Utilizing Unsupervised Machine Learning to Derive and Analyze Therapeutically Insightful Regions of Bacteriophage Genomes

Rosenberg, William (School: Episcopal School of Jacksonville)

While bacteriophages show strong potential as therapeutics for bacterial infections, too little about their genomic structure and content is understood for widespread implementation. The purpose of this investigation was to leverage machine learning in order to identify meaningful patterns in the phage genome–specifically, shared regions (clusters) of genes–and utilize them to gain a better understanding of it. These shared regions are very telling as they are potential operons, providing both a more accurate and more efficient framework for gaining therapeutic insight. Data was derived through web scraping the government database Genbank, and a machine learning environment was created with Python and Conda. Pandas data science tools, partitional clustering, dimensionality reduction, and hierarchical clustering (all unsupervised) were leveraged to pull out clusters. K-means clustering was used to create optimized subsets of the training examples, PCA was used to optimize the number of clusters, and feature agglomeration, combined with Pandas functionality, was used to pull out the clusters. All clusters were saved into a dynamically-updating dictionary. Clusters were then analyzed by I-TASSER to predict structure, and then cross-referenced with BLAST data to predict operon function. Subsets of 4-15 phages across a variety of hosts were inputted into the algorithm; the outputs matched hand-clustering test results with expected accuracy. Single phages were also inputted and broken down into clusters that were saved in the dictionary. Clusters ranged in size from 3 to 15 genes, and were present in both the left and right arms of the phage genome. This both indicates the evolutionary significance of the clusters and warrants further genomic exploration with this approach.