Dhirubhai Ambani Institute of Information and Communication Technology

SC 435

INTRODUCTION TO COMPLEX NETWORKS

Large-scale analysis of disease pathways in the human interactome

November 22, 2020

Assigned by:
Prof. Mukesh Tiwari

Submitted by:
Niharika Dalsania (201701438)
Darshan Patel (201701436)
Ruchit Shah (201701435)

ABSTRACT

Discovery of disease pathways rely on Protein-Protein Interaction (PPI) networks which have a few known disease associated proteins that aid in finding the rest of the pathway. But such prediction methods have limited success and reasons for failure are not well understood. The study here tries to suggest that a plausible reason could be their unreliable assumption that pathways are densely clustered. In fact, in reality, majority of pathways do not correspond to single well-connected components in the PPI network. Instead, proteins associated with a single disease form many separate connected components in the network, which helps concluding that network connectivity structure alone may not be sufficient for disease pathway discovery. So one needs to look at the higher order PPI network structure to extract useful information for improving the performance of prediction models.

INTRODUCTION

Disease pathways are sets of genes(proteins) corresponding to a given disease. Each node in this **biological network** represents a protein, some of which are disease proteins, and each edge represents an interaction. Major challenge for analysing these networks to identify disease pathways comes from the inter connectivity of a pathways' constituent proteins, changing which the impact is just local. It spreads throughout the PPI network via the links and affect activities of other proteins. Although different methods predict disease proteins, how arrangement of disease pathways affects the performance of these algorithms is less known. The immense importance of these strategies for developing timely measures for disease prognosis, discovery and healing makes it critical to identify conceptual and methodological limitations of current approaches. Hereby we explore basic characteristics of this network, along with the higher order connectivity of proteins in network, which will throw light on the issue (1).

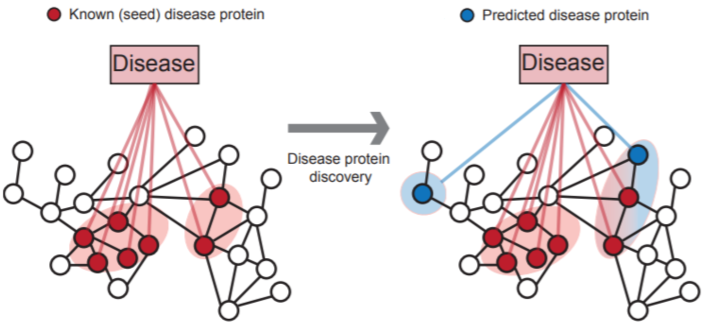
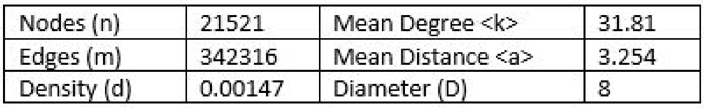

● Known (seed) disease protein
Disease
Disease protein discovery
● Predicted disease protein
Disease

Figure 1: Network based discovery of disease proteins

DATA CHARACTERISATION

Here, a PPI network structure of 519 diseases is studied. The datasets used for this study are downloaded from SNAP (2). It mainly comprises of Human PPI Network and Protein Disease Associations. The human PPI network used here is unweighted and undirected with the following characteristics:



Nodes (n)	21521	Mean Degree <k>	31.81
Edges (m)	342316	Mean Distance <a>	3.254
Density (d)	0.00147	Diameter (D)	8

Preliminary analysis shows that the disease nodes and their direct neighbours constitute majority of the graph that gives the first intuition that pathways are not well-knit community separated from the rest of the network, which will be proved formally in this study. The overall topology of the network characterised by highly heterogeneous degree distribution follows a **power law** (3) where vast majority of proteins have only few connections along with some highly connected proteins (hubs).

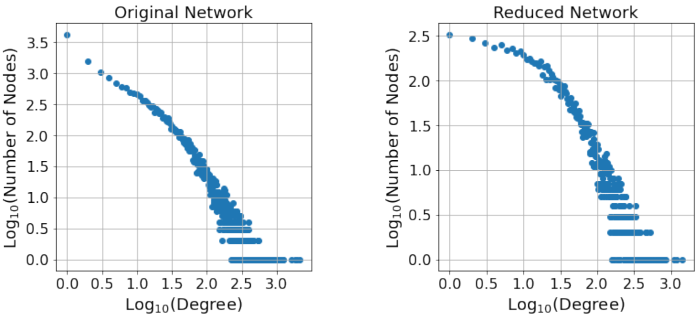


Figure 2: Power law property in PPI network

The network reduction such that it preserves all the components of all the disease pathways and the shortest paths between them, where the shortest paths are calculated by making the node with highest **betweenness centrality** in each disease pathway component as the representative node for distance calculation of that component preserves the basic characteristics while essentially removing low degree nodes as can be seen here. The diameter and mean distance of the network suggests that it follows the **small world property**.

DISEASE PROTEIN DISCOVERY

With the background knowledge of disease pathways and the associated data, we now use the PPI network and known disease genes of a given specific disease and identify most probable new disease proteins belonging to that disease pathway. A disease centric 10-fold prediction is used where all proteins is randomly split into 10 groups, all of which have same number of disease nodes. Using this prediction algorithm 10 times, the aim is to identify most probable disease proteins in one group while using the knowledge of disease proteins in other 9 groups. Each method attributes a score to each node - the probability of the protein being associated with the disease. For evaluation, Recall-at-100 quantifies the percentage of disease proteins identified correctly in the top 100 identified proteins, where a higher score indicates better performance. Approaches for the prediction task are majorly divided into the following two categories:

Clustering based methods

These methods assume that proteins belonging to same network community are likely to be a part of same disease pathways. In direct **Neighborhood** based scoring, each protein is attributed a number which depends on the fraction of its adjacent nodes which are disease proteins. These scores are then used to assign ranks to the nodes. Following equation governs the score assignment using Neighbourhood scoring for a particular disease:

$\mathbf{Scores} = \boldsymbol{A}\cdot\boldsymbol{P}$, where A is the adjacency matrix and P is a vector that has 1s corresponding to disease nodes of a particular disease, 0 otherwise. We take the nodes with top 100 scores into further consideration for finding recall-at-100.

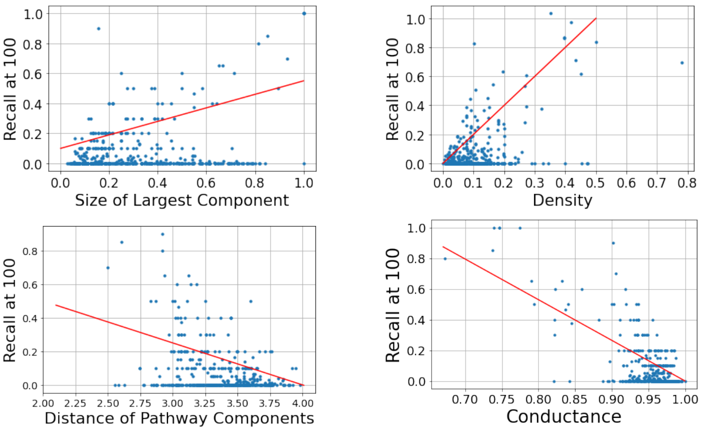


Figure 3: Prediction quality versus PPI connectivity using Neighbourhood based scoring

Above figure shows performance of neighbourhood based scoring method as a function of different distance and concentration measures of disease proteins (which will be explained in the following section).

Diffusion based methods

These methods use the known disease proteins as a starting point for a random walker. In **Random Walk** scoring, we use random walk with restart which will assign higher scores to the nodes closer to the starting point. This is because of the assumption that disease pathways are clustered in the PPI network. Here, at every iteration, the walker moves to a randomly selected neighbour protein with some probability α and returns to its initial position with probability 1 - α, with an α as high as 0.75. The more the number of times the nodes in the network are visited, the higher the scores assigned to it. As the scores converge after multiple iterations, they are used to rank the corresponding proteins. Following equation governs the score assignment using Random Walk based scoring:

$\mathbf{Scores[i]} = (\mathbf{1} - \alpha)\boldsymbol{A}\cdot\mathbf{Scores[i\text{-}1]} + (\alpha)\mathbf{Scores[0]}$, where A is the adjacency matrix and Scores[i] are the scores assigned to each nodes at i^{th} iteration. We stop the iteration when scores saturate and then consider the nodes with top 100 scores into further consideration for finding recall-at-100.

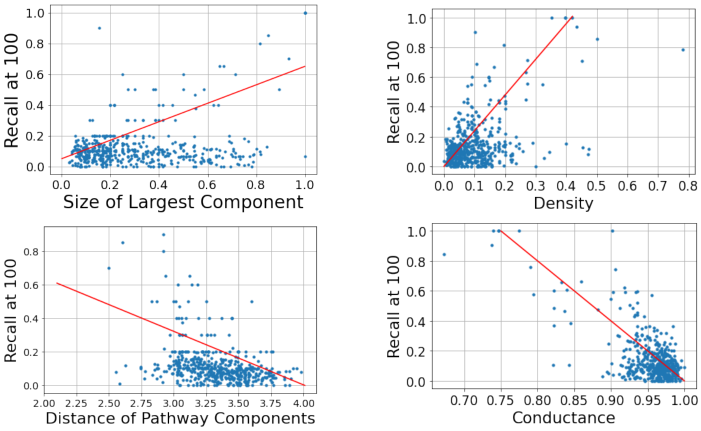


Figure 4: Prediction quality versus PPI connectivity using Random walk based scoring

Above figure shows performance of random walk based scoring method as a function of different distance and concentration measures of disease proteins (which will be explained in the following section).

We observe that Neighborhood scoring performs poorer than Random Walk based scoring in terms of overall distribution of recall-at-100 values. The improved performance of Random Walk based method over Neighbourhood based method suggests that it is too narrow to take into account just the immediate disease proteins while identifying potential disease nodes locally. However, even the Random Walk based method is dependent on the fraction of disease proteins in the pathway component from where the random walker starts the walk and this is problematic since the method becomes dependent on properties that are not typical of all disease pathways. However, in general for both methods, performance correlates positively with density and size of largest pathway component and negatively with distance between pathway components and conductance. This suggests that stronger the clustering of disease proteins within the PPI network, higher is the recall-at-100 of prediction algorithms.

PPI NETWORK CONNECTIVITY

Following are some network measures that can characterise connectivity of disease proteins with the disease pathway as well as the outside network and gives quantitative arguments to support the fact that pathways are indeed disconnected.

Size of largest pathway component

It is the ratio of disease proteins in the largest pathway component to the total number of disease protein in that particular disease. For this network, a large number of diseases have small size of largest pathway component, indicating that disease nodes are spread across many components rather than being densely clustered. Loosely, half of the disease pathways have above 16 connected components. Nearly, only half have above 0.21 fraction of protein nodes in the largest pathway component. Roughly, only 12% of pathways have more than 0.6 fraction of their proteins in the largest pathway component.

Density of the pathway

It is the ratio of number of edges to the number of maximum possible edges in pathway i.e. $|E_d| \ / \ \binom{|V_d|}{2}$. A lower density suggests that the ratio of edges out of all possible edges that can appear between nodes in disease pathway is quite small. This means all disease nodes in disease pathways are not well-connected internally with a median density of 0.076. Nearly 87.5% of diseases have a density below 0.17.

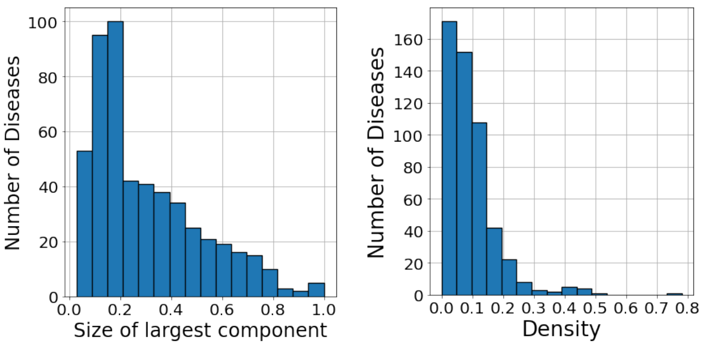


Figure 5: The distribution of (A) Size of largest pathway component and (B) Density of pathway

Conductance

It is given by $|B_d| \ / \ (|B_d|+2|E_d|)$ where pathway boundary B_d is the set of all pairs of nodes, exactly one of which belongs to pathway. A high conductance indicates the pathway is internally a loosely knit community, rather connected to the outside network. Here, around 50% of disease pathways have conductance above 0.96, which indicates that there are quite many edges pointing outside the pathway to the rest of the PPI network as compared to just edges lying inside the pathway.

Distance of pathway components

It is the average over distance between all pairs of components, which in turn is the mean of minimum path length between all protein pairs in a particular pair of components. For majority of the diseases, the distance between pathway components is high because the pathways are spread out in the network rather than being present as a community. Nearly 50% of disease pathways have a distance of approximately 2.9 between the pathway components.

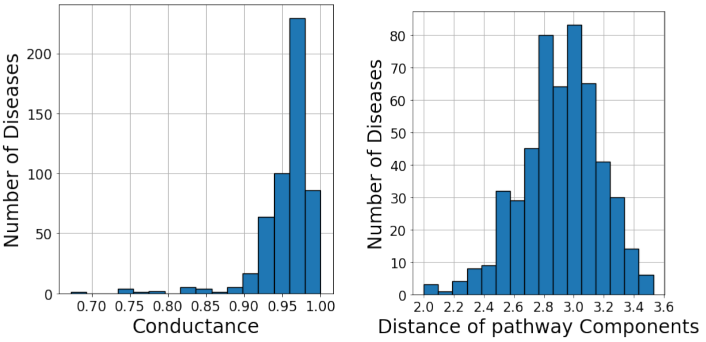


Figure 6: The distribution of (A) Conductance and (B) Distance of pathway components

Spatial Network Association

It measures the localization of disease nodes within the network by intuitively quantifying the clustering of disease nodes in the PPI network. It is given as $K_d(s) = 2/(\bar{p}n)^2 \sum_i p_i \sum_j (p_j - \bar{p}) I(l_G(i,j) < s)$, where p_i is a binary indicator indicating if node i belongs to the disease pathway, $\bar{p} = 1/n \sum_i p_i$ and $I(l_G(i,j) < s)$ is 1 if the minimum distance between i and j is within s, otherwise it is 0. If majority disease genes are grouped in a single area within the network, majority will have distance less than s between them for negligible values of s and hence higher $K_d(s)$ even for smaller s which will result in higher area under $K_d(s)$ curve (AUK), while for uniformly spread disease proteins in the PPI network, $K_d(s)$ achieves large values only for large values of s which will result in relatively lower AUK. An empirical p-value for these observed AUK values is calculated from their probability distribution curve, where the p-value is the area under this probability curve past the observed AUK. Majority p-values being large implies that majority AUK values lie towards the left of probability distribution curve which correspond to lower AUK values. To statistically check the localisation of disease pathways in the PPI network, a spatial analysis of the PPI network shows that there is no evident pathway clustering for majority of the diseases suggesting that these diseases have pathways disconnected into multiple regions of disease proteins in different parts of the PPI network.

Network Modularity

It is the portion of edges on average within a specific disease pathways if the edges were assigned arbitrarily removed from the actual fraction of edges within the pathway. This is given by $Q_d = 1/(2m)\sum_{ij}(I((i,j) \in E) - k_i k_j/2m)\delta(pi,pj)$, where k_i is the degree of node i, and $\delta(pi,pj)$ is 1 if p_i and p_j take the same value, else 0. A smaller value indicates that the pathways are well knit communities within themselves, without major interactions with the surrounding network. About 50% of disease pathways have modularity lower 8.6×10^{-5} (-4.06 on log scale), demonstrating the absence of noticeable number of edges within disease pathways.

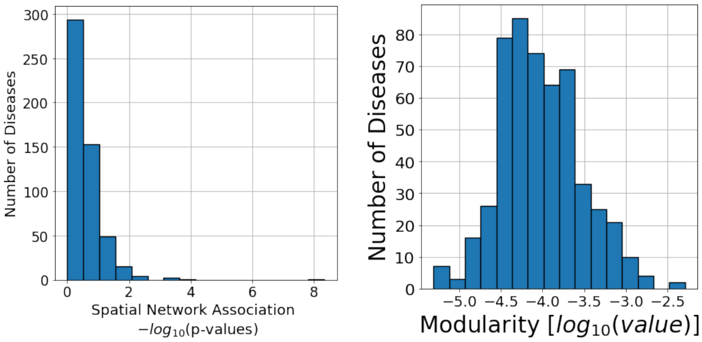


Figure 7: The distribution of (A) Spatial network association and (B) Network modularity

Thus, the quantification of above properties provides an explanation for the performance of the prediction methods seen earlier. It highlights, as expected, that the underlying assumptions made by these methods do not precisely capture the protein's connection with a specific disease. These assumptions not being in agreement with the physical networks lead to failure in giving good predictions.

HIGHER ORDER ANALYSIS

As the closeness of disease proteins in the PPI network is not enough for finding new disease protein, we need to look beyond just edge connectivity for disease nodes prediction into higher-order PPI network structure characterised by motifs (subgraphs recurring within a larger network). Taking into account the symmetry of motif, 73 different orbit signatures exist for 2–5-node motifs. Studies have shown that a specific kind of higher order network exists around disease proteins, indicating that disease proteins are heavily a part of some specific orbits, which hint at the underlying mechanisms they participate in. The signature similarity (4) between all pairs of nodes in the network shows that majority of the diseases nodes pertaining to a particular disease have orbit signatures different from other proteins. Thus, inspite of disease proteins not being close in the PPI network, the diseases significantly represent certain orbit positions indicating that proteins in disease pathway have similar positional roles, inspite of being non-adjacent in the PPI network. One can incorporate this structural information to enhance prediction capability of current methods.

CONCLUSION

Upon studying the PPI network structure of disease pathways, we discovered that disease pathways are loosely knit in the PPI network and clustering within them is not significant. To better understand the limitations of current disease protein discovery methods, we analysed their performance and discovered that the flaw lies in their underlying assumptions about disease pathways being densely clustered; an idea that does not correctly capture the PPI network structure. A significant higher-order PPI network structure detected around disease proteins can be leveraged to boost the performance of current prediction algorithm. These results can form a base for advances in disease protein discovery.

References

[1] M. Agrawal, M. Zitnik, and J. Leskovec, "Large-scale analysis of disease pathways in the human interactome," 2017.

[2] SNAP, "Disease pathways in the human interactome," 2018.

[3] M. Caldera, P. Buphamalai, F. Müller, and J. Menche, "Interactome-based approaches to human disease," 2017.

[4] T. Milenkovic and N. Przulj, "Uncovering biological network function via graphlet degree signatures," 2008.