**Restriction Enzyme Mapper**

The web application titled “Restriction Enzyme Mapper” serves as a bioinformatics annotation tool and restriction digest mapper. During microbiological research, organisms and their genes are studied for their mechanisms of action and associated phenotypes. The driver of this research involves the use of plasmids which are circular DNA constructs that can be manipulated to amplify a phenotype of interest. The laboratory process involves dissecting this circular DNA and inserting, deleting, or replacing genes within the plasmid itself. Once this process has occurred, colonies are grown and the colonies which show successful integration/deletion of a particular gene are harvested further for research.

In order to manipulate the plasmid, the DNA itself needs to be dissected. This is achieved with the use of restriction enzymes which “digest” the DNA at locations with specific nucleotide patterns, referred to as “digestion sites” with respect to this project. There are many different restriction enzymes, each which recognize and cut at a specific nucleotide pattern. The goal of this application is to allow the user to enter a plasmid accession number and return the plasmid’s sites of digestion, as well as which enzymes would be most compatible at cutting between genes.

Some restriction enzymes are more ideal than others when analyzing particular genomes. An enzyme might cut between genes, but there might also be other sites on the plasmid which occur right in the middle of a gene, interfering with that gene’s function and thereby interfering with the results of the experiment (is the resulting phenotype a product of the inserted gene or a product of a gene made inactive through digestion?). This tool filters out these enzymes with the analysis portion. All “compatible sites” or digestion sites which occur between genes (no interference) are displayed to the user. Optimal sites or sites where the enzymes involved only cut at compatible sites are also displayed to the user. This allows users to view the effectiveness of using certain enzymes over others depending on the optimal and compatible sites. The application also lists gene information for upstream and downstream genes of a given digestion site, as well as a gene table with annotated information from NCBI.

The development of this project involved a lot of python modules and was very resource intensive. At first, it seemed more prudent to parse the FASTA files line by line as opposed to creating a large string and then analyzing the string. However, wrapping the lines of the file and trying to manage the location of matching sites became more difficult and inaccurate than expected. It became more prudent and effective to store the FASTA data as a string and use the power of python’s regex module to locate all occurrences of specific digestion sites. This project also used several tables in the MySQL database. This involved an organism table, a gene table which links to the organism table, and a site table specific to each organism which also links to both the gene and organism tables via the GI number.

Many connections and commits are performed through the web application. In an effort to improve efficiency, the restriction enzyme information was queried once and stored in a dictionary (as opposed to making multiple connections within the main loop). The dictionary was used as a reference when iterating over the genes. In another effort to improve efficiency, a digestion site table was created for each individual organism as opposed to storing all sites in one large table. The large table became problematic as more organisms were uploaded to the server and this approach helped cut down analysis time. Genes were also selected instead of digestion sites as the driver of the main iteration.

There were also many utility methods which were created to help the development process. This involved a script which allowed a restriction enzyme list to be copy and pasted into a text file from addgene.org and then formatted in a way which could be easily used by the application (“cleanup.py”). This generated a file which could be uploaded to the database by another script. During development, debugging scripts with insert statements became difficult to manage. A script called “deleteOrg.py” was created to clear out and reset the database and came in handy quite often. Of course, all of the main scripts such as uploading FASTA and GenBank information and conducting the analysis had their own testing versions. These versions wrote to text files and printed information to the terminal. The use of these scripts aided in the debugging process substantially.

The overall limits of the application in terms of how much it can take before generating a Gateway Timeout Error in the browser vary. Originally, 96 restriction enzymes were used in the application. However, even with the slight improvements made in the code mentioned previously, this limited the size of plasmids to around 300kbp. Therefore, amount of enzymes was decreased three fold, with the amount of enzymes in the database currently at 37. This improved the load capacity by nearly eight fold. Now the application can handle plasmids with 2.5-3 million base pairs.

Overall, I learned how to modularize my code more effectively and look for ways of increasing efficiency when dealing with extremely large sets of data. I also learned how to utilize the power of a database and how to build sub-scripts to help automate and provide support in the development process. I also learned how frequently a lot of restriction enzyme patterns occur in the plasmids listed in NCBI’s nucleotide database. The amount of enzymes which cut in the middle of genes present a unique problem where the right enzymes must be used to ensure that a quality experiment is performed and no genes are interfered with. This also makes sense of why so many organisms have different versions of plasmids and vectors. Many vectors contain gene inserts or deletions which serve as phenotype markers when trying to grow colonies. These vectors also contain unique and distinct restriction sites which allow for optimal digestion.