Part 3.1 - Network Analysis of Personal Genomes

James Diao (*Writing*), Zhaolong Yu (*Coding*), Dingjue Ji (*Pipeline*)

May 8, 2017

1 Instructions

Propose a tool that calculates the degree centrality and betweenness centrality of proteins containing and not containing SNPs in Carl's genome using a PPI file. PPI data can be downloaded from DIP, BIND, MIPS, MINT, and InAct databases.

Writing: Discuss the difference of degree centrality and betweenness centrality you observed. How are these centralities measurement helpful for us to understand different mutations and the protein-protein network.

Coding: Calculate the degree centrality and betweenness centrality of proteins containing and not containing SNPs in Carl's genome using a PPI file.

Pipeline: Use Cytoscape or other softwares to visualize the protein-protein network. Check the centrality calculations with the software and demonstrate one or two examples. Perform hierarchical network analysis and determine if there is enriched or depleted mutation in each hierarchy.

2 Introduction

For the past decade, protein networks have provided valuable insights into molecular evolution and function. New methods of measuring protein interaction have provided the data for bringing network theory and analysis to this important area. Protein networks and their properties are critical for the study of disease classification, personalized medicine, and pharmacology. As available information grows in coverage and quality, protein networks will play an ever-increasing role in the clinical interpretation of genetic variation. In the context of this project, protein networks allow us to better understand the importance of the Carl's mutant coding genes, and evaluate the deleteriousness of such variants.

2.1 Network Theory

Application to Protein Interations: In the theory of protein-protein interaction (PPI) networks, each protein is considered to be a node. Evidence of interaction is encoded as an edge between two nodes.

Degree Centrality: the normalized number of edges, or links, connected to a node. In the context of protein networks, this value encodes how many other proteins a certain protein is connected to. The degree centrality of a single node is its degree (number of edges) divided by (n-1), where n is the total number of nodes. Alternatively, the degree centralities can be computed as the row means (or column means) of an adjacency matrix.

Betweenness Centrality: quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Node with higher betweenness centrality have more control over the network, because more information passes through that node. In the context of protein networks, this value encodes how often a certain protein mediates the closest interaction between two unconnected proteins.

Betweenness centrality is computed as:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(v)$ is the number of those paths that pass through v.

Hierarchical Network Analysis: Hierarchical models treat a network as a nested series of sub-networks. In other words, the lowest level is the original network, and every level above it collapses clusters of the network into individual nodes. These models enable the study of properties between successively larger clusters of nodes. Hierarchical models are intended for scale-free networks, which are the norm in PPI networks.

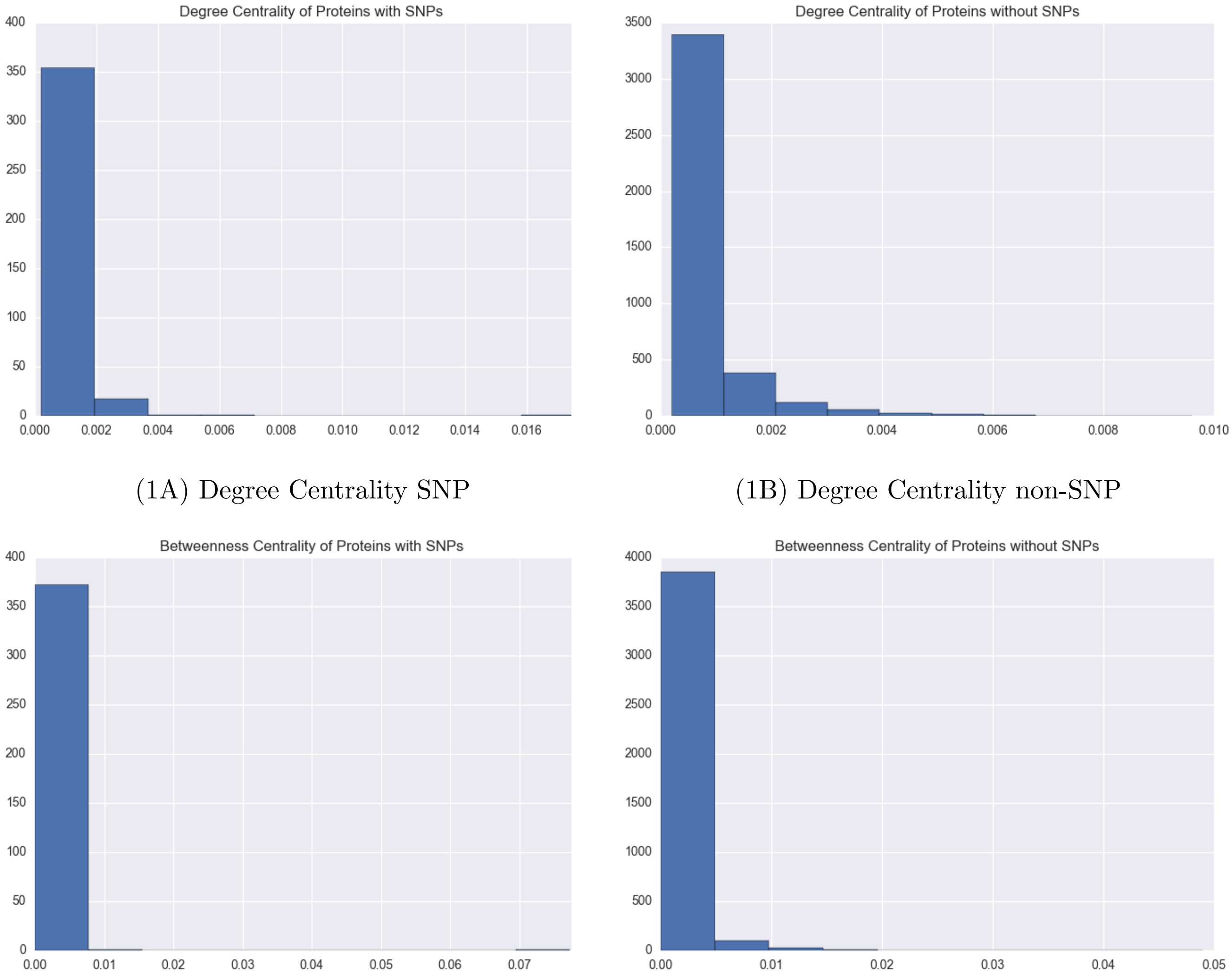
2.2 Databases

There exist many databases that document experimentally determined and theoretically predicted protein-protein interactions. Some commonly used PPI databases include:

- DIP: Database of Interacting Proteins
- BIND: Biomolecular Interaction Network Database
- MIPS: Munich Information Center for Protein Sequences
- MINT: Molecular INTeraction Database
- IntAct: Molecular Interaction Database

3 Coding

We have computed the (1) degree centrality and (2) betweenness centrality of proteins in (A) proteins with Carl's SNPs and (B) proteins without Carl's SNPs. Data was collected from the DIP database including 4,679 proteins. Proteins containing SNPs were extracted from `Z.3DStruct_annotation.txt`.



(1A) Degree Centrality SNP

(1B) Degree Centrality non-SNP

(2A) Betweenness Centrality SNP

(2B) Betweenness Centrality non-SNP

Figure 1: Network properties of proteins with and without Carl's SNPs

It is clear that proteins with SNPs are shifted left on both measures of centrality, meaning that they are both less likely to be connected to many other nodes, and less likely to help connect other nodes. SNPs that interrupt network interactions are more likely to be deleterious and therefore less common in a healthy individual like Carl.

4 Pipeline

We have chosen to use Cytoscape to visualize the protein-protein network. In each plot, the genes with Carl's SNPs are marked in red, and the genes without SNPs are marked in green. The node size is proportional to its degree.

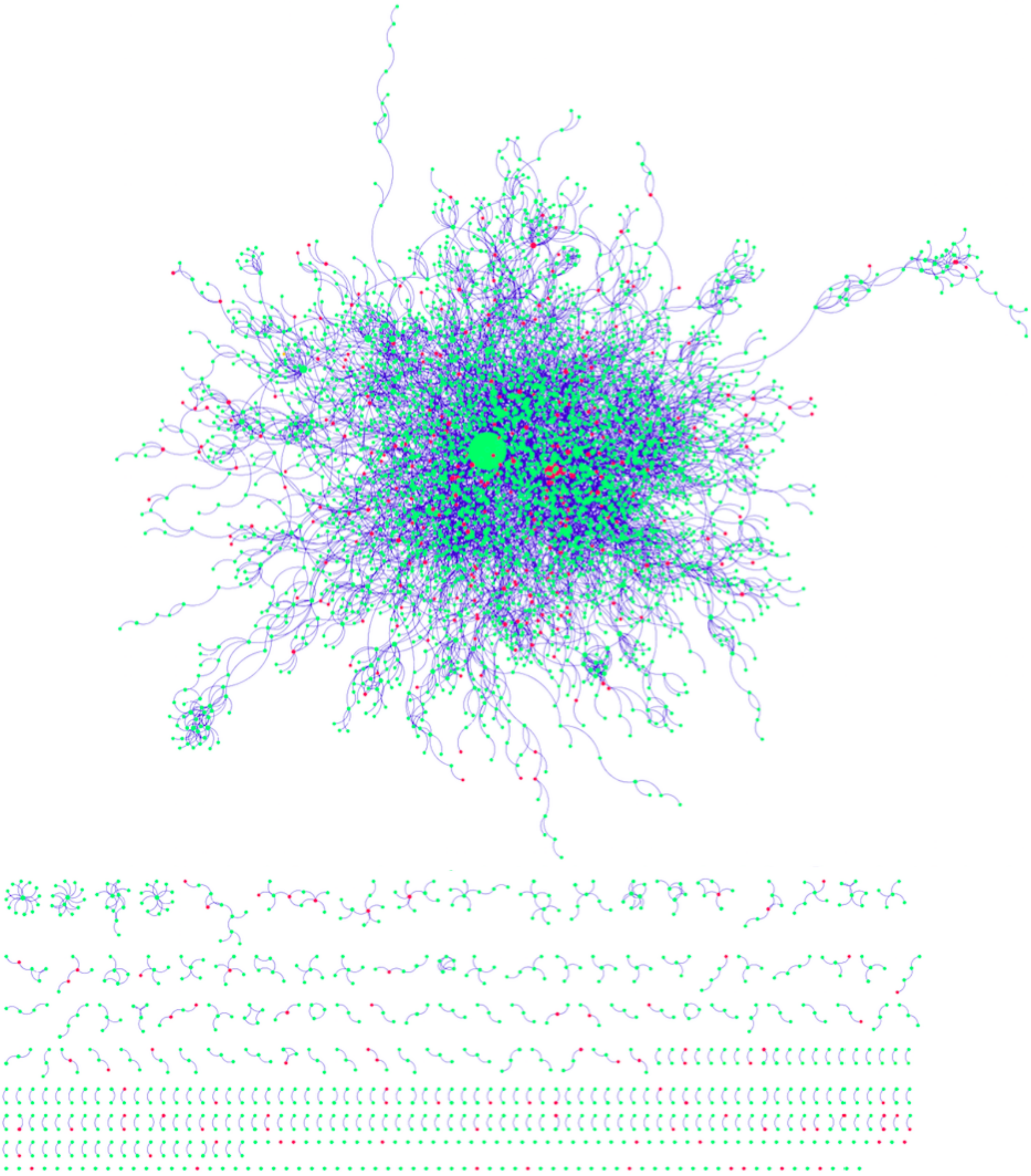


Figure 2: Network from DIP

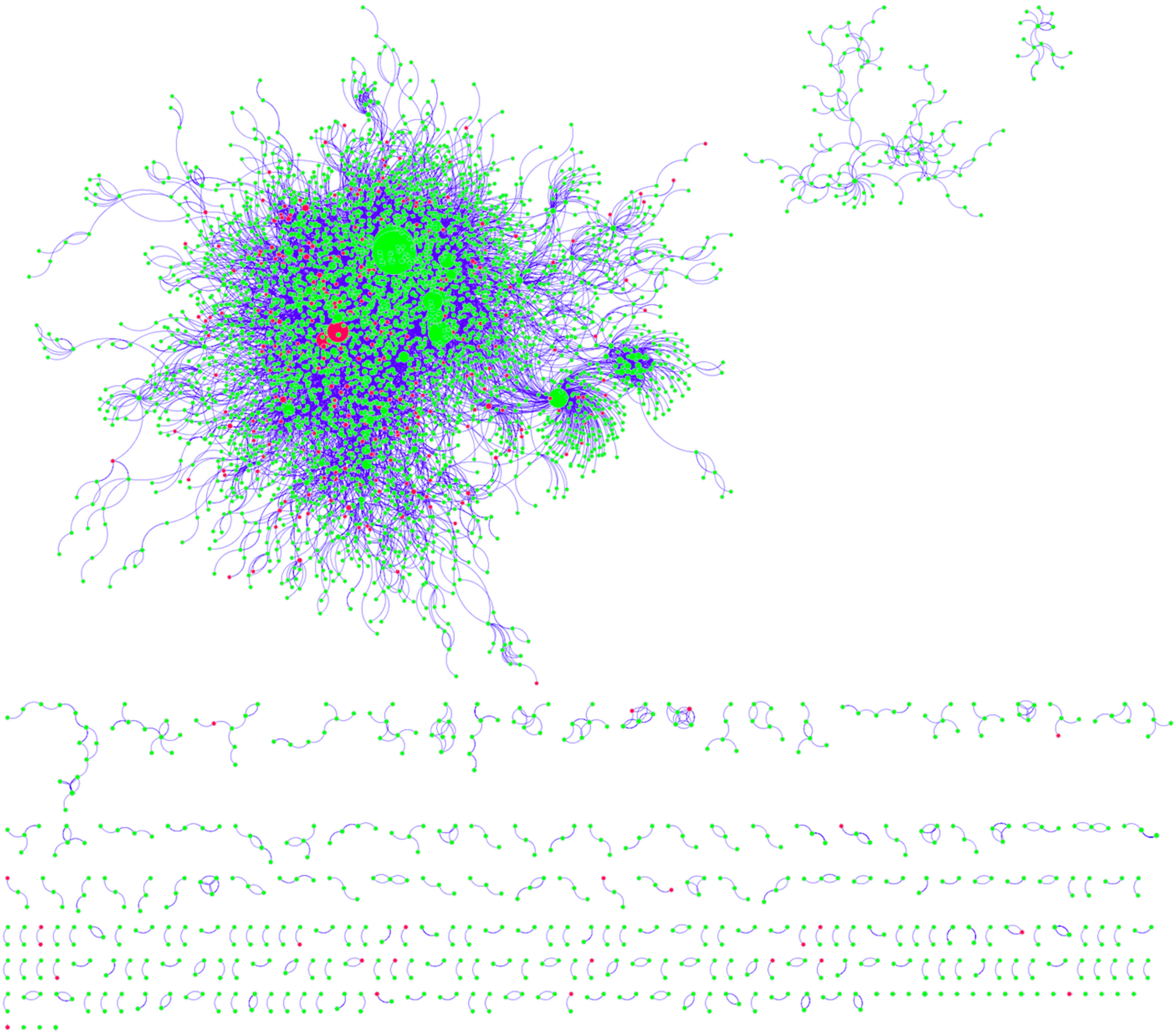


Figure 3: Network from MINT

Because of the large number of nodes, it is not immediately obvious whether the red genes (with SNPs) are enriched at the borders and in smaller nodes, as would be expected. However, by inspecting the details below each map, such a pattern may be deduced.

Cytoscape was used to replicate the histograms for degree centrality and betweenness centrality. This time, paired barplots are used for easy comparison.

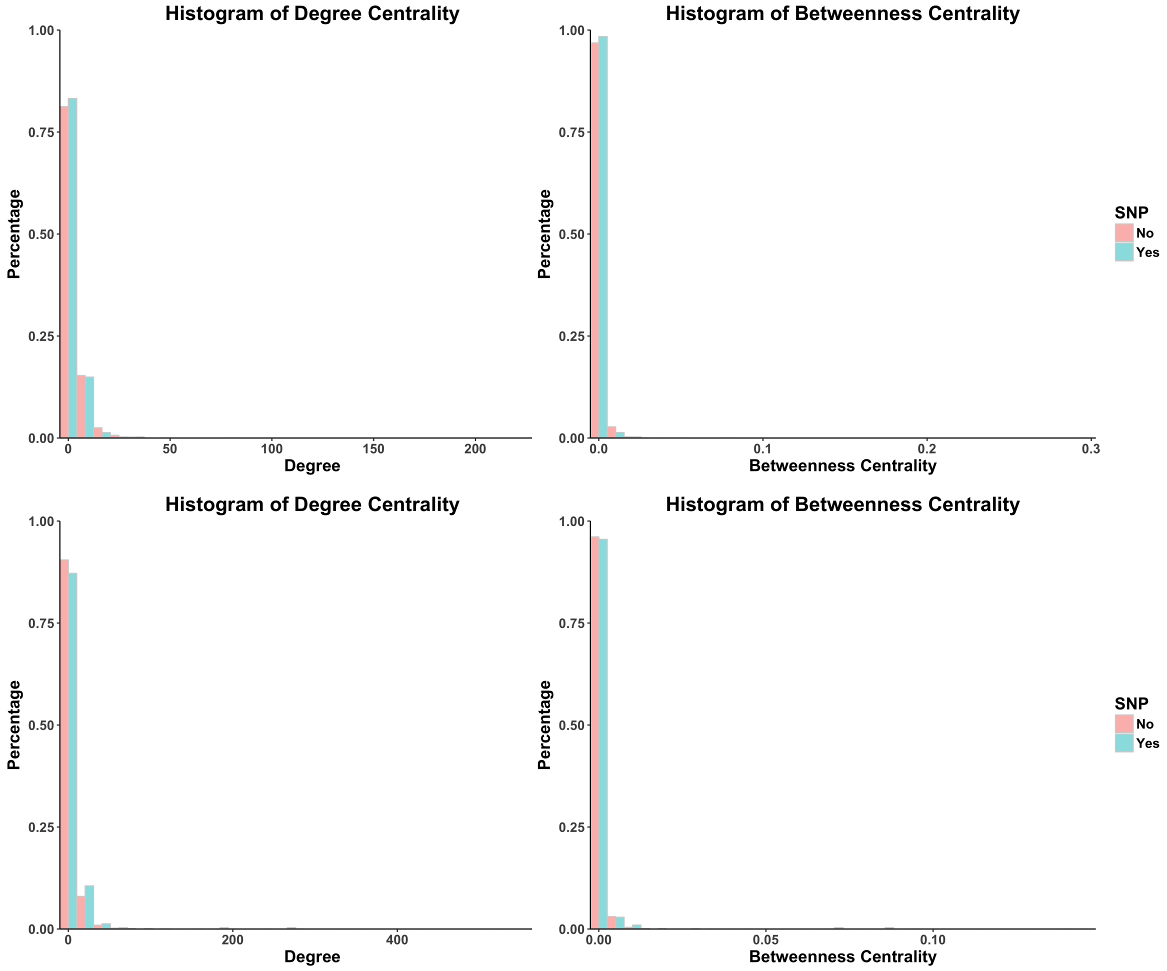


Figure 4: Network properties of genes with and without Carl's SNPs (with Cytoscape)

Just as before, it is clear that proteins with SNPs are more likely to have low values for both degree centrality and betweenness centrality (as expected).

Lastly, hierarchical network analysis was performed to look for enrichment or depletion of mutations in different hierarchies ...

*Note: The gene name mapping is based on ENSEMBL transcript ID - uniprotKB ID pairs. The number of genes recorded in Carl's coding region SNPs file in DIP is 375 and the number in MINT is 306. All the analyses are done by Cytoscape. Histograms are plotted in R ggplot2.

5 Conclusion

A network analysis of Carl's personal genome reveals some properties of his genes with SNPs. By two separate metrics, these genes are quantitatively demonstrated to have lower influence on the overall protein network. The Cytoscape visualization further helps demonstrate this pattern. These network plots also reveal a few genes (the large, red nodes) that are more likely to be problematic based on their positioning (between other nodes) and degree (high). None of these findings can definitely demonstrate pathogenicity, but are valuable for identifying and validating candidate mutations to report.

References

[1] Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101-113. doi:10.1038/nrg1272.

[2] Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18(4):644-652. doi:10.1101/gr.071852.107.

[3] Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-2504. doi:10.1101/gr.1239303.