Fanconi anemia (FA) is a rare genetic disorder characterized by various physical abnormalities and a high predisposition to cancer due to progressive deficiency of bone marrow production[1]. Currently, 23 genes have been associated with FA, all of which are involved in the DNA repair pathway[2]. Creating and visualizing subnetworks have the potential to reveal potential biomarkers and mechanisms related to FA.

To approach this method, a given network and set of query genes associated with FA will be utilized in order to filter the data and determine the subnetwork of FA associated genes. The subnetwork will be composed of nodes (genes) and their edges (significant correlations) and visualized using Cytoscape.

It was critical to define the network and the query genes as two separate variables. The first two elements of each row in the query genes array were removed as the locus number and locus description was not necessary to visualize the associations between the genes in the subnetwork. The subnetwork should identify all FA genes irrespective of whether they are in the same or different loci. The query genes were iterated over the network by checking if the string matches and appended to a list that only included the nodes and edge attributes that match the query genes. If the string did not match the row, the loop will continue until it finds a match to add to the empty list. Lists were chosen as opposed to dictionaries due to the ease of mutating data by splicing, appending and removing as well as it allows duplicates in the array. Dictionaries may have been the better approach to this project as they are more efficient than lists to traverse through the data for a specific element.

Three empty lists were created as a placeholder for the data to be processed during downstream data wrangling. Both the “input.gmt.txt” and “string1.txt” files were defined as separate objects and read in via the with open() function in contrast with open() function where keeping track of closing each file is critical for processing. Due to the fact that “input.gmt.txt” is a tab-delimited file, the strings from each line in the file had to be split by its specified separator, which then would return the object as a list with string elements. Processing the data in this manner would allow each string element to be iterated over the network file later. The new list without the first two elements would then be appended to the empty list *FA_genes* as described above. The “string1.txt” file was also tab-delimited and processed in the same manner as the “input.gmt.txt” file and appended to the empty list *all_genes*. *FA_genes* was then iterated by row through each row in *all_genes* to find matches in string elements. If a match is found, the row from *all_genes* is appended to the empty list *filtered_list*. This list was then exported to a .txt file for visualization in Cytoscape.

Pseudocode:

```
INPUT_ONE data_set
INPUT_TWO string_set

f <- empty list
a <- empty list
s <- empty list

OPEN data_set
    READ data_set
    SPLIT line by tab
    APPEND line to f
CLOSE data_set

OPEN string_set
    READ string_set
    SPLIT line by tab
    APPEND line to a
CLOSE string_set

FOR i in a
    FOR j in f
        IF j is in i
            APPEND i to s
        END IF
    END FOR
```

1. Mehta, P. A., & Ebens, C. (2002). Fanconi Anemia. In M. P. Adam (Eds.) et. al., *GeneReviews*®. University of Washington, Seattle.

2. Velleuer, E., & Carlberg, C. (2020). Impact of Epigenetics on Complications of Fanconi Anemia: The Role of Vitamin D-Modulated Immunity. *Nutrients*, *12*(5), 1355. https://doi.org/10.3390/nu12051355