Probabilistic Hierarchical Clustering for Biological Data

Data Science for Biology: Hierarchical Clustering of Adenovirus Codon Usage

Adenoviruses are a family of viruses that infect a wide range of vertebrate hosts, including humans. Understanding the genetic and genomic features of adenoviruses is an active area of research in computational biology and bioinformatics. In this report, we explore the application of hierarchical clustering to analyze the codon usage patterns in adenovirus genomes.

Codon usage refers to the frequency with which different codons (triplets of nucleotides that encode amino acids) are used in the protein-coding regions of genomes. Analyzing codon usage patterns can provide insights into the evolutionary pressures and constraints shaping the genome composition of viruses.

Data and Methods

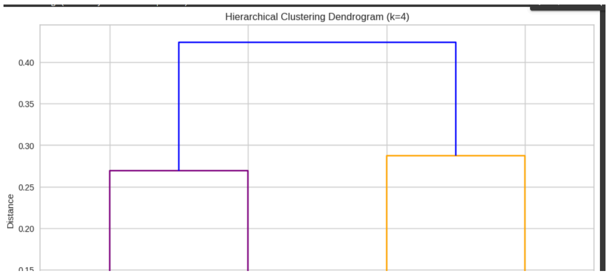
We used a dataset of adenovirus codon usage obtained from the Codon Usage Database (CUD). The dataset contains the codon usage frequencies for 64 different adenovirus species. We performed the following analysis steps:

1.Data Preprocessing: We loaded the adenovirus codon usage data into a pandas DataFrame and performed log-transformation to normalize the values.
2. Hierarchical Clustering: We applied hierarchical clustering to the preprocessed data using the Euclidean distance metric and the Ward's linkage method. This allowed us to identify groups of adenovirus species with similar codon usage patterns.
3. Cluster Visualization: We generated a dendrogram plot to visualize the hierarchical clustering results. We also created a heatmap to represent the codon usage patterns across the different adenovirus species.
4. Cluster Evaluation: We assessed the validity of the identified clusters by calculating the inter-cluster and intra-cluster distances. This provided insights into the compactness and separation of the clusters.

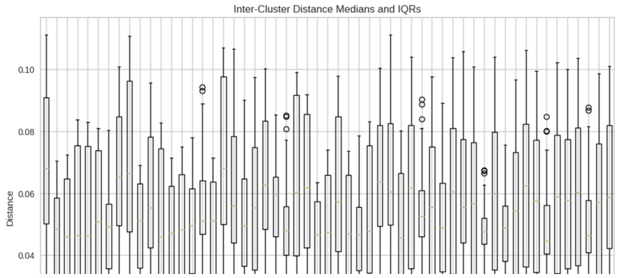
Results

The hierarchical clustering of the adenovirus codon usage data resulted in the identification of four distinct clusters. The dendrogram and heatmap visualizations revealed the following insights:

1. Cluster Structure: The four clusters showed clear separation, indicating distinct codon usage patterns among the adenovirus species.



2. Cluster Distances: The analysis of inter-cluster and intra-cluster distances confirmed the validity of the identified clusters. The inter-cluster distances were larger than the intra-cluster distances, suggesting that the clusters were well-separated.



3. Biological Interpretation: The clustering patterns could be related to factors such as host specificity, tissue tropism, and evolutionary divergence among the adenovirus species. Further investigation would be needed to establish these connections.

Discussion and Conclusion

The hierarchical clustering of adenovirus codon usage data provided a data-driven approach to explore the genetic diversity within this virus family. The identified clusters suggest the existence of distinct codon usage signatures that may be linked to the biological characteristics and evolutionary history of the adenovirus species.

This study demonstrates the utility of data science techniques, such as hierarchical clustering, in the field of computational biology. By leveraging the wealth of genomic data available, researchers can uncover patterns and relationships that can inform our understanding of viral biology and evolution.

Future research directions could include integrating additional genomic and phenotypic data to further elucidate the factors underlying the observed codon usage patterns. Additionally, applying other clustering algorithms or incorporating phylogenetic information could provide complementary insights.

References

1. Codon Usage Database: https://hive.biochemistry.gwu.edu/cgi-bin/codon/cgi-bin/getcodon.cgi
2. Hierarchical Clustering in Python: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering