**Report**

Motivating Problem From Domain

Fanconi anemia (FA) is a rare disease that affects how DNA is translated and repaired. This leads to abnormalities in bone marrow and skeletal structure as well as an increased risk for cancer. Due to causing low red and white blood cells, patients are at an increased risk for anemia, infections, and excessive bleeding.[1] There are 22 known genes that will lead to FA if mutated. By evaluating the density of connections between random FA genes and completely random unlinked genes, it can be determined if FA genes are significantly strongly linked, or if any genes would experience the same strength of connections.

Computational Problem Formation

Compare alternative hypothesis and null hypothesis gene connection data sets using p-value.

Specific Approach

Using the list of FA genes, create 5,000 subnetworks with 1 gene from each loci. Based on the initial subnetworks, null subnetworks can be created using random genes from the STRING input file that have the same number of connections as the original gene. The densities of connection strengths between each set of subnetworks can be compared using a p-value.

Specific Implementation of Approach

Starting with a set of FA genes from OMIM[2] and a STRING input of all gene linkages, every gene from the STRING input is sorted into bins based on their total number of connections. Using randomization, 5,000 unique subnetworks of just FA genes using 1 gene from each loci, 12 genes total, can be generated. Another set of 5,000 unique subnetworks are created by swapping each gene from the original subnetwork with a random gene that is in the same connection bin as the original. After both sets of subnetworks are created, the average densities of each subnetwork is determined using the edge weight of each connection. The average densities between the two sets are then compared using a permutation test to determine if the strengths of the connections are unique to FA linked genes or would happen randomly.

**Pseudo-Code**

1. *Sort genes into connection bins:*

```
Connection_bins = {}
Unique_connections = set()
For gene_connection in STRING:
        Add gene_connectin to unique_connections
All_genes = [list of all genes]
Gene_counts = []

For each gene in unique_connections:
        Add gene to gene_counts

For gene in all_genes:
        Num_connections = Count occurrences in gene_counts
        If num_connections key exists in connection_bins:
                append gene to bin
        Else:
                Create new bin
```

2. *Make 5,000 FA subnetworks:*

```
Initial_subnetwork = {}
For i in range(5,000):
        Network = []
        For locus in FA_gene_file:
                Randomly select a gene from locus
                Append gene to network
        Make sure network is unique and add to initial_subnetwork dictionary
```

3. *Make 5,000 null subnetworks*

```
Null_subnetwork = {}
For subnetwork in all_FA_subnetworks:
        New_subnetwork_genes = []
        For gene in subnetwork:
                Find gene in connection_bin dictionary
                Select random gene from connection_bin from same key value
                new_subnetwork_genes.append(new_random_gene)

        Null_subnetwork[subnetwork_number] = new_subnetwork_genes
```

4. *Find p-value*

```
Null_subnetwork_densities = {}
FA_subnetwork_densities = {}
For subnetwork in FA_subnetworks:
        Densities = [list of edge weights for each connection in subnetwork; if both genes
                        from line in STRING is in subnetwork]
        Avg_density = sum(densities)/12
        FA_subnetwork_densities[subnetwork_number] = avg_density

Repeat for loop for null_subnetworks

All_densities = [combine FA densities and null densities]

FA_mean = mean of all 5,000 mean FA densities
Null_mean = mean of all 5,000 mean null densities
Actual_diff = FA_mean - null_mean

Permutation_values = []
For i in range(10,000):
        Randomize all_densities
        Split up all_densities into two groups
        Find mean of group1 and group2
        Perm_diff = Mean_group1 - mean_group2
        permutation_values.append(perm_diff)

Extremes = [Count value for value in permutation_values if value >= actual_diff]
P_value = len(extremes) / 10,000
```

## Discussion

Results

5,000 random subnetworks of 12 genes with 1 FA gene from each FA loci were successfully created. An additional null set of 5,000 random genes from the same connection bins[3] as the initial 5,000 FA subnetworks was generated. The alternative hypothesis was that average edge weight densities of the FA subnetworks will be significantly different than the average edge weight densities of the null subnetworks. The null hypothesis was that the average edge weight densities of the FA subnetworks will not be significantly different than the average edge weight densities of the null subnetworks as the edge weight densities are due to random chance.

To test these hypotheses, the average edge weight density of each 10,000 subnetwork was found and the two sets were compared through a permutation test. Using 10,000 permutations, a p-value of 0.0 was calculated. A p-value of 0.0 shows that there is no other combination of data that gives a more extreme t-statistic than the original data. This supports the alternative hypothesis that the densities of the FA subnetworks are significant and specific to the FA genes, and would not occur due to random chance. This also theoretically makes sense as FA genes have been shown to be linked, so would be more likely to have a higher density of edge weights compared to random and potentially unlinked genes.

Alternatively, another permutation test was done using the average number of total edges as opposed to the edge weight. Using 10,000 permutations, a p-value of 0.0011 was calculated. This shows that there was a 0.11% probability that a permutation would result in a more extreme difference in mean between the null and FA set. This p-value also supports the alternative hypothesis. This p-value being larger than the p-value of the edge weights is consistent with the understanding that while a set of subnetworks can have the same number of edges, the strength of those edges could highly differ. In FA subnetworks, it is likely the subnetworks will have a higher proportion of both total edges and edge strength, which could result in a larger difference between the null set mean and the FA set mean. Using the edge weight takes both these factors into consideration. Using just edge count increases the possibility a permutation would have the same or more extreme difference.

Limitations, assumptions, and improvements

The assumption that is being tested is that FA genes are more likely to have more and stronger connections compared to random genes from the STRING input file. To account for this assumption while making a null set, a way to normalize the number of connections was used. Every FA and non-FA gene was sorted based on their total number of connections into “bins”. When generating the null subnetworks, an FA subnetwork was used as the template in which the FA genes were swapped with random genes within the same bin as the original FA gene. This ensures that each set of subnetworks have the same possibility for connections. While making

the connection bins, it was also assumed each FA gene would appear at least once in the STRING file. In testing, there were FA genes that had no connections to any other genes, in which case this FA gene was replaced with a random gene from bin 1. To further account for this, a check to make sure the FA gene used in the initial subnetwork is also present in the STRING file. However, this would result in disregarding some genes.

Assumptions about the data were made that could potentially affect results. It was attempted to take duplicate gene pairs from the STRING file into consideration. Exact duplicate lines from the STRING file were sorted out. However, it is possible there were instances where gene pairs would be repeated with different edge weights. As the different edge weight would cause the line to be unique, those values were kept. When finding the average densities of each subnetwork, it was assumed each unique gene pair would show up once. If there was a case where a gene pair connection was in the STRING file with two different edge weights, both edge weights would be counted resulting in a higher average density. This could be avoided by combining the two edge weights by either using the higher or lower of the two densities, or the average.

While creating permutations, a randomization function was used to randomly shuffle all the data values 10,000 times before being re-assigned into two groups. It was assumed that each of these permutations were unique, there was no check to ensure that the permutations were unique due to repeating values making each permutation difficult to distinct. In the future, a more advanced permutation algorithm that iterates through each possible permutation should be implemented.

**Sources**


1. National Center for Advancing Translational Sciences |. rarediseasesinfonihgov. https://rarediseases.info.nih.gov/diseases/6425/fanconi-anemia.

2. OMIM - Online Mendelian Inheritance in Man. omimorg. https://omim.org/

3. Taşan, M.; Musso, G.; Hao, T.; Vidal, M.; MacRae, C. A.; Roth, F. P. Selecting Causal Genes from Genome-Wide Association Studies via Functionally Coherent Subnetworks. *Nature Methods* **2014**, *12* (2), 154–159. DOI:10.1038/nmeth.3215.