Classification of Burkholderia Species Using Random Forest for Detection of Pathogenicity

Lee, Dennis (School: Starkville High School)

The bacterial species of Burkholderia are renown by scientists to be extremely difficult to distinguish. As a result, melioidosis, which is one of the many diseases caused by Burkholderia, has 165,000 cases and 90,000 deaths each year. All bacteria have proteins, which are composed of unique sequences of amino acids. The goal of this study is to develop an automated way to classify a Burkholderia strain as harmful to humans or not by looking at its recombinase A protein sequence. The Random Forest machine learning model was used to analyze a protein's amino acid sequence and find patterns that determine whether the particular strain is pathogenic to humans. The dataset consisted of protein sequences from 23 different species of Burkholderia which were inputted into the random forest algorithm. After 150 training sessions, a number of unique patterns were found. For example, if the amino acid serine is at the 229th position of the amino acid sequence, this strain likely pathogenic to humans. This novel method allows for a new and efficient way to classify harmful Burkholderia strains and could be implemented in a way that helps prevent people from contracting harmful Burkholderia related infections like Melioidosis.