Supplementary to "Genome-wide Analysis of 10664 SARS-CoV-2 Genomes to Identify Virus Strains in 73 Countries based on Single Nucleotide Polymorphism"

Nimisha Ghosh[a,g], Indrajit Saha[b,g,*], Nikhil Sharma[c,g], Suman Nandi[d], Dariusz Plewczynski[e,f]

[a]*Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Orissa, India*
[b]*Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India*
[c]*Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India*
[d]*Department of Computer Science and Engineering, National Institute of Technical Teachers' Training and Research, Kolkata, West Bengal, India*
[e]*Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
[f]*Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland*
[g]*Equally contributed*

S1. Preliminaries

S1.0.1. Hierarchical Clustering

Hierarchical clustering algorithms (Tou & Gonzalez, 1974, Devijver & Kittler, 1982) organize data into a hierarchical structure according to the proximity matrix. The results of hierarchical clustering are usually depicted by a binary tree or a dendrogram. The branches of a dendrogram not only record the formation of the clusters but also indicate the similarity between the clusters. By cutting the dendrogram at some level, we can obtain a specified number of clusters. Hierarchical clustering algorithms can be further divided into agglomerative approaches and divisive approaches based on how the hierarchical dendrogram is formed. Agglomerative algorithms (bottom-up approach) initially consider each data object as an individual cluster. Subsequently at each step, it merges the closest pair of clusters until all the groups are merged into one cluster. For this purpose, different measures, such as Average Linkage (AL) (Tou & Gonzalez, 1974, Devijver & Kittler, 1982) and Complete Linkage (CL) (Tou & Gonzalez, 1974) of cluster proximity are used. They are described in below.

- **Average Linkage:** The average linkage (AL) clustering algorithm, also known as the unweighted pair-group method using arithmetic averages (UPGMA) (Devijver & Kittler, 1982), is one of the most widely used hierarchical clustering algorithm. The average linkage algorithm is obtained by defining the distance between two clusters to be the average distance between a point in one cluster and a point in the other cluster. Formally, the

*Corresponding author: indrajit@nitttrkol.ac.in

distance between two clusters can be defined as follows:

$$\mathcal{D}_{AL}(C_i, C_j) = \frac{1}{|c_i||c_j|} \sum_{x_i \in C_i, x_j \in C_j} D(x_i, x_j) \tag{1}$$

- **Complete Linkage:** In complete linkage (CL) clustering the distance between two clusters is calculated as the greatest distance between members of the relevant clusters. Formally, the distance between two clusters can be defined as follows:

$$\mathcal{D}_{CL}(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} D(x_i, x_j) \tag{2}$$

S1.0.2. Distance Function

The distance function is one of the most important factors for clustering of data objects. The performance of above average and complete linkage algorithms depends on the distance functions as well. In this study, the Jaccard and Hamming distance functions (Tou & Gonzalez, 1974, Devijver & Kittler, 1982) are used as we have binary data of SNPs of SARS-CoV-2 genomes to compute the distance between two virus genomes.

- **Jaccard:** Jaccard distance (Tou & Gonzalez, 1974, Devijver & Kittler, 1982) is useful in determining the similarity between the binary data objects of two different clusters, C_i and C_j. It is defined as:

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} \tag{3}$$

where $x_i \in C_i$, $x_j \in C_j$. Moreover, pairwise dissimilarity between two binary data objects can be computed as:

$$D_J(x_i, x_j) = 1 - J(x_i, x_j) = \frac{|x_i \cup x_j| - |x_i \cap x_j|}{|x_i \cup x_j|} \tag{4}$$

- **Hamming:** Hamming distance (Tou & Gonzalez, 1974, Devijver & Kittler, 1982) is another well-known similarity measure between two binary data objects through calculating the total number of matches in values at different positions or attributes. Let $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$, and $x_j = [x_{j1}, x_{j2}, \ldots, x_{jm}]$ be two binary data objects described by m attributes. The Hamming distance between x_i and x_j , $D_H(x_i, x_j)$, can be defined by the total number of matches of the corresponding attribute of the two data objects. Formally,

$$D_H(x_i, x_j) = \sum_{k=1}^{m} \delta(x_{ik}, x_{jk}) \tag{5}$$

where

$$\delta(x_{ik}, x_{jk}) = \begin{cases} 1 & \text{if } x_{ik} = x_{jk} \\ 0 & \text{if } x_{ik} \neq x_{jk} \end{cases} \tag{6}$$

It is worth mentioning that Euclidean distance will not work for Average and Complete Linkage methods. Hierarchical clustering like Average and Complete Linkage work with binary dataset and Euclidean distance does not work well with binary data.

S1.0.3. Silhouette Index

Silhouette Index (Rousseeuw, 1987) reflects the compactness and separation of the clusters. Given a set of n data objects $X = \{x_1, x_2, \ldots, x_n\}$ and a clustering of such data objects into $C = \{C_1, C_2, \ldots, C_K\}$, the silhouette width $S(x_i)$ for each data object x_i belonging to cluster C_j denotes a confidence measure of belongingness, and it is defined as follows:

$$S(x_i) = \frac{b(x_i) - a(x_i)}{max\{a(x_i), b(x_i)\}} \tag{7}$$

Here $a(x_i)$ denotes the average distance of the data object x_i from the other data objects of the cluster to which x_i is assigned, and $b(x_i)$ represents the minimum of the average distances of x_i from the data objects of the clusters C_k, $k = 1, 2, \ldots, K$, and $k \neq j$. The value of $S(x_i)$ lies between - 1 and 1. Large value of $S(x_i)$ (approaching 1) indicates that the x_i is well clustered. Overall silhouette index $S(C)$ of a clustering C is defined as

$$S(C) = \frac{1}{n} \sum_{i=1}^{n} S(x_i) \tag{8}$$

Greater value of $S(C)$ (approaching 1) indicates that most of the data objects are correctly clustered and this in turn reflects better clustering solution. This is to be noted that Jaccard and Hamming distance functions are used while computing Silhouette values.

References

Devijver, P. A., & Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. London: Prentice Hall.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. doi:https://doi.org/10.1016/0377-0427(87)90125-7".

Tou, J. T., & Gonzalez, R. C. (1974). *Pattern Recognition Principles*. Reading: Addison-Wesley.

Experiment Data and Files	Web Link
Details of 107 SNPs	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/Unique-107-SNPs.csv
All Clustering Results with varying clusters 2 to 100 with Labels	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/All-ClusteringResults-2-100-Labels.xlsx
Refined Clustering Results of Average Linkage, Complete Linkage as 5 Clusters with Labels	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/Refined-ClusteringResults-5-Labels.xlsx
Cluster wise Distribution of SARS-CoV-2 strains in different Countries	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/ClusterwiseDistributionCountries.xlsx
Common clusters as virus strains present in different Countries	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/CommonClusters.csv
All VAT plots for Average Linkage, Complete Linkage and Single Linkage	http://www.nitttrkol.ac.in/indrajit/projects/COVID-Mutation-10K-Clustering//downloads/supplementary/All-VATs-All-Linkages.zip

Table S1: Various results of clustering on SNPs data of SARS-CoV-2

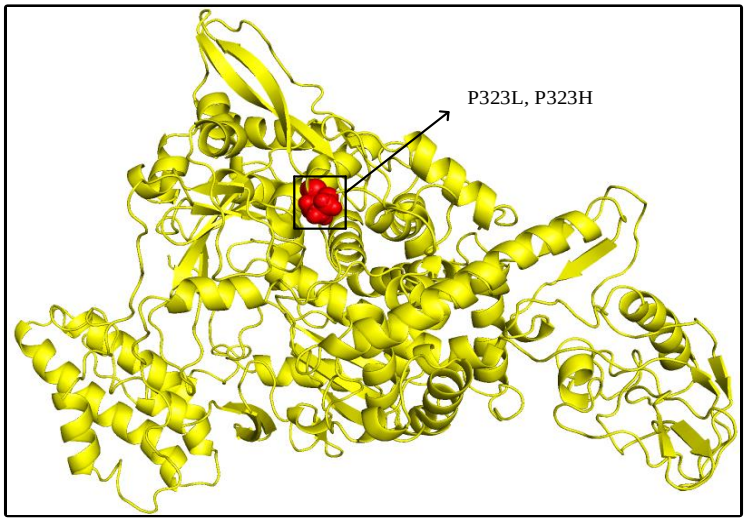

Figure S1: SNPs highlighted in the structure of RdRp

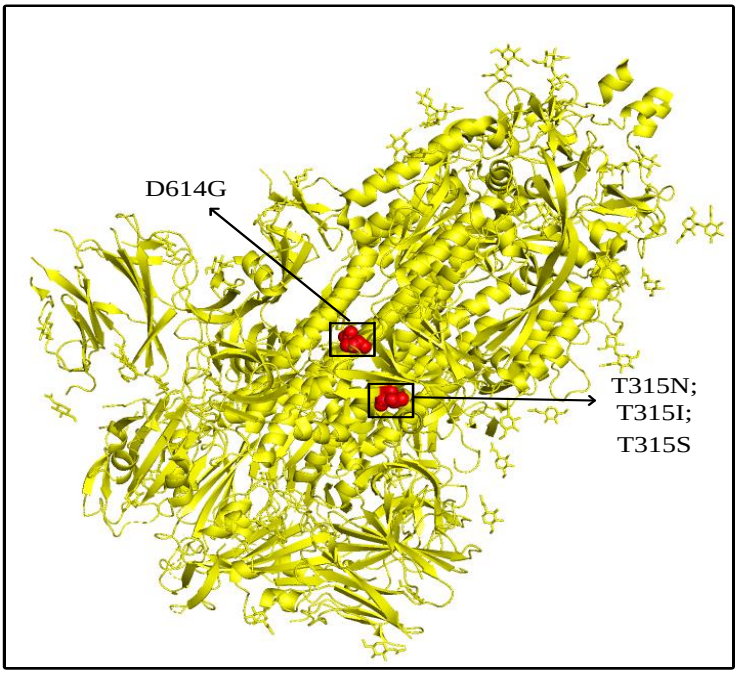

Figure S2: SNPs highlighted in the structure of Spike