ISSN 2816-8089

RESEARCH ARTICLE

ProCbA: Protein Function Prediction based on Clique Analysis

Mohammad Hossein Olyaee[1], Soudeh Behrouzinia[2], Mohammad Bagher Ghajehlo[2], Alireza Khanteymoori[3*]

[1]*Department of Computer Engineering, University of Gonabad, Gonabad, Iran.*
[2]*Department of Computer Engineering, University of Zanjan, Zanjan, Iran.*
[3]*Neurozentrum, Universitätsklinikum Freiburg, Freiburg, Germany.*

Abstract

Protein function prediction based on protein-protein interactions (PPI) is one of the most important challenges of the post-Genomic era. Due to the fact that determining protein function by experimental techniques can be costly, function prediction has become an important challenge for computational biology and bioinformatics. Some researchers utilize graph- (or network-) based methods using PPI networks for unannotated proteins. The aim of this study is to increase the accuracy of the protein function prediction using two proposed methods. To predict protein functions, we propose a Protein Function Prediction based on Clique Analysis (ProCbA) and Protein Function Prediction on Neighborhood Counting using functional aggregation (ProNC-FA). Both ProCbA and ProNC-FA can predict the functions of unknown proteins. In addition, in ProNC-FA which does not include a new algorithm; we attempt to solve the essence of incomplete and noisy data of the PPI era in order to achieve a network with complete functional aggregation. The experimental results on MIPS data and the 17 different explained datasets validate the encouraging performance and the strength of both ProCbA and ProNC-FA on function prediction. Experimental result analysis demonstrates that both ProCbA and ProNC-FA are generally able to outperform all the other methods.

Key Words: *Proteomics; Complex network; Protein function prediction; PPI networks; Clique analysis; Functional integration*

***Corresponding Author**: Alireza Khanteymoori, Neurozentrum, Universitätsklinikum Freiburg, Freiburg, Germany; E-mail: khanteymoori@gmail.com*

***Received Date**: January 13, 2023, **Accepted Date**: February 21, 2023, **Published Date**: February 27, 2023*

***Citation**: Olyaee MH, Behrouzinia S, Ghajehlo MB, et al. ProCbA: Protein Function Prediction based on Clique Analysis. Int J Bioinfor Intell Comput. 2023;2(1):99-125.*

1. Introduction

Proteins are large, complex, essential, and the most important molecules of life. They are the main constituents in all living organisms and associate the second cell weight after water to themselves. Proteins are also responsible for some of the most important functions in an organism. Defending the body from antigens, being involved in muscle contraction and movement, facilitating biochemical reactions, and helping to coordinate certain body activities are some of the protein functionalities.

In this regard, Protein Function Prediction is one of the most important fields of study in system biology. Also, it is one of the major challenges in the Post-Genomic era. There are different methods for achieving good function prediction. These methods use distinct approaches such as sequence and structural similarity and also gene expression profiles.

Though some percent of proteins can be expected to work in relative isolation, it is not valuable to study a protein in isolation [1]. Protein interactions play key roles in their functionality and perform a specific function. The interaction of proteins can have a structure such as a network. This structure is called Protein-Protein Interaction Network (PPI). In this network, nodes consider proteins, and edges represent the interactions between proteins when two proteins interact with each other. The position of each protein in the interaction network plays an important role in understanding cell activities. According to this fact, the graph-based methods attempt to determine the function of unknown proteins by discovering their interaction with a known protein target having a known function. Therefore, it is critical to develop graph-based methods to predict protein functionality [2-12].

The recent availability of protein data and important interaction between proteins (exploiting the protein similarity) led to the development of various methods based on interaction networks for predicting protein functions. Several progress reports have been published in this area that take advantage of network-based methods as well as machine learning to capture the interaction between proteins and employ them to predict protein functions. In recent years, employing deep learning-based methods to predict protein function prediction is providing higher performance [13-21].

NG *et al.* [12] considered the so-called protein function pair approach, which is carried forward from the protein domain pair approaches. Their approach is based on Kim *et al.* [22,23] by incorporating a randomization procedure in order to assign function–function correlation score for a protein function pair, which could facilitate protein function prediction [21]. In [24], Zhu *et al.* proposed a Semantic and Layered Protein Function Prediction (SLPFP) framework. SLPFP is an unknown protein functions predictor and a new clustering-based function prediction algorithm at different functional layers within the Function Catalogue (FunCat) Scheme and also from different clusters rather than from just one.

To address the issues of protein similarity measurement and prediction domain selection, Zhu *et al.* proposed an innovative approach to predict functions of unknown proteins iteratively from a PPI dataset. The iterative approach of [25] considers the semantic similarity of protein interactions is based on the multi-layered information carried by protein functions as dynamic features of protein structure [25]. Hou *et al.* [26] take into account aggregating the functional correlations among relevant proteins to predict protein functions from PPI data.

This functional aggregation considers the positive impact of each relevant and negative repeated protein function on the final prediction results [26]. Differently, Zhu *et al.* [27] used a functional connectivity feature to represent the strength of a protein's impact on its neighbor's functions. The functional connectivity approach of [27] is a PPI network-based method.

According to the above approaches, we try to propose an effective approach to predict protein functions. To this end, we focus on protein function prediction and in this regard develop a method, called Protein Function Prediction based on Clique Analysis (ProCbA). ProCbA composes a set of interacting proteins that each protein is described by its structure and purpose and expressed in its functioning. The main step of our proposed method is the Data Pre-processing that removes the additional protein-protein interactions. This step processes the downloaded data from MIPS before the proposed method can use them. After the pre-processing stage, the proposed method applies to the data. In this method, PPI network partitions to the different sizes of a clique will be used to extract specification and the average number of adjacency proteins. In the following, these cliques are analysed in order to function the prediction of unknown proteins.

Addition of ProCbA, we propose Protein Function Prediction on Neighborhood Counting using functional aggregation (ProNC-FA). ProNC-FA is based on Neighborhood Counting with no additional contribution. In order to achieve complete protein-protein interactions data, this method only uses an integration of different scientific literature and databases for functional aggregation. The evaluation results of ProCbA and ProNC-FA demonstrate the effectiveness of the proposed method and the capability of the method in providing better prediction results compared to the well-known methods.

The remainder of this paper is organized as follows. Section 2 describes some related research in different aspects of our work. This section includes protein function prediction and details of the proposed method. Section 3 presents the obtained results. Finally, Section 4 concludes this paper and provides some direction for improving this method.

2. Materials and Methods

2.1. Protein function prediction based on clique analysis (ProCbA)

In order to protein function prediction, this study develops ProCbA method which utilized the graph theory concept. The structure and main steps of ProCbA are illustrated in Figure 1. In this figure, each part denotes the key functionalities of the proposed method, and each arrow corresponds to relations between the parts. Preprocessing is an important step in the ProCbA that processes its input data to produce appropriate output. The next part is Clique Extraction which identifies and extracts all the maximal cliques of the input network and then feeds them as input to the part named Clique Evaluation. Finally, necessary processing and evaluation of the extracted clique in the protein network are performed.

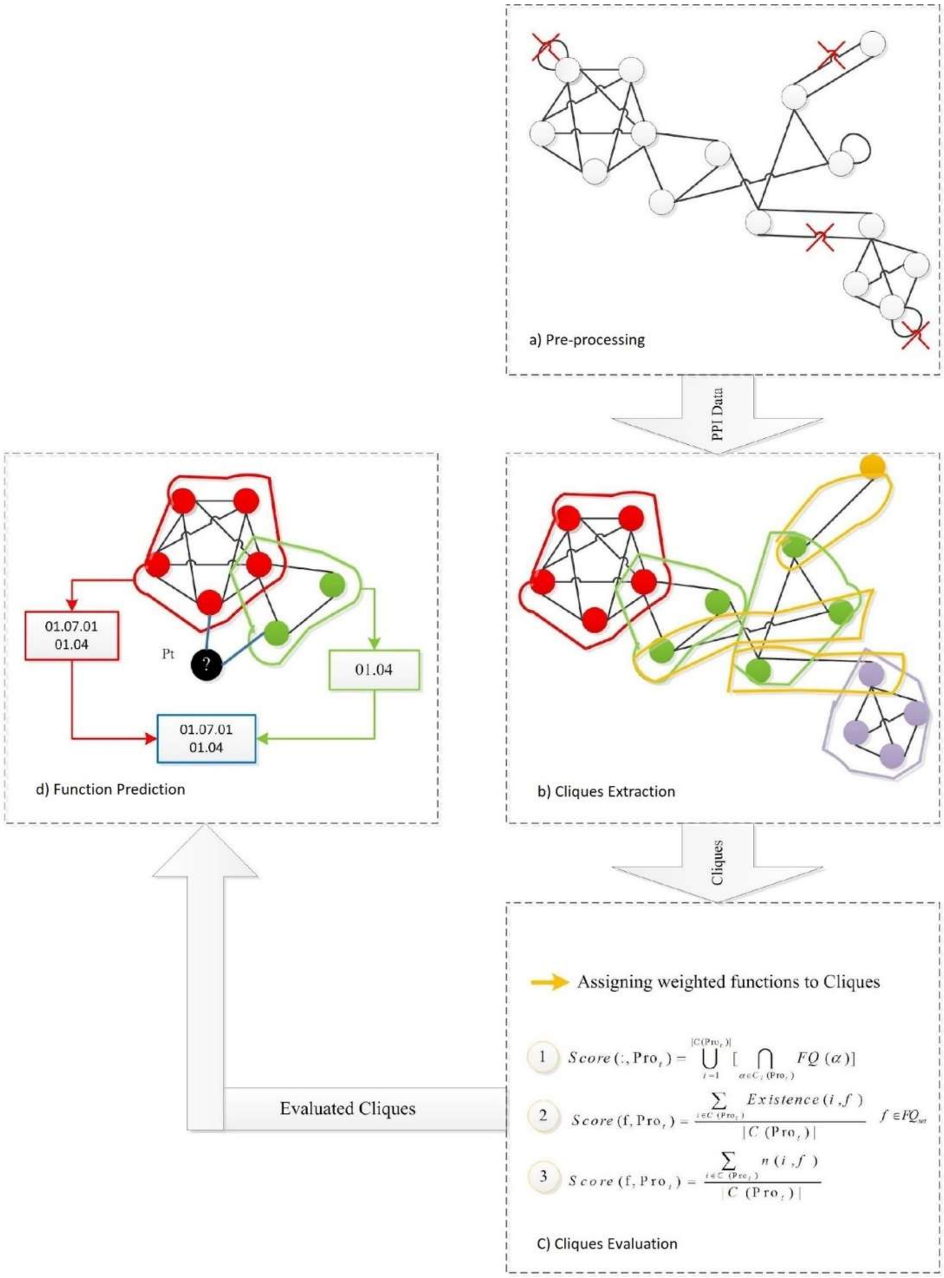


Figure 1: *Overview of the main steps of ProCbA method. a) Preparing the graph by removing additional edges, b) Extracting maximal cliques, c) Evaluating of the obtained cliques, and d) Predicting the functions of unknown proteins.*

The last and most important phase of ProCbA is Function Prediction. This part uses two steps to choose the correct protein function. In the first step of Function Prediction, the number of possible functions for each protein is identified. In the second step, this component uses two important parameters as inputs to protein function prediction: the first parameter is the number of the functionality of each protein and the other is functionality frequency. In the following, each section gives more details about what components do during method execution. The ProCbA algorithm is described in Algorithm 1.

Algorithm 1: *ProCbA: Protein Function Prediction with Clique-based Analysis.*

Input: Protein-Protein Interaction Network

Output: Predicted unknown protein functions

1: **Do** Net Pre-processing and remove some interaction of PPI network

2: **Extract** PPI network cliques

3: **Evaluate** Cliques functions using Equation 1 or 2 or 5

4: **for** t = 1 to Protein_test_size //(The number of unknown protein)

  Do Predict,pt. functions

6: **end for**

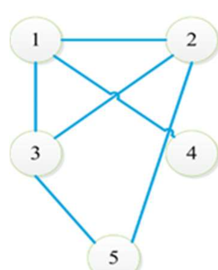
7: **return**

2.1.1. Pre-processing

Main goal of this step is network pre-processing that removes some of the protein-protein interactions under certain conditions. For this aim, it processes the input data that was downloaded from MIPS. In this component, all of the interactions that satisfy one of the below conditions are removed from the protein-protein interaction list. These conditions are generally as follows:

1. Remove duplicate transactions.
2. Remove the protein-protein interactions which proteins are the same in interactions, or in the other words, each protein interacts herself.
3. Remove the protein-protein interactions in which at least one of the proteins has not any functions.

By applying the above rules, the output of this step is a binary and symmetric adjacency matrix that is named PPI-graph or PPI-Network. PPI-graph is a Protein-Protein Interaction graph based on this adjacency matrix that each element of the matrix indicates whether pairs of vertices are adjacent or not in the graph. The PPI network is an undirected graph. Figure 2 shows an example of an adjacency matrix.



$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Figure 2: *The extraction of adjacency matrix from PPI Network.*

2.1.2. Clique extraction

A clique in undirected graph G, is a subset of vertices that are pair-wise adjacent in the graph. Clique is one of the fundamental concepts in graph theory. The maximum clique of graph G, is a clique if and only if it has maximum cardinality among all possible cliques of graph G. Finding the Maximum clique is an NP-hard problem [28,29], and for this reason, several exact algorithms have been developed for solving this problem [30].

Having the same functions between proteins can be concluded protein-protein interactions with high probability. As a result, it is concluded that, the clique is a meaningful concept from a biological perspective [31]. The used algorithm to extract the clique consists of four phases as illustrated in Figure 3. After the initialization is performed, the algorithm iterates phase 3 and 4 until stopping conditions are met.

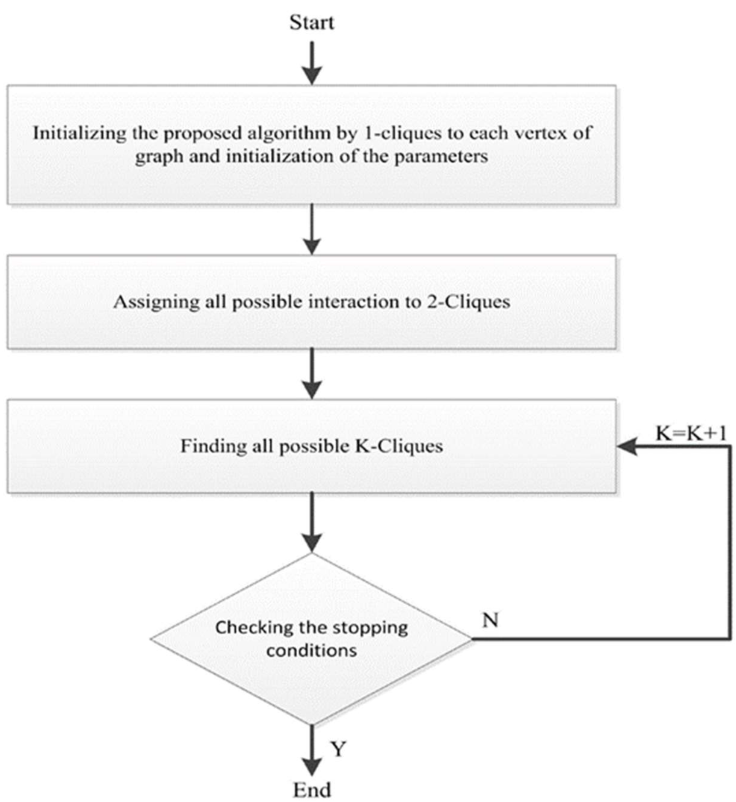


Figure 3: *Flowchart of clique extraction.*

In the initialization phase, since proteins can only interact with one protein, each protein is assigned to 1-clique. In phase two, all possible interaction is assigned to 2-clique. In phase 3, each protein that interacts with proteins in 2-cliques, is searched to find 3-clique. This search continues until the creation of all 3-cliques in the graph. Phase 3 is repeated to produce all possible 4, 5, …, and K-cliques. Finally, the algorithm stops if the stop conditions are met, and the algorithm cannot create any other maximal clique.

2.1.3. Clique evaluation

Clique Evaluation component investigates and collects the set of adjacency cliques of each protein with unknown functions. The collected cliques set for each protein can be repeated for different cliques sets but they keep without removing repeated cliques due to their importance. Although this approach increases the running time of the method to achieve the highest prediction, improving the accuracy is more important compared to the running time. In the following, the goal is a calculation of function frequency. Due to this goal, all identified clique sets are used to assign appropriate scores to all possible clique functions using one of the following three strategies *S1*, *S2*, and *S3*:

S1: Analysis of common functionalities to be considered,

$$Score(:, Pro_t) = \boxtimes_{i=1}^{|C(Pro_t)|} [\boxtimes_{\alpha \in C_i(Pro_t)} FQ(\alpha)] \quad [1]$$

where $C(Pro_t)$ is set of all adjacency cliques to test protein.

S2: In the second strategy, one of the local scoring schemes is applied to create scored functions. This schema considers the presence probability of all functions in each clique. Continue union of all high-scored functions of cliques assigned to test protein. One reason for using this strategy is to overcome the limitation of the first's strategy in noisy interaction. For example, when in a specific clique, all proteins have similar functions except one protein (that has different functions), the first strategy removes different functions. Removing alone functionality can be a mistaken strategy in some cases. The scoring function of this strategy is defined as:

$$Score(f, Pro_t) = \frac{\sum_{i \in C(Pro_t)} Existence(i, f)}{|C(Pro_t)|} \qquad f \in FQ_{set} \quad [2]$$

where Pro_t is unknown protein, $C(, Pro_t)$ is a set of all cliques that are connected to Pro_t where FQ is a set of possible functions of all cliques which as follows:

$$FQ_{set} = \boxtimes_{i \in C\,(Pro_t)} S_i \quad [3]$$

And S_i is the union of functions of, i^{th} clique as follows:

$$S_i = \boxtimes_{\alpha \in C_i(Pro_t)} FQ_{set}(\alpha) \quad [4]$$

S3: The final strategy is somewhat similar to the second strategy. The third strategy counts the number of evidence for each possible function in each clique. Continue union of all high-scored functions of cliques assigned to test protein. The scoring function of this strategy define as:

$$Score(f, Pro_t) = \frac{\sum_{i \in C(Pro_t)} n(i, f)}{|C(Pro_t)|} \quad [5]$$

Different three explained strategies are shown in Figure 4.

2.1.4. Function prediction

Assigning relative functions from candidate functions is one of the most important problems in proteomic study and protein function prediction. There is not any knowledge of the number of protein functions in Biological Theory. According to this fact, several methods have been developed to select the number of functions and assign functionalities to an unknown protein.

Selection of fixed numbers for all functions [32] and average numbers of functions of all network proteins are different strategies that have been used to determine number of functionalities.

To measure the number of protein functions, ProCbA uses Equation 6 which is defined by,

$$NumberofFunction_{Pro_t} = \frac{\sum_{\alpha \in N(Pro_t)} |FQ_{set}(\alpha)|}{|N(Pro_t)|} \quad [6]$$

which $N(Pro_t)$ denotes the number of test protein neighbors. Forasmuch as Equation 6 uses average numbers of functions of all neighbor proteins, it can be the best strategy for determining the number of functions.

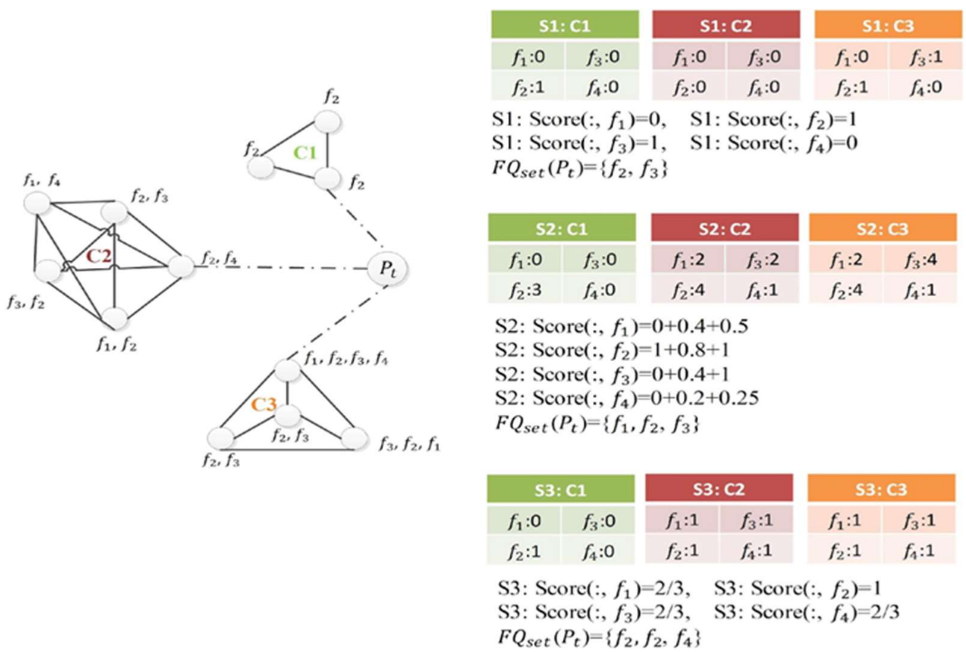


Figure 4: *The example of demonstrating the effect of the three proposed strategies to clique evaluation.*

2.2. Protein function prediction on neighborhood counting using functional aggregation (ProNC-FA)

Due to the tentative identification of data in the proteomic study, some of the major important limitations of protein-protein interaction networks are incomplete and noisy data. To address this issue, this study introduces ProNC-FA, which takes advantage of integrated data as the functional aggregation. This functional aggregation investigates the impact of integrated data on the performance and accuracy of protein function prediction and the impact of integrated data in performance reduction of the previous approach. The functional aggregation feature of ProNC-FA reduces the impact of repeated functional information on the prediction.

Figure 5 shows an overview of the ProNC-FA method. As illustrated in Figure 5, ProNC-FA uses a functional aggregation component based on the Neighborhood Counting algorithm that is proposed by Schwikowski [32]. The goal of this component is aggregation and combination of all interactions in the Protein-Protein Network. The improvement of speed and performance is the best result of ProNC-FA.

The main part of ProNC-FA is the functional aggregation that uses the output of the pre-processing step. ProNC-FA uses 17 different PPI datasets as illustrated in Figure 5.

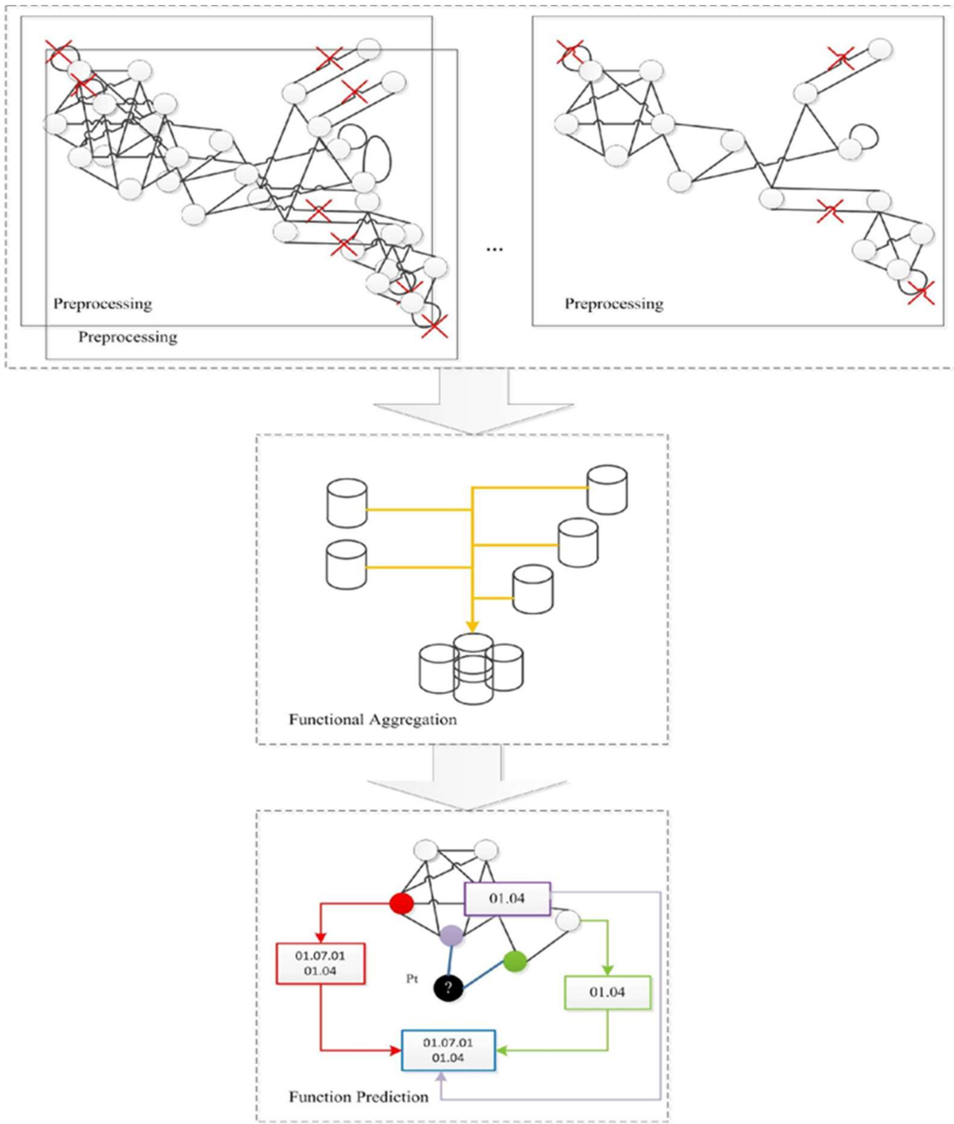


Figure 5: *The illustration of the function prediction process of ProNC-FA. The output of pre-processing phase is fed to the function aggregation step. Finally, the integrated data is used to extraction the maximal cliques.*

First, each dataset is processed to produce appropriate output by pre-processing components. Second, all preprocessed data integrates as an output of functional aggregation. Finally, this functional aggregation data is sent as input to Clique Extraction. The evaluation result of ProNC-FA demonstrates the effectiveness of the proposed method and the capability of the method in providing better prediction results compared with existing methods.

3. Result and Discussion

3.1. Datasets

Generally, protein function prediction methods use protein-protein interaction and functional annotation scheme. We evaluate the ProCbA by testing their performance on the tasks of predicting protein functions on MIPS [42-43, 51-52] and FunCat3 [53].

MIPS is a database for genomes and protein sequences that is provided genome-related information by the Munich Information Center in Germany. Further information on the MIPS Dataset is shown in Table 1.

Table 1: *MIPS data specifications.*

Dataset	#Protein		#Interaction	
	Before Preprocessing	After Preprocessing	Before Preprocessing	After Preprocessing
MIPS	4554	3483	15456	10204

FunCat3 [53] is a functional annotation scheme which has wide coverage and standard hierarchical structure. The 28 existing functions in FunCat3 are organized in a hierarchical tree structure. In this study, the most informative functions of FunCat are used for testing.

In addition to what has been said, we use 17 different datasets to achieve the complete network of Protein-Protein Interaction. The details of this user data within ProNC-FA are presented in Table 2.

3.2. Evaluation metrics

This section presents evaluation metrics to test the effectiveness and estimate the expected accuracy of ProCbA to protein function prediction. Various evaluation metrics have been developed for evaluating the effects of different prediction methods. To evaluate the proposed method, we use three evaluation metrics, namely, Precision, Recall, and F-Measure.

- **Precision**: In the field of function prediction, precision is the fraction of predicted functions that are relevant to the protein. Precision criteria is defined as,

$$Precision = \frac{N_c}{N_p} \quad [7]$$

where N_p is the number of predicted functions of protein and N_c is the number of correct predicted functions.

- **Recall**: Recall in the function prediction field is the fraction of the functions that are relevant to the protein that is successfully predicted. Recall criteria is defined as,

$$Recall = \frac{N_c}{N_r} \quad [8]$$

where, N_r is the number of identified functions of protein and N_c is the number of predicted functions.

- **F-Measure**: The traditional F-measure or balanced F-score is a measure that combines precision and recall as the weighted harmonic mean of precision and recall. The balanced F-Measure criteria is defined as:

$$F - Measure = \frac{(2 * balanced\ F - score * Recall)}{balanced\ F - score + Recall} \quad [9]$$

All the experiments reported in this section were performed on a system with an Intel Core i7-2410M 2.3 GHz processor with 6 GB RAM.

Table 2: *The specification of 17 used different datasets in ProNC-FA.*

Dataset	Before Preprocessing		#Protein (Without interaction)	#Interaction (After redundancy removal)	After Preprocessing	
	#Interaction	#Protein			#Interaction	#Protein
Alexei [33,34]	2238	1827	2238	244	1931	1519
Shin [35]	22571	5496	22568	1157	19057	4199
Tong2004 [36]	7941	2262	7175	394	6171	1812
yeastHighQuality [34]	2455	988	2455	32	2408	947
Yeast_data_2007 [34]	17481	4931	17194	974	14686	3873
DIP_MMIPS_iPfam [37]	3201	1681	2857	44	2806	1541
Gavin2006 [38,39]	6531	1430	6531	60	6340	1363
Krogan [40]	7123	2708	7084	317	6291	2316
MINT [41]	48321	5341	24421	1120	20527	4158
MIPS [42, 43]	15456	4554	12319	942	10204	3483
Se2012 [44]	112331	6012	112010	1395	93190	4612
DIP2013 [45]	22995	5004	22493	987	19697	3934
SGD [46]	338246	5999	222230	1289	192766	4710
Utez [47]	1033	1003	1033	165	700	764
Ito2001 [48]	4038	2937	3959	539	2984	2252
STRING v9.1(2013) [49]	1660496	6397	830248	1692	674911	4704
BIOGRID2013 [50]	338904	6234	31201	1523	192836	4711

3.3. ProCbA evaluation

Given the importance of the strategy that is used for Clique Evaluation, more specifically ProCbA prediction based on the three proposed strategies can be concluded different evaluation results by using the evaluation criteria expressed. Table 3 shows the results of ProCbA for various strategies. The obtained results demonstrate that the applied strategy can have an important role in the performance of the approach. As can be seen in the table, the second strategy provides the best results in function prediction in three evaluation criteria. Moreover, it should be noted that for this assessment, both explained datasets are employed.

Table 3: *Dependency of prediction results on different strategies.*

Strategy	ProCbA		
	Precision	**Recall**	**F-Measure**
S1	42.93594	52.73269	43.8063
S2	52.81869	57.5718	52.04874
S3	25.80969	28.87197	25.85506

To test the ProCbA's accuracy, we adopt k-fold cross validation. In this process, all the Protein-Protein interactions are randomly divided into k subsets. Each time of *k*, one subset of *k* subsets is selected as testing data and the rest subsets are used as the training set. In this paper, *k* is 10. In the following, experimental results of ProCbA on *k* cross-validation subsets are shown in Table 4.

Table 4: *Overall precisions, recalls and F-Measure of ProCbA method on the MIPS dataset based on k-cross validation.*

k	Precision	Recall	F-Measure
1	55.38	61.88	55.02
2	56.02	60.54	55.67
3	56.31	55.76	53.03
4	45.69	55.61	47.45
5	51.88	62.57	53.70
6	52.62	55.39	51.38
7	49.47	56.01	48.80
8	51.53	52.79	49.88
9	52.46	51.24	48.75
10	56.80	63.90	56.77
Average	52.81869	57.5718	52.04784

As can be seen, the mean results obtained for the different *k*, is 52% that is the best result with different train data in each 10 steps. The reason is that ProCbA is not sensitive to input data or input interaction networks. In other words, the results show that the proposed algorithm is resistant or robust to network change.

Table 5 gives the comparisons with the other algorithms such as GM [31], GMV, χ^2, FF, and FCML [56]. All reported results in Table 5 are based on the presented results by [56]. According to the obtained results, ProCbA outperforms the GM, GMV, χ^2, FF and FCML on F-Measure and outperform all of the on Precision. The most current method of function prediction focuses on the improvement of one of the F-Measure or Precision criteria, but ProCbA can achieve the minimum difference between these two criteria. This may be because ProCbA uses Equation 6 to measure the functions number of proteins. Equation 6 can present the best estimation of the number of allocable functions to proteins.

Table 5: *The comparison of ProCbA with five well-known algorithms.*

Algorithms	The Mean of Precision	The Mean of F-Measure
GM	30.69	29.04
GMV	31.13	22.41
χ^2	14.80	07.60
FF	28.01	27.05
FCML	54.83	43.74
ProCbA	52.81	52.04

By comparing the results of the assessment, according to what is shown in the table, ProCbA has a better result in both Precision and F-Measure. For example, the reason is that the GM algorithm only considers directed neighbors, but ProCbA does not limit itself to considering directed neighbors. The neighborhood or the protein degree includes the number of protein neighbors. In other words, the neighborhood degree of a protein *p* in *N* as a protein network, is the number of subnets of *N* consisting of all proteins adjacent to *p*. The most important goal of this study is the proof of the relationship between the neighborhood degrees of proteins with the number of proteins functions. Figure 6 shows the performance of ProCbA based on neighborhood degrees of proteins.

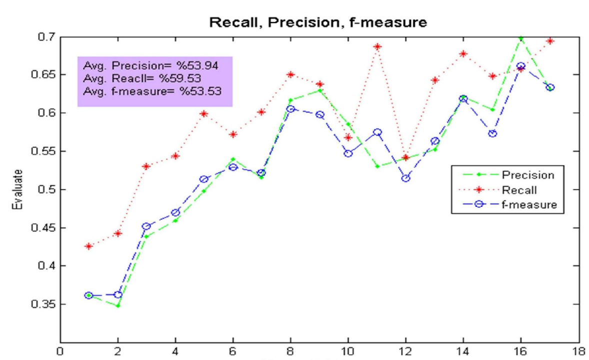


Figure 6: *Performance of ProCbA based on neighborhood degrees of proteins.*

According to increasing the number of protein neighbors, the growth of ProCbA performance in Precision, Recall, and F-Measure will be more significant. In Protein-Protein Networks, with increasing and decreasing the neighborhood degree, the number of proteins gradually decreases and increases respectively. Due to this specification of the network, the accuracy, and performance of ProCbA are not good.

An incomplete functional annotation scheme is another consequence that can be concluded from this evaluation. Due to Figure 6, it is expected that the performance of the algorithm increases with increasing the number of neighborhood degrees. But, as can be seen in the results of the evaluation, the performance of ProCbA has oscillated from the neighborhood degree of 2. In the other words, the Accuracy of ProCbA has not only not improved but also has been reduced.

The interaction of proteins with proteins that have similar specific functions is an important general hypothesis in the protein-protein interaction network. In this regard, two probabilities can be considered. First, the general hypothesis is not correct, and second, data related to the functional annotation scheme is not complete. The second probability is more powerful than the first. It is expected, if the protein's interacting partners identify, not only the accuracy of the proposed algorithm but also the accuracy of most existing algorithms will increase. To prove this theory, the second proposed method named ProNC-FA is presented. According to this theory, ProNC-FA uses functional aggregation for protein function prediction.

3.4. ProNC-FA evaluation

Since using the functional aggregation data is the main novelty of ProNC-FA, in this section the performance of this method using one of the first presented methods in this scope is evaluated. Table 6 demonstrates the results of ProNC-FA with 17 different datasets. Each row of the table demonstrates the result of PPI Network made using the intersection from first row's data till this row's data. Column 2 and 3 in table represent the number of interactions and the number of proteins, respectively and the column 4 and 5 in table represent the min and max neighborhood degree of protein, respectively. For example, there are proteins that have interaction with one protein or 3188 proteins in the first row. The difference between the neighborhood degree of column 4 and 5 concludes two main results. First, the network is still incomplete and second, there are many pseudo interactions in networks which the difference causes an increase in neighborhood degree of some of proteins.

According to the obtained results, a small percentage of the interactions, about 0.0004% is in the different dataset and this suggests that interactions have different degree of reliability. Considering different degree of reliability as weighted protein interaction network can lead to future research.

Furthermore, the proposed method is compared with some popular methods such as PClustering[54] and PRODISTIN[55]. These methods were evaluated by Ashish in 2013 [54]. Table 7 shows the obtained results from the comparison of PClustering, PRODISTIN and ProNC-FA methods on the 27 explained proteins in [54]. In general, as can be seen in the results, the ProNC-FA shows an appropriate and high accuracy compared with other methods. For example, ProNC-FA can predict all functionality of protein 19 and 27. In the most rows of Table 7, the prediction accuracy of ProNC-FA is equal to or greater than other

ISSN 2816-8089

algorithms. Using the simplest approach without any math complexity to predict protein function is one of most important features of ProNC-FA.

Table 6: *The integration results of ProNC-FA on the 17 different datasets that are used to functional aggregation.*

∩Data	#Interaction	#Protein	Min Neighborhood degree	Max Neighborhood degree
=<1	740145	4722	1	3188
=<2	212375	4719	1	2816
=<3	164623	4707	1	2112
=<4	61486	4522	1	1584
=<5	31820	4358	1	214
=<6	21921	4161	1	203
=<7	14939	3874	1	184
=<8	10489	3573	1	123
=<9	5660	2485	1	85
=<10	3375	1975	1	45
=<11	1855	1427	1	21
=<12	1027	1007	1	20
=<13	481	584	1	14
=<14	203	294	1	7
=<15	76	114	1	6
=<16	19	30	1	3
=<17	3	6	1	1

Since using the neighborhood counting method is the function prediction component of ProNC-FA, in this section, the low performance of neighborhood counting using two important proteins is evaluated and compared with ProNC-FA. Table 8 demonstrates the results of the neighborhood counting algorithm and ProNC-FA with 2 different proteins. According to the obtained results, ProNC-FA has more complete data than the neighborhood counting algorithm and can predict YBL072c protein functions with high accuracy. Although initial information for YAL012w protein in FunCat is complete, ProNC-FA does not have acceptable result in function prediction.

Figure 7 gives the comparisons results of different algorithms such as ProCbA, ProNC-FA(1) and ProNC-FA(2). According to the obtained results, both ProNC_FA(1) and ProNC–FA(2) outperform ProCbA with high accuracy in protein interaction prediction. It should be noted that ProNC-FA has complete information of neighbors of proteins.

Table 7: *Evaluation results of ProNC-FA compared to PClustering and PRODISTIN. (ProNC-FA(1) is related to the original Neighbourhood Counting algorithm with functional aggregation, ProNC-FA(2) is related to using Neighbourhood Counting algorithm based on Equation 6 to measure the function number of proteins.*

ID	**Protein**	**FunCat**	**Description**	**PClustering [54]**	**PRODISTIN [55]**	**ProNC-FA (1)**	**ProNC-FA (2)**
1	YBR103w	10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05
		10.03.02	Meiosis	x	x	x	x
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		14.07.04	Modification by acetylation, deacetylation	x	x	x	x
		32.01	Stress response	x	x	x	x
2	YDR092w	10.01.05.01	DNA repair	10.01.05.01	x	10.01.05.01	10.01.05.01
		14.07.05	Modification by ubiquitination, deubiquitination	x	x		14.07.05
		14.1	Assembly of protein complexes	x	x	x	14.01
		14.13.01.01	Proteasomal degradation	x	x		14.13.01.01
		16.01	Protein binding	16.01	x	x	x
3	YDL042c	10.01.03	DNA synthesis and replication	x	x		
		10.01.05.01	DNA repair	x	x		
		10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		14.07.04	Modification by acetylation, deacetylation	x	x		
		16.01	Protein binding	x	x		
		40.2	Cell aging	x	x		

		34.11.03.07	Pheromone response, mating-type determination, sex-specific proteins	x	x		
4	YEL056w	10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.06
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		14.07.04	Modification by acetylation, deacetylation	x	x		
5	YGL133w	10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
6	YLR086w	1.04	Phosphate metabolism	x	x	x	1.04
		10.03.01.01.11	Mitosis M phase	x	10.03.01.01.11	x	x
		10.03.04.03	Chromosome condensation	x	10.03.04.03	10.03.04.03	10.03.04.03
		16.19.03	ATP binding	x	x	x	x
		42.10.03	Organization of chromosome structure	x	x	x	x
7	YIL150c	10.01.03.03	Ori recognition and priming complex formation	x	x	10.01.03.03	10.01.03.03
		10.01.03.05	extension/polymerization activity	10.01.03.05	10.01.03.05	10.01.03.05	10.01.03.05
		10.03.01	Mitotic cell cycle and cell cycle control	10.03.01	x	x	x

8	YNL330c	1.04	Phosphate metabolism	x	x		
		10.01.05.03.03	Somatic/mitotic recombination	x	x	x	x
		10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		14.07.04	Modification by acetylation, deacetylation	x	x	14.07.04	14.07.04
		40.2	Cell aging	x	x	x	x
		43.01.03.09	Development of ascobasidio-zygospore	x	x	x	x
		34.11.03.07	Pheromone response, mating-type determination, sex-specific proteins	x	x	x	x
9	YFR031c	1.04	Phosphate metabolism	x	x	1.04	1.04
		10.03.01.01.11	Mitosis M phase	10.03.01.01.11	x	x	x
		10.03.04.03	Chromosome condensation	10.03.04.03	x	10.03.04.03	10.03.04.03
		16.03.01	DNA binding	x	x	16.03.01	16.03.01
		16.19.03	ATP binding	x	x		16.19.03
10	YKL108w	10.01.03.05	Extension/polymerization activity	10.01.03.05	x	x	x
11	YKR010c	10.01.02	DNA topology	10.01.02	x	x	x
12	YAR007c	10.01.02	DNA topology	x	x	x	x
		10.01.03.05	Extension/polymerization activity	x	x	x	x
		10.01.05.01	DNA repair	10.01.05.01	x	x	x

		10.01.05.03	DNA recombination	10.01.05.03	x	x	10.01.05.01
		10.03.01.03	Cell cycle checkpoints	x	x	x	x
		16.03.01	DNA binding	16.03.01	x	16.03.01	x
		32.01.09	DNA damage response	x	x	x	x
		34.11.03.07	Pheromone response, mating-type determination, sex-specific proteins	x	x	x	x
13	YNL107w	10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
14	YJL173c	10.01.02	DNA topology	10.01.02	x	x	x
		10.01.03.05	Extension/polymerization activity	10.01.03.05	x	x	x
		10.01.05.01	DNA repair	10.01.05.01	x	x	x
		10.01.05.03	DNA recombination	10.01.05.03	x	x	X
		16.03.01	DNA binding	16.03.01	x	16.03.01	16.03.01
15	YML069w	10.01.03	DNA synthesis and replication	x	x	x	x
		10.01.05	DNA recombination and DNA repair	x	x	x	x
		10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.03	mRNA synthesis	x	x	x	x
16	YGL037c	1.07	Metabolism of vitamins, cofactors, and prosthetic group	x	x	x	x

		10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	x	x
		40.2	Cell aging	x	x	x	x
		01.03.04	Pyrimidine nucleotide metabolism pyrimidinenucleotide/nucleoside/ nucleobase metabolism	x	x	x	x
17	YMR127c	10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05
		10.03.01	Mitotic cell cycle and cell cycle control	x	x	x	x
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		14.07.04	Modification by acetylation, deacetylation	14.07.04	x	14.07.04	14.07.04
		34.11.03.07	Pheromone response, mating-type determination, sex-specific proteins	x	x	x	x
18	YPL001w	10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05
		14.07.04	Modification by acetylation, deacetylation	x	x	x	x
		42.10.03	Organization of chromosome structure	42.10.03	x	x	x
19	YAR003w	10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05

		11.02.03.04.01	Transcription activation	x	x	x	11.02.03.04.01
		14.07.09	Posttranslational modification of amino acids	x	x	14.07.09	14.07.09
		42.10.03	Organization of chromosome structure	x	x	42.10.03	42.10.03
20	YPR052c	10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.02	tRNAsynthesis	x	x	x	x
		11.02.03.04	Transcriptional control	11.02.03.04	11.02.03.04	11.02.03.04	11.02.03.04
		40.01	Cell growth/morphogenesis	x	x	x	x
		43.01.03.05	Budding, cell polarity and filament formation	x	x	x	x
21	YGL090w	10.01.05.01	DNA repair	10.01.05.01	10.01.05.01	10.01.05.01	10.01.05.01
		10.03.02	Meiosis	x	x	x	x
		16.07	Structural protein structural protein binding	16.07	x	x	x
		43.01.03.09	Development of asco- basidio- or zygospore	x	x	x	x
22	YNL031c	10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	11.02.03.04	11.02.03.04	11.02.03.04
		16.03.01	DNA binding	x	16.03.01	16.03.01	16.03.01
23	YNL312w	10.01.02	DNA topology	10.01.02	x	x	x
		10.01.03.05	Extension/polymerization activity	10.01.03.05	x	x	x

		10.01.05.01	DNA repair	10.01.05.01	x	10.01.05.01	10.01.05.01
		10.01.05.03	DNA recombination	10.01.05.03	x	x	x
		16.03.01	DNA binding	16.03.01	x	16.03.01	16.03.01
24	YML127w	10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05
		11.02.01	rRNAsynthesis	x	x	x	x
		11.02.03.04	Transcriptional control	11.02.03.04	11.02.03.04	11.02.03.04	11.02.03.04
25	YDR191w	1.06	Lipid, fatty acid and isoprenoid metabolism	x	x	1.06	1.06
		10.01.09.05	DNA conformation modification	10.01.09.05	10.01.09.05	10.01.09.05	10.01.09.05
		11.02.03.04	Transcriptional control	11.02.03.04	x	11.02.03.04	11.02.03.04
		16.03.01	DNA binding	x	x	16.03.01	16.03.01
		34.11.03.07	Pheromone response, mating-type determination, sex-specific proteins	x	x	x	x
26	YML062c	10.01.05.03	DNA recombination	10.01.05.03	10.01.05.03	x	10.01.05.03
		11.02.03.01.04	Transcription elongation	x	11.02.03.01.04	11.02.03.04	11.02.03.04
		14.04	Protein targeting, sorting and translocation	x	x	x	x
		16.19	Nucleotide binding	x	16.19	x	x
		20.01.21	Nucleotide/nucleoside/nucleobase binding	20.01.21	20.01.21	20.01.21	20.01.21

27	YJL081c	20.09.04	RNA transport mitochondrial transport	x	x	x	x
		10.01.09.05	DNA conformation modification	10.01.09.05	x	10.01.09.05	10.01.09.05

Table 8: *The benefits and drawbacks of using ProNC-FA(1) to function prediction of two known protein namely, YAL012w and YBL072c.*

Protein	#Neighbors	FunCat	Prediction Based on ProNC-FA(1)	%Frequency
YAL012w	572	01.01.06.05.01.01 01.01.09.03.01	01.04 01.07.01 01.05	8 7 7
YBL072c	606	12.01.01	12.01.01 11.04.01 16.03.03	26 12 11

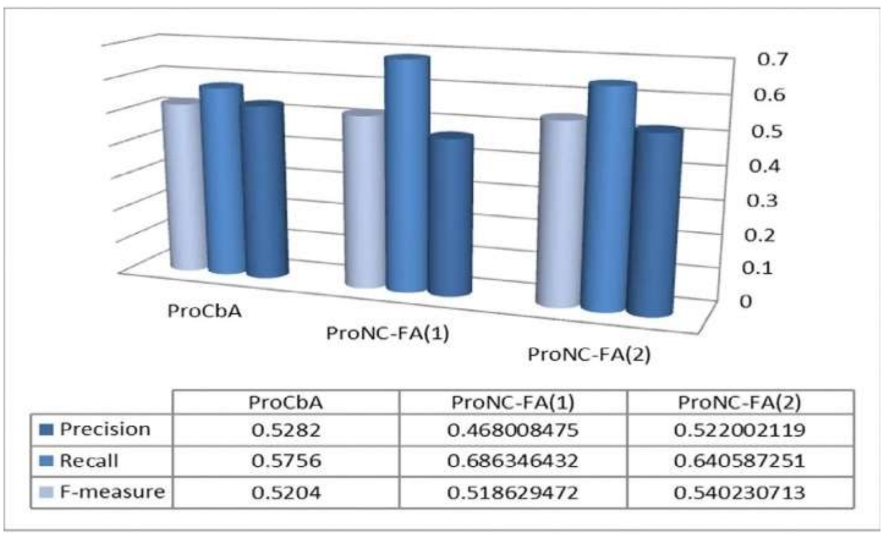


Figure 7: *The evaluation results of proposed methods.*

4. Conclusion

Though numerous studies and many interaction detection methods have presented protein function prediction, extracting useful knowledge can still be important and one of the challenging tasks on function prediction in the bioinformatics field. Clique based analysis is one of the most important methods. ProCbA which is presented in this study is Clique based

Analysis method for protein function prediction. The evaluation results of ProCbA suggest that the cliques found by ProCbA algorithm are consistent with biological knowledge.

One of the important problems in Protein dataset is the presence of noise in data. Thus, affects reduction of noise in a dataset may conclude the best result in the function prediction process. In this paper, a method based on functional aggregation data using Neighbourhood Counting algorithm named as ProNC-FA is proposed to solve this problem. In order to analysing ProNC-FA, 17 datasets as a popular benchmark are used to functional aggregation. Comparison with other algorithms is shown that the proposed method prepares promising results for function prediction.

References

1. Weaver RF. Molecular Biology. (2ndedn), McGraw-Hill, USA. 2002.

2. Li S, Wu S, Wang L, et al. Recent advances in predicting protein-protein interactions with the aid of artificial intelligence algorithms. Curr Opin Struct Biol. 2022;73:102344.

3. Elloumi M, Zomaya AY. Biological knowledge discovery handbook: preprocessing, mining and postprocessing of biological data. John Wiley & Sons, USA. 2013.

4. Shenoy SR, Jayaram B. Proteins: sequence to structure and function - current status. Curr Protein Pept Sci. 2010;11:498-514.

5. Li M, Lu Y, Wang J, et al. A topology potential-based method for identifying essential proteins from PPI networks. IEEE/ACM Trans Comput Biol Bioinform. 2015;12:372-83.

6. Wang J, Zhong J, Chen G, et al. ClusterViz: a cytoscape APP for cluster analysis of biological network. IEEE/ACM Trans Comput Biol Bioinform. 2015;12:815-22.

7. Wang J, Liu B, Li M, et al. Identifying protein complexes from interaction networks based on clique percolation and distance restriction. BMC Genom. 2010;11:S10.

8. Wang J, Chen G, Liu B, et al. Identifying protein complexes from interactome based on essential proteins and local fitness method. IEEE Trans Nanobioscience. 2012;11:324-35.

9. Ding X, Wang W, Peng X, et al. Mining protein complexes from PPI networks using the minimum vertex cut. Tsinghua Sci Technol. 2012;17:674-81.

10. Tang X, Feng Q, Wang J, et al. Clustering based on multiple biological information: approach for predicting protein complexes. IET Syst Biol. 2013;7:223-30.

11. Wang J, Ren J, Li M, et al. Identification of hierarchical and overlapping functional modules in PPI networks. IEEE Trans Nanobioscience. 2012;11:386-93.

12. Ng KL, Ciou JS, Huang CH. Prediction of protein functions based on function-function correlation relations. Comput Biol Med. 2010;40:300-5.

13. Hu S, Zhang Z, Xiong H, et al. A tensor-based bi-random walks model for protein function prediction. BMC Bioinform. 2022;23:199.

14. Zhao B, Zhao Y, Zhang X, et al. An iteration method for identifying yeast essential proteins from heterogeneous network. BMC Bioinform. 2019;20:355.

15. Yu B, Chen C, Zhou H, et al. GTB-PPI: predict protein-protein interactions based on l1-regularized logistic regression and gradient tree boosting. Genom Proteom Bioinform. 2020;18:582-92.

16. Yu B, Chen C, Wang X, et al. Prediction of protein–protein interactions based on elastic net and deep forest. Expert Syst Appl. 2021;176:114876.

17. Wang X, Zhang Y, Yu B, et al. Prediction of protein-protein interaction sites through eXtreme gradient boosting with kernel principal component analysis. Comput Biol Med. 2021;134:104516.

18. Li X, Han P, Wang G, et al. SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction. BMC Genom. 2022;23:474.

19. Yu B, Wang X, Zhang Y, et al. RPI-MDLStack: Predicting RNA–protein interactions through deep learning with stacking strategy and LASSO. Appl Soft Comput. 2022;120:108676.

20. Mahapatra S, Gupta VR, Sahu SS, et al. Deep neural network and extreme gradient boosting based hybrid classifier for improved prediction of protein-protein interaction. IEEE/ACM Trans Comput Biol Bioinform. 2022;19:155-65.

21. Gao H, Chen C, Li S, et al. Prediction of protein-protein interactions based on ensemble residual convolutional neural network. Comput Biol Med. 2023;152:106471.

22. Kim WK, Park J, Suh JK. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. Genome Inform. 2002;13:42-50.

23. Han DS, Kim HS, Jang WH, et al. PreSPI: a domain combination based prediction system for protein-protein interaction. Nucleic Acids Res. 2004;32:6312-20.

24. Zhu W, Hou J, Chen YP. Semantic and layered protein function prediction from PPI networks. J Theor Biol. 2010;267:129-36.

25. Zhu W, Hou J, Chen YP. Exploiting multi-layered information to iteratively predict protein functions. Math Biosci. 2012;236:108-16.

26. Hou J, Chi X. Predicting protein functions from PPI networks using functional aggregation. Math Biosci. 2012;240:63-9.

27. Zhu W, Hou J, Chen YP. Semantically predicting protein functions based on protein functional connectivity. Comput Biol Chem. 2013;44:9-14.

28. Bomze IM, Budinich M, Pardalos PM, et al. The maximum clique problem. In: Du DZ, Pardalos PM (eds), Handbook of Combinatorial Optimization. Springer, Boston, MA. 1999.

29. Engebretsen L, Holmerin J. Clique is hard to approximate within n 1-o(1). In: Montanari U, Rolim JDP, Welzl E (eds), Automata, Languages and Programming. ICALP 2000. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2000.

30. Prosser P. Exact algorithms for maximum clique: a computational study. Algorithms. 2012;5:545-87.

31. Butenko S, Wilhelm WE. Clique-detection models in computational biochemistry and genomics. Eur J Oper Res. 2006;173:1-7.

32. Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. Nat Biotechnol. 2000;18:1257-61.

33. Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol. 2003;21:697-700.

34. Sun S, Zhao Y, Jiao Y, et al. Faster and more accurate global protein function assignment from protein interaction networks using the MFGO algorithm. FEBS Lett. 2006;580:1891-6.

35. Shin CJ, Wong S, Davis MJ, et al. Protein-protein interaction as a predictor of subcellular location. BMC Syst Biol. 2009;3:28.

36. Tong AH, Lesage G, Bader GD, et al. Global mapping of the yeast genetic interaction network. Science. 2004;303:808-13.

37. Yip KY, Gerstein M. Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. Bioinformatics. 2009;25:243-50.

38. Gavin AC, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006;440:631-6.

39. Zhang XF, Dai DQ, Ou-Yang L, et al. Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinform. 2014;15:186.

40. Krogan NJ, Cagney G, Yu H, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006;440:637-43.

41. Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40:D857-61.

42. Güldener U, Münsterkötter M, Kastenmüller G, et al. CYGD: the comprehensive yeast genome database. Nucleic Acids Res. 2005;33:D364-8.

43. Mewes HW, Amid C, Arnold R, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 2004;32:D41-4.

44. http://cbg.garvan.unsw.edu.au/pina/download/Saccharomyces%20cerevisiae-20121210.txt

45. Xenarios I, Salwínski L, Duan XJ, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30:303-5.

46. Cherry JM, Hong EL, Amundsen C, et al. Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40:D700-5.

47. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature. 2000;403:623-7.

48. Ito T, Chiba T, Ozawa R, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci. 2001;98:4569-74.

49. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808-15.

50. Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535-9.

51. Güldener U, Münsterkötter M, Oesterheld M, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. 2006;34:D436-41.

52. Mewes HW, Frishman D, Güldener U, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 2002;30:31-4.

53. Ruepp A, Zollner A, Maier D, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 2004;32:5539-45.

54. Saini A, Hou J. Progressive clustering based method for protein function prediction. Bull Math Biol. 2013;75:331-50.

55. Brun C, Chevenet F, Martin D, et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol. 2003;5:R6.

56. Wang H, Huang H, Ding C. Function-function correlated multi-label protein function prediction over interaction networks. J Comput Biol. 2013;20:322-43.