Description of Project/ Significance

The construction of bacterial phylogenetic trees is extremely useful in showing evolutionary relationships and genomic diversity among bacterial strains and species. Given the high evolutionary rate of bacteria, phylogenetic tree construction can be a helpful tool in genetically understanding bacterial pathogenesis and virulence strategies of healthcare-adapted bacterial pathogens. *Pseudomonas aeruginosa* (*P. aeruginosa*), a common opportunistic pathogen within the healthcare setting, has been shown to rapidly evolve new mechanisms of virulence, depending on host condition (**Figure 1**). *P. aeruginosa* isolates extracted from patients can have large genomic diversity from isolate to isolate potentially contributing to why some strains have increased infectious outcomes. This project aims to 1) construct a phylogenetic tree of publicly available patient-derived *P. aeruginosa* isolates, and 2) show the evolutionary relationship and genomic diversity between each isolate.

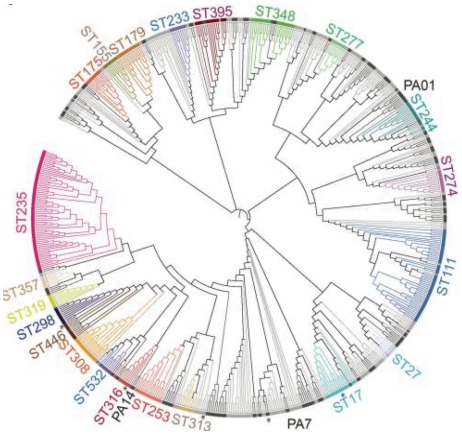

Figure 1 : Phylogenetic distribution of difference sequence types (ST) of *Pseudomonas aeruginosa*.
Reference: Belkum *et al.* 2015, PMID: 26604259

Overview of features/stages/components to be implemented

The major focus of this project is to implement key coding and python strategies utilized into stages: 1) using publicly available *P. aeruginosa* genomic sequences, 2) aligning bacterial genomic sequences (using multiple sequence alignments), and 3) constructing and visualizing a bacterial phylogenetic tree (similar to that described in **Figure 1**). Most of these stages involve using Biopython import packages that help in phylogenetic tree construction and visualization.

Code workflow

For the majority of this project, python function coding would be helpful (but not necessary). After obtaining *P. aeruginosa* whole genome FASTA files (typically 6.3 Mb in size/ea and should be stored in their own directory), the first goal would be to align genomic sequences with help from MUMer (a conda-derived multiple alignment software package; https://github.com/mummer4/mummer). After successful installation, you can align desired genomes (i.e., different *P. aeruginosa* patient-derived isolates) using python scripting (opening, reading, and parsing through .fasta files, etc). Results should include SNP differences or core-genome alignments which will be used for phylogenetic tree construction. Tree construction can be done with IQ-Tree, an optimized software package that can be called in python and used for better developing phylogenetic trees of whole genomes.

Desired Inputs

The main inputs for this project include publicly available *P. aeruginosa* strains (obtainable through 'The Pseudomonas Database', sponsored by the Cystic Fibrosis Foundation; www.pseudomonas.com). *P. aeruginosa* clinical strains can be searched using the "current strain" toolbar, by searching *'Pseudomonas aeruginosa'*. Different *P. aeruginosa* strain genomes (i.e., FASTA files) can be accessed using their corresponding accession number. Laboratory adapted *P. aeruginosa* strains can be used as a control (ex: PAO1 and/or PA14). Other desired inputs include Biopython import packages (such as SeqIO and Phylo). Whole genome alignment packages include MUMer and NUCmer, both of which are extremely fast at aligning whole bacterial genomes.

Desired Outputs

By the end of this coding project, outputs should be 1) generating a successful alignment of n=20 *P. aeruginosa* hospital-acquired genomes, 2) creating a successful visualization and description of genomic diversity and evolution among bacterial pathogenic isolates, and 3) building upon python scripting tools and workflows.

Potential Challenges

Each stage of this project may have challenges. For example, the output from one stage is needed for the input in the next stage. Thus, errors in scripting and output quality can significantly impact project progression.