Int. Jrl. of **Human Computations & Intelligence**
Vol. 02, Issue 02, 2023
ISSN: 2583-5696

RESEARCH ARTICLE

Approaches for Network Analysis in Protein Interaction Network

S S Patil[1]. Anooja Ali[2]. A Ajil[2]. Meenakshi Sundaram A[2]

[1] Department of Agricultural Statistics University of Agriculture Sciences, Bengaluru, India
[2] School of CSE, REVA University, Bengaluru, Karnataka, India.

Received: 28 March 2023 / Revised: 8 April 2023/ Accepted: 22 April 2023

Abstract – Analysis of protein interaction is important for detailing the cell physiology and predicting disease conditions and drug optimizations. The detection of the crucial proteins in Protein Protein Interaction (PPI) networks is made easier by the accession of these interaction data. The revelation of essential protein nodes in PPI networks is possible using a variety of centrality methods. The hub nodes are decisive in a biological structure because these nodes adjoin profoundly and operate as regulatory hub. The majority of techniques, however, focus on the topological characteristics of PPI. For determining essential proteins, topology and gene annotation are rarely combined. Graph-theoretic methods are used to infer this biological framework in PPI networks. The protein, their interconnections, and the subnetworks are the main subjects of the topological study. In this study, we examine the standard centrality metrics. In order to identify the PPI's prominent nodes and the influence of topological features on centrality metrics, we carefully examined each node's centrality aspect. In this research, we consider Mammalian Protein Database (MIPS) and Biological General Repository for Interaction Networks (BioGRID) datasets and the empirical analysis of individual centrality measures are performed on PPI networks The experimental interpretation shows the behavior of centrality measures on the datasets.

Index Terms – Centrality, Interaction Network, Protein, Topology

I. INTRODUCTION

Proteins are crucial for an organism's cellular and reproductive processes. The area of biological science that focuses on the investigation of proteins is known as proteomics. The systematic examination of protein structure, relationships, and functionality is the main focus of proteomics research [1]. The rapid discovery and study of proteins was made possible by proteomics. Proteins can be divided into categories that are necessary and unnecessary. The conservation of essential categories will withstand biological evolution [2]. The evaluation of essential proteins can be done in two ways. The first group includes gene knockout and ribonucleic acid interfaces, which are conventional methods for predicting important genes [3]. Because this method involves a series of experimental steps, the temporal complexity rises. Analyzing the molecular network's characteristics is included in the second category. This

encouraged the dissemination of functional groups of protein and gene, essential protein, protein complexes, and gene prediction [4]. The availability of such a large volume of omics data expanded the possibilities for determining important proteins. Machine-learning algorithms can effectively recognize the genes responsible for a disease and forecast the necessary proteins. Yet, compared to negative samples or unpredicted genes, the positive samples or training set of genes are quite few. As a result, classification algorithms' predictions become significantly unbalanced. All of the current databases have evidences to support the significant genes that direct to a specific syndrome. Hence, the accuracy of prediction is similarly decreased when a dataset for unknown genes is not available.

Static and dynamic PPI networks are used to apply protein predictions. The dynamic changes cannot be replicated by static interaction networks. The architecture of the PPI network or biological characteristics are the main focuses of methods for the discovery of critical proteins. Any experiment's data could have outliers and unintentional interactions. The popular topological techniques for centrality include Degree Centrality (DC) [5], Betweenness Centrality (BC) [6], Centrality based on Subgraph (SC) [7], and Closeness Centrality (CC) [8]. DC is the most widely used measurement. The number of edges that occur on a node determines how it is appraised. The influential hub nodes are eminent for a network and are crucial. This is evident from the network monitoring of various species, including Escherichia coli, Mus musculus, and Drosophila melanogaster [9][21].

Some centrality measurements focus on the network's nodes, whereas others incorporate both nodes and edges. The centrality calculation for weighted PPI also considers the edge weights. The degree of a node, mainly in and out degree must be assessed using different strategies. Consequently, in this study, we concentrate on the feasibility of several topological methods for calculating centrality on PPI. On the dataset, we implement the widely used distance metrics and assess their effectiveness. The remaining paper is arranged as follows- the next session is literature survey, followed by the different methods for centrality calculation, evaluation of these methods on the datasets and the paper is concluded.

II. LITERATURE SURVEY

The theme of centrality can be applied to all the categories of network. In social media network, it stands for a person's influence over a group's decision-making. Eventually, researchers developed a connection between graph theory and centrality measurements on social network, and later, accomplished community detection [10]. To identify the most effective or competent hub node in the network, this technique is used to biological networks. A protein interaction network is created when proteins interact with one another. A PPI network is constructed by connecting functionally equivalent proteins in a graph. These networks illustrate biological operation. Any undirected graph can be visualized as a PPI. Proteins operate as nodes in PPI, whereas interactions act as edges. G = (V, E) describe a PPI with V vertices and E edges. Each of these edges, represented as (u, v) establish the interactions. Each vertex node is an interacting protein, u Є U, and v Є V where u and v are vertices.

Edges represent a node's capacity for forming connections with other nodes or its proximity to other nodes in the network [11][22]. We take into consideration a PPI network of N nodes and A, an adjacency matrix to make it easier to describe similarity measurements. Adjacency matrix have an

inherent symmetric nature. An edge associating the two nodes has a value of one in the adjacency matrix. 0 will be inserted if there is no connecting edge. We now give examples of common centrality measures that have been used in the literature to identify hub nodes in distinct biological networks.

The relevance of a node in the network is ranked according to centrality. We use centrality measurements to identify the influential node in the PPI network, because it is significantly interlinked with majority of the nodes, and hence, the hub node must be identified [12]. They are dominant in assessing the global network structure. In biological networks, the standard measurements of degree, proximity, betweenness, and subgraph are used. The DC calculates the node's degree. There will be a maximum degree for hub nodes. Even network disruptions can result from the arbitrary removal of a hub node. The shortest path measured or allocated through CC indicates the hub node's successful establishment of communication with the other nodes in the network. The length of the route and CC are inversely correlated [13].

The network components, often called as features can be ranked through EC. This score called prestige score to the node is approximated based on centrality. The same methods of EC are followed in Google's page rank algorithm [14][23]. BC is a geodesic calculation between nodes. BC is often evaluated through mammalian regulatory networks [15]. An alternate for BC is suggested as k betweenness for a shortest path length with less than k . The BC, a quantifier measurement tracks the count of shortest paths that pass through each node. The propensity of a node to form clusters is known as the cluster coefficient [16]. NC calculates how close a given edge is to its corresponding nodes. It also counts the total count of triangles present in a network. Any eventful alterations to a network's edges have the potential to alter the clustering coefficient. A dynamic clustering coefficient will exist in every complex network.

III. EVALUATION METHODS

The different centrality measurements and methods for calculating centrality are described in this section. The centrality metrics being taken into account include DC, BC, NC, EC and CC. The methods concentrate on the nodes, adjacent, edges, subgraph, loops or shortest path.

Degree Centrality

The number of edges linked or associated to a node corresponds to degree of a vertex and it is double for the looped vertices. A graph can be expressed as G = (V, E) with |V| vertices and |E| edges, centrality interpreted as DC is mentioned as in (1)

$$DC(v) = \deg(v) \tag{1}$$

Betweenness Centrality

The shortest path quantifier, BC of a node v, indicated by BC (v) is a median count of shortest paths that pass across node v. Between two nodes, a and node b, the shortest path, ρ (a, b) and ρ (a, v, b) list all possible shortest routes between points a and b that pass-through node v. Betweenness centrality is expressed as mentioned in (2)

$$BC(v) = \sum_a \sum_b \frac{\rho(a,v,b)}{\rho(a,b)} \tag{2}$$

Closeness Centrality

The total measure of the shortest distance between node v and every other node is known as CC. The ability to compare graphs of different sizes is made possible by normalizing closeness centrality. Eq (3), closeness is calculated.

$$CC(v) = \frac{N-1}{\sum_u d(v,u)} \quad (3)$$

Subgraph Centrality

The number of closed loops is counted by SC. It assesses the network subgraphs that node v is a part of. The count of closed pathways with length l is shown in (4), v A closed pathway in number theory is one whose edge could be traversed precisely once, with the path beginning and ending at a single location.

$$SC\ (v) = \sum_{l=0}^{\infty} \frac{\mu(v)}{l!} \quad (4)$$

Eigen Vector Centrality

Impact of nodes on a network is measured by eigen vector centrality (EC) [17]. The network nodes are evaluated, and those with better scores are given priority for use in subsequent calculations. The Eigen vector associated with the maximum of Eigen value is denoted by αmax (v) in (5). The amount of interconnected or densely coupled important proteins or hub proteins with other network nodes will depend on a variety of factors.

$$EC\ (v) = \alpha\, max(v) \quad (5)$$

Edge Clustering Co-efficient

Edge clustering (NC) is defined as the addition of edge features along with node features [18]. NC is a method for edge or node clustering, that appraise centrality by taking into account a node's neighbours' centrality as well. Edges are interactions between nodes, and the edge clustering coefficients are taken into account for each node. The adaptability or flexibility of proteins in a network is the foundation of NC. N (v) in (6) denotes the node v's neighbours. The number of triangles in the network with edges is represented by Z (u, v) (u, v). Nodes u and v have degrees of du and dv, respectively.

$$NC(v) = \sum_{V\ \in\ \mathrm{N(v)}} \frac{Z(u,v)}{\min(dv-1)(du-1)} \quad (6)$$

Table 1. The number of protein nodes and interactions in Saccharomyces cerevisiae dataset

Database	Proteins	Protein Interactions	Essential Proteins
MIPS	4545	12317	1016
BioGRID	5615	52823	1194

It is critical in PPI to identify the major hub node or any other node that have a significant impact on topology. networks. These nodes will interact with the majority of other proteins in a network. The examination of pathways benefits from these core nodes. According to the rules of lethality- centrality

hypothesis, hub nodes participate in most routes and these pathways are conserved. In a scale-free network like Wireless network, the centrality-lethality principle is valid [18]. As a result, we compare the different centrality metrics for biological networks in this research. MIPS and BioGRID dataset details are shown in Table 1. These benchmark datasets are used to assess any global PPI as well as protein and genetic relationships.

IV. RESULTS AND DISCUSSION

We concentrate on identifying Saccharomyces cerevisiae's necessary proteins. This is because Saccharomyces cerevisiae has access to the entire information on the important genes' interactions with proteins. The species is trustworthy for experimental analysis because of this. Saccharomyces cerevisiae is a good example of the centrality-lethality concept. Integration of prospective proteins is the main aim of centrality calculation. MIPS, DEG [19], and SGDP are a few of the datasets used for integration. The MIPS and BioGRID databases were used to find the protein interactions. Table 1 shows the dataset's executive summary. There are 12317 interactions and 4545 proteins in the MIPS database. There are 52823 interactions between the 5615 proteins in the BioGRID database. Proteins are separated into categories that are necessary and optional. 1014 and 1192, respectively, of the essential protein were found in the MIPS and BioGRID datasets. Unknown or non-essential proteins make up the remaining proteins.

Using the MIPS and BioGRID datasets, we assess the NC, DC, EC, CC and BC centrality measures. We chose the best proteins from each dataset and used them for the experimental assessment. The proteins are isolated in 5% intervals. Considered are the top 5%, 10%, and 15% of proteins. In MIPS dataset, 5% of proteins are essential. Although DC extracts 55 proteins, NC extracts roughly 60 proteins. With maximal proteins ranging from 10 to 20, SC and EC operate poorly. With 120 proteins extracted for the top 10% and 240 proteins in the top 15%, NC exhibits a consistent performance. The performance of the different centrality measurements differs noticeably at 15%. Thus, NC is the appropriate measure for determining the necessity of MIPS nodes.

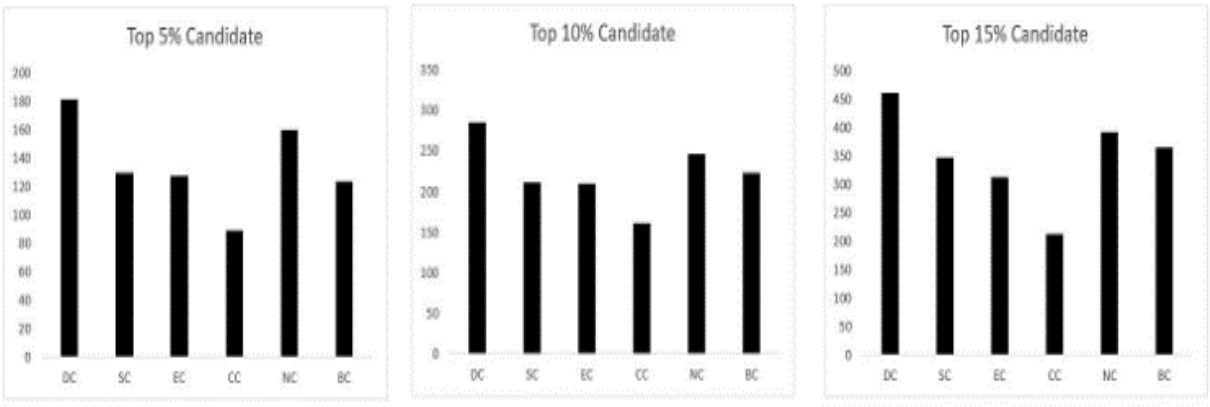

Fig 1. Analysis of critical proteins found in the BIOGRID dataset in comparison using centrality measures

Using the MIPS dataset, it is visible that edge-clustering performs better than alternative centrality measures. In the BioGRID dataset, DC performs better than other centrality measures. The outcomes

shown in Figures 1 and 2 illustrate the performance. Edge clustering exhibits a 20% improvement for the MIPS dataset, because NC determines centrality by taking a node's relationship to its close neighbors into account. Eq. (6) predicts that nodes with stronger linkages have higher values for NC and the proteins are important. As compared to other competing metrics for BioGRID, DC performs well, detecting up to 5%, 10%, and 15% of the proteins.

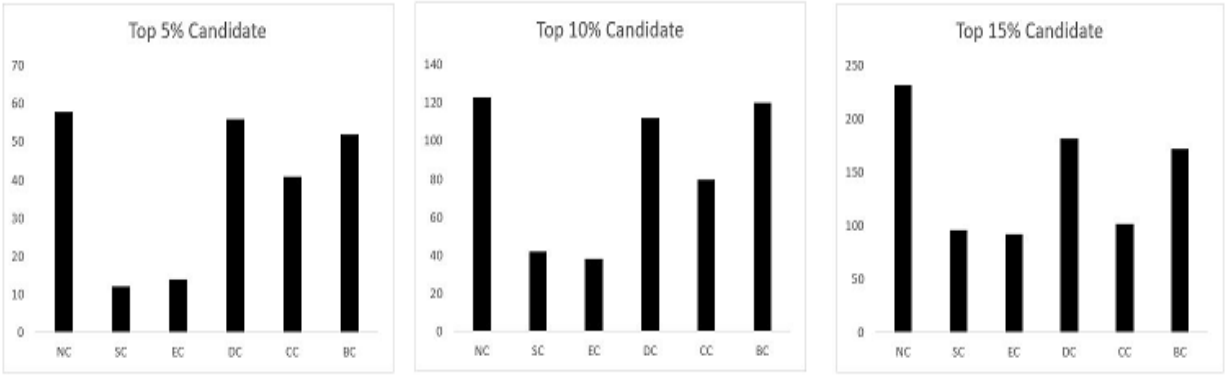

Fig 2. Analysis of critical proteins found in the MIPS dataset in comparison using centrality measures

The amount of time required to calculate the NC coefficient for a degree-n network will be O (n2). An average species' PPI will have a maximum degree of eight., so the complexity cannot exceed O (n2). EC and SC display decreased MIPS performance. When compared to other competitive approaches, SC and EC found less important nodes. However, in the BioGRID dataset, they are demonstrating a significant improvement in the performance for EC and SC. For both datasets, CC performs worse than expected. Figures 1 and 2 demonstrate that BC exhibits a consistent and generally superior performance for both networks.

Other centrality measures will perform better in a different dataset, according to the evidence. This might be brought on by variations in these datasets' node counts and connectedness [20]. Even the false positives and false negatives that can occur while identifying the crucial proteins in Saccharomyces cerevisiae are reflected in this. The effectiveness of centrality measurements in the datasets are assessed under the presumption that false positive and false negative results are equally spaced and limited in number. The finite/infinite sequence that joins the edges on their vertices is known as a "walk". Eigen centrality can estimate walks of any length, degree centrality can only count walks of a single length.

The in and out-degree of a node need special consideration for social media network. While attempting to define terms like association or friendship, popularity and camaraderie are in ascending order of importance. A few centrality analysis limitations were noted. Measures of centrality are application-specific, and the centrality of one application may not be ideal for another. Second, the relative influence of each node varies depending on how the vertices are ranked based on centrality values. The disparity in how centralization scores are evaluated must be addressed immediately. The characteristics that were used to find the crucial node do not apply to the remaining vertices.

V. CONCLUSION

Finding the significant nodes of a PPI network aids in finding drugs, cancer medication candidates, biomarker candidates, and illness prognosis. The literature provides a variety of computational methods for identifying and prioritizing the influential nodes in biological networks. Identifying prominent nodes and hub nodes frequently involves using centrality measurements. Network-based approaches comprehend the intricate reconstruction and analysis of biological networks. MIPS and BioGRID datasets are under consideration. In this study, the top 5%, 10%, and 15% of the important proteins were extracted after evaluating the widely used topological centrality metrics DC, NC, EC, CC, and BC. The experimental findings describe the properties of datasets in relation to different centrality measurements. In MIPS dataset edge clustering outperforms the other centrality measures and in BIOGRID degree centrality is better than other measures.

REFERENCES

1. Graves, P. R., & Haystead, T. A. (2002). Molecular biologist's guide to proteomics. *Microbiology and molecular biology reviews*, *66*(1), 39-63.
2. Ali, A., Hulipalled, V. R., & Patil, S. S. (2022). A NOVEL SEMANTIC SIMILARITY SCORE FOR PROTEIN DATA ANALYSIS. *Computing Technology Research Journal*, *1*(1), 1-4.
3. Cullen, L. M., & Arndt, G. M. (2005). Genome-wide screening for gene function using RNAi in mammalian cells. *Immunology and cell biology*, *83*(3), 217-223.
4. Ali, A., Viswanath, R., Patil, S. S., & Venugopal, K. R. (2017, May). A review of aligners for protein protein interaction networks. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1651-1655). IEEE.
5. Joy, M. P., Brock, A., Ingber, D. E., & Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, *2005*(2), 96.
6. Wuchty, S., & Stadler, P. F. (2003). Centers of complex networks. *Journal of theoretical biology*, *223*(1), 45-53.
7. Estrada, E., & Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, *71*(5), 056103.
8. Ali, A., Hulipalled, V. R., & Patil, S. S. (2020, December). Centrality Measure Analysis on Protein Interaction Networks. In *2020 IEEE International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)* (pp. 1-5). IEEE.
9. McClure, R. S., Overall, C. C., Hill, E. A., Song, H. S., Charania, M., Bernstein, H. C., ... & Beliaev, A. S. (2018). Species-specific transcriptomic network inference of interspecies interactions. *The ISME journal*, *12*(8), 2011-2023.
10. Jere, S., Jayannavar, L., Ali, A., & Kulkarni, C. (2017, February). Recruitment graph model for hiring unique competencies using social media mining. In *Proceedings of the 9th International Conference on Machine Learning and Computing* (pp. 461-466).
11. Memoria, M., Shah, S. K., Anandaram, H., Ali, A., Joshi, K., Verma, P., ... & Akram, S. V. (2023). An internet of things enabled framework to monitor the lifecycle of Cordyceps sinensis mushrooms. *International Journal of Electrical & Computer Engineering (2088-8708)*, *13*(1).

12. Ali, A., Hulipalled, V. R., Patil, S. S., & Abdulkader, R. (2021). DPEBic: detecting essential proteins in gene expressions using encoding and biclustering algorithm. *Journal of Ambient Intelligence and Humanized Computing*, 1-8.
13. Rashmi, C., & Kodabagi, M. M. (2017, August). A review on overlapping community detection methodologies. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 1296-1300). IEEE.
14. Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-Pour, H. (2020). MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. *Expert Systems with Applications*, *142*, 113024.
15. Ruan, Y., & Li, A. (2015). Influence of dynamical change of edges on clustering coefficients. *Discrete Dynamics in Nature and Society*, *2015*.
16. Ahmed, S. T., Sreedhar Kumar, S., Anusha, B., Bhumika, P., Gunashree, M., & Ishwarya, B. (2020). A generalized study on data mining and clustering algorithms. *New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India*, 1121-1129.
17. Ali, A., Ajil, A., Meenakshi Sundaram, A., & Joseph, N. (2023). Detection of Gene Ontology Clusters Using Biclustering Algorithms. *SN Computer Science*, *4*(3), 217.
18. Lü, L., Chen, D., Ren, X. L., Zhang, Q. M., Zhang, Y. C., & Zhou, T. (2016). Vital nodes identification in complex networks. *Physics reports*, *650*, 1-63.
19. Sharon Priya, S., & Ali, A. (2016). Localization of WSN using IDV and Trilateration Algorithm. *Asian Journal of Engineering and Technology Innovation*, *4*(7).
20. Ahmed, S. T., & Basha, S. M. (2022). *Information and Communication Theory-Source Coding Techniques-Part II*. MileStone Research Publications.
21. Mewes, H. W., Frishman, D., Mayer, K. F., Münsterkötter, M., Noubibou, O., Pagel, P., ... & Stümpflen, V. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic acids research*, *34*(suppl_1), D169-D172.
22. Negre, C. F., Morzan, U. N., Hendrickson, H. P., Pal, R., Lisi, G. P., Loria, J. P., ... & Batista, V. S. (2018). Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, *115*(52), E12201-E12208.
23. Ahmed, S. T. (2017, June). A study on multi objective optimal clustering techniques for medical datasets. In *2017 international conference on intelligent computing and control systems (ICICCS)* (pp. 174-177). IEEE.