Hierarchical Clustering and Dimensionality Reduction for SARS-CoV-2 Genome Analysis Across Highly Affected Nations

Venkataramanan V[1*], Srinivasan J[2], Ramadevi K[3], Dillibabu M[3]
[1]Department of Information Technology, K J Somaiya School of Engineering, Somaiya Vidyavihar University, Mumbai 400076., India
[2]Department of Computer Applications, Madanapalle Institute of Technology and Science MITS, Madanapalle 517325, Andhra Pradesh, India
[3]Department of Information Technology, Panimalar Engineering College, Anna University, Chennai 600025, Tamil Nadu, India
*E-mail: venkataramanan@somaiya.edu
*Corresponding author

Keywords: COVID-19, SARS-CoV-2, clustering, dimensionality reduction

Received: December 19, 2024

*The global pandemic caused by the novel coronavirus SARS-CoV-2 has prompted extensive research into its genetic diversity to support drug development and vaccination strategies. In this study, we analyze the genetic similarity patterns of SARS-CoV-2 genome sequences from six severely affected nations: USA, Italy, Spain, France, Germany, and the UK. A total of 359 complete human host SARS-CoV-2 genome sequences, ranging from 29,538 to 29,987 base pairs, are processed using k-mer representation, with k = 2 (dinucleotides) and k = 3 (codons). These representations are converted into 50-dimensional feature vectors. To identify intrinsic patterns within this high-dimensional dataset, we apply agglomerative hierarchical clustering using average linkage. A Silhouette score of **0.48** and a Hopkins statistic of **0.85** indicate moderate clustering tendency and structure. Four primary clusters are identified, highlighting notable genomic similarities. Specifically, sequences from the USA, Spain, and Italy predominantly group together, suggesting shared genetic traits. To further aid interpretation, we apply dimensionality reduction techniques—Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE)—which project the high-dimensional feature vectors into 2-dimensional space. Visualizations confirm the clustering structure, with USA, Spain, and Italy forming a distinct and tight cluster, while sequences from France, Germany, and the UK show more dispersed patterns. This study provides a quantitative and visual understanding of SARS-CoV-2 genetic diversity across heavily impacted nations. The combination of k-mer-based feature encoding, hierarchical clustering, and dimensionality reduction offers actionable insights that may inform more targeted therapeutic and vaccine design strategies.*

Povzetek: Študija primerja mehki adaptivni krmilni sistem s tradicionalnim PID krmilnikom za večdimenzionalno regulacijo temperature v električni opremi. Novi adaptivni sistem je zmanjšal temperaturna nihanja in izboljšal energetsko učinkovitos, saj je ohranjal temperaturo znotraj intervala kljub motnjam.

1 Introduction

Coronaviruses (CoVs) belong to the Orthoradial Kingdom, Class Pisoniviricetes, Order Nidovirales, and Family Coronaviridae. Among the four subfamily Orthocoronavirinae genera, alpha(α)-CoV and beta(β)-CoV affect animals, while gamma(γ)-CoV and delta(δ)-CoV target fowls. Within beta(β)-CoV, four lineages (A, B, C, and D) exist. Recent decades have witnessed outbreaks caused by these β-CoVs [1,2]. In 2002, severe acute respiratory syndrome (SARS), linked to β-CoV lineage B, emerged in China's Guangdong area, affecting 29 countries with a fatality rate of 11% [3,4]. SARS-CoV likely originated in Chinese horseshoe bats, possibly transmitted to humans through palm civets [5,6]. In 2012, Middle East Respiratory Syndrome (MERS), caused by β-CoV lineage C, emerged in Saudi Arabia, resulting in a 37% mortality rate across 27 countries [7,8,9]. MERS-CoV possibly originated in bats, transmitted through dromedary camels [10,11].

The novel coronavirus SARS-CoV-2 belongs to β-CoV lineage B. Responsible for Coronavirus Disease 2019 (COVID-19), it was first identified in Wuhan, China, in December 2019 [12,13,14]. Due to its rapid transmission and severity, the World Health Organization declared COVID-19 a pandemic and a global health emergency on January 30, 2019 [15,16]. By May 5, 2020, it had spread to 180 countries and 33 territories, causing millions of cases, deaths, and recoveries [17,18]. SARS-CoV-2's genome shares similarities of 96% with bat CoVs and 92.4% with pangolin CoVs [19,20]. Despite these

similarities, the exact source of infection remains unknown [21].

In light of these past outbreaks and the ongoing global COVID-19 pandemic, understanding the genetic diversity and evolution of SARS-CoV-2 is paramount. The identification of mutation patterns and genetic similarities among different regions can provide critical insights into the virus's transmissibility, virulence, and potential drug and vaccine targets. Such knowledge is pivotal for devising effective strategies to combat the spread of the virus and mitigate its impact on public health.

Genome sequencing, a cornerstone of virology research, deciphers nucleic acid sequences. Our study focuses on SARS-CoV-2 genome sequences from profoundly affected countries: USA, France, UK, Germany, Spain, and Italy. These sequences, obtained from the GISAID gene bank, are represented as *k*-mers—subsequences of '*K*' length with adenine (A), guanine (G), cytosine (C), and thymine (T) bases [22,23].

The *k*-mer-represented genome sequences are transformed into 50-dimensional numeric vectors. These vectors undergo unsupervised machine learning via agglomerative hierarchical clustering to group similar genomes [24,25]. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), further transform the clustered high-dimensional genome sequence vectors into a more manageable 2-dimensional space [26-28]. This visualization enables an understanding of genome similarities among different countries, which vary due to mutations affecting SARS-CoV-2's severity and spread [29,30].

High-dimensional genetic data poses challenges in visualization and interpretation. To address this, we employ dimensionality reduction techniques, notably Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA, a linear technique, reduces data while maximizing captured variance across uncorrelated principal components. This simplifies visualization and uncovers hidden patterns in lower-dimensional space.

PCA's reduction of complex data into informative components enables the identification of regions with shared genetic traits. It aids in recognizing potential clusters and insights into virus transmission and evolution. Additionally, agglomerative hierarchical clustering groups genome sequences with similar attributes, enhancing our grasp of genetic relationships among nations.

Identifying patterns in SARS-CoV-2 genetic diversity could inform vaccine and drug development targets. Our study aims to uncover clusters of genome sequence similarity, potentially aiding decisions on medication and attenuated vaccines [31]. The global impact of the COVID-19 pandemic underscores the urgency of understanding the genetic diversity of the SARS-CoV-2 virus to devise effective interventions. Genome sequencing has provided us with a wealth of data, offering insights into the virus's evolution and transmission patterns. However, interpreting these extensive datasets is challenging due to their high dimensionality.

This study's motivation lies in harnessing advanced data analysis techniques to decipher the intricate genetic relationships within SARS-CoV-2 genomes from severely affected nations. By identifying clusters of similar genome sequences, we aim to achieve two critical goals: aiding in the development of targeted drugs and vaccines, and providing actionable insights for public health strategies.

Clustering results hold immense significance in guiding our response to the pandemic. These results translate into tangible strategies for combatting COVID-19. Identifying clusters reveals groups of genome sequences that share distinct genetic characteristics, shedding light on how the virus evolves and spreads. Such insights can inform the development of tailored medications and vaccines that target specific clusters, optimizing their efficacy.

Moreover, the clustering outcomes provide vital information for public health officials. By pinpointing regions with shared genetic traits, we can trace the virus's path and understand how it has been transmitted between different nations. This knowledge aids in devising targeted containment measures and travel restrictions, ultimately curbing the virus's spread. By focusing on regions profoundly impacted by the pandemic, we aim to enhance our understanding of the virus's spread within these contexts. Ultimately, we strive to offer actionable insights to inform researchers, healthcare professionals, and policymakers in crafting targeted strategies to control the pandemic and mitigate its societal and public health repercussions.

The characterization and clustering of SARS-CoV-2 genomic data require a multidisciplinary approach, integrating methods from bioinformatics, genomics, and data science. Recent advancements in sequencing technologies, such as next-generation sequencing (NGS), have enabled rapid and high-throughput decoding of viral genomes, significantly advancing the understanding of viral evolution and mutation patterns (Behjati & Tarpey, 2013). Studies on coronavirus origin and transmission have highlighted the zoonotic potential of bats and pangolins, supporting the theory of cross-species transmission events that contributed to the emergence of SARS-CoV-2 (Banerjee et al., 2019; Lam et al., 2020).

For effective analysis and categorization of genomic data, various computational and statistical methods have been adopted. Dimensionality reduction techniques like Principal Component Analysis (PCA) (Hotelling, 1933; Jolliffe & Cadima, 2016) and visualization approaches such as t-SNE (van der Maaten & Hinton, 2008) facilitate the interpretation of high-dimensional genetic datasets. Clustering methods, particularly hierarchical clustering (Bouguettaya et al., 2015) and its validation metrics like the Hopkins statistic (Banerjee & Davé, 2004) and ClValid package (Brock et al., 2008), are crucial in uncovering patterns and structure within viral genome sequences. Moreover, embedding models such as dna2vec (Ng, 2017) and word2vec (Mikolov et al., 2013) provide meaningful vector representations of nucleotide sequences for advanced data-driven analysis.

Comprehensive analysis of coronavirus genomes has also revealed important features related to codon usage bias (Hershberg & Petrov, 2008), dinucleotide patterns (Karlin, 1998), and structural properties of viral spike proteins that influence host specificity and viral entry mechanisms (Lu et al., 2015). The emergence of recurrent mutations and genomic diversity across variants further underscores the importance of continuous phylogenetic and comparative studies (Van Dorp et al., 2022). Table 1 summarizes key studies addressing methodologies and biological insights relevant to SARS-CoV-2 genome analysis, clustering, and evolution.

Table 1: Summary of key contributions

Reference	**Methodology/Focus**	**Key Findings/Applications**
Banerjee A, Davé RN (2004)	Hopkins Statistic for cluster validation	Introduced a method to validate clusters based on spatial data distributions, crucial for genomic data clustering.
Banerjee A, Kulcsar K, Misra V, et al (2019)	Bats and Coronaviruses	Examined the zoonotic transmission of coronaviruses from bats, providing insight into SARS-CoV-2's origin.
Behjati S, Tarpey PS (2013)	Next-generation sequencing (NGS)	NGS technologies enable comprehensive analysis of viral genomes, aiding the detection of mutations.
Bouguettaya A, Yu Q, Liu X, et al (2015)	Hierarchical Clustering	Provided efficient clustering algorithms for large datasets, applicable to genomic data grouping.
Brock G, Pihur V, Datta S, Datta S (2008)	ClValid package for cluster validation	Introduced a method for validating clustering methods with a focus on genomic data, ensuring reliable analysis results.
Compeau PEC, Pevzner PA, Tesler G (2021)	De Bruijn Graphs for Genome Assembly	Discussed the application of de Bruijn graphs for efficient genome assembly, an essential step in viral genome sequencing.
Hotelling H (1933)	Principal Component Analysis (PCA)	PCA for reducing dimensionality in large genomic datasets, enhancing the interpretation of complex data.
Karlin S (1998)	Dinucleotide Signatures	Identified dinucleotide biases in genomes, aiding in the study of viral genomic features and evolutionary patterns.
Lu G, Wang Q, Gao GF (2015)	Host Jump Mechanism in Coronaviruses	Examined the spike protein features in coronaviruses, explaining the host jump from bats to humans.
Van Dorp L, Acman M, Richard D, et al (2022)	Genomic Diversity and Mutations	Studied the recurrent mutations and diversity of SARS-CoV-2, contributing to the understanding of viral evolution and resistance.

. The primary objective of this study is to identify and analyze clusters of genetically similar SARS-CoV-2 genomes from highly affected countries using advanced machine learning and data visualization techniqu**es.** This is achieved by transforming viral genome sequences into k-mer-based numerical vectors, which are then subjected to unsupervised learning through agglomerative hierarchical clustering. To address the challenge of high dimensionality in genetic data, we employ Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to project the data into two dimensions for easier interpretation and visual cluster identification.

By linking our research goals to experimental procedures, this study seeks to:

- Uncover distinct clusters of SARS-CoV-2 genome sequences across regions such as the USA, France, UK, Germany, Spain, and Italy.
- Understand the evolutionary and transmission-related genetic similarities among these clusters.
- Generate actionable insights that can guide targeted drug and vaccine development by identifying common mutation patterns.
- Provide public health authorities with genomic-level evidence to support region-specific pandemic response strategies.

2 Materials and methods

2.1 Genome sequences and their *K*-mers

In this study, we collected genome sequences of the human host SAR-CoV-2 virus from various countries highly affected by the COVID-19 pandemic, including the United States of America (USA), France, United Kingdom (UK), Germany, Spain, and Italy. The lengths of the collected genome sequences range from 29,538 to 29,987 base pairs. The distribution of genome sequences obtained from each country as of April 16, 2020, is summarized in Table 2 below.

Table 2: Country-wise distribution of collected genome sequences

Countries	Number of Genome Sequences obtained
USA	85
France	72
UK	67
Germany	59
Spain	43
Italy	33

Genome sequence of length x will have $x - k + 1$ k-mers. The following Table 3 describes the k-mers sequences corresponding to the genome sequence.

Table 3: Sample genome sequence and its *k*-mers

Sample Genome Sequence	*k*-values	*k*-mers
ATTAAAGGTT	2	AT, TT, TA, AA, AA, AG, GG, GT, TT
	3	ATT, TTA, TAA, AAA, AAG, AGG, GGT, GTT

We have considered 2-mer and 3-mer sequence for each genome sequence in our study. *K*-mers sequence generated with *k*=2 reveals the dinucleotide biases which remain constant throughout the genome. Since we look for similarity in genome sequences, dinucleotide biases act as a distance measure between phylogenetically alike genomes. The genomes of organisms that are related to each other, shares more alike dinucleotide biases, than more differently related organism [32]. An amino acid can be uniquely represented by 64 distinct 3-mers present in DNA. Codons are non-overlapping 3-mers present in a genome sequence. Each codon or 3-mer maps itself to only one amino acid. Multiple codons are required to represent each amino acid [33]. Therefore, *K*-mer sequence generated with *k*=3 expresses the internal codons present in the genome sequence. The steps for the implemented K-mers sequence generation is given in algorithm1:

Algorithm 1: *K*-mers sequence generation.
Input: Genome Sequence (seq),
Input: *k*-value (k),
Output: *k*_mers
Begin:
x = length(seq)
n = *x*-*k*+*1*
for $i = 0$ to n do
j = *i*
j = *j*+*k*
temp = seq.substring(*i*,*j*)
*k*_mers = *k*_mers + temp
end
return k_mers
end

The methodology involves generating *k*-mers sequences from collected genome data, representing these sequences as numerical vectors using a skip-gram neural network, and ultimately obtaining gene2vec representations for further analysis. The methodology flow diagram (Figure 1) visually summarizes the steps outlined.

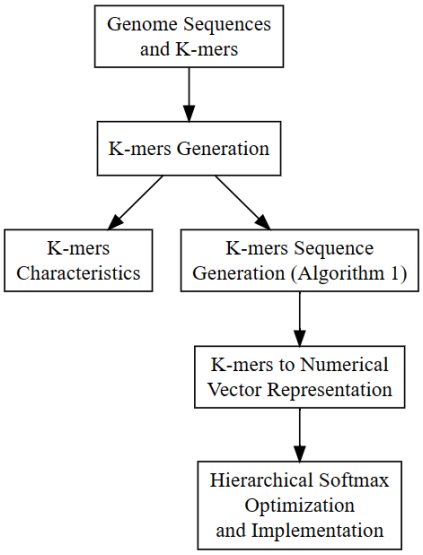


Figure 1: Flow diagram

2.2 *K*-mers sequences to numerical vector representation

Genome sequence to numeric vector representation (gene2vec) uses a shallow 2-layer neural network to train *k*-mers of the genome sequence. There are two options to perform gene2vec. The first procedure is the continuous bag of words (CBOW), which deduce the focus word given the surrounding terms, while the second procedure called skip-gram auspicate the surroundings terms given the focus word. Skip-gram performs better, even with fewer data and infrequent words [34]. We use skip-gram procedure in our experimentation.

The neural network takes one hot encoded *2*-mer (dinucleotide) or 3-mer (codon) into a 50-dimensional hidden layer with linear activations. The hidden layer is fully connected to the softmax output layer, which gives a numeric vector for each dinucleotide or codon. Finally, gene2vec of the genome sequence is given as average gene2vec of each dinucleotide or codon. We have selected hierarchical softmax optimization, which uses a binary tree to represent all 2-mer or 3-mer in the sequence. The 2-mer or 3-mer are leaves in the binary tree. The unique path from the root to each leaf is used to calculate the probability of the 2-mer or 3-mer represented by the leaf. We have implemented the skip-gram model using gensim [35] python library.

2.3 Evaluating clustering tendency

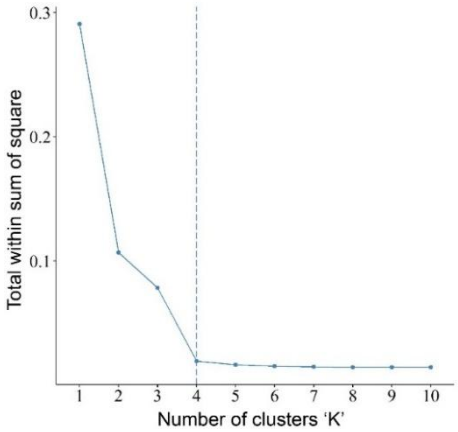
The clustering analysis process starts with assessing the clustering tendency of the dataset. It is essential to check whether the data comprises of meaningful clusters or not, before applying the clustering techniques to ensure the feasibility for performing cluster analysis. We have used Hopkins statistic [36,37] to assess the clustering tendency of both 2-mers and 3-mers numeric vector dataset (D containing n datapoints). For every datapoint $r_i \in D$, obtain its nearest neighbor r_j, then calculate the distance between r_i and r_j, which is denoted as $X_i = dist(r_i, r_j)$. Generate a simulated observations or dataset ($simulated_D$) drawn from a random uniform distribution with n points ($s_1, s_2, s_3, \ldots, s_n$) with the same variations as D (original dataset). For every datapoint $s_i \in simulated_D$, obtain its nearest neighbor s_j in D, then calculate the distance between s_i and s_j, which is denoted as $Y_i = dist(s_i, s_j)$. The Hopkins statistic (H) is computed as the mean nearest neighbor distance in the $simulated_D$ dataset upon the sum of the mean nearest neighbor distances in D and across the $simulated_D$ dataset i.e., given in equation 1

$$H = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i} \quad (1)$$

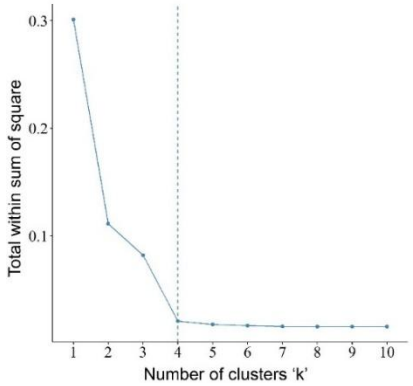
The value of H about 0.5 indicates that there no intrinsic clusters present in the dataset because the value of $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n} X_i$ become close to each other. If the value H is close to 0, then the dataset contains significant clusters. The Hopkins statistic value H = 0.01031 for 2-mers dataset and H = 0.01608 for 3-mers dataset concludes that there are useful clusters present in our data.

2.4 Optimal number of clusters (*k*)

The total intra-cluster disparity or total within-cluster sum of square (WSS) assesses the density of the agglomeration and it is recommended to be as low as possible. The Elbow method considers the total WSS as a tool to determine the number of clusters. The value for which the total WSS does not improve by adding another cluster is regarded as the number of clusters value '*k*'. The position of a bend (elbow) in the graph (Figure 2) usually indicates the appropriate number of clusters.

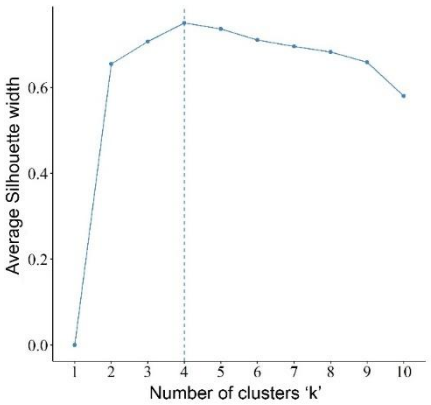


(a) Elbow method plot on 2-mers numeric vector dataset indicates k=4.

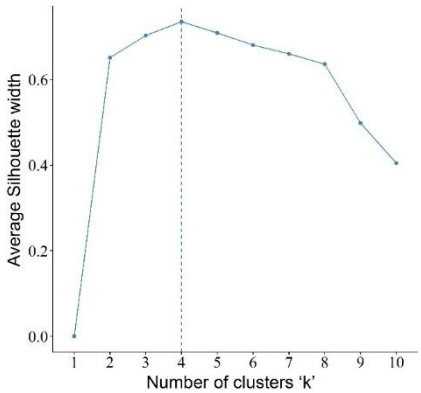


(b) Elbow method plot on 3-mers numeric vector dataset shows k=4.

Figure 2: Optimal '*k*' using Elbow method



(a) AS method plot on 2-mers numeric vector dataset indicates *k=4.*



(b) AS method plot on 3-mers numeric vector dataset shows *k=4*.

Figure 3: Optimal '*k*' using the AS method

The Average Silhouette (AS) approach is used to ascertain how well each datapoint is positioned within its group. A high AS width value suggests a useful agglomeration/clustering. The AS method calculates the AS of data points for dissimilar '*k*' values. The preferred number of clusters '*k*' is the value that has a high AS among a scope of possible '*k*' values, as shown in Figure 3.

2.5 Selecting best clustering algorithm

The clustering algorithms such as hierarchical, kmeans and Partition Around Medoids (PAM) are selected. Their internal measures such as connectivity, Dunn index and Silhouette coefficient are computed. Stability measures such as Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM) and Figure of Merit (FOM) are also calculated (Brock et al. 2008). On the basis of values obtained from above measures, a suitable clustering algorithm is preferred. Connectivity explains the extent of datapoints positioned in the same group as their nighest neighbors in the data space. The connectivity has a value lies between 0 and ∞ . Minimized connectivity value is preferred. The Dunn index is computed using equation 2.

$$D = \frac{min.separation}{max.diameter} \quad (2)$$

Where 'D' is the Dunn index, min.separation indicates the minimal pairwise distance between the data points of inter-cluster. The value max.diameter specifies the maximal pairwise distance between datapoints of intra-cluster. Maximized 'D' value is preferred because the dataset, which comprises compact and well set-apart clusters, has small max.diameter value and large min.separation value. Silhouette width S_i of the datapoint 'i' is given by equation 3.

$$S_i = \frac{(q_i - p_i)}{max\,(p_i, q_i)} \quad (3)$$

Where p_i is the average dissimilarity between i and other data points present in the same group and q_i is the minimum average dissimilarity between i and data points belonging to different group. The datapoint with S_i value close to 1 are correctly clustered, negative S_i value indicates wrongly clustered data point and S_i value around 0 indicates that the data point halts between two clusters. Table 3 shows the calculations of all the measures mentioned above on 2mers and 3mers numerical vector datasets. The optimal scores in Table 3 conclude hierarchical clustering as a suitable clustering technique for our datasets.

.

Table 4: Calculation of internal and stability clustering measures

Clustering algorithms	Measures	Cluster Size					
		K=4		*K*=5		*K*=6	
		2-mers Dataset	3-mers Dataset	2-mers Dataset	3-mers Dataset	2-mers Dataset	3-mers Dataset
Hierarchical Clustering	Connectivity	**6.2496**	**6.1829**	8.6980	9.2091	14.0222	14.8468
	Dunn	**0.2014**	**0.2379**	0.0260	0.0253	0.0349	0.0347
	Silhouette	0.6732	0.6726	0.7043	0.6991	0.6350	0.6329
	APN	**0.0000**	**0.0000**	0.0000	0.0327	0.0017	0.1009
	AD	0.0213	0.0217	0.0089	0.0099	0.0083	0.0091
	ADM	**0.0000**	**0.0000**	0.0000	0.0014	0.0000	0.0022
	FOM	0.0022	0.0021	0.0010	0.0010	0.0009	0.0008
Kmeans Clustering	Connectivity	9.9143	11.6202	10.9143	12.6202	12.5075	12.9976
	Dunn	0.0232	0.0208	0.0264	0.0234	0.0295	0.0314
	Silhouette	0.0295	0.7045	0.7034	0.6988	0.7487	0.7332
	APN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AD	0.0090	0.0094	0.0089	0.0092	0.0048	0.0053
	ADM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	FOM	0.0010	0.0010	0.0010	0.0010	**0.0006**	**0.0006**
PAM Clustering	Connectivity	8.6484	10.0341	18.4726	18.8817	12.2706	17.7635
	Dunn	0.0067	0.0076	0.0046	0.0037	0.0034	0.0070
	Silhouette	**0.7521**	**0.7355**	0.7236	0.7039	0.7112	0.6849
	APN	0.0000	0.0000	0.0001	0.0025	0.0002	0.0009
	AD	0.0054	0.0059	0.0045	0.0049	**0.0039**	**0.0044**
	ADM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	FOM	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008

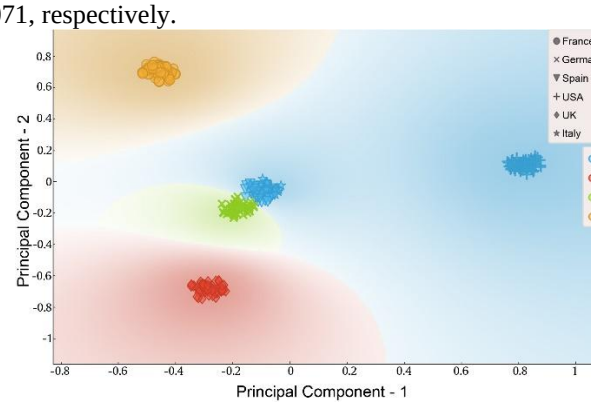
2.6 Agglomerative hierarchical clustering

Agglomerative Nesting (AGNES), commonly known as agglomerative hierarchical clustering, is used to cluster data points on the basis of their similarity. It operates in a bottom to top manner, i.e., a data point at first taken as a single-element group. At every move of the algorithm, couple of most similar groups are merged into a new large cluster. The first step is to calculate (dis)similarity information (Euclidean distance) between each pair of data points in the dataset, followed by joining together the data points or clusters that are nearest in proximity using the linkage function. Average or mean linkage function is used in our implementation. In the case of average linkage function, the interval between two groups is determined as the average distance interval between the data points in batch 1 and the data points in batch 2. After connecting the data points in a data set into a hierarchical cluster tree, it is significant to evaluate that the distances (heights) in the tree resemble the original distances precisely. This is achieved by computing the correlation between the cophenetic distance interval and real distance interval. The clustering is valid only if the connecting data points in the cluster tree has a firm correlation with the distances interval between data points in the original distance matrix [38-43]. The cophenetic correlation coefficient value close to 1 reflects better clustering of the data. The correlation between cophenetic distance and the original distance for 2mers and 3mers datasets are found to be 0.9069 and 0.9071, respectively.

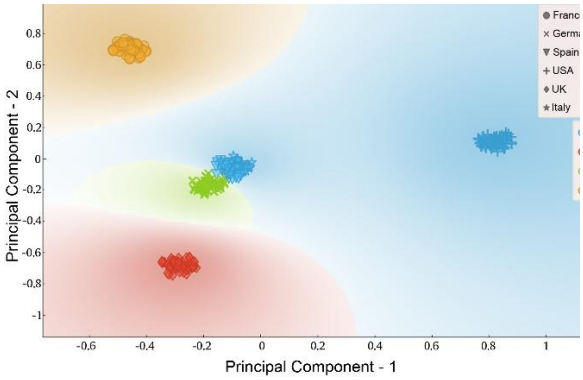
3 Results analysis

The hierarchically clustered 2mers and 3mers data contains a 50-dimensional numeric vector for each genome sequence, which is beyond the scope of visualization due to its high dimensions. Visualization of analyzed data is essential to interpret and derive knowledgeable insights. Transformation or dimensionality reduction should be applied to the data to ensure visualization in 2-dimensional space. PCA is an unsupervised linear dimensionality reduction technique, which is based on eigen vectors. PCA tries to reduce the dimensions of feature space by preserving the utmost amount of variance of the given data. It is computed using eigen values. The objective of PCA is to recognize a set of uncorrelated characteristics or features called Principal Components (PCs). PC1 retains the maximum amount of diversity of the given data, whereas PC2 reserves the second maximum measure of diversity and so forth. The first few 'm' PCs maintain the most substantial measure of the variance of the given data and thus reduces the dimensions of data from 'n' to 'm'. Only salient PCs retaining maximum information are used to project data in 2-dimension or 3-dimension (low dimensional space). Figure 4 shows a 2-dimensional projection of clustered numerical vectors using PCA.

.



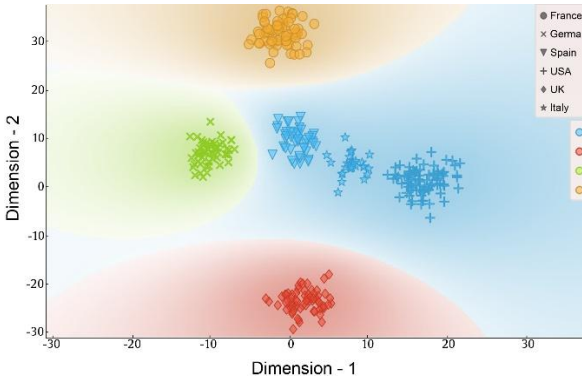
(a) Visualization of 2mers numerical vectors dataset using PCA



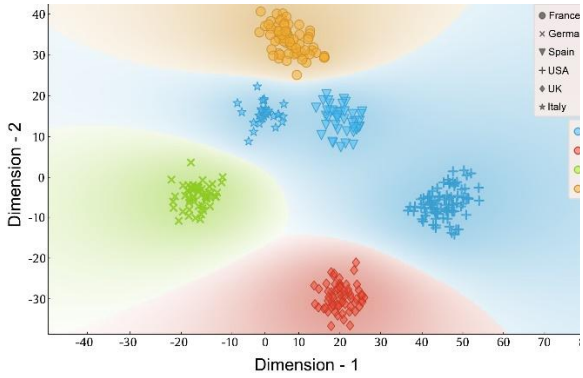
(b) visualization of 3mers numerical vectors dataset using PCA

Figure 4: 2-Dimensional projection using PCA

t-SNE is a nonlinear dimensionality reduction algorithm appropriate for embedding high-dimensional data into 2-dimensional or 3-dimensional space, making data feasible for visualization. The most significant characteristic of *t*-SNE is that it specifically transforms data points of high dimensions into low-dimensional space in such a way that similar and dissimilar data points in high-dimensional space are also reflected and retained in low-dimensional space. This transformation is achieved by assigning a high probability for similar data points and a very low probability for dissimilar data points. Further, it minimizes the Kullback–Leibler divergence among high and low dimensional space with respect to positions of the data points. Since it deals with high-dimensional data, it leads to crowding problem, which is skillfully handled by employing Cauchy distribution or Student t-distribution. Perplexity is the most essential hyperparameter, which is defined as an effectual number of neighbors for a data point. The recommended value for perplexity lies between 5 and 50. Figure 5 depicts the 2-dimensional visualization of clustered numerical vectors using *t*-SNE.



(a) Visualization of 2mers numerical vectors dataset using t-SNE



(b) visualization of 3mers numerical vectors dataset using t-SNE

Figure 5: 2-Dimensional projection using t-SNE with perplexity = 42

4 Discussion

The clustering results in this study have been compared with the state-of-the-art (SOTA) methods from recent literature. In particular, the application of dimensionality reduction techniques like PCA and t-SNE, followed by clustering methods such as K-means, has shown promising results for the analysis of genomic data. Notably, Hozumi et al. (2021) employed UMAP-assisted K-means clustering for analyzing large-scale SARS-CoV-2 mutation datasets, achieving meaningful insights into viral evolution. Similarly, Achtman et al. (2022) utilized hierarchical clustering for bacterial genomes, demonstrating the power of such methods in structuring vast amounts of genomic data into relevant species, subspecies, and populations. In comparison to these studies, the clustering results in this paper exhibit some differences. For example, our study employed traditional PCA and t-SNE dimensionality reduction techniques before performing K-means clustering, whereas Hozumi et al. (2021) utilized UMAP, a non-linear dimensionality reduction technique, which may offer more accurate representations of complex, high-dimensional datasets. While both PCA and UMAP are effective in reducing dimensionality, UMAP tends to preserve local structures better, which could be beneficial in certain contexts. However, PCA remains a powerful and efficient technique, especially when computational resources are limited. The choice of dimensionality reduction method could explain some differences between our results and those observed by Hozumi et al. (2021). The novelty of our work lies in the application of these well-established clustering and dimensionality reduction techniques in virology and genomics, particularly in genomic sequence analysis. This approach is crucial for visualizing high-dimensional data from genomic sequences, such as 2mers and 3mers, which are commonly encountered in viral and bacterial genomics. As demonstrated by Achtman et al. (2022), hierarchical clustering has been instrumental in analyzing large genomic datasets, providing insights into the hierarchical relationships between bacterial genomes. By using PCA and t-SNE for dimensionality reduction and K-means for clustering, we further refine these established techniques for the visualization and interpretation of viral genome sequences. One of the key advantages of our methodology is its simplicity and accessibility for researchers with limited computational resources. While UMAP may offer more complex representations, the combination of PCA and t-SNE provides a clear, interpretable view of the clustered data, with minimal computational overhead. Moreover, our method is applicable to a wide variety of genomic studies, including the analysis of SARS-CoV-2 mutations and bacterial genomics, as shown by the cited works.

5 Conclusion

In our comprehensive analysis of SARS-CoV-2 genome sequences from multiple nations, we employed advanced computational techniques to unravel genetic diversity patterns with significant implications for virology, drug development, and vaccination strategies. We compiled genome sequences from heavily impacted nations including the USA, France, UK, Germany, Spain, and Italy. These sequences were transformed into 50-dimensional numerical vectors through our gene2vec approach, facilitating quantitative genetic analysis. The crux of our analysis lay in agglomerative hierarchical clustering, unveiling hidden relationships within genome sequences. By rigorously assessing clustering tendency, identifying optimal cluster numbers using the Elbow and Average Silhouette methods, and employing internal and stability measures, we identified hierarchical clustering as the most effective algorithm. Utilizing dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), we visualized the high-dimensional data. The PCA projection showed 2-dimensional cluster patterns, while t-SNE revealed intricate similarities in a reduced dimensionality. Numerical results indicated clustering tendencies with Hopkins statistic values of H = 0.01031 for 2-mers and H = 0.01608 for 3-mers, signifying meaningful clusters. Both Elbow and Average Silhouette methods suggested optimal cluster numbers of k = 4 for both datasets. In conclusion, our analysis demonstrated the genetic diversity of SARS-CoV-2 across nations.

Clustering patterns suggested shared genetic features among countries, impacting vaccine and therapeutic strategies. This study underscores computational methodologies in understanding complex biological phenomena, contributing to preparedness against emerging infectious diseases and advancements in genomics research.

References

[1] Banerjee A, Davé RN (2004) Validating clusters using the Hopkins statistic. IEEE Int Conf Fuzzy Syst 1:149–153. https://doi.org/10.1109/FUZZY.2004.1375706

[2] Banerjee A, Kulcsar K, Misra V, et al (2019) Bats and coronaviruses. Viruses 11:7–9. https://doi.org/10.3390/v11010041

[3] Behjati S, Tarpey PS (2013) What is next generation sequencing? Arch Dis Child Educ Pract Ed 98:236–238. https://doi.org/10.1136/archdischild-2013-304340

[4] Bouguettaya A, Yu Q, Liu X, et al (2015) Efficient agglomerative hierarchical clustering. Expert Syst Appl 42:2785–2797. https://doi.org/10.1016/j.eswa.2014.09.054

[5] Brock G, Pihur V, Datta S, Datta S (2008) ClValid: An R package for cluster validation. J Stat Softw 25:1–22. https://doi.org/10.18637/jss.v025.i04

[6] Compeau PEC, Pevzner PA, Tesler G (2021) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29:987–991. https://doi.org/10.1038/nbt.2023

[7] De Wit E, Van Doremalen N, Falzarano D, Munster VJ (2016) SARS and MERS: Recent insights into emerging coronaviruses. Nat Rev Microbiol 14:523–534. https://doi.org/10.1038/nrmicro.2016.81

[8] Dorp L Van, Acman M, Richard D, et al (2022) Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83:104351. https://doi.org/10.1016/j.meegid.2020.104351

[9] Graham RL, Donaldson EF, Baric RS (2013) A decade after SARS: Strategies for controlling emerging coronaviruses. Nat Rev Microbiol 11:836–848. https://doi.org/10.1038/nrmicro3143

[10] Gralinski LE, Menachery VD (2020) Return of the coronavirus: 2019-nCoV. Viruses 12:1–8. https://doi.org/10.3390/v12020135

[11] Hershberg R, Petrov DA (2008) Selection on Codon Bias. Annu Rev Genet 42:287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442

[12] Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441

[13] Hui DS, Azhar EI, Kim YJ, et al (2018) Middle East respiratory syndrome coronavirus: risk factors and determinants of primary, household, and nosocomial transmission. Lancet Infect Dis 18: e217–e227. https://doi.org/10.1016/S1473-3099(18)30127-0

[14] Jolliffe IT, Cadima J, Cadima J (2016) Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci 374:

[15] Karlin S (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol 1:598–610

[16] Khailany RA, Safdar M, Ozaslan M (2022) Gene Reports Genomic characterization of a novel SARS-CoV-2. 19:

[17] Kuehn BM (2013) More evidence emerges that bats may have spread SARS. JAMA - J Am Med Assoc 310:2138. https://doi.org/10.1001/jama.2013.283495

[18] Lam TTY, Shum MHH, Zhu HC, et al (2020) Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature. https://doi.org/10.1038/s41586-020-2169-0

[19] Li H, Liu SM, Yu XH, et al (2020a) Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int J Antimicrob Agents 2019:105951. https://doi.org/10.1016/j.ijantimicag.2020.105951

[20] Li Q, Guan X, Wu P, et al (2020b) Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 382:1199–1207. https://doi.org/10.1056/NEJMoa2001316

[21] Lu G, Wang Q, Gao GF (2015) Bat-to-human: Spike features determining "host jump" of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol 23:468–478. https://doi.org/10.1016/j.tim.2015.06.003

[22] Maaten L van der, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605

[23] Mikolov T, Corrado G, Chen K, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA.

[24] Ng P (2017) dna2vec: Consistent vector representations of variable-length k-mers. CoRR 1–10

[25] Rehurek R, Sojka P (2010) Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp 45–50

[26] Reusken CBEM, Haagmans BL, Müller MA, et al (2013) Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: A comparative serological study. Lancet Infect Dis 13:859–866. https://doi.org/10.1016/S1473-3099(13)70164-6

[27] Saraçli S, Doğan N, Doğan I (2013) Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequalities Appl 2013:1–8. https://doi.org/10.1186/1029-242X-2013-203

[28] Song Z, Xu Y, Bao L, et al (2019) From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses 11:. https://doi.org/10.3390/v11010059

[29] Su S, Wong G, Liu Y, et al (2015) MERS in South Korea and China: a potential outbreak threat? Lancet 385:2349–2350. https://doi.org/10.1016/S0140-6736(15)60859-5

[30] University JH COVID-19 Dashboard by the Center for Systems Science Engineering. https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6. Accessed 5 May 2020

[31] Wang C, Liu Z, Chen Z, et al (2020) The establishment of reference sequence for SARS-CoV-2 and variation analysis. J Med Virol 92:667–674. https://doi.org/10.1002/jmv.25762

[32] WHO COVID-19 Global Situation Data - 106. In: World Heal. Organ. https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed 5 May 2020

[33] Wong JEL, Leo YS, Tan CC (2020) COVID-19 in Singapore - Current Experience: Critical Global Issues That Require Attention and Action. JAMA - J Am Med Assoc 323:1243–1244. https://doi.org/10.1001/jama.2020.2467

[34] Wu YC, Chen CS, Chan YJ (2020) The outbreak of COVID-19: An overview. J Chinese Med Assoc 83:217–220. https://doi.org/10.1097/JCMA.0000000000000270

[35] Xu X, Chen P, Wang J, et al (2020) Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci 63:457–460. https://doi.org/10.1007/s11427-020-1637-5

[36] Yang D, Leibowitz JL (2015) The structure and functions of coronavirus genomic 3' and 5' ends. Virus Res 206:120–133. https://doi.org/10.1016/j.virusres.2015.02.025

[37] Zhou P, Yang X Lou, Wang XG, et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7

[38] Zumla A, Hui DS, Perlman S (2015) Middle East respiratory syndrome. Lancet 386:995–1007. https://doi.org/10.1016/S0140-6736(15)60454-8

[39] Khan, G.A., Hu, J., Li, T. et al. multi-view data clustering via non-negative matrix factorization with manifold regularization. Int. J. Mach. Learn. & Cyber. 13, 677–689 (2022). https://doi.org/10.1007/s13042-021-01307-7

[40] Diallo, B., Hu, J., Li, T. et al. multi-view document clustering based on geometrical similarity measurement. Int. J. Mach. Learn. & Cyber. 13, 663–675 (2022). https://doi.org/10.1007/s13042-021-01295-8

[41] Khan, G.A., Hu, J., Li, T. et al. multi-view low rank sparse representation method for three-way clustering. Int. J. Mach. Learn. & Cyber. 13, 233–253 (2022). https://doi.org/10.1007/s13042-021-01394-6

[42] Hozumi, Y., Wang, R., Yin, C., & Wei, G.-W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. Computers in Biology and Medicine, 131, 104264. https://doi.org/10.1016/j.compbiomed.2021.104264

[43] Achtman, M., Zhou, Z., Charlesworth, J., & Baxter, L. (2022). EnteroBase: Hierarchical clustering of 100,000s of bacterial genomes into species/subspecies and populations. Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1849), 20210240. https://doi.org/10.1098/rstb.2021.0240.