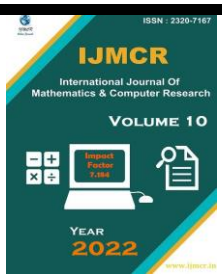
International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 10 Issue 05 May 2022, Page no. – 2685-2689

Index Copernicus ICV: 57.55, Impact Factor: 7.362

DOI: 10.47191/ijmcr/v10i5.04



Clique Based Approach to Predict Complexes from Protein Interaction Network

Sonali[1], Prashanta Kumar Parida[2]

[1,2]C. V. Raman Global University, Bhubaneswar, Odisha, India

ARTICLE INFO

Published Online:
18 May 2022

Corresponding Author:
Sonali

ABSTRACT

The ample availability and importance of large-scale protein-protein interaction (PPI) data demand a flurry of research efforts to understand cells' organization, processes, and functioning by analyzing these data at the network level. In the bioinformatics and data mining fields, network clustering requires a lot of attraction to discover clusters of interacting proteins. Clustering proteins in a PPI network has been an excellent method for discovering functional modules, disclosing functions of unknown proteins, and other tasks in numerous research over the last decade. In this research, a unique graph mining approach is proposed to detect dense neighborhoods (highly connected regions) in an interaction graph, including protein complexes. Our technique first finds size-3 cliques and then expands these size-3 cliques based on their affinity to produce maximal dense regions. To highlight the efficiency of our suggested strategy, we present experimental results using yeast and human protein interaction data. Our predicted complexes match or overlap much better with the gold standard protein complexes in the CYC-2008 and CORUM benchmark databases than other existing approaches.

KEYWORDS: Biological network protein complex neighbourhood expansion.

I. INTRODUCTION

Biological networks are used to describe biological processes and understand their functioning [1, 2]. Identification of genes and proteins linked to diseases, network-based disease classification, protein function annotation, protein superfamily classification, prediction of protein complexes, prediction of new interactions, drug design, and so on are some of the important applications of these networks. Graph data structures represent biological networks in the context of network analysis. PPI networks are usually represented as undirected graphs, with nodes representing proteins and edges representing protein interactions in an organism [3]. In such networks, there is no direction associated with interactions [4]. The features of biological networks can be divided into two categories: global and local [4]. Global properties are used in network modelling and characterization, including small-world property, scale-free network characteristics, power-law degree distributions, and clustering coefficients [5]. Clustering and network motifs are examples of local properties that can be used to represent an extensive, complicated network as a collection of small subgraphs.

Clustering is the process of grouping data in which the data in each group is very similar to each other [6]. Potential protein complexes can be considered as subgraphs in a PPI network with high structural and functional cohesion [7] that can be found by searching high-density regions [8]. Fig. 1 depicts high-density regions in an E. coli PPI network obtained from the DIP database [9].

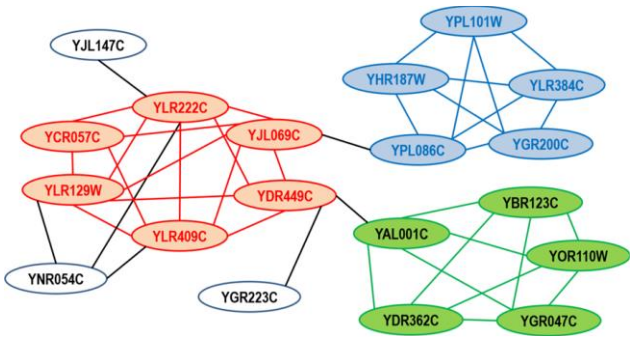


Figure 1: Some of the high-density regions of the PPI network of S. cerevisiae taken from the DIP database [Xenarios et al. (2000)] (marked with red, yellow and blue color)

Density search in the local neighborhood (LD) is an essential category of graph-based algorithms. A local neighborhood density search (LD) aims to optimize cluster density by adjusting a few parameters. Gary Bader and C. Hogue proposed the first graph-based clustering algorithm, MCODE (Molecular COmplex DEtection), in 2003 [10]. This technique examines the dense region of a vast PPI network to detect protein complexes. This approach starts with a seed node and then uses a local neighborhood search strategy to extend it. MCODE solely considers the network topology and ignores the protein organization within the network.

Palla et al. created the CFinder (Complex Finder) algorithm in 2005 to locate and analyses overlapping clusters using the clique percolation idea [11]. Amin et al. suggested DPClus (Density-Periphery-based Clustering) in 2006 to predict protein complexes using the LD method [12]. DPClus is based solely on network topology and ignores the internal network organizations. In 2004, King et al. presented the Restricted Neighborhood Search (RNSC) algorithm to predict protein complexes using a cost-based local search (CL) strategy [13]. Mokhtarul et al. introduced the PROCODE (PROtein COmplex DEtection) algorithm in 2018 [14] that uses the LD method. To anticipate protein complexes, PROCODE considers the dense portions of the PPI network and the inherent organization of proteins within the network.

We offer an effective and efficient algorithm for predicting protein complexes from protein-protein interaction graphs in this study using the basic clique expansion technique. First, we use our suggested polynomial-time technique to locate size-3 cliques (a clique consisting of three vertices) in an interaction network. Our technique builds the fundamental cliques into more dense graphs for protein complex identification. It's worth noting that we use dense graphs rather than cliques to anticipate complexes. Our methodology is less susceptible to incomplete protein interaction data than traditional clique recognition methods since the dense graphs do not need to be fully connected. We tested our method using yeast protein interaction data and human interaction data. We discovered that the F-measures predicted by our technique are much higher than those detected by previous computational methods.

II. PROPOSED METHOD

In this research, we propose to employ a Subgraph Expansion Technique (SET) to locate maximal dense subgraphs in the input PPI network for predicting protein complexes. Our SET approach has two fundamental phases to identify maximal dense subgraphs in the network. The basic cliques of size-3 for all the vertices in the network are computed in the first phase. This is because a maximal dense region covering vertices in the network must contain a basic clique of size-3 by definition. We next expand these basic cliques to build maximal dense graphs in the second stage.

A. Size-3 Cliques Extraction

The first step of our SET algorithm is to find the size-3 cliques in the input graph G. For each edge from graph G, we extract the size-3 clique consisting of that edge and then remove all the edges of this clique from the graph. The pendant edges are also removed after the removal of the discovered clique. The details of our SET algorithm to mine size-3 cliques (LC) are shown in Algorithm-1.

Algorithm-1: SET algorithm phase-1 (Clique extraction): Finding Size-3 cliques from input graph
INPUT: G: Input network
OUTPUT: S3: Size-3 clique set
1: BEGIN
2: Initialize an empty set S3
3: Remove all pendant edges from G
4: Take an edge (u, v) from input graph G
5: Choose a common Neighbor(w) of both u and v
6: Add (u, v, w) to the clique set S3
8: Remove the edges (u, v), (v, w) and (u, w) from graph G
9: Remove all pendant edges corresponding to the pendant vertices u, v, and w
10: Go to step-4 for next untraversed edge
11: Return S3
12: END

B. Subgraph Expansion Technique (SET)

A large protein complex is more likely to be displayed as a maximal dense neighborhood consisting of a size-3 clique as the center of an interaction graph with incomplete interaction data. After the size-3 cliques have been identified, the SET algorithm undertakes an extension phase to detect dense graphs that better match the larger complexes. Our SET algorithm aims to expand the subgraph starting from size-3 cliques with a cluster density greater than 50% of the maximum density feasible. The expansion step is repeated iteratively to update the partially generated dense subgraph. The details clique expansion step of our SET method is described in Algorithm-2.

Algorithm-2: SET algorithm phase-2 (Subgraph Expansion): Expansion of the subgraphs to mine protein complexes in a PPI network
INPUT: G: Input network., S3: Set of size-3 cliques
OUTPUT: C: List of predicted protein complexes in the input PPI network G
1: BEGIN
2: Initialize an empty complex set C
3: Take a size-3 clique from the list S3
4: Set Maximum Cluster Density MCD = 1
5: Initialize Partially form Complex PC = size-3 clique
6: Set Current Cluster Density CCD = 1

7: Set Vertex Count of partially form Complex VC = 3
8: Set Edge Count of partially form Complex EC = 3
9: Go to step-20 if CCD < (1/2) MCD
10: Take an untraversed vertex v neighboring to any one vertex u ∈ PC
11: Set the PC connectivity count of v as Cnt =1
12: Increase connectivity count (Cnt) of v for each internal vertex w ∈ PC connected to v
13: If Cnt < (1/2)VC then go to step-10
14: PC = PC U {v}
15: Increment VC and EC
16: Go to step-10 for next untraversed vertex
17: Update MCD = VC*(VC-1)/2
18: Update CCD = EC
19: Go to step 9
20: Update the complex set C= C U {PC}
21: Go to step-3 for next clique to be expanded
22: END

III. PERFORMANCE EVALUATION

For assessing the complicated prediction tools and algorithms, various PPI network datasets are accessible. Some algorithms may only use a single dataset, whilst others may use a large variety of datasets. However, almost all tools and algorithms make use of at least one PPI network dataset. In this part, we test our proposed SET algorithm on a real-world interaction graph derived from yeast and human protein-protein interaction data. We selected to infer protein complexes using yeast and human interaction data since it is currently the most publicly available organism with the most comprehensive experimental datasets. We will compare our results to those of existing algorithms MCODE [10], RNSC [13], DPClus [12], and CFinder [11], using the same datasets to demonstrate the usefulness of our technique.

A. Dataset for Protein Interaction Graph

The performance of the proposed algorithm is tested on two extensively utilised species: yeast and humans. To build our protein interaction graph for mining complexes, we leverage the DIP and BioGRID datasets [15, 16]. The DIP dataset for yeast has 24574 experimentally determined protein-protein interactions between 5038 yeast proteins, while the BioGRID dataset contains 137675 protein-protein interactions between 18270 human proteins.

B. Protein Complex Gold Standard and Evaluation Metric

The collection of known yeast protein complexes acquired from the CYC2008 Database [17] serves as the benchmark against which we test our method. After filtering the predicted protein complexes from the dataset, we acquired a final set of 408 yeast complexes with 1627 proteins as our gold standard for evaluation. We analyse our proposed approach using the gold standard CORUM dataset [18] for the human data. There are 1843 gold standard protein complexes in this dataset. The efficiency of the proposed SET method is validated using several validation metrics, including recall, precision, F-measure, sensitivity, PPV, and accuracy.

C. Experimental Results

The proposed complex prediction algorithm is evaluated by using four PPI networks of two species. The results shown in Table-1 and Table-2 indicate the relative performance of the proposed algorithm as compared to others. The percentage of significant complexes represents the overall performance of the algorithms as compared to the total number of identified complexes. For example, out of the 316 complexes predicted by the SET in DIP datasets for S. cerevisiae, 296 complexes are found to be significant, which is 93.67%. Similarly, SET predicts 551 protein complexes from the BioGRID dataset for human and out of those, 505 complexes are significant. Table-1 and Table-2 show the matching score and coverage value of the above algorithms in the DIP and BioGRID datasets.

Table 1. Experimental results on DIP dataset. (Performance on S. cerevisiae data)

Algorithm	Number of Significant Complexes	Number of Predicted complexes	Parentage of Significant Complexes
MCODE	89	106	83.96
DPClus	1106	1217	90.88
RNSC	2352	2556	92.02
CFinder	848	942	90.02
SET	296	316	93.67

Table 2. Experimental results on BioGRID dataset (Performance on Homosapiens data)

Algorithm	Number of Significant Complexes	Number of Predicted complexes	Parentage of Significant Complexes
MCODE	224	276	81.16
DPClus	699	781	89.5
RNSC	3080	3415	90.19
CFinder	1988	2249	88.39
SET	505	551	91.65

D. Evaluation measures Recall, Precision and F-measure

Fig. 2 shows the comparison of the above evaluation measures among SET, MCODE, RNSC, DPClus and CFinder. The evaluation metrics like precision and F-measure of the SET are comparable with the state-of-art algorithms. For the DIP dataset, the recall of the CFinder is highest, but SET has the highest recall value for S.

cerevisiae in the DIP dataset. The precision of MCODE is highest for humans in the DIP data. However, SET has the maximum precision value for S. cerevisiae in the DIP data. SET achieves the highest F-measure value for both datasets. For example, the evaluation metrics of the SET for Human interaction networks in the BioGrid data are 0.395, 0.481 and 0.434, respectively. These experimental results signify the superior performance of the SET as compared to others in the interaction network of S. cerevisiae and Homosapiens.

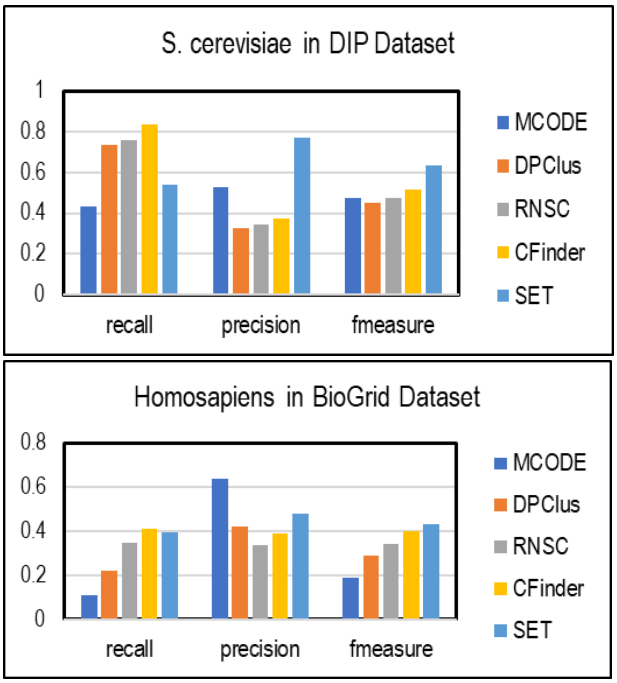


Figure-2: Evaluation metrics Precision, recall, and F-measure for a) S. cerevisiae in DIP dataset, (b) Homosapiens in BioGRID dataset.

E. Evaluation measures PPV, Sensitivity and Accuracy

Fig. 3 shows the comparison of sensitivity, PPV, and accuracy among SET, MCODE, RNSC, DPClus and CFinder for the BioGRID and DIP datasets. For example, the evaluation metrics sensitivity, PPV and accuracy of the SET for S. cerevisiae in the DIP dataset are 0.308, 0.554 and 0.413. The PPV of the proposed algorithm is also highest in humans in the DIP data. The proposed algorithm has the maximum value of sensitivity and accuracy for Homosapiens in the BioGRID dataset, indicating the excellent coverage of the predicted protein complexes. The sensitivity and PPV value of the proposed algorithm is comparable with the state-of-art algorithms. However, the accuracy of the SET is at par with other competing algorithms for both datasets. The higher value of sensitivity and accuracy of the SET indicates its superior performance.

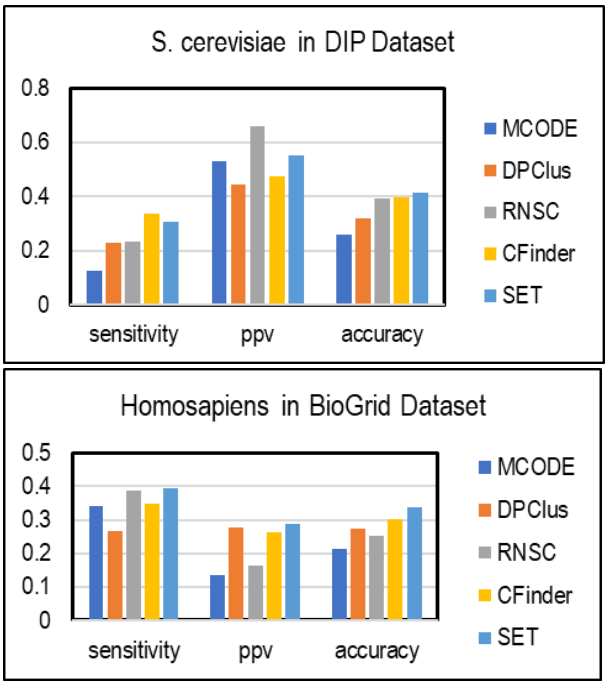


Figure-3: Evaluation metrics Sensitivity, PPV, and accuracy for (a) S. cerevisiae in DIP dataset, (b) Homosapiens in BioGRID dataset.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

In recent years, protein complex prediction methods and algorithms have increasingly considered various biological aspects. As a result, current prediction algorithms are more efficient than prior methods that rely solely on network structural knowledge. Using the subgraph expansion technique, we describe an efficient and successful technique for mining protein complexes from protein interaction graphs. Our proposed algorithm uses a bottom-up approach that considers the size-3 cliques for each edge in the interaction graph and then expands the overlapping local cliques for maximal dense neighbourhoods. Because many fundamental biological processes in the cell are carried out through the formation of protein complexes, identifying protein complexes is critical for biological knowledge discovery. However, there is a significant data gap between protein complexes and technology for identifying pairwise protein-protein interactions. Our projected complexes matched or overlapped well with known protein complexes in the CYC2008 and CORUM benchmark databases; the unmatched complexes may be actual complexes. Thus, our method can be utilised to detect novel protein complexes from the CYC2008 and CORUM benchmark databases, with the unmatched complexes having the potential to be actual complexes.

REFERENCES

1. E.Wong, B. Baur, S. Quader, and C.H. Huang. Biological network motif detection: principles and practice. Briefings in Bioinformatics, 13(2):202-215, 2011.

2. N.T. Tran, S. Mohan, Z. Xu, and C.H. Huang. Current innovations and future challenges of network motif detection. Briefings in Bioinformatics, 16(3):497-525, 2014.
3. Guimin Qin and Lin Gao. An algorithm for network motif discovery in biological networks. International journal of data mining and bioinformatics, 6(1):1-16, 2012.
4. Bjorn H Junker and Falk Schreiber. Analysis of biological networks, volume 2. John Wiley & Sons, 2011.
5. Martin G Grigorov. Global properties of biological networks. Drug discovery today, 10(5):365-372, 2005.
6. Anil K Jain and Richard C Dubes. Algorithms for clustering data. Englewood Cliffs: Prentice Hall, 1988.
7. Liang Yu, Lin Gao, Kui Li, Yi Zhao, and David KY Chiu. A degree-distribution based hierarchical agglomerative clustering algorithm for protein complexes identification. Computational biology and chemistry, 35(5):298-307, 2011.
8. Saeed Jalili, Sayed-Amir Marashi, et al. Camwi: Detecting protein complexes using weighted clustering coefficient and weighted density. Computational biology and chemistry, 58:231-240, 2015.
9. Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins. Nucleic acids research, 28(1):289-291, 2000.
10. Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics, 4(2), 2003.
11. Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. nature, 435(7043):814{818, 2005.
12. Md Altaf-Ul-Amin, Yoko Shinbo, Kenji Mihara, Ken Kurokawa, and Shigehiko Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC bioinformatics, 7(207), 2006.
13. Andrew D King, Natasa Przulj, and Igor Jurisica. Protein complex prediction via cost-based clustering. Bioinformatics, 20(17):3013-3020, 2004.
14. Mokhtarul Haque, Rosy Sarmah, and Dhruba K Bhattacharyya. A common neighbour based technique to detect protein complexes in ppi networks. Journal of Genetic Engineering and Biotechnology, 16(1):227-238, 2018.
15. Salwinski L, Miller CS and Smith AJ, et al., The database of interacting proteins: 2004 update, Nucleic acids research, vol 32, number suppl_1, pp. D449–D451, 2004.
16. Oughtred R, Stark C and Breitkreutz BJ, et al., The BioGRID interaction database: 2019 update, Nucleic acids research, vol 47, number D1, pp. D529–D541, 2019.
17. Pu S,Wong J and Turner B, et al., Up-to-date catalogues of yeast protein complexes, Nucleic acids research, vol 37, number 3, pp. 825–831, 2009.
18. Giurgiu M, Reinhard J, Brauner B, et al., CORUM: the comprehensive resource of mammalian protein complexes—2019, Nucleic acids research, vol 47, number D1, pp. D559–D563, 2019.