Angela Sofia Burkhart Colorado

7711 2021, Module 3

Day 3 Written Report

Motivating Problem:

Understanding the relevance of a gene’s impact on a phenotype is extraordinarily important when studying a particular trait. Not all genes associated with a particular trait are created equally. Some genes, should they be eliminated (in other words, knocked out), might cause a complete loss of the observed phenotype, while others may only dimmish its expression. Consequently, quantifying the relevance of potential genes of interest with regards to the phenotype could be useful for further investigation and future hypothesis testing. Subsequently, being able to visualize such rankings is particularly important as it would allow a quick and succinct method of determining potentially more relevant genes, while eliminating others that may have less of an impact.

Computational Problem:

To address this problem, we used a networks-based approach, generating subnetworks of genes that were located in 12 different loci associated with Fanconi anemia (FA), a disease caused by genomic instability. Using multiple subnetworks, we can determine how connected a gene can be, in other words, how important they are to the connectivity of the network. This can then be quantified (in this case, we averaged all edges connected to a gene in each subnetwork) to provide each gene associated with the disease a score. The higher the gene score, the more likely its impact on the disease phenotype it might be significant. This is under the assumption that these genes are more connected to one another than they are to genes outside of the FA loci.

Specific Approach:

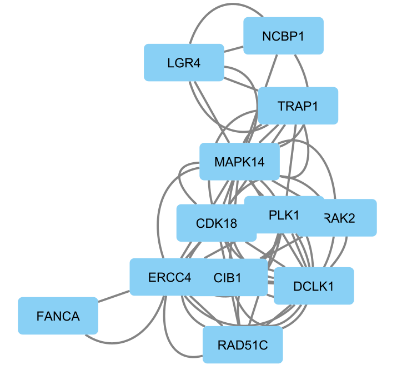
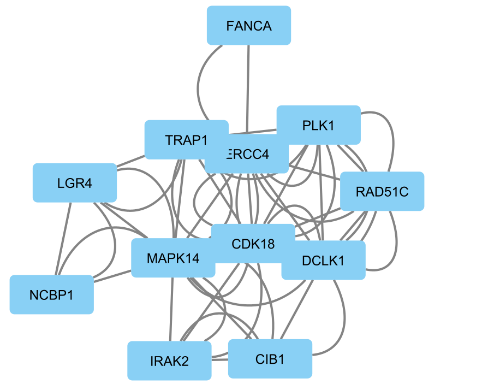
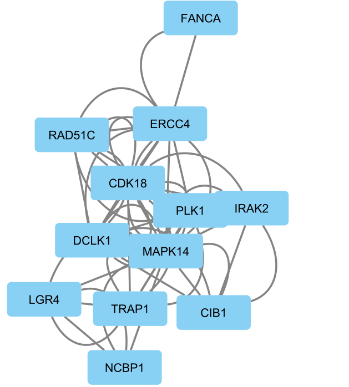
This script has one main function wherein a score of each gene will be calculated for every gene in all 12 FA loci, and the genes with the highest scores from each locus will be visualized. In order to do this n number of subnetworks containing one randomly chosen gene from each FA locus were generated. The number of subnetworks is defined by the user, but the default is 5000 [4], therefore the show results were derived from 5000 subnetworks. For each subnetwork, the number of connections each gene associated with the disease phenotype was calculated. In other words, for each locus all genes at said locus were inserted into the subnetwork and the number of edges that connected to them were counted [1]. One caveat is that Tasan et al. calculated this by counting the total number of edges in the subnetwork, then counting the total edges without the gene of interest (say gene one in locus one) and then subtracting the two totals. However, this is the equivalent of counting the number of edges that are associated with one node [1]. This process was then repeated for all 5000 subnetworks generated. What resulted was a gene associated with 5000 integers representing the number of edges connected to it. These 5000 integers were then averaged to give the gene one score. It must be noted that an assumption is that these networks are directed which may lead to duplicated edges. However, this accounts for the fact that not all genes may have a reversed association (for example, gene one might have a connection with gene two but gene two may not have that connection to gene one). While this may appear unlikely as functional similarity is what is determining the presence of an edge between the two gene nodes, as we are averaging the individual subnetwork scores, proportionally the scores will all be the same. Furthermore, another limitation that may arise from this is the fact that the database from which the connections between the genes were derived only represent the known connections. As a result, genes that may have higher scores, might be those genes that are more well studied. In addition, the weight of the interaction does not play a role in the analysis as well. Some genes are more functionally similar than others, but this is not represented in the code. Consequently, this data is lost because each edge is weighted the same. Finally, once gene scores were calculated the genes with the highest score from each locus were found and a SIF file was written wherein each gene was associated with the genes connected to them in the network. Visualization was then performed in Cytoscape [3].

Specific Implementation:

One consequence of this approach was the extremely slow run time. 5000 trials take around two minutes to calculate. One of the reasons for this is due to the fact that the same dictionary of loci (which are the keys) and the list of genes at each locus (the values) is initialized in multiple functions. If I had had the time to optimize this, I would have generated this dictionary only once and passing it to the functions that needed them, and as a result, this probably would have reduced the run time significantly. This was, in fact, done with another dictionary. However, the usage of dictionaries is extremely useful as they do not need to be sorted or indexed. One dictionary was initialized that contained all genes (keys) in the STRING database and a list of all the genes that were connected (values). When needing to reference the list of connected genes the big O value would be O(N0). Nevertheless, the need to replace every gene at every locus for one subnetwork requires nested for loops which can increase run time significantly.

Furthermore, when generating subnetworks from randomly chosen genes at each FA loci, resulted in very few connections. When looking at edge counts for each of the genes for only ten trials most showed only one connection or none per trial. When looking at the averages over 5000 trials some gene scores were floats less than one. This is interesting because when visualizing genes with the highest scores per locus it might be that those genes had a score only 0.01 greater than the gene with the second highest score. However, when running the algorithm multiple times, I did notice that some genes appeared more frequently, which means that their score remains the highest over more trials. This can be seen in figures 1A, 1B, and 1C where the gene in locus 0 was always PLK1. The ability to see these patterns was one of the reasons why I chose to visualize these particular subnetworks. While it may have been beneficial to see the top 15 genes with the highest overall scores, I would not have been able to confirm whether some genes had consistently higher scores within a locus. Unfortunately, there were two things that I noticed which I was unable to examine further. The first point is that I did not record the gene scores for each gene used in the visualizations. This leads me to the second realization which was that the same genes in all loci were the ones being visualized. While this did not happen initially (suggesting that these genes do indeed appear to score higher than other genes within their respective locus), it is concerning that it did happen four out of seven times. As a result, I cannot deny that there might be a bug I have not caught.

Figures 1: Three subnetworks of the highest scoring gene from each locus associated with Fanconi anemia.

C

B

A

Citations:

1. Discussed possible pseudocode and implementation of problem with Katherina Cortes.
2. Dr. Mazen CPBS7711 lecture. Analysis and Computational design.
3. Talon Arbuckle visualized SIF files in Cytoscape.
4. Taşan, Murat, et al. “Selecting Causal Genes from Genome-Wide Association Studies via Functionally Coherent Subnetworks.” *Nature News*, Nature Publishing Group, 22 Dec. 2014, [www.nature.com/articles/nmeth.3215](http://www.nature.com/articles/nmeth.3215).