Application of K-Means and Hierarchical Agglomerative Machine Learning Algorithms to Cluster Wolbachia Genomes based on Host Organism's Phylum

Lee, Sean (School: Loomis Chaffee School)

Wolbachia, a species of endosymbiotic intracellular bacteria found in a diverse set of species, is most notable for its role in vector control strategies to reduce the spread of diseases that are transmitted by mosquitoes, such as malaria and Dengue fever. Despite this significance in disease prevention and decades of research, the exact mechanism of Wolbachia’s parasitism at the genomic level is not fully understood. Searching for patterns in Wolbachia’s genome sequence with machine learning algorithms may offer insight into its mode of action. This project utilizes both the k-means and hierarchical agglomerative clustering machine learning algorithms to explore how Wolbachia’s genome sequence is influenced by its host organism’s phylum. From the NCBI database, Wolbachia genomes from both arthropod and nematode hosts were collected then preprocessed to create a composition matrix of 5-mer frequency counts. Cluster analysis was performed on this data and was visualized using Voronoi diagrams and dendrograms, both of which showed stark clusters of genomes organized by the host organism’s phylum with a high accuracy of 88.5%. These findings indicate a strong correlation between Wolbachia genomes and its host, suggesting that Wolbachia species adapt their genetic sequence based on its host organism’s phylum, and thus play differing symbiotic roles. This result is promising in enhancing the understanding of Wolbachia’s parasitic mechanism and also offers potential future studies, such as searching for the specific genes that differ for Wolbachia in the various hosts and discovering other beneficial applications of Wolbachia outside the field of disease prevention.