Project Background

In the midterm assignment, we had to select a single prokaryotic gene prediction tool. I used Prodigal as I read an article comparing multiple tools and it seemed to be one of the better ones in terms of speed and accuracy. However, I have been very curious about the relative performance as I did not compare them myself.

Tool Functionality:

I am proposing a tool that will accept a user input of an *Escherichia coli* plasmid sequence (limited to E. coli plasmids to avoid excessive file size) and accession id. The sequence will be evaluated by Glimmer, Prodigal, GeneMark, and MetaGeneAnnotator. The predictions will then be compared to the GenBank annotation for the given plasmid to determine the accuracy of the results. The output to the user will provide metrics including total genes predicted by each tool, number of exact matches, number of predictions with either 5' or 3' agreement, number of predictions with no agreement, accuracy as determined by performance on the previously stated metrics, and time taken for analysis. These metrics will be displayed graphically.

Technical Specifications:

To start, I am somewhat unsure if this is actually feasible as I will need to run shell commands from my CGI script. This is possible with the Python subprocess module, but I am unsure if the server will allow it. I will proceed with this section under the assumption that this will be possible.

The database schema will consist of six tables. The first table will contain the reference annotations split into columns of accession number of associated plasmid, protein ID, gene start position, and gene end position. This will primarily be used for retrieving information necessary for the comparison metrics. The second table will consist of the plasmid sequences and the final metrics related to each tool. It will have columns for accession number and sequence. Additionally, the following columns will exist for each tool: total genes predicted, number of exact matches, number of predictions with 5' or 3' agreement, number of predictions with not agreement, accuracy, and runtime. The first two columns will be prepopulated, but the remaining columns will populate over time as users run each sequence. This table will be used to confirm user input and store prediction data to avoid repeat runs. The third, fourth, fifth, and sixth tables will all serve the same purpose of storing data related to the predictions obtained from each gene prediction tool. They will

consist of columns for associated accession number, predicted gene start, predicted gene end, and confidence score.

The CSS/HTML/JavaScript GUI will accept user input of a plasmid sequence and the accession number associated with it. This information is sent to the python code which will run the gene prediction tools through the subprocess module which can run shell commands. The output data will be read from each tool's output file and parsed for the necessary metrics. Additionally, runtime will be collected at this point. Finally, accuracy will be calculated then all metrics will be returned to the GUI via a JSON message. I intend to write custom JavaScript code to visualize the metrics as graphs as the raw numbers will be somewhat overwhelming.

Resources:

I used https://ccb-microbe.cs.uni-saarland.de/plsdb/plasmids/ to obtain the GenBank annotations for the E. coli plasmids.