sets of clusters. Similar types of proteins are grouped in the same cluster.
After completion of clustering, next task is to predict functions of proteins. The output set got by the first task (clustering) has been used as the input sets for this case. Specifically some specific proteins have been targeted as input proteins which are considered as a protein of unknown characteristics. And other proteins have been considered as annotated proteins. A complete code has been implemented to conduct this experiment. Each time, protein information is input, and the characteristics of the target protein are obtained as output.
To assess the reliability of the obtained results, the dataset sourced from www.uniport.org [33] was utilized. By iteration, the characteristics derived from this experiment have been scrutinized against the dataset. Each protein has been individually queried on the website www.uniport.org [33] to ascertain its functions and validate the accuracy of the findings.

6. METHOD DESCRIPTION

The proposed method includes two main steps. In the first step, from the available protein protein interaction networks, a protein protein interaction (PPI) network has been chosen and clustered. Markov clustering algorithm has been used to cluster a chosen PPI network.
Then, after clustering, in the second step, Majority rule [3] has been applied to predict the function of a protein. This two-step process is then sequentially applied to all available PPI networks.
Majority rule says [3, 36] that possible functions can be assigned to uncharacterized proteins based on the known functions of their direct neighbor proteins.
For example, in figure 2 [36], there are eight proteins. Among those proteins, the proteins in gray boxes are unclassified (their functions are not known). The rest five proteins are classified (their functions

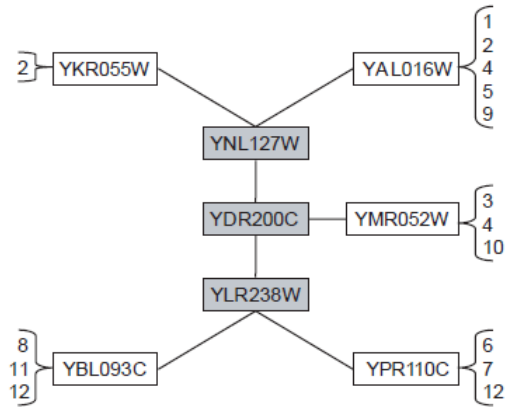

Fig. 2. Protein Interaction Network [36]

are given in brackets) [36]. The functions are labeled according to the following criteria [36]:

1 Cell growth
2 Budding, cell polarity, filament formation
3 Pheromone response, mating type determination, sex-specific protein
4 Cell check pointing proteins
5 Cytokinesis
6 rRNA synthesis
7 tRNA synthesis
8 Transcriptional control
9 Other transcriptional activities
10 Other pheromen response activities
11 Stress response
12 Nuclear organization

Given one of these proteins of unknown function, if we take as a prediction the function that appears more often in the neighbor proteins of known function, then the following classification is obtained [36].

—YNL127W (2)
—YDR200C (3, 4, 10)
—YLR238W (12)

Here, according to Majority rule [3, 36], protein YNL127W gets the functionality of Budding, cell polarity, filament formation. Protein YDR200C gets the functionality of Pheromone response, mating type determination, sex-specific protein (3), Cell check pointing proteins (4) and other pheromen response activities (10) [36]. And protein YLR238W gets the functionality of nuclear organization [36].

7. PERFORMANCE ANALYSIS

A complete process for this research has been developed by implementing a comprehensive code using C programming language. This process takes protein protein interaction network sets as inputs and clusters them. It clusters a protein protein interaction network set by using MCL algorithm. After clustering, it takes the proteins of unknown functions as inputs and it annotates the functions of the input proteins. The function annotation is done by using Majority rule [3, 36].

This method has been applied on a protein interaction data (of Saccharomyces cerevisiae) which has been collected from STRING [8]. This network contains 1550 proteins and 2505 edges. The data set has been divided into 10 parts and the experiment has been done sequentially 10 times for each part. The accuracy of protein function prediction for each part has been presented on table 1. The average accuracy of the proposed method (in this research) is 73.4%. The performance of Majority [36], Neighborhood [4], GenMultiCut [18, 36], and this method (used in this research) on the weighted graph have been compared for the same data sets of Saccharomices cerevisiae protein collected from string-db.org [34].

Table 1. Accuracy of Protein Function Prediction By the Proposed Method

Part	Correctly Predicted Functions
Part 1	64%
Part 2	68%
Part 3	65%
Part 4	66%
Part 5	81%
Part 6	70.5%
Part 7	82%
Part 8	76.5%
Part 9	78.5%
Part 10	82%
Average	73.4%

Table 2. Success Rate of Various Protein Function Prediction Methods

Methods	Success Rate
Neighborhood1(radius=1)	57.7%
Neighborhood2 (radius=2)	61.71%
Neighborhood3 (radius=3)	70.02%
GenMulticut	62.15%
Majority	71.79%
Proposed Method (in this research)	73.4%

Table 2 and figure 3 shows the accuracy rates of various protein function prediction methods using protein protein interaction networks. From this graph and figure, a comparison of the performance of various methods can be comprehended.

From the table 2, it is seen that the success rate of this method (the presented method in this research paper) is higher than other methods.

8. CONCLUSION

Protein is considered one of the most critical components of the human body. Within our body, a multitude of proteins exists, each assigned with distinct tasks. Specific types of proteins are designated for specific functions, thereby contributing significantly to bodily processes. Alterations in their structures can potentially impact their functionalities, subsequently posing significant risks to human health. The comprehension of protein functions facilitates the identification of anomalies within the human body or other biological systems, attributing them either to protein modifications or alternative factors. However, without knowledge regarding protein functions, the detection and understanding of such matters remain unattainable.

When protein functions are understood, there exists a high probability that any biological problem will be quickly comprehended,

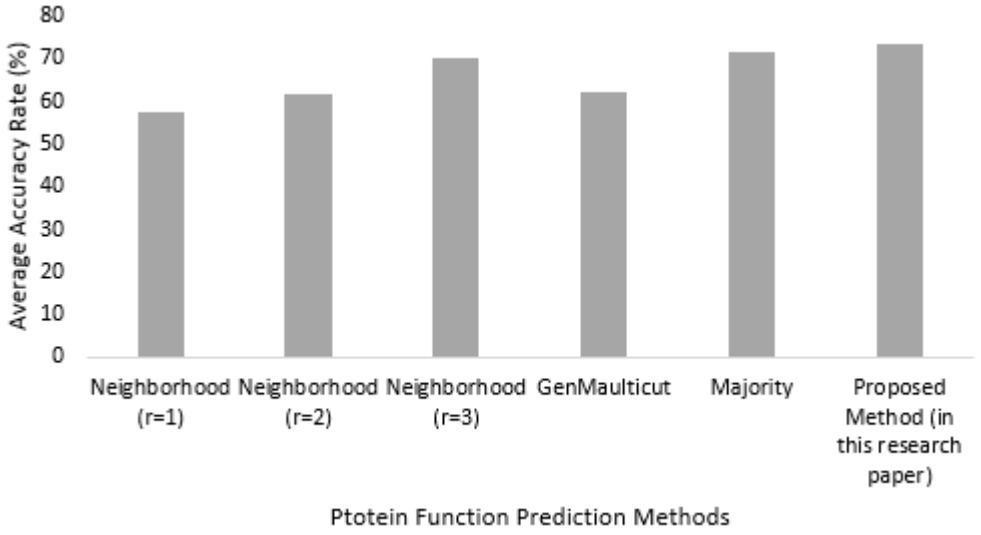

Fig. 3. Accuracy of Protein Function Prediction Rate (%) in Saccharomyces cerevisiae by the Proposed Method

and solutions will be feasible. Consequently, numerous new approaches can be developed for any biological process. The prediction of protein functions based on experiments necessitates extensive experimental and human resources for the analysis of a single protein. Therefore, efforts have been made to devise a new approach for predicting protein functions utilizing computational techniques. Biological research is facilitated with the aid of bioinformatics, as it contributes to time and cost reduction while diminishing human labor. Various methods are available for predicting protein functions, one of which involves predicting protein function through protein-protein interaction networks. This approach is chosen due to the tendency of a protein to exhibit similar characteristics to its neighboring proteins.

The two specific parts of our approach primarily involve protein clustering and protein function prediction. A fast, effective, and adaptable algorithm has been chosen to ensure the optimal suitability of protein clusters. Subsequently, the ”Majority” rule has been applied. Under this rule, identical characteristics are assigned to a protein based on those of its direct neighbor proteins. It is often observed that proteins within a certain area exhibit highly similar characteristics, thus justifying the application of the majority rule in our methodology to achieve accurate function predictions. The performance of our approach is notable, evidenced by the attainment of a good percentage of accuracy.

9. REFERENCES

[1] Pearson, W. R. (1996). [15] Effective protein sequence comparison. In Methods in enzymology (Vol. 266, pp. 227-258). Academic Press.

[2] Stephen, F. A. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

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

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

[5] Deng, M., Sun, F., Chen, T. (2002). Assessment of the reliability of protein-protein interactions and protein function prediction. In Biocomputing 2003 (pp. 140-151).

[6] Rives, A. W., Galitski, T. (2003). Modular organization of cellular networks. Proceedings of the national Academy of sciences, 100(3), 1128-1133.

[7] Vazquez, A. (2011). Protein interaction networks.

[8] 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.

[9] Chua, H. N., Sung, W. K., Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics, 22(13), 1623-1630.

[10] Pandey, G., Kumar, V., Steinbach, M. (2006). Computational approaches for protein function prediction: A survey.

[11] Ahmad, A., Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. Data Knowledge Engineering, 63(2), 503-527.

[12] Chen, S. H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., ... Hsu, F. C. (2008). A support vector machine approach for detecting gene-gene interaction. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 32(2), 152-167.

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

[14] 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.

[15] Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. PloS one, 8(10), e76339.

[16] 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.

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

[18] 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.

[19] Yu, F., Chen, M. H., Kuo, L., Talbott, H., Davis, J. S. (2015). Confident difference criterion: a new Bayesian differentially expressed gene selection algorithm with applications. BMC bioinformatics, 16, 1-15.

[20] Smyth, P. (1996). Clustering sequences with hidden Markov models. Advances in neural information processing systems, 9.

[21] Marcotte, E. M., Xenarios, I., Eisenberg, D. (2001). Mining literature for protein–protein interactions. Bioinformatics, 17(4), 359-363.

[22] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., ... Legrain, P. (2001). The protein–protein interaction map of Helicobacter pylori. Nature, 409(6817), 211-215.

[23] Chen, Y., Xu, D. (2003). Computational analyses of high-throughput protein-protein interaction data. Current protein and peptide science, 4(3), 159-180.

[24] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

[25] Pearson, W. R. (1996). [15] Effective protein sequence comparison. In Methods in enzymology (Vol. 266, pp. 227-258). Academic Press.

[26] Pandey, G., Kumar, V., Steinbach, M. (2006). Computational approaches for protein function prediction: A survey.

[27] Bogdanov, M., Heacock, P., Guan, Z., Dowhan, W. (2010). Plasticity of lipid-protein interactions in the function and topogenesis of the membrane protein lactose permease from Escherichia coli. Proceedings of the National Academy of Sciences, 107(34), 15057-15062.

[28] Cao, M., Zhang, H., Park, J., Daniels, N. M., Crovella, M. E., Cowen, L. J., Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. PloS one, 8(10), e76339.

[29] Zhao, X. M., Chen, L., Aihara, K. (2008). Protein function prediction with high-throughput data. Amino Acids, 35(3), 517-530.

[30] 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.

[31] Horvatovich, P., Lundberg, E. K., Chen, Y. J., Sung, T. Y., He, F., Nice, E. C., ... Hancock, W. S. (2015). Quest for missing proteins: update 2015 on chromosome-centric human proteome project. Journal of proteome research, 14(9), 3415-3431.

[32] Ning, Q., Ma, Z., Zhao, X., Yin, M. (2020). A novel succinylation sites prediction method incorporating K-means clustering with a new semi-supervised learning algorithm. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(1), 643-652.

[33] www.uniport.org . [Accessed May 2024]

[34] https://string-db.org/ . [Accessed May 2024]

[35] Jiang, J. Q., Wu, M. (2012, June). Predicting multiplex subcellular localization of proteins using protein-protein interaction network: a comparative study. In BMC bioinformatics (Vol. 13, pp. 1-15). BioMed Central.

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

[37] Bonetta, R., Valentino, G. (2020). Machine learning techniques for protein function prediction. Proteins: Structure, Function, and Bioinformatics, 88(3), 397-413.

[38] https://europepmc.org/article/MED/23476125

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