Module 1: Day 3 Assignment
Report

Pseudocode

Create a list to store STRING network call it network_edges
read in the network data from STRING1.txt
create a set to store FA genes call it fa_genes
read the input.gmt.txt file and extract FA genes

extract subnetwork of FA genes
create empty list call it fa_subnetwork
for every edge in network_edges
    # network edge structure ('node1', 'node2', interaction weight = 'edge')
    The three statements below are written to see if firs the FA genes are associated with both node1 and node 2, only node1, and only node2. We check node1 and node 2 separately to determine edges, where one gene is an FA gene and the other is a non-FA gene itself but is still associated with FA genes.
    # if both node1 and node2 are FA genes append to subnetwork. This means both genes are associated with FA
    # elif check if node1 is FA gene append to subnetwork. , if only node1 is an FA gene from the gene pair then the other is a non-FA gene but still associated wit an FA gene.
    # elif check if node2 is FA gene append to subnetwork, if only node2 is an FA gene from the gene pair then the other is a non-FA gene but still associated with an FA gene.
    # else create a list that contains the non-FA genes.

Convert the non-FA genes list into a dictionary that removes repeats and converts them into counts. This dictionary will be used when looking at the edges of the subnetwork. The counts will be used to determine how many connections the non-FA genes have to FA genes to determine which non-FA genes are more important to the subnetwork in terms of how many connections they have and how they connect FA genes to each other.

Check length of FA subnetwork and print subnetwork

Check length of non-FA genes and print dictionary with counts

Save the subnetwork to a text file in a 3-column format.
Save non-FA genes dictionary to a text file

Motivation

Fanconi anemia (FA) is a rare genetic disease that mostly affects bone marrow, this leads to a decrease in all types of blood cell production. This disease is caused by genetic instability.

Computational problem

Given a network (STRING1.txt) and an input file of FA genes (input.gmt.txt), visualize a subnetwork with all the nodes and edges joining FA genes.

Specific approach

Given a network of functional linkages and a set of FA genes, implement an algorithm between query nodes and visualize all nodes and edges that are FA genes or genes that are associated with FA genes.

Specific implementation

The specific implementation used was to take the given FA genes and see if they first match both nodes from the network (node1 and node2) and if they do to add them to the subnetwork. Next check node1 and node2 separately, if the genes match at either one of the nodes add them to the subnetwork. Finally create a dictionary that contains the genes that that do not match node1 or node2 (non-FA associated genes).

Results

I was able to create a subnetwork that is very large. First I checked if the given FA genes matched node1 and node2, when I did this I got 2,048 matched. Next, when I also added only node1 and only node2 to my subnetwork it increased in size to 123,364. This is a much larger network than 2048 that only contains FA genes but it also does not account for the edges of the subnetwork that may be important overall, which was the purpose of the dictionary to account for the non-FA genes.

Discussion

The subnetwork is very large but it contains both FA and non-FA genes. This is because the non-FA genes are associated to one or more FA genes and may be the connection between FA genes that are not associated with each other. These non-FA genes also become the edges of the subnetwork that connect non associated FA genes to each other. This subnetwork can always be made smaller by removing the genes that

correspond to only node1 or node2. This new subnetwork would only show genes that were matched with both node1 and node2 but may not show all the possible associations between the FA genes. Overall, subnetworks are important in hypothesis testing, reducing complexity of a very large network, as well as many other purposes.