Keegan Shudy
Nick Mulvay

Final Project Progress Report

Since our project is heavily rooted in coding, the majority of this project report will consist of the algorithm design and accomplishments thus far in the project. Below is a representation of the algorithm design of the system as a whole. At this point, there is the potential for the system to transform from a query-result type of system to a input-display type of system. This input-display architecture would take human gene's as an input and display all possible diseases that we test for as well as their associated risk factor. If this change is implemented, it would take out the mechanization portion of finding a clean FASTA file for comparison. This change would make for a less dynamic system but would be easier to implement in the long run making this a viable substitute for the current system if the need were to arise.

As for the code written, the entire mechanization (collecting the clean FASTA from DisGeNET) is almost at a completed stage but requires more testing in order to assess its accuracy. The next step is to work on scrapping the links from DisGeNET to NCBI and then obtaining the FASTA from NCBI. Also, the transcription algorithm is finished and the alignment (will most likely be taken from biopython) is handled with the potential to a small adaptation.

```
                    ┌─────────────────────┐
                    │ Take User Disease   │
                    │ Query               │
                    └─────────────────────┘
                              │
                              ▼
                         ╱────────╲
                        ╱  Query   ╲         N
                        ╲ indexed? ╱ ──────────────────┐
                         ╲────────╱                    │
                              │ Y                       │
                              ▼                         │
                    ┌─────────────────────┐             │
                    │ Query DisGeNET for  │             │
                    │ disease umls        │             │
                    └─────────────────────┘             │
                              │                         │
                              ▼                         │
                    ┌─────────────────────┐             │
                    │ Link follow from    │             │
                    │ DisGeNET to ncbi    │             │
                    └─────────────────────┘             │
                              │                         │
                              ▼                         │
                    ┌─────────────────────┐             │
                    │ Obtain Healthy Gene │             │
                    │ FASTA_C             │             │
                    │ Index it            │             │
                    └─────────────────────┘             │
                              │                         │
                              ▼                         │
                    ┌─────────────────────┐             │
                    │ Obtain Gene FASTA_I │ ◄───────────┘
                    │ from user           │
                    └─────────────────────┘
```

Obtain Healthy Gene $FASTA_C$ Index it

Obtain Gene $FASTA_I$ from user

$Align(FASTA_C , FASTA_I)$

```
        ╱────────╲         N      ┌──────────────┐        ╱────────╲    Y    ┌──────────────────┐
       ╱          ╲ ──────────►   │ Transcribe   │ ────► ╱ Mutants  ╲ ─────► │ Calculate and    │
       ╲ Score==100╱               │ both FASTA   │       ╲          ╱        │ display risk     │
        ╲────────╱                 └──────────────┘        ╲────────╱         │ Score            │
            │ Y                                                 │ N           └──────────────────┘
            ▼                                                   │
   ┌──────────────┐                                             │
   │ Alert no risk│ ◄───────────────────────────────────────────┘
   └──────────────┘
```