Application of Clustering Methods for Genetic Analysis of SARS-CoV-2

Areeba Qazi, Divya Gollapalli, Rushi Surampudi, Shriya Jejurkar, Bryant Hou, Dr. Anjum Chida

Abstract

The SARS-CoV-2 pandemic has affected many individuals and completely changed the way the world functions. The development of a vaccine for this virus has been streamlined more than any other because of how widely it was spreading. As with any virus, the SARS-CoV-2 virus has mutated into many variants, two of which have taken over the world. This project attempts to help scientists develop future vaccines by comparing all discovered variants and finding similarities between them and the most widely found strains. Doing this allows scientists to predict the variants' response to the vaccine, which was developed based on the original Wuhan strain. Our project also aims to assist public health officials target specific strains of any virus and develop further preventative measures.

Introduction

Current genetic comparison software like MEGA utilize the Maximum Likelihood Phylogeny algorithm to compare genetic sequences and generate phylogenetic trees. While the method is reliable, it is very slow and computationally taxing. We aim to test clustering methods to determine whether the same accuracy can be achieved faster and with less processing power. Clustering is an example of unsupervised machine learning that utilizes a specific distance metric to measure space between all points and then group these by smallest distances. There are many clustering methods available that differ on how they group the data.

Approach

Two different clustering methods were chosen. Method 1 uses K-Means, which randomly selects centroids and then assigns clusters based on distance to the centroids. This algorithm is iterative in that each time it adjusts the centroids to reduce distance and re-clusters until the centroids no longer move. This method was picked as it is the simplest and least computationally heavy clustering algorithm and could be used as a benchmark. Method 2 is Agglomerative clustering. This algorithm works hierarchically, using a bottom-up approach. Each point starts as its own cluster and is sequentially grouped with neighboring points until the specified number of clusters is reached. Due to the constraints of these predefined algorithms, the approach was slightly different for each. Method 1 compares all sequences against each other while Method 2 compares each variant to the UK (suspected to be vaccine resistant) and Wuhan (reference) strains to determine which strain it is closer to as depicted in the scatter plot below.

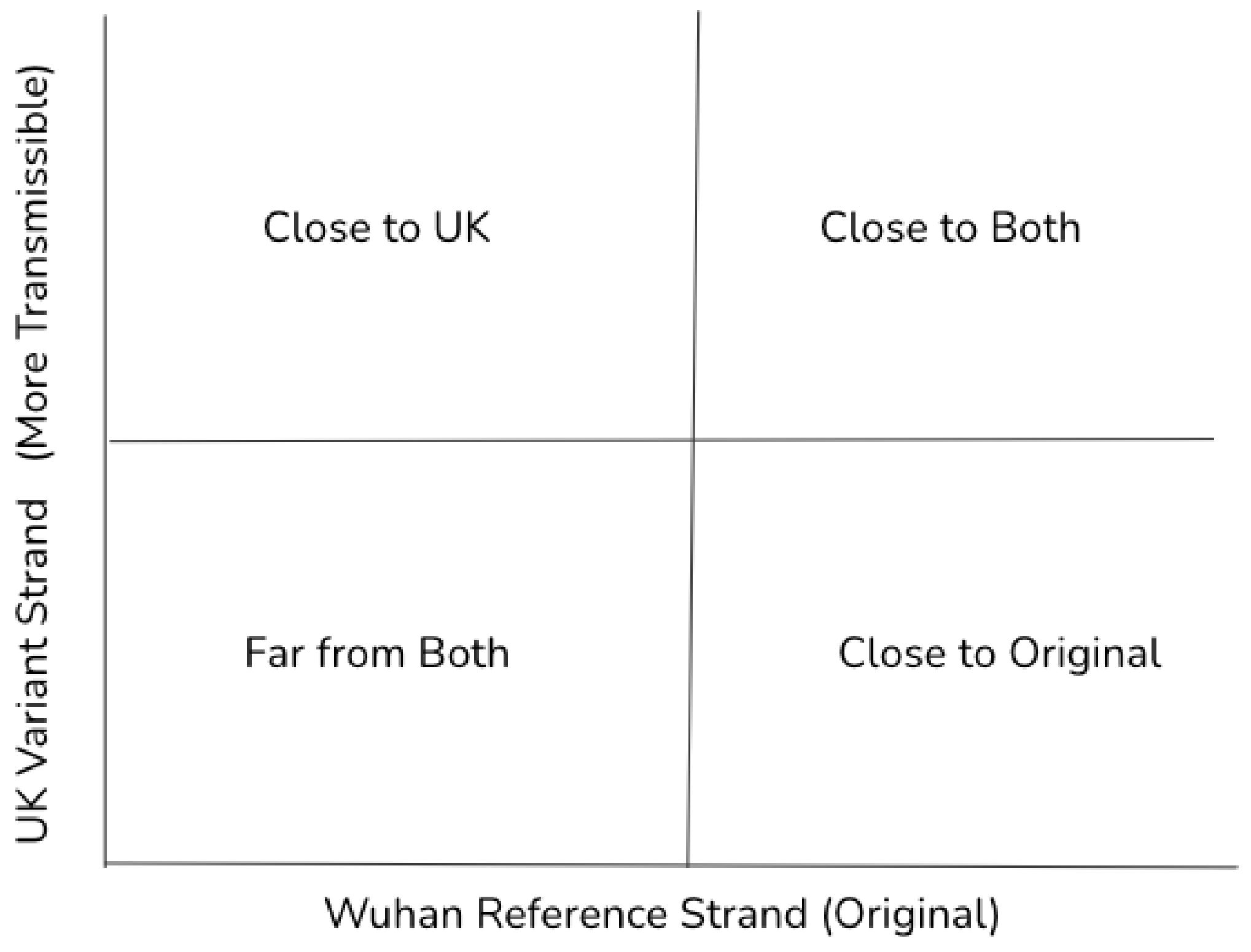

UK Variant Strand (More Transmissible)
Close to UK
Close to Both
Far from Both
Close to Original
Wuhan Reference Strand (Original)

Fig. 1: Interpreting Scatter Plots

Data

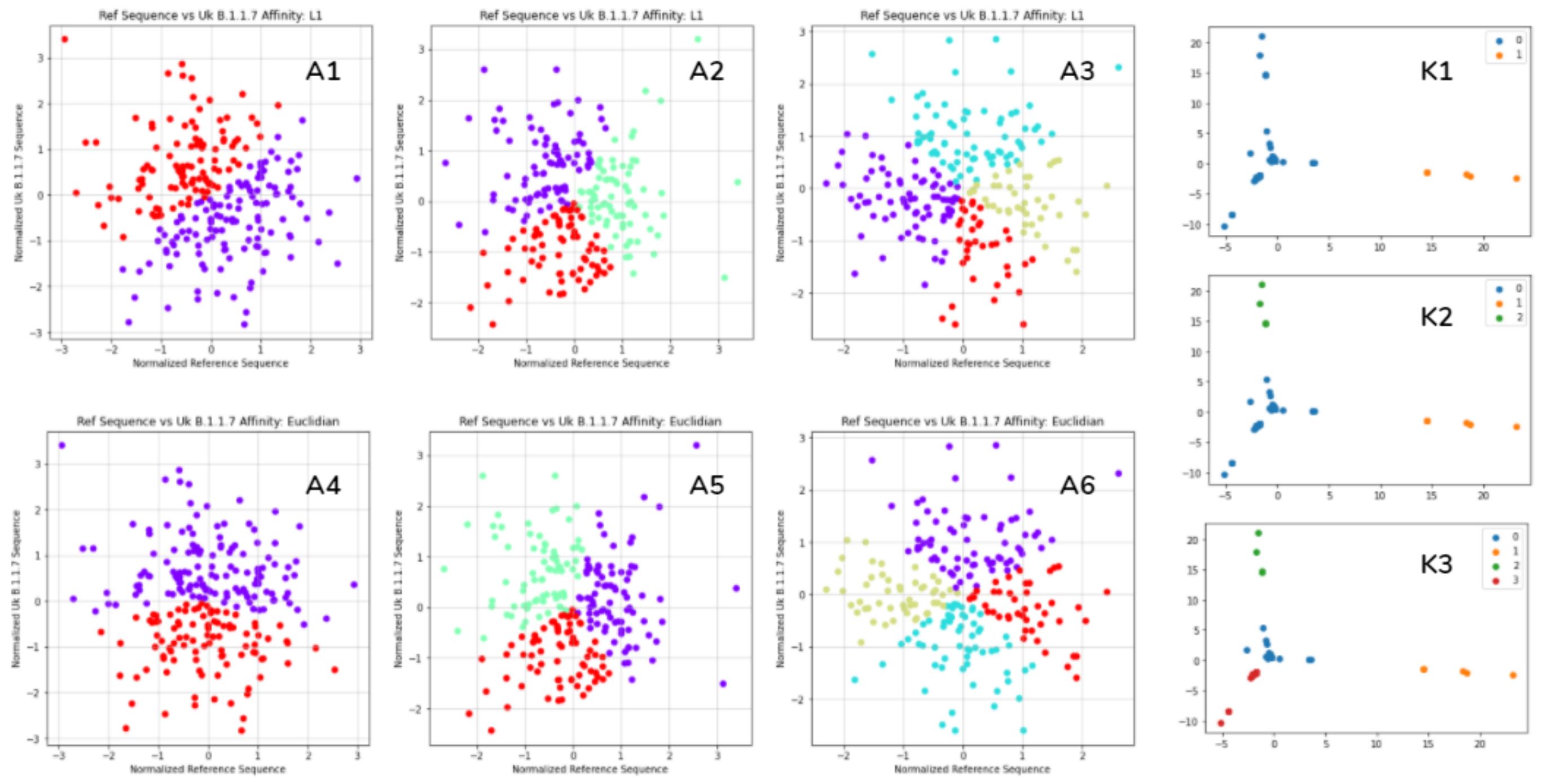


Fig. 2: Visualization of Clustering:
A1, A2, A3 = Agglomerative Clustering, L1 Affinity, 2, 3, 4 clusters
A4, A5, A6 = Agglomerative Clustering, Euclidean Affinity, 2, 3, 4 clusters

K1, K2, K3 = K-Means Clustering, 2, 3, 4 clusters

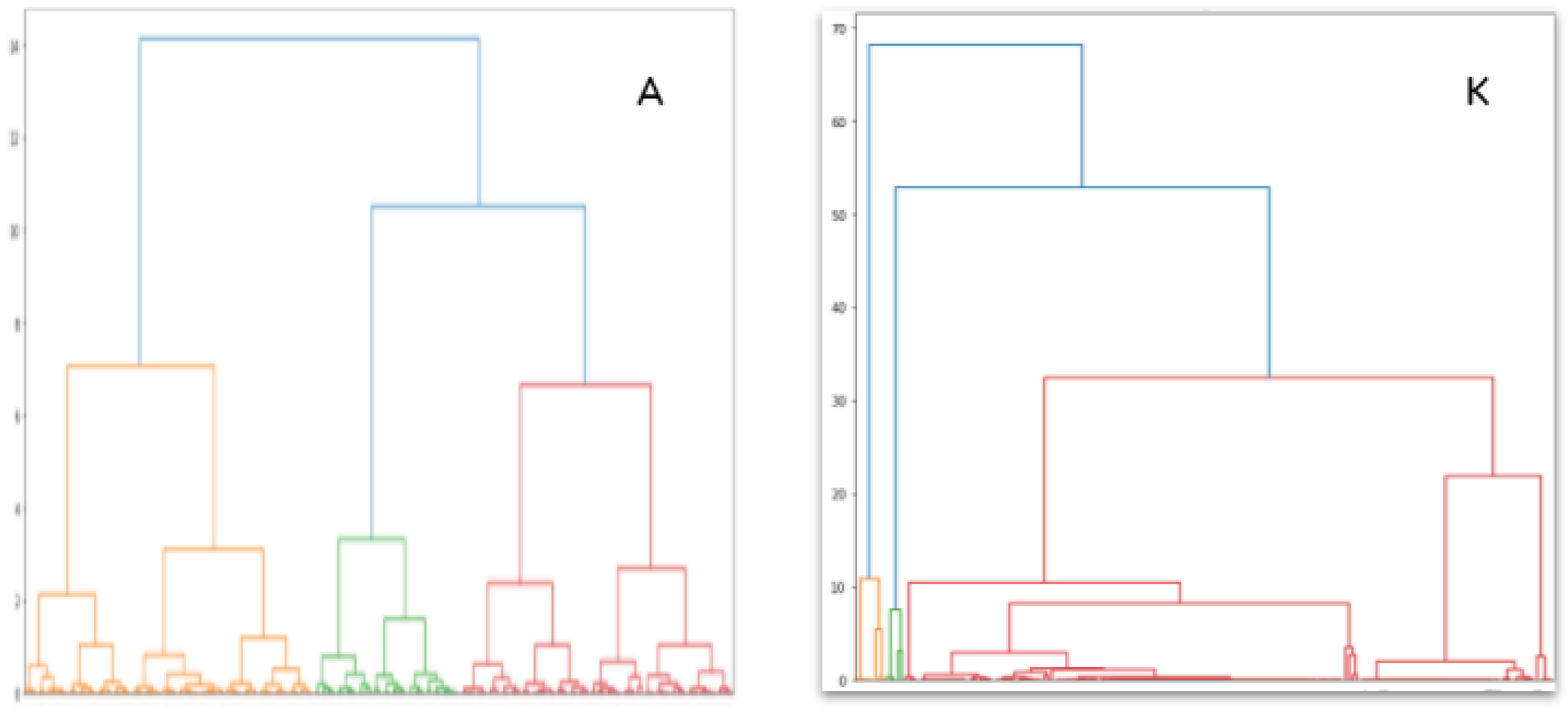


Fig. 3: Dendrograms for Agglomerative (left) and K-Means (right)

Evaluation Methods

To evaluate the accuracy and quality of our clusters as compared to the phylogeny generated by MEGA, we used the accuracy score and purity score metrics provided by Scikit Learn. The accuracy score method (calculated using Eq. 1) takes the predicted labels and true labels as inputs, and does a 1-to-1 comparison of each object to see if the predicted labels match the true ones. The data sets can be split into different buckets: false negatives (FN), true negatives (TN), true positives (TP), and false positives (FP). The accuracy score (A) is calculated by dividing the sum of the true positives and true negatives, which were the labels predicted correctly, by the overall data set.

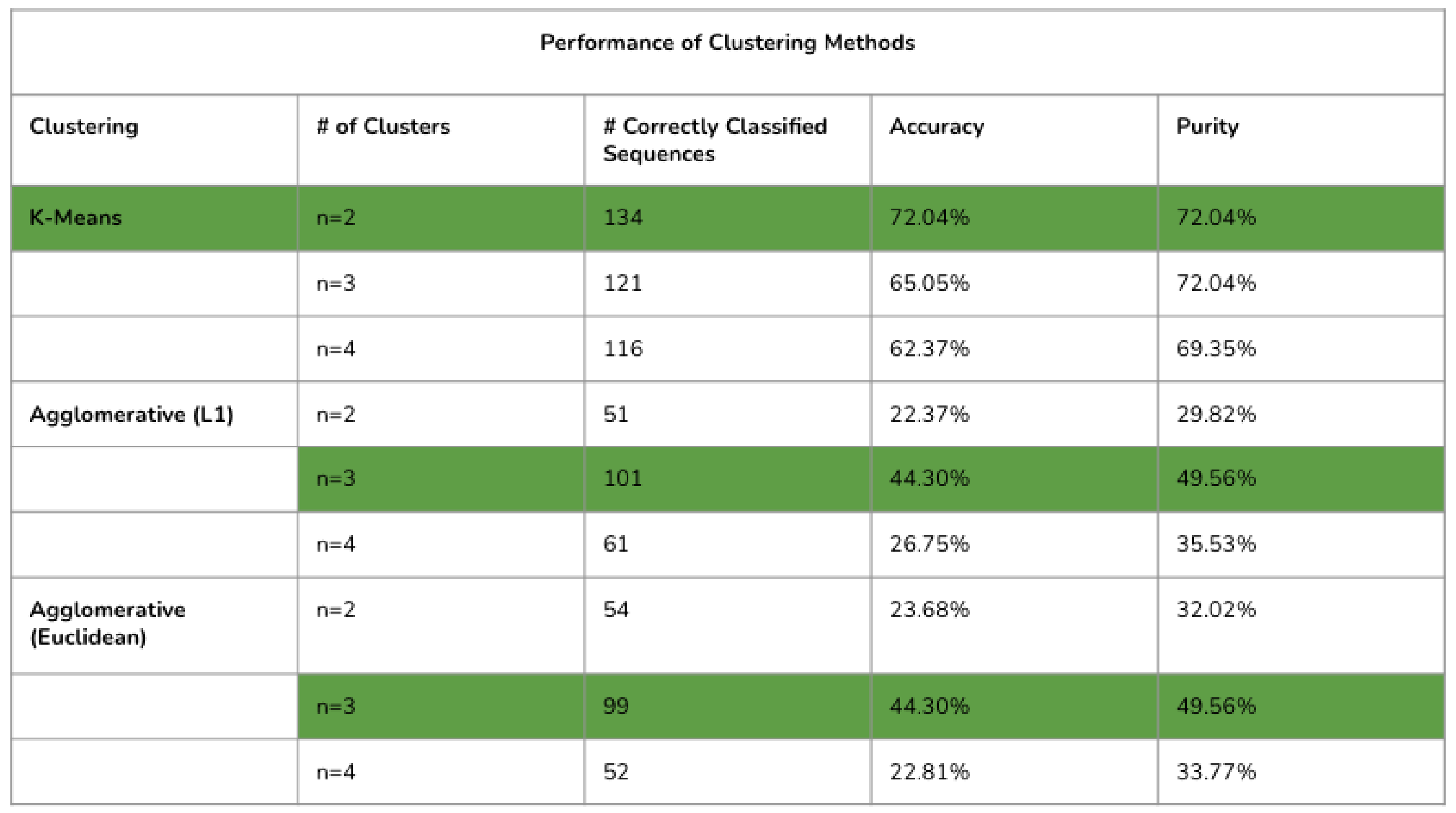
$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The purity score (P) is calculated using Eq. 2 shown below. The quantity P for each cluster is the fraction of the majority class in the cluster over the total number of classes in the cluster. P is then multiplied by $\frac{N_i}{N}$, which is the fraction of the total number of classes in each cluster divided by the total number of classes in all clusters. In this way, the purity score allows us to see how many classes were clustered accurately.

$$P = \sum \frac{N_i}{N}(P_i) \tag{2}$$

Analysis

Our results show that out of the K-Means models, 2 clusters performed the best, as it correctly identified 134 samples and grouped them with an accuracy and purity of 72.04%. Both Agglomerative models, L1 and Euclidean, performed best with the 3 cluster approach, correctly identifying 101 and 99 samples respectively, with an accuracy of 44.30% and purity of 49.56% for both. The table highlights the most accurate results in green. While these were the best models, there are a few other note-worthy models as well. K-Means with 4 clusters had a purity of 69.35%, which is 2.5 times better than random chance. Agglomerative had a purity of 35.53% (L1), 33.77% (Euclidean) for 4 clusters as well, meaning all three models performed better than random chance. This leads us to believe that with improvements to the models, the performance of these methods will grow significantly.



Performance of Clustering Methods

Clustering	# of Clusters	# Correctly Classified Sequences	Accuracy	Purity
K-Means	n=2	134	72.04%	72.04%
	n=3	121	65.05%	72.04%
	n=4	116	62.37%	69.35%
Agglomerative (L1)	n=2	51	22.37%	29.82%
	n=3	101	44.30%	49.56%
	n=4	61	26.75%	35.53%
Agglomerative (Euclidean)	n=2	54	23.68%	32.02%
	n=3	99	44.30%	49.56%
	n=4	52	22.81%	33.77%

Fig. 4: Performance of Cluster Methods

There are two main areas we plan on improving upon. The first step is to eliminate the two-level distance metric used for the agglomerative models. This may have squared any error and caused the unexpectedly low scores. The next step is to re-evaluate our distance metric. Levenshtein was built for basic string comparison but does not take into account the genetic alphabet, varied significance of mutations, and importance of location of the mutation. By utilizing Genetic Distance metrics, we can improve the models.

Conclusion

While it seems that there is no strong conclusion to be made from our preliminary results, we strongly believe there is promise to this methodology. Our first iteration shows that our models are better than random chance and with the proper adjustments, can be made to be highly accurate and efficient. It is also important to note that 9 different models were run simultaneously and produced results in under a minute, approximately 4.5 times faster than MEGA. Ideally, these models can lead to efficient and correct predictions of vaccine efficacy, resistance and more on not only COVID but other fast-mutating viruses as well.

References

[1] "NCBI Virus," National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi//virus?SeqType$_s = Nucleotideamp; VirusLineage_s s = SARS - CoV - 2, \%20taxid : 2697049amp; ProtNames_s s = surface\%20glycoprotein.[Accessed : 30 - Apr - 2021]$.

[2] A. Kassambara, "Agglomerative Hierarchical Clustering," Datanovia, 20-Oct-2018. [Online]. Available: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/. [Accessed: 30-Apr-2021].

[3] A. Eshaghi, "Unsupervised Learning," Ludwig Maximilian University of Munich. Available: https://www.mathematik.uni-muenchen.de/ deckert/teaching/WS1819/ATML/arman$_unsupervised_learning.pdf.[Accessed : 30 - Apr - 2021]$.