Protein-Protein Interaction Network Analysis

Authors
Hager Samir & Youssef Ahmed

Supervisor
Dr.Ibrahim Youssef

Abstract

Now the biological processes are more complex, where proteins serve as the major molecule guiding a specific biological pathway. Proteins are long chains of amino acids, which are folded in a particular configuration. It is this specific configuration that enables a protein to physically interact with other proteins to form protein complexes and serve in downstream processes. Since proteins play a principal role in determining the molecular mechanisms and cellular responses, understanding the protein interaction networks is becoming a salient subject in research. In the following sections, we are going to consider a PPI of human beings to get some hands-on experience in extracting insights about such data to help us in understanding cellular processes, disease mechanisms, and potential therapeutic targets.

1 Introduction

Discovering the functional interdependencies among molecular components is critical because it sheds light on the structure–function relationships. Generally, most important biological activities are not the result of a single molecule but depend on the coordinated effects of multiple molecules interacting with others. This is confirmed by the fact that the polygenic disorders result from various biological processes that interact in a complex network, rather than from an abnormality in a single effector gene product. Thus, studying biology under the context of networks is very essential and promising.

The data under examining is a PPI of human being. It consists of 4 columns, the first 2 columns representing the tail and head protein of each edge in the network. The third one contains the probability of interaction between each of the proteins composing the edge. Lastly, the edge type that formed between the two nodes.

Tail	Head	Edge_weight	Edge_type
Q8TBF5	Q9UKB1	0.311133	MI:0004 (affinity chromatography technology)
Q8TBF4	Q15717	0.311133	MI:0004 (affinity chromatography technology)
Q8TBF4	P08865	0.311133	MI:0004 (affinity chromatography technology)
Q8TBF4	Q02539	0.311133	MI:0004 (affinity chromatography technology)
Q8TBF4	Q96J01	0.201461	MI:0401 (biochemical)

Table 1: Sample Data

2 Methods & Results

2.1 Graph Visualization

The PPI data is extremely large. It contains 17168 nodes and 612516 edges. In addition, the network has a somehow a larger value of transitivity nearly 0.0645, compared to random graph of the same number of nodes and edges. transitivity is a global measure of the tendency of the nodes to cluster together. Such value of transitivity value indicates that the neighbors of any node in the graph are more likely to be directly connected to each other, resulting in a more dense network structure, making the visualization output as a hard stem to get insights from

The color of each node represents the connectivity measure of the node (degree), which is considered as a local measure, providing information about a node's immediate neighbors. The size of each node reflects the ability of the node to control or influence communication paths in the network (betweenness centrality), which is more of a global measure.

To further clarify the difference between degree and betweenness centrality, the degree is used to identify well-connected nodes, while betweenness centrality is used to identify which nodes play a critical role in maintaining efficient communication in the network. So, they differ from each other.

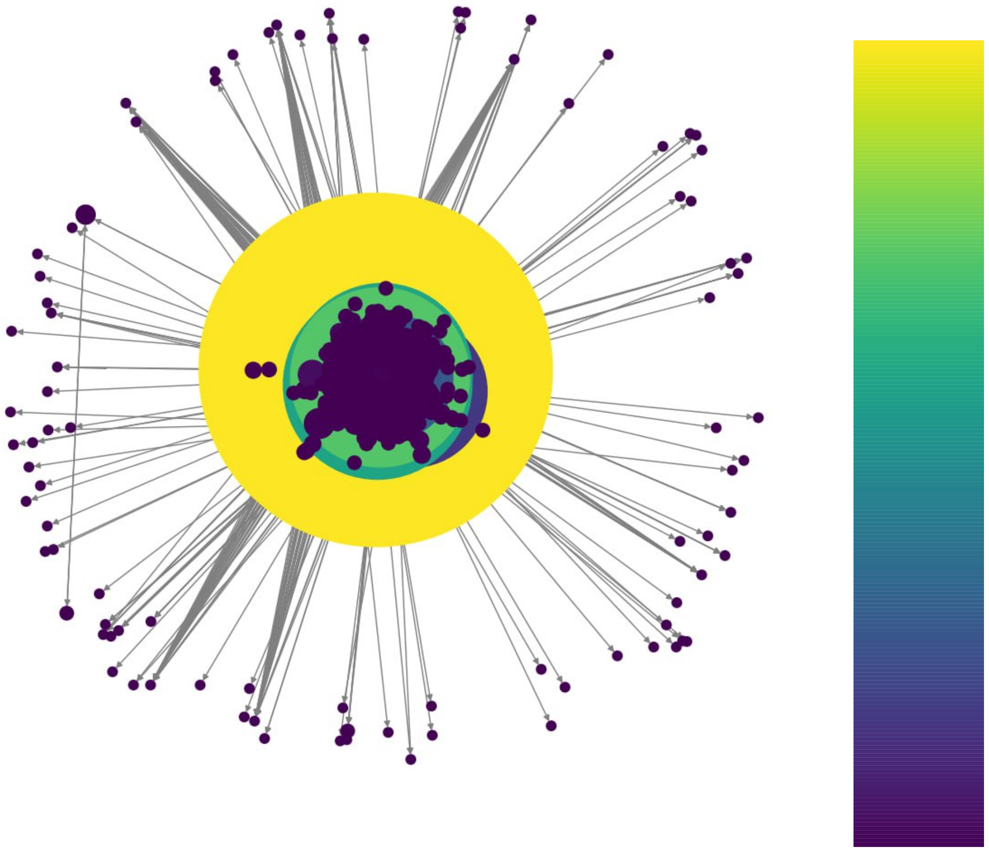


Figure 1: Colored Visualization Graph

2.2 K-Acyclic Shortest Paths

Interactome (totality of PPI) represents the medium through which signaling pathways occur. It helps in conveying biological signals in the cell. And since nature tends towards low energy consumption (shortest paths). It has become crucial to predict and construct such shortest paths that convey these signals to understand the cellular function under normal and abnormal conditions. In what is known as PathLinker, it's a technique to construct the k-shortest paths of a signaling pathway giving only a set of receptors and another set of transcriptional regulators for that pathway.

There are two types of biological paths, namely cyclic and acyclic paths. Cyclic paths can represent feedback loops in PPI networks. Feedback loops are important for regulatory mechanisms where the output of a process influences its own regulation. They play a role in maintaining homeostasis and regulating cellular responses. Acyclic paths in PPI networks often represent sequences of protein interactions that occur in a linear and ordered manner. These pathways are essential for processes like signal transduction, where information flows from one protein to another in a specific order. Acyclic paths can highlight functional modules or complexes within the PPI network. Identifying acyclic paths helps in understanding how groups of proteins collaborate without forming cycles, providing insights into specific biological functions.

For such purpose, we have applied Yen's algorithm to get k-shortest paths, which is an extension of Dijkstra's algorithm by iteratively removing edges from the initially found shortest path to explore alternative paths. This process continues until K-shortest paths are identified. However, the direct application of Yen's algorithm is statistically inappropriate, since the weights of the PPIN are actually the probability of interaction between the proteins composing each edge, meaning that the scoring scheme used in determining the shortest path is the overall probability of the series of interactions composing the path. Phrased differently, we need to multiply the weights of the edges not to add them as to be in case of Yen's algorithm. Therefore, we have had to mathematically manipulate the weights of the graph to be applicable to Yen's algorithm and also to be reversible.

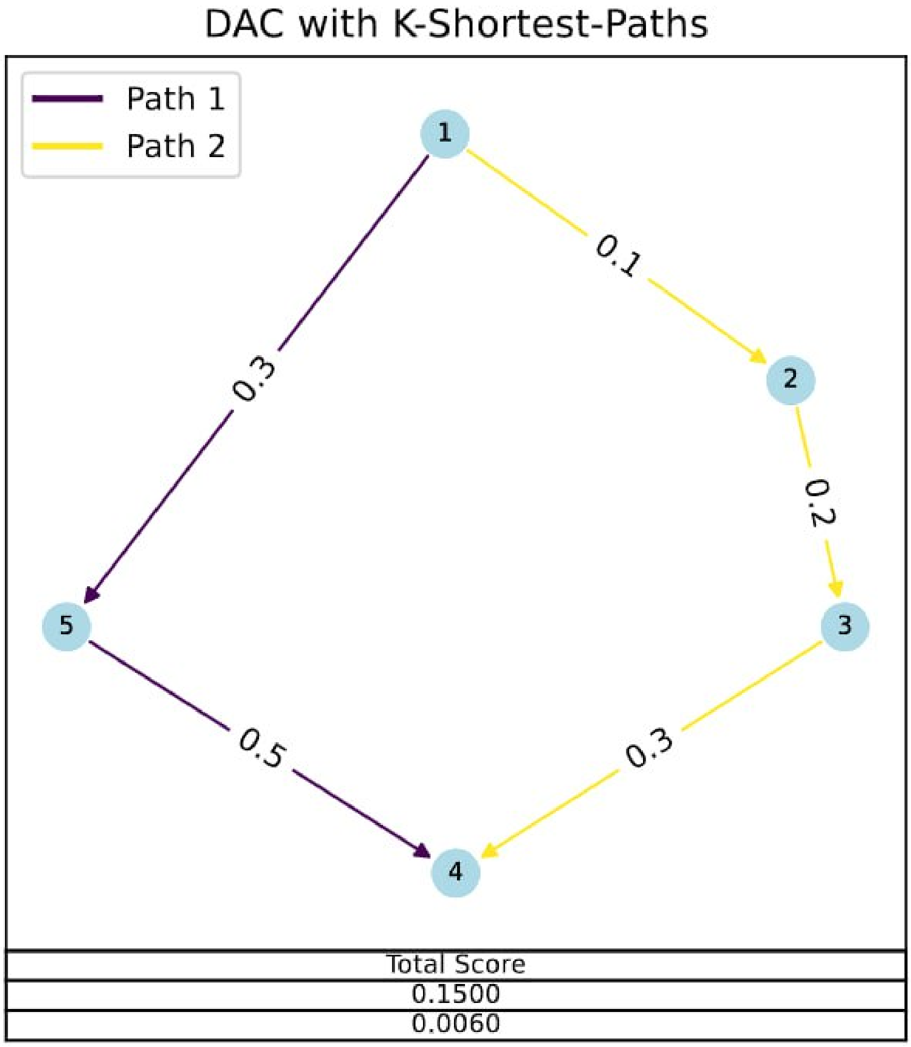
Given the weights of an arbitrary shortest path $\{P_1, P_2, P_3, P_4\}$. The overall score of such path is $(P_1 \times P_2 \times P_3 \times P_4)$. Yen's algorithm computes the score as the following $(P_1 + P_2 + P_3 + P_4)$. So, to maintain the addition rule followed in Yen's algorithm and at the same time maintain the reversibility of the score calculation. We have to transform the weight of each edge in the PPIN by taking the negative logarithmic of base 10 of each weight and then apply Yen's algorithm.
Note: Yen's algorithm deals with positive weights.

$$\textbf{Path Score} = \textbf{-}\ (log_{10}P_1 + log_{10}P_2 + log_{10}P_3 + log_{10}P_4)$$

$$\textbf{Actual Score} = 10^{\textbf{ - Path Score}} = \textbf{Overall Probability}$$

Yen's algorithm does not enforce any restrictions regarding the existence of cycles in the graph, which means that the output of Yen's algorithm can contain cyclic paths. So, we have had to filter out such cyclic paths, and that is exactly what we did. We have considered each edge in each shortest path given by the algorithm, and temporarily remove it and check if there is a path back from the target to the source proteins composing the current edge. If that happens to be, then there is a feedback loop in this shortest path. Therefore, we exclude such path.

For the sake of testing, we have developed some synthetic paths, by randomly generating source and target protein, which do not have direct edge between each other, but have at least a one simple path that connects them. After applying the K-acyclic shortest path algorithm on a small portion of PPI data with the help of the source and target proteins synthetically generated, we have found that most of the Acyclic shortest paths have a maximum length of **3-4 steps**, that does not change dramatically by increasing the data size considered. In addition, the diameter of the strongest connected component in the whole graph is of **8 steps**, which means that the largest shortest path is of **8 step**s, which is small comparted to the whole data size.



Total Score
0.1500
0.0060

Figure 2: Sub-network Graph

2.3 Degree Analysis

The PPIN is a directed graph, where the influence direction matters. So, examining both in-degree and out-degree of each protein in the network is crucial for comprehensively characterizing the role of a protein within a network and provides insights into the direction of information flow within the network. It helps in understanding how signals are received and transmitted, contributing to network dynamics. Proteins with high degree (Hubs) may participate in signaling cascades, transmitting information to downstream effectors. If this protein function has any issue, this will affect large number of proteins functions could know which Protein affect the functions of the other proteins.

After analyzing the PPIN of interest, we found out that the average degree of the network is **71.36**. and the most connected hub is **’P05067’** of **4170** edges. The figure below shows the degree distribution of the Network.

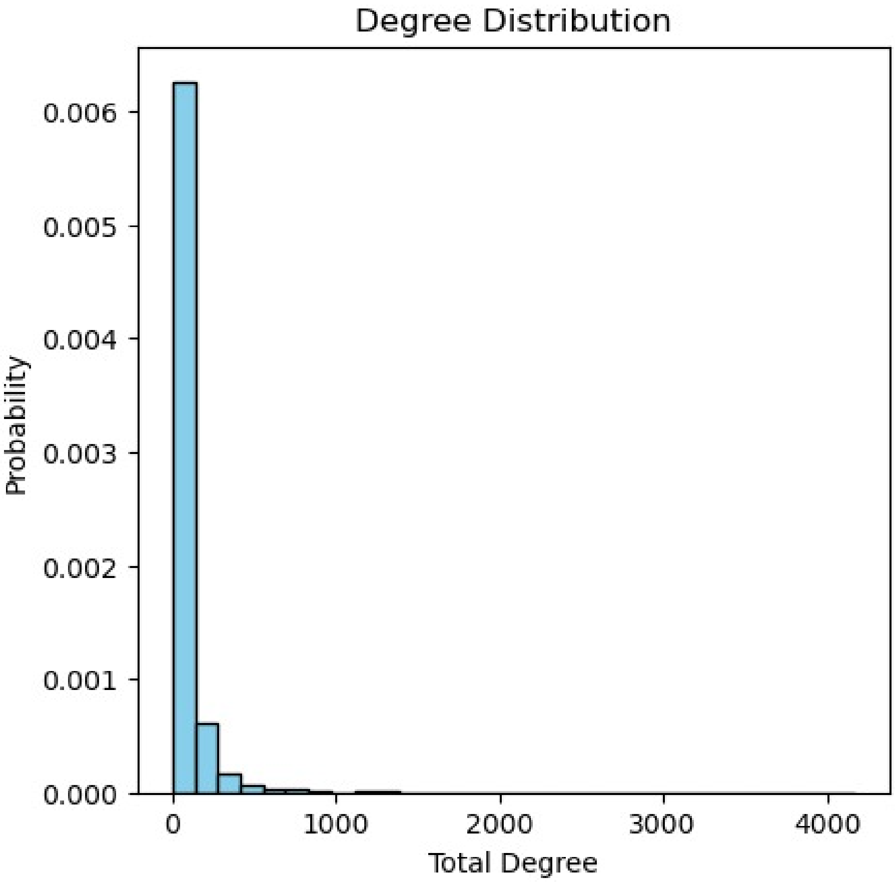


Figure 3: Histogram Degree

3 Conclusions

3.1 Small World Effect

After examining the results of k-acyclic shortest paths, we can deduce that PPIN exhibits a small-world effect, indicating significant connectivity among proteins. This phenomenon implies that the network's diameter, representing the maximum number of steps between any two nodes, remains small regardless of the network's size. In essence, it suggests that the distance between any two nodes is typically less than six steps,more or less, aligning with the well-known concept of "six degrees of separation" popularized in social sciences.

As we have seen in the result of analysis, the diameter of the strongly connected component in the whole interactome is **8**, which means that the maximum shortest path in length is equal to **8**. Compared to the huge number of edges in the network. We can prove the verification of small world effect in the interactome.

3.2 Scale-free Property

AS We can notice from the degree distribution of the PPIN, such networks are scale-free networks, meaning the majority of nodes (proteins) in scale-free networks have only a few connections to other nodes, whereas some nodes (hubs) are connected to many other nodes in the network. If we represent the degree distribution of a scale-free network in a logarithmic scale, we can see how it fits with a line (they fit a power-law), having a small number of nodes with high degree (the hubs) and a large number of nodes with a low degree.
The scale-free nature of protein-protein interaction networks gives them a number of important features:

1. Stability
   - If failures occur at random, and the vast majority of proteins are those with a small degree of connectivity, the likelihood that a hub would be affected is small.
   - If a hub-failure occurs, the network will generally not lose its connectedness, due to the remaining hubs.
2. Invariant to changes of scale
   - No matter how many nodes or edges the network has, its properties remain stable.
   - The presence of hubs is what allows for the small-world effect to be present regardless of the size of the network.
3. Vulnerable to targeted attack
   - If we lose a few major hubs from the network, the network is turned into a set of rather isolated graphs.
   - Hubs are enriched with essential/lethal genes. For example, many cancer-linked proteins are hub proteins.

4 Transitivity

As we have seen, the value of transitivity, which is a global measure of the graph nodes tendency to cluster together, is larger compared to a random graph of same number of nodes and edges, which gives an indication that the whole interactome might contains functional units that are specific.

5 Packages

- networkx
- pandas
- matplotlib
- numpy
- copy
- math
- bioservices
- collections
- random

6 References

- https://link.springer.com/chapter/10.1007/978-3-030-02634-9_12
- https://academic.oup.com/bfg/article/11/6/434/238834
- https://www.nature.com/articles/s41598-018-23672-0
- https://www.ebi.ac.uk/training/online/courses/network-analysis-of-protein-interaction-data-an-introduction/protein-protein-interaction-networks/