Predicting Mycobacteriophage Gene Start Sites Using Artificial Intelligence

Lohano, Sarisha (School: The Carol Martin Gatton Academy of Mathematics and Science in Kentucky)
Norman, Bella (School: The Carol Martin Gatton Academy of Mathematics and Science in Kentucky)

When exposed to antibiotics for long enough, a bacterial population can evolve to become resistant. This prompts the creation of innovative therapeutics such as bacteriophage therapy. 'Phage Therapy', using doses of bacteriophage to combat infections, may be the new alternative to antibiotics. This research project aims to contribute to the genomic analysis of mycobacteriophages, a group of bacteriophages that infect Mycobacterium. Using Mathematica, a computational software created by Wolfram, GenFind - a gene prediction program - was made. To make predictions, a Support Vector Machine trained with the introns and exons of Mycobacterium tuberculosis and Mycobacterium smegmatis was implemented. A scoring matrix was created using codon frequency in coding versus non-coding regions in M. tuberculosis and M. smegmatis. A query sequence, normalized by the scoring matrix, can be fed into a Support Vector Machine, which returns a numerical confidence score in both 'gene' and 'non gene' predictions. Another Support Vector Machine is used to score and classify the Shine-Dalgarno ribosomal binding site upstream of the start codon of an open reading frame. A full genome analysis finds every open reading frame in all 6 reading frames and displays only those that have a confidence score of 97.5% or greater. A single gene analysis does not include elimination of open reading frames with low scores. Instead, a display of the coding capacity and confidence scores are shown. This research has potential to aid in the discovery of cytotoxic genes and annotation of novel bacteriophage that can be used in phage therapy.