Graph Theoretical Analysis of Protein Networks

RSI – C SP 2018

S. V. Shanmugha Balan, SBOA Anna Nagar and S. Arjun, PSBB Nungambakkam

Under the Guidance of:-

Dr. Karthik Raman, Department of Biotechnology, IIT Madras and Aarthi Ravikrishnan

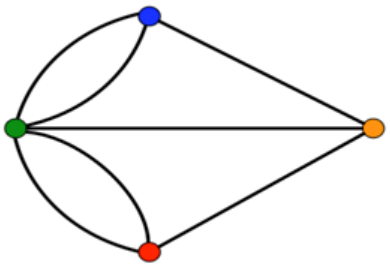
Introduction

Ever since Alexander Fleming discovered penicillin, humans have used antibiotics to fight pathogenic invasions. Broad-spectrum antibiotics tend to work by disrupting enzymes involved in biosynthesis, nucleic acid metabolism, protein synthesis and membrane structure. In recent years, abuse of antibiotics has led to antibiotic resistant 'superbugs'. This is due to the process of divergent evolution. Therefore, we must make the transition to narrow spectrum antibiotics, which are more personalised and efficient in their action.

In this project, we used graph theory, to analyze protein networks and to identify important proteins. The effects of removal of these proteins on the network were computed. We interpreted our results and discussed the practical removal of these proteins. We also listed a few ways in which we can improve our analysis.

Graph Theory

Graphs are mathematical structures used to model pairwise relations between objects. A graph in this context is made up of nodes (points) which are connected by edges (lines). Graphs can be directed (following and followers in Twitter) or undirected (friends in Facebook). The edges in a graph may also be weighted with numerical values. For example, in a graph representing the road networks in a city, wider roads will have a higher weight than a narrow road.



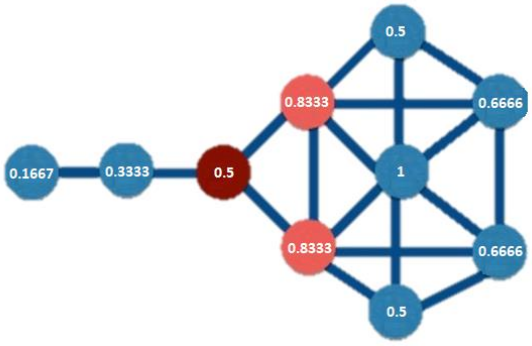
1

[1] The degrees centralities of nodes in a graph Source: https://en.wikipedia.org

Centrality Measures

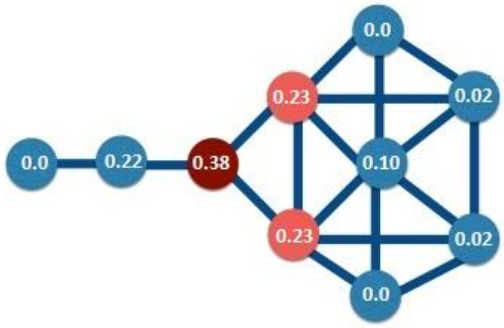
Centrality measures are the identifiers of the most central nodes of a graph. This project used the following centrality measures.

1. Degree Centrality: The degree of a node refers to the number of edges connected to it. Here, degree centrality of a node is the fraction of the total nodes of a graph to which a node is connected to.



[2]

2. Betweenness Centrality: The betweenness centrality of a node is the ratio of the number of shortest paths between any pair of nodes to the total number of shortest paths between them. This is normalised by dividing by the number of pairs of nodes.



[3]

Some of the other factors that were considered in this study are:-

1.) Average Shortest Path Length: The mean of the length of the shortest path between each pair of nodes.
2.) Degree Distribution: A plot of $P(k)$ versus k, where $P(k)$ is the probability that a node has a degree k.

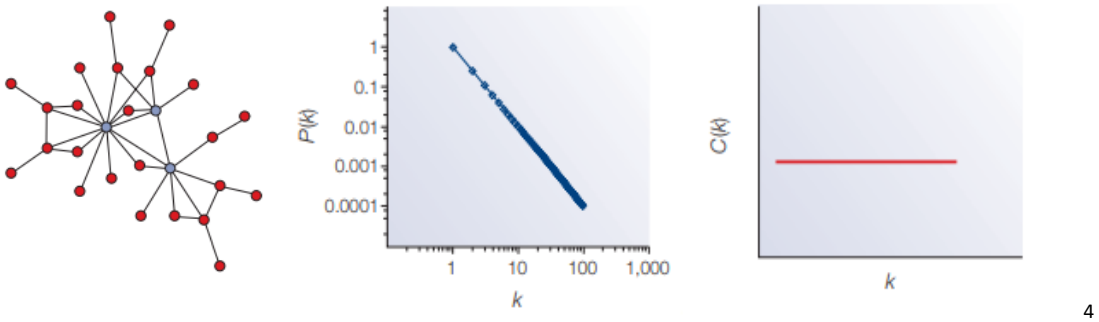
[2] The degree centralities of nodes in a graph, edited, source: https://image.slidesharecdn.com/networkanalysislecture-150320110618-conversion-gate01/95/network-analysis-lecture-14-638.jpg?cb=1426867653
[3] The betweenness centralities of nodes in a graph, edited, source: https://image.slidesharecdn.com/networkanalysislecture-150320110618-conversion-gate01/95/network-analysis-lecture-14-638.jpg?cb=1426867653

Protein Networks

Proteins are bio macromolecules which are vital to the survival of living organisms. They perform various functions such as catalysing reactions, transporting molecules, DNA replication and providing cellular structure. Recent research shows that many processes are not carried out by proteins themselves but rather because of interaction between various proteins.

These highly specific interactions are regulated by electrostatic forces and can occur both in and out of cells. They can be represented using graphs. Research shows that these graphs have very specific properties.

Most protein networks tend to follow a power-law degree distribution. This means that $P(k)$ is proportional to $k^{-\gamma}$ where γ is called the degree exponent. A highly connected node is called a hub. A scale free distribution tends to have a few hubs to which other nodes are connected to.



[4]

Most biological networks are scale free and have $2<\gamma<3$ with increase in γ leading to decrease in the degree centrality of a hub. The log-log plot of the degree distribution is a straight line. Clustering coefficient reflects the tendency of the node to form groups or clusters. For biological networks, average clustering coefficient of a node with degree k is independent of k.

[4] A representation of a scale free network (hubs are blue), its degree and clustering coefficient distributions. Source : Network Biology-Understanding the Cell's Functional Organization (Paper by Albert-László Barabási & Zoltán N. Oltvai)

Creating a Model

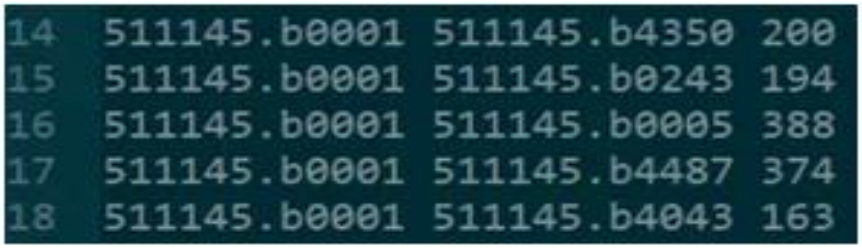
The Centrality-Lethality Rule states that hubs in a biological network tend to be essential. On the basis of this assumption the following model was made:

Biochemical	Graph Theoretical
Proteins	Nodes
Interactions	Edges
Strength of Interaction	Edge-Weights
Number of interactions the protein is involved in	Degree Centrality
Importance of the protein	Betweenness Centrality
Fitness of the organism	Shortest Path Length

[5]

Methodology

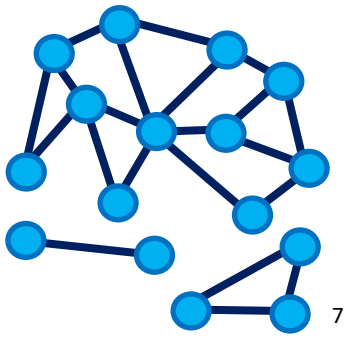
- The protein-protein interaction data was obtained for *E. coli* from the STRING database. The NetworkX package of Python was used.
- For computational purposes, only the nodes connected to edges with weights greater than 700 were considered.
- The degree centralities and betweenness centralities of all the nodes and, the average shortest path length for the graph was computed.
- The nodes with the highest values of those centralities were removed separately. The changes to the shortest path length were computed.



[6]

[5] Graph Theoretical Equivalents of Protein Networks and Interactions
[6] A sample of Interacting proteins and their strength of interaction, taken from String Database.

- As the graph produced was disconnected, the largest sub graph was used.

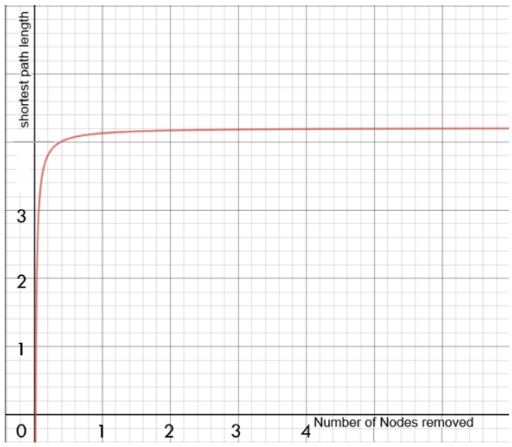


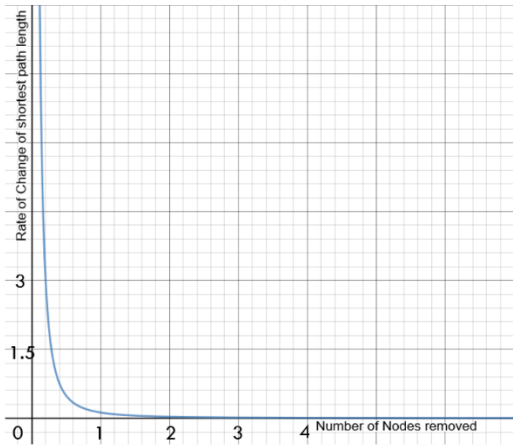
[7]

- To interpret the data a plot indicating the path length upon removal of the nodes was made using regression analysis and by matching the correlation coefficient.[8] The derivative of this graph was also plotted.

Results and Interpretation

Upon removal of the node with highest degree, the path length increased. Upon subsequent removal of nodes, the path length continued to increase; however the rate of increase slowed down. This is goes on to prove the fact that protein networks have a few highly connected proteins which are essential to the life of the organism.

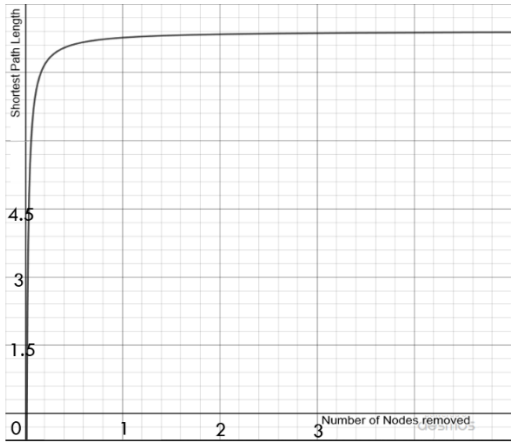


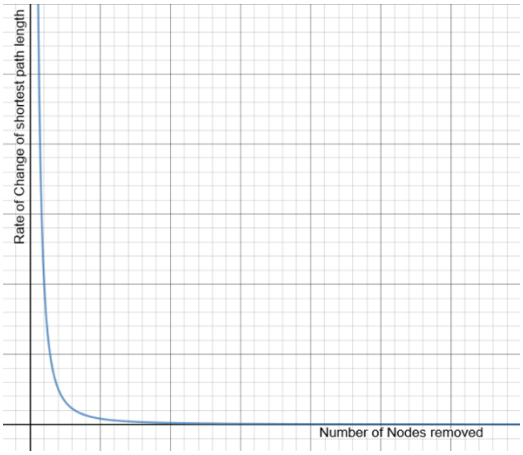


[9]

The same trend was observed in betweenness centrality. In addition, many of the nodes that had a high degree also had a high betweenness. This further strengthens the idea of a few important nodes vital to the life of the organism.

[7] A Graph with many disconnected sub graphs
[8] The degree of association of the best fit line with individual data points
[9] Shortest path length upon removal of nodes based on decreasing value of degree centrality. |r|=0.99
The graph is in red and the derivative is in black. Source: https://www.desmos.com/





10

Two of the most influential proteins were:-

1.) Bacterial alkaline phosphatase (involved in dephosphorylation)
2.) Glutamate Synthase (an oxidoreductase)

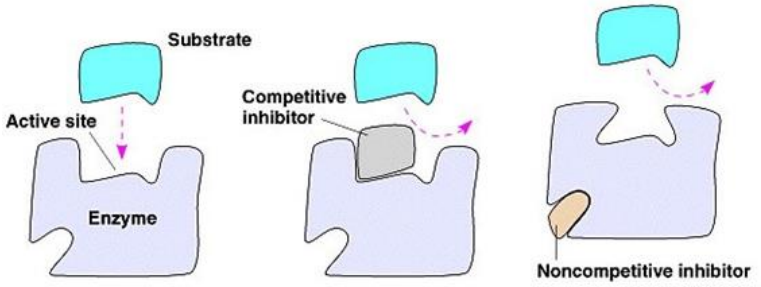
Another observation was that when the threshold value for edge-weights was set to 600, the shortest path length reduced by 14.7784 %. This is because of an increase in links between proteins. However, as the interaction between these proteins is weak, they have a lesser chance of being essential to the organism.

Applications

This method of analysis will allow more specific targeting of pathogens and will help combat against the increasing antibiotic resistance. Once the most essential protein is identified:

1) A molecular model of the enzyme involved in the activation of that protein can be made. Induced Fit hypothesis states that enzymes have active sites that are highly dynamic and alter the structure of the substrate to produce product molecules. Inhibitors are compounds that either bind to the active site and prevent substrate from binding or side sites (allosteric sites) and conformationally change the enzyme. Using the model and by calculating electrostatic potentials, various inhibitory compounds can be made and activation of the protein will be stopped.

[10] Shortest path length upon removal of nodes based on decreasing value of degree centrality. |r|=0.99
The graph is in blue and the derivative is in black. Source: https://www.desmos.com/



[11]

2) The gene responsible for the protein can be silenced using methods like mRNA interference. A RNA molecule can be used to inhibit selected mRNA molecules and this affects the expression of genes and hence the production of proteins. [12]

Limitations and Scope for Improvement

- The most connected node may not actually be part of an essential pathway for the organism.
- This model does not take into account the highly dynamical nature of protein-protein interactions.
- Only two centrality measures (betweenness and degree) were considered here. Others like eigenvalue and closeness centrality, clustering coefficient were not considered.
- Different centrality measures were considered separately. They should be combined together using regression analysis into an overall centrality measure and that should be used for finding the most influential node.
- Detailed protein-protein interaction data may not be available for all organism.
- This method serves no purpose in the analysis of fast evolving pathogens or in the analysis of viruses.
- The protein produced by this method may already be targeted by a drug to which a pathogen has evolved against.
- The enzyme deduced by this method may be essential to humans and can't be targeted because of lethal consequences for us.

[11] Action of Inhibitors on Enzymes Source: https://i2.wp.com/tuitiontube.com/wp-content/uploads/2016/10/competitive-and-noncompetitive-inhibitors.jpg?fit=637%2C379&ssl=1
[12] One Gene – One Enzyme Hypothesis

Outlook

Through graph theoretical analysis it is possible to identify the most important proteins in organism. These can then be specifically targeted. This approach can be used to counter the increasing menace of drug resistant pathogens (like XDR TB). The authors of this report will like to conclude with a word of caution. The prevalence of superbugs is due to the abuse of antibiotics. Similarly any artificial method of treatment will lead to problems. There must be a focus on more natural and organic ways of immunity, with an increased reliance on the body's inbuilt defence capability. Drug designing should be more of a last resort and should be used sparingly.

Acknowledgements

- Dr Karthik Raman – Our Mentor
- Ms Aarthi Ravikrishnan – Our Guide

References

- Wikipedia (https://en.wikipedia.org)
- Embryo Project Encyclopaedia (https://embryo.asu.edu)
- NetworkX Documentation (https://networkx.github.io)
- String Database (https://string-db.org)
- Desmos Graphing Calculator (https://www.desmos.com)
- Keisan Online Regression Calculator (https://keisan.casio.com/exec/system/14059929550941)
- Network Biology-Understanding the Cell's Functional Organization (Paper by Albert-László Barabási & Zoltán N. Oltvai)
- Trueman's Elementary Biology Volume I & II
- Anaconda, Spyder, Cytoscape

Code

```python

"""
RSIC Project - Analysis of Protein Networks with Graph Theory and NetworkX

Author: Balan and Arjun
"""

import networkx as nx

i = 0
x = 0
G = nx.Graph()
C = nx.Graph()
deg_cen_list = list()
deg_cen_invert = dict()
bet_cen_list = list()
bet_cen_invert = dict()

# Making the graph
with open('E:ecoli.txt', 'r') as f:
    for line in f.readlines():
        words=line.split()
        if ((int(words[2]))>=700):
            G.add_edge(words[0], words[1], weight = int(words[2]))

# Extracting the largest subgraph
for comp in nx.connected_component_subgraphs(G):
    if comp.number_of_nodes()>i:
        C = comp
        i = comp.number_of_nodes()

# Getting the centrality measures
deg_cen = nx.degree_centrality(C)
bet_cen = nx.betweenness_centrality(C)

# For inverting the dictionaries
for keys, values in deg_cen.items():
    deg_cen_invert[values] = deg_cen.get(values, [])
for keys, values in deg_cen.items():
    deg_cen_invert[values].append(keys)

for keys, values in bet_cen.items():
    bet_cen_invert[values] = bet_cen.get(values, [])
for keys, values in bet_cen.items():
    bet_cen_invert[values].append(keys)


# Extracting highest values of centrality
deg_cen_list = list(deg_cen_invert.keys())
deg_cen_list.sort()
bet_cen_list = list(bet_cen_invert.keys())
bet_cen_list.sort()

# Receiving average shortest path length
x = nx.average_shortest_path_length(C)
print("Initial average shortest path length", str(x))


# Average shortest path based on betweenness centrality
i = 0
for i in range(1, 11):
    for j in range(len(bet_cen_invert[bet_cen_list[-i]])):

        #Remove Node
        C.remove_node(bet_cen_invert[bet_cen_list[-i]][j])
        print('Node Removed : ', bet_cen_invert[bet_cen_list[-i]][j])

        # Re-Checking Shortest Path Length
        try:
            x = nx.average_shortest_path_length(C)
            print('Avg Path Length after removal : ', x)

            print('\n')
        except nx.NetworkXError:
            print('ERROR')
            pass

# Average shortest path based on degree centrality
i = 0
for i in range(1, 11):
    for j in range(len(deg_cen_invert[deg_cen_list[-i]])):

        #Remove Node
        C.remove_node(deg_cen_invert[deg_cen_list[-i]][j])
        print('Node Removed : ', deg_cen_invert[deg_cen_list[-i]][j])

        # Re-Checking Shortest Path Length
        try:
            x = nx.average_shortest_path_length(C)
            print('Avg Path Length after removal : ', x)
            print('\n')
        except nx.NetworkXError:
            print('ERROR')
            pass
```