International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 6
ISSN: 2088-5334

Analyzing Kinship in Severe Acute Respiratory Syndrome Coronavirus 2 DNA Sequences Based on Hierarchical and K-Means Clustering Methods Using Multiple Encoding Vector

Evander Banjarnahor [a], Alhadi Bustamam [a,*], Titin Siswantining [a], Patuan Tampubolon [a]

[a] *Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Indonesia, Depok, 16424, Indonesia*
*Corresponding author: *alhadi@sci.ui.ac.id*

Abstract—**Based on the World Health Organization data obtained in mid-April 2021, Coronavirus or Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has already infected more than 134.9 million people worldwide. The virus attacks human breathing, which can cause lung infections and even death. More than 2.9 million people worldwide have died due to coronavirus infection. Meanwhile, more than 1.5 million people in Indonesia have been infected, and 42.5 thousand died because of this coronavirus. Based on this data, carrying out a kinship analysis of the coronavirus is important to reduce its spread. Identifying the kinship of the COVID-19 virus and its spread can be done by forming a phylogenetic tree and clustering. This study uses the Multiple Encoding Vector method in analyzing the sequences and Euclidean distance to determine the distance matrix. This research will then use the Hierarchical clustering method to determine the number of initial centroids, which will be used later by the K-Means clustering method kinship in the SARS-CoV-2 DNA sequence. This study took samples of DNA sequences of SARS-CoV-2 from several infected countries. From the simulation results, the ancestors of SARS-CoV-2 came from China. The analysis results also show that the closest ancestors of COVID-19 to Indonesia came from India. The SARS-CoV-2 DNA sequence also consisted of nine clusters, and the sixth cluster has the greatest number of members.**

Keywords—**Sequence alignment; bioinformatics; clustering; DNA kinship, phylogenetic analysis.**

Manuscript received 24 Jun. 2021; revised 24 Sep. 2021; accepted 17 Jan. 2022. Date of publication 31 Dec. 2022.

I. INTRODUCTION

Bioinformatics is a branch of science that applies computational science in the world of biology [1]. Bioinformatics is an interdisciplinary science that aims to collect, store, and analyze biological data [2]. Bioinformatics is an interdisciplinary field that is very useful in research related to the biology of macromolecules such as DNA, RNA, and proteins that accelerate analysis using computations. This field of science covers the areas of Mathematics, Biology, Statistics, and Computer Science. Polanski [3] explain that Bioinformatics is a scientific discipline emerging for practical purposes that introduces order into large datasets generated by new molecular biology technologies. This technique stems from large-scale DNA alignment and the need for tools for sequence assembly and sequence annotation, determining the location of protein-coding regions in the DNA. The development of bioinformatics has been in a broader area. Bioinformatics is an interdisciplinary area that studies mathematical and computational biology, molecular evolution, genetics, and molecular and cell biology.

In this study, the researchers conducted a genomic analysis using computational technique from DNA sequences of SARS-CoV-2. This research also aimed to decrease the spread of the virus. Therefore, the identification of the virus's kinship is essential. The most familiar way to see the kinship of a virus is by building a phylogenetic tree or through clustering. The researcher used the Multiple Encoding Vector (MEV) method to analyse the DNA's sequences, hierarchical clustering to form a phylogenetic tree, and K-Means to group them into a cluster. There are several studies that analyse DNA sequences. A study conducted by Crochemore et al. [4], for example, used a linear-time and a linear-space algorithm to compare two sequences by considering all the minimal words from the missing words in the DNA sequence. This study aimed to increase efficiency in sequence analysis through the information that is not in the sequence.

A study by Bustamam et al. [5] used K-Means clustering method in grouping hepatitis B virus (HBV) DNA sequences.

Furthermore, a study by Bustamam et al. [6] used the tribe Markov clustering (Tribe-MCL) algorithm to classify and analyze the protein sequences of the herpes simplex virus. Furthermore, Bustamam et al. [7] used hierarchical clustering with a sparse matrix to conduct phylogenetic analysis on the MERS-CoV (Middle East respiratory syndrome-related coronavirus) DNA sequence. The results of the study indicate that the MERS-CoV ancestor originated in Egypt. A study by Li et al. [8] used MEV to perform sequence alignment and DNA data analysis on mammals, bacteria, and viruses.

A study by Qian and Luan [9] also conducted phylogenetic analysis on DNA sequences using the Fractional Fourier transform. Furthermore, a study by Huang & Girimurugan [10] also in his research uses the method of converting DNA into 12-dimensional numerical vectors in real-time. This method shows accurate and effective results in analyzing sequences and grouping data. A study by Criscuolo [11] used the fast alignment-free method in analyzing DNA sequences and showed that the method can accurately form phylogenetic trees and has a fast time. [12] in his research using the K-Mer natural vector method in analyzing protein sequences. This study demonstrates a precise method of demonstrating the evolutionary relationship of influenza viruses. A study by Muflikhah et al. [13] used the clustering method for the hepatitis B virus DNA sequence. This study employed the Hierarchical K-Means Algorithm, which aimed to improve the K-Means clustering method by defining the initial cluster center using the average results of the hierarchical clustering method. According to the findings of this study, the proposed method outperforms the traditional K-Means clustering algorithm in terms of performance.

A study by Gao et al. [14] used an alignment-free method in analyzing the 2019-nCoV DNA sequence and informing its phylogenetic tree using the neighbor-joining method. The results of his research indicate that COVID-19 may have spread before it started in Wuhan. Furthermore, a study by Ma et al. Furthermore, a study by Ma et al. [15], used the Position-Weighted K-Mers method in analyzing the DNA sequences of the HIV-1 virus. The results of this study indicate that this method is simple and computationally fast. A study by Gamage et al. [16] used the K-Mer Forest method in analyzing sequences and phylogenetic trees built based on distances calculated with the help of the k-medoid clustering algorithm. The results of this study indicate that the method has significant efficiency and accuracy compared to traditional methods. A study by Das et al. [17] employed the tri-nucleotide representation method in a sequence analysis that accounts for their biochemical properties. This research employed the UPGMA method with a distance matrix approach to construct its phylogenetic tree. This study compared several methods that achieve good accuracy and have less time complexity.

Furthermore, He et al. [18] used the positional correlation natural vector to analyze and calculate the bacteria and viruses. The results of this study indicated that the techniques used are accurate and fast in the phylogenetic analysis. A study by Maio [19] used a cumulative indel model in his alignment analysis, and this study demonstrated that the proposed model is fast and accurate in the phylogenetic analysis.

Moreover, a study by Amiroch et al. [20] used Needleman Wunsch Algorithm to analyze the DNA sequence and used the Neighbor Joining method with the Felsenstein model to construct a phylogenetic tree in Influenza virus type A/H9N2 in Indonesia. A study by Das et al. [21] used Codon Feature-based Amino acid Sequence Analyzer (CoFASA) to analyze the protein sequence's similarity, resulting in a 20-dimensional vector. The results of this study showed accurate and similar results when compared to other alignment methods using ClustalW, Clustal, MAFFT, and MUSCLE. A study by Banjarnahor et al. [22] used hierarchical clustering and multiple sequence alignment (MSA) to analyze the SARS-CoV-2 DNA sequence. According to the findings of this study, the ancestors of SARS-CoV-2 originated in China. State of the art from this research is presented in Fig.1.

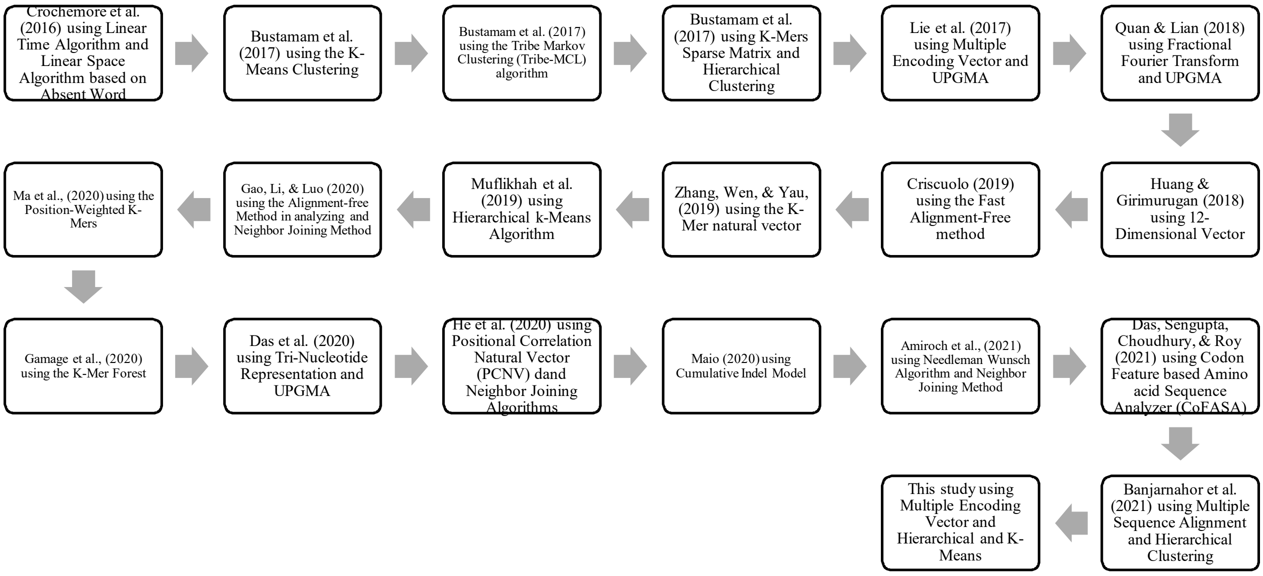

Crochemore et al. (2016) using Linear Time Algorithm and Linear Space Algorithm based on Absent Word
Bustamam et al. (2017) using the K-Means Clustering
Bustamam et al. (2017) using the Tribe Markov Clustering (Tribe-MCL) algorithm
Bustamam et al. (2017) using K-Mers Sparse Matrix and Hierarchical Clustering
Lie et al. (2017) using Multiple Encoding Vector and UPGMA
Quan & Lian (2018) using Fractional Fourier Transform and UPGMA
Huang & Girimurugan (2018) using 12-Dimensional Vector
Criscuolo (2019) using the Fast Alignment-Free method
Zhang, Wen, & Yau, (2019) using the K-Mer natural vector
Muflikhah et al. (2019) using Hierarchical k-Means Algorithm
Gao, Li, & Luo (2020) using the Alignment-free Method in analyzing and Neighbor Joining Method
Ma et al., (2020) using the Position-Weighted K-Mers
Gamage et al., (2020) using the K-Mer Forest
Das et al. (2020) using Tri-Nucleotide Representation and UPGMA
He et al. (2020) using Positional Correlation Natural Vector (PCNV) dand Neighbor Joining Algorithms
Maio (2020) using Cumulative Indel Model
Amiroch et al., (2021) using Needleman Wunsch Algorithm and Neighbor Joining Method
Das, Sengupta, Choudhury, & Roy (2021) using Codon Feature based Amino acid Sequence Analyzer (CoFASA)
Banjarnahor et al. (2021) using Multiple Sequence Alignment and Hierarchical Clustering
This study using Multiple Encoding Vector and Hierarchical and K-Means

Fig. 1 State of the art from this research

From the above description, it can be concluded that the researchers analyze the kinship of a virus DNA sequence by constructing phylogenetic trees or clustering. Researchers used MEV in this study owing to its fast computation time, as referenced in the study [8]. Therefore, researchers analyze kinship in DNA sequences of SARS-CoV-2 based on the hierarchical and K-Means clustering methods using MEV.

The COVID-19 (coronavirus disease 2019) pandemic is an outbreak of pneumonia that began in Wuhan, Hubei Province, Central China [23]. This disease causes respiratory problems in humans, and SARS (severe acute respiratory syndrome) is a form of the acute respiratory syndrome. SARS-CoV-2 is a modern coronavirus responsible for the global pandemic of COVID-19, a human respiratory disease, and it poses a danger to the public's health and safety. Below is an image of the RNA-dependent RNA polymerase (RdRp) from SARS-CoV-2.

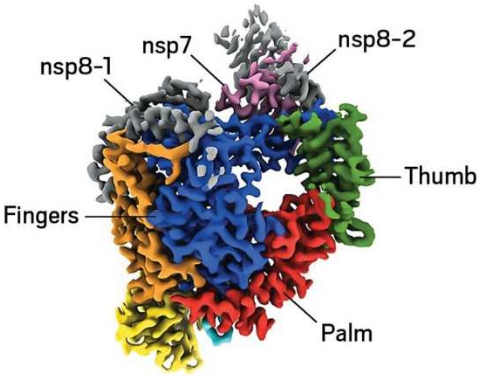


Fig. 2 Structure of SARS-CoV-2 RNA-dependent RNA polymerase, which includes the small proteins nsp7 and nsp8 [24]

The new coronavirus's RdRp is complicated, with the polymerase nsp12 linked to two smaller proteins called nsp7 and nsp8. This RdRp, among others, takes the form of a right hand, complete with fingers, thumb, and palm domains [24]. SARS-CoV-2 is related to the Middle East respiratory syndrome-related coronavirus (MERS-CoV) and SARS-CoV in terms of genetic tree analysis [25]. Table I presents the differences between SARS-CoV-2, SARS-CoV, and MERS-CoV.

TABLE I
THE MAIN DIFFERENCES BETWEEN SARS-COV-2, MERS AND SARS

	SARS-CoV-2	MERS	SARS
Time of origin	December 2019	June 2012	November 2002
Place of origin	Wuhan, China	Jeddah, Saudi Arabia	Fushan, China
The transmission	Animal to human, then human to human	Animal to human, then human to human	Animal to human, then human to human
The main transmission	Droplets or direct contact	Air and direct contact	Air and direct contact
Incubation period	4–7	2–15	2–14
Main symptoms	Fever, cough, exhaustion, and shortness of breath	Fever, cough, tiredness, shortness of breath, and acute renal failure	Fever, cough, exhaustion, and shortness of breath

COVID-19, formerly known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been renamed 2019-nCov. While preliminary research suggested a correlation between fish and wild animal markets, most infection cases suggested that transmission from animals to humans was likely. SARS-CoV-2 can be transmitted from human to human through droplets or direct touch, according to an increasing number of studies. SARS-CoV-2 can spread between humans and has a high risk of becoming a pandemic [25]. The symptoms of COVID-19 include fever, cough, shortness of breath, nausea, and loss of taste and smell.

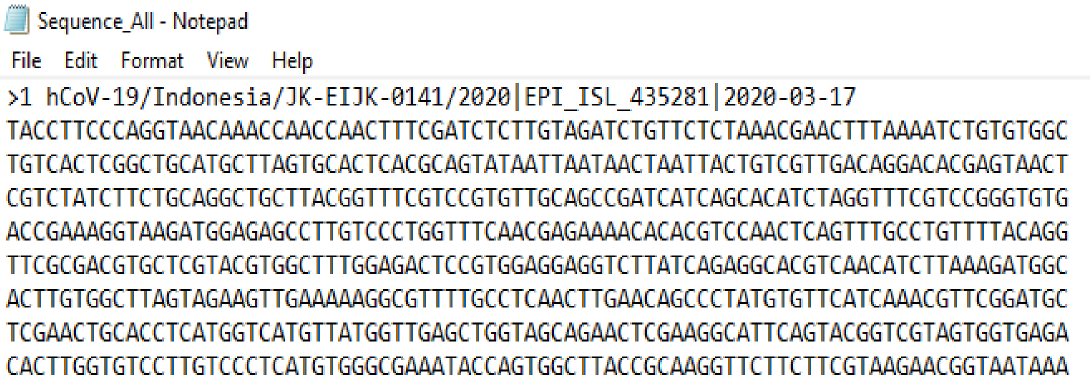
The COVID-19 infection is rapidly spreading worldwide (Corona Virus Disease 2019). This infection originally started in Wuhan, Hubei Province, China, and has since spread worldwide. COVID-19 poses a great danger to the well-being of humans, especially when intense respiratory contamination occurs. COVID-19 has already infected more than 134.9 million people worldwide as of mid-April 2021, according to World Health Organization (WHO) reports. This virus targets the respiratory tract, thus resulting in lung infections and even death in humans. Over 2.9 million people around the world have already died due to COVID-19. Fortunately, vaccines have already been created to manage COVID-19, including the Oxford–AstraZeneca COVID-19 Vaccine, Sinopharm, Moderna, Pfizer-BioNTech, Sinovac, and Novavax.

II. Materials and Method

A. Research Data

This study will look at the relationship between 2305 DNA sequences of SARS-CoV-2 from different countries obtained from GenBank *via* https://www.gisaid.org. Gisaid, established in May 2008, is a site that provides open access to genome data of the influenza virus and coronavirus, and the COVID-19 virus. Information about influenza viruses and coronaviruses that cause COVID-19 are quickly shared *via* Gisaid. To help researchers understand how viruses grow and spread during epidemic and pandemic, the site provides genetic sequence data, clinical and epidemiological data of human viruses, and geographic- and species-specific data of avian and other animal viruses. Since its launch, Gisaid has played an essential role in sharing data among the WHO collaborating centers and National Influenza Centres for the biennial influenza virus vaccine recommendations by the WHO Global Influenza Surveillance and Response System.

Gisaid contains SARS-COV-2 data obtained from all countries worldwide. People can access the data by registering an account in advance or an email address. After writing a statement, the user can search for the needed data and download it. The results of the data download will be in the form of FASTA. Fig. 3 below is an illustration of the SARS-CoV-2 DNA. In this study, each data of 2305 and SARS-CoV-2 sequences from various countries is presented in Table 2:



```
>1 hCoV-19/Indonesia/JK-EIJK-0141/2020|EPI_ISL_435281|2020-03-17
TACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGC
TGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACT
CGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTG
ACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGG
TTCGCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGC
ACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAAACGTTCGGATGC
TCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTAGTGGTGAGA
CACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAA
```

Fig. 3 Illustration of the DNA SARS-CoV-2

TABLE II
2305 CODES FOR DNA SEQUENCES OF SARS-COV-2 AND COUNTRY OF ORIGIN

Code	Country	Code	Country
1-309	Indonesia	1117-1216	Australia
310-314	Brunei	1217-1316	South Africa
315-350	Philippines	1317-1416	Nigeria
351-356	Cambodia	1417-1516	USA
357-457	Malaysia	1517-1608	Brazil
458-558	Thailand	1617-1708	Italy
559-599	Myanmar	1709-1807	France
600-699	Singapore	1808-1907	Russia
700-717	Timor Leste	1908-2006	England
718-816	Vietnamese	2007-2106	German
817-916	China	2107-2205	Spanish
917-1016	Japan	2206-2305	Turkey
1017-1116	India		

B. Multiple Encoding Vector

Let Ł be a set of four bases, namely, {A, C, G, T}, and $Q = (S_1, S_2, \ldots, S_n)$ is a DNA sequence of length n, that is, $S_i \epsilon L$, I = 1,2,…,n. The four bases are split into two groups based on their three types of chemical and physical characteristics. Bases A and G are purines stood by the letter R. Bases C and T are pyrimidines represented by Y. In another grouping, nucleotides A and C are amines represented by M, G and T are keto indicated by K, G and C have a strong H bond and are represented by the letter S. Bases A and T contain weak H bonds and are represented by W. Three numerical values are specified in this system for each of the letters R, Y, M, K, S and W. To describe the R and Y distributions in the order Q, the letters A and G are replaced by R, and the letters C and T are replaced by Y in the order Q. Then, there are only two types of letters in the sequence: R and Y. For R, defined as $W_R(.):\{R, Y\} \rightarrow \{0,1\}$, so that $W_R(S_i) = 1$ if $S_i = R$ and 0 otherwise. For R, we define n_R, μ_R, D_2^R to explain the number of R, the mean of R's position and the difference in R's position in sequence Q:

- Suppose $n_R = \sum_{i=1}^{n} W_R(S_i)$ shows the number of letter R that occurs in line Q.
- Suppose $\mu_R = \sum_{i=1}^{n} i.\frac{W_R(S_i)}{n_R}$ is the average position in which the letter R appears.
- Suppose $D_2^R = \sum_{i=1}^{n} \frac{(i-\mu_R)^2 W_R(S_i)}{n_R n}$ is the 2nd moment scaled from the position of the letter R.

Likewise, for Y, we define $W_Y(.):\{R, Y\} \rightarrow \{0,1\}$, so that $W_Y(S_i) = 1$ if $S_i = Y$ and 0 otherwise. Then, we get three characteristics for Y: n_Y, μ_Y and D_2^Y. Six values were chosen to indicate the distribution of four bases associated with the chemical characteristic in this nucleotide categorization.

As a result, $(n_R, \mu_R, D_2^R \ldots, n_W, \mu_W, D_2^W)$ defines the 18-dimensional vector of the DNA sequence Q. The MEV method uses three letters to encode DNA sequences [8].

C. Distance Matrix

The Euclidean distance is the most commonly used matrix for continuous features. The Euclidean distance equation is calculated as follows:

$$d_e = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{1}$$

where index i loops all the values in the vector. The Euclidean matrix distance is a measure of the geometric distance between two components based on the size (Matthiesen in [22]). The Euclidean distance is used to calculate each distance from each sequence after the 18-dimensional MEV data is found.

D. Hierarchical Clustering

The hierarchical clustering method has two approaches: agglomerative hierarchical clustering (AHC) and divisive hierarchical clustering (DHC). AHC takes a bottom-up approach to clustering problems, whereas DHC takes a top-down method. In the AHC method, each observation started with its cluster and then followed by merging and combining several clusters into a larger one [26]. In AHC method, the computational complexity is decreased by the distance function that could utilize the clustering concept through segregating the dataset into ta total amount of cluster until it becomes a solitary attribute[27]. This will be counted as one step in the hierarchy.This method is continuously used until only one group remains or until the minimum number of groups is reached. The steps in AHC are as follows [28]:

- Calculate the similarity in all pairs of clusters, i.e., calculating the similarity matrix that goes into i and giving the similarity between the *i*-th and *j*-th groups.
- Combine the two clusters that are the most related (closest).
- Update the similarity matrix to represent the new cluster's paired similarities to the original cluster.
- Steps 2–3 should be repeated until only one cluster remains.

A hierarchical tree can be produced using a variety of AHC techniques. A distance difference between two clusters is used in this approach. The AHC method is as follows:

The average linkage method is used to count the average distance between all pairs in a group [29]. The equation of the average observed length is as follows:

$$d_{c_u,c_v} = \underset{(u,v)}{argmin}\left(\frac{1}{|c_u|}\frac{1}{|c_v|}\sum_{x\in c_u}\sum_{y\in c_v} D(x,y)\right) \quad (2)$$

where $|c_u|$ and $|c_v|$ denote the number of each object in cluster c_u and c_v.

The complete linkage method is commonly known as the farthest-neighbor technique. This method has the farthest observation distance in the cluster, as expressed in the following equation:

$$d_{c_u,c_v} = \underset{(u,v)}{argmin}\left(\underset{x\in c_u, y\in c_v}{max} D(x,y)\right) \quad (3)$$

As an optimization problem, the main goal of this approach is to minimize the maximum distance between clusters [29].

E. K-Means Clustering

Owing to its high computational speed, K-Means clustering using the top-down approach is one of the most well-known clustering methods. This method aims to place the data points into several partition to decrease the within-cluster sum of squares so that the pairwise squared deviations of points are decrease in the same cluster until the centroids are stable[30]. This method also requires the user to use three parameters: the number of cluster K, cluster initiation, and distance metric[31]. The cluster number is required as the initial cluster center in this process, which is randomly calculated. As a result, the clustering results vary depending on which version of the K-Means algorithm is used. The K-Means algorithm seeks to minimize the total intra-cluster variation while increasing the total variation within the clusters. As expressed in the following equation, the total intra-cluster variance (W) is the sum of the squares of the distance between the object data and the corresponding cluster core [13]:

$$W(C_k) = \sum_{x_i \in c_k}(x_i - \mu_k)^2 \quad (4)$$

where

x_i= is the data object under cluster C_k

μ_k denotes the average value of data object points allocated to cluster C_k

Each object (x_i) is assigned to the cluster in such a way that the sum of square (SS) distances between objects, and the centroid is as small as possible, as expressed in the equation below:

$$\text{Total withinss} = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k}\sum_{x_i \in c_k}(x_i - \mu_k)^2 \quad (5)$$

This study combines two methods: hierarchical and K-Means clustering. The hierarchical clustering method aims to obtain the number of centroids from the grouping process and to develop a hierarchy of clusters in trees [32]. The result of this grouping will be the initial number of clusters or cluster centers in the K-Means clustering method.

F. Research Flow

Overall, the general flow of this research is presented in Fig. 4. This research uses the R program to input the SARS-CoV-2 DNA data and then performs the alignment process using MEV, which produces an 18-dimensional matrix (TABLE III). From this data, the results of the distance matrix using the Euclidean distance were searched (TABLE IV). The distance matrix data are stored in the Newick (NWK) format. Using the Dendroscope programming, the distance matrix esteem recreates the arrangement of a phylogenetic tree using the average linkage method, which produces a circular cladogram (Figs. 5 and 6), and the complete linkage method, which produces a phylogenetic tree circular cladogram (Figs. 7 and 8).

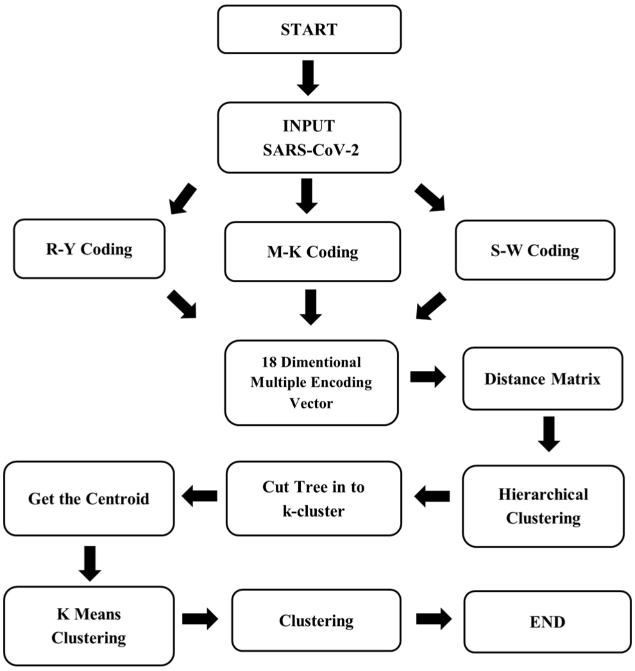


Fig. 4 Flowchart of the Method

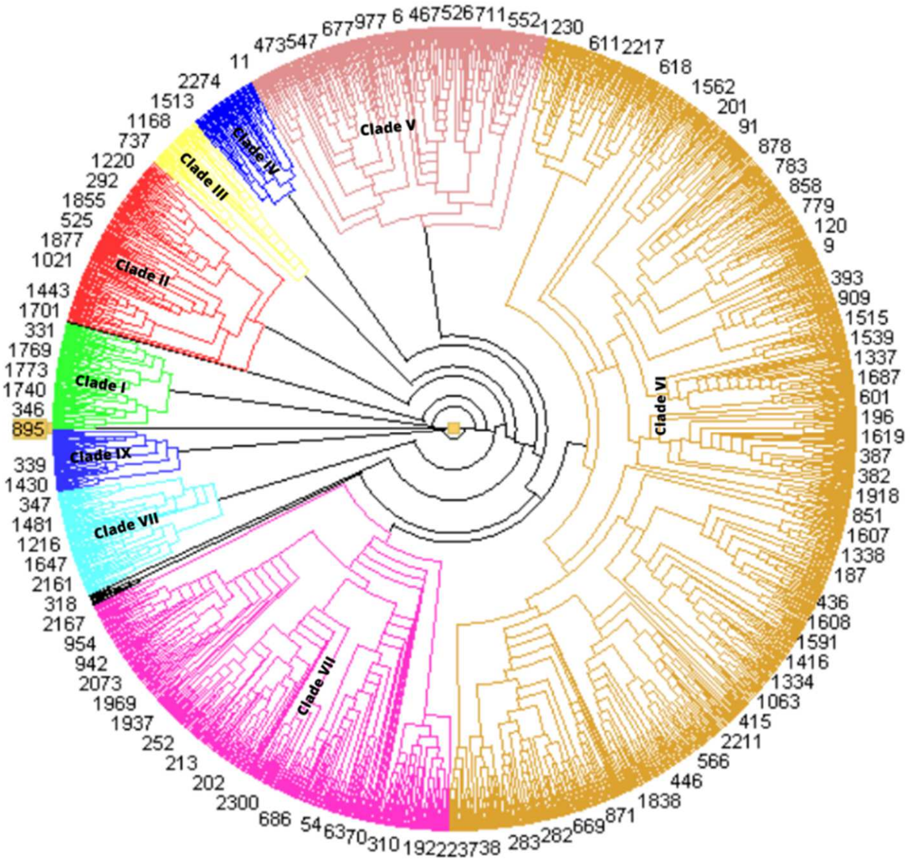


Fig. 5 Circular cladogram of 2305 SARS-CoV-2 DNA data obtained using the average linkage method using the MEV

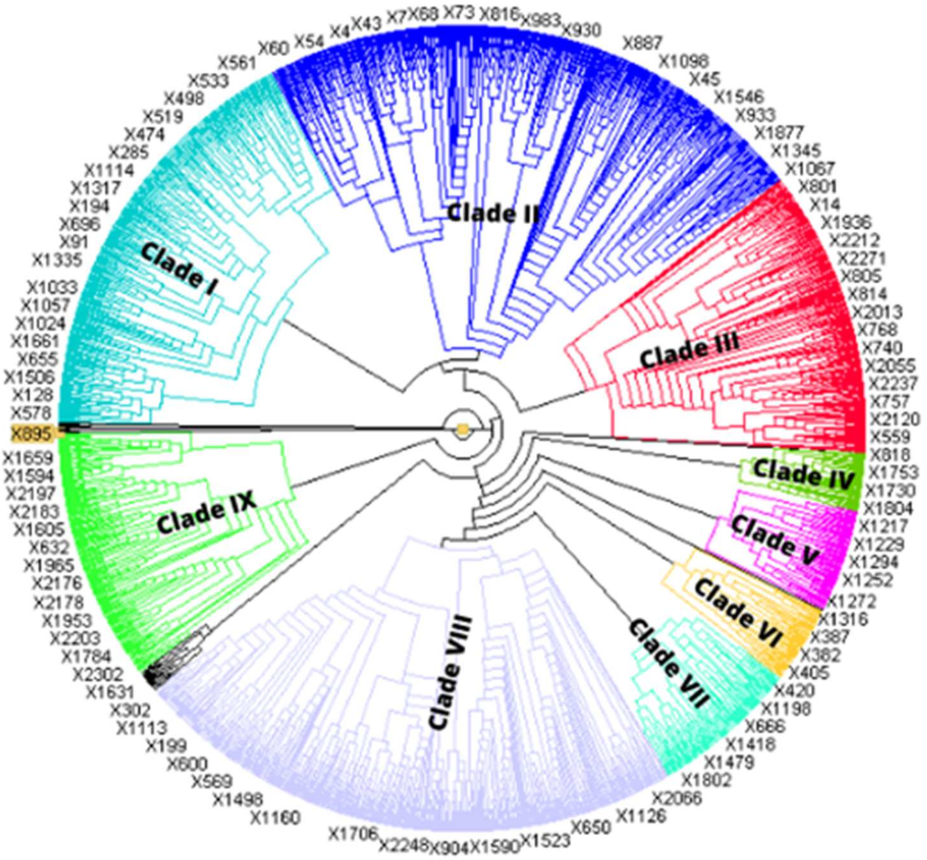

Clade I
Clade II
Clade III
Clade IV
Clade V
Clade VI
Clade VII
Clade VIII
Clade IX
X895
X60 X54 X4 X43 X7 X68 X73 X816 X983 X930 X887 X1098 X45 X1546 X933 X1877 X1345 X1067 X801 X14 X1936 X2212 X2271 X805 X814 X2013 X768 X740 X2055 X2237 X757 X2120 X559 X818 X1753 X1730 X1804 X1217 X1229 X1294 X1252 X1272 X1316 X387 X382 X405 X420 X1198 X666 X1418 X1479 X1802 X2066 X1126 X650 X1523 X1590 X904 X2248 X1706 X1160 X1498 X569 X600 X199 X1113 X302 X1631 X2302 X1784 X2203 X1953 X2178 X2176 X1965 X632 X1605 X2183 X2197 X1594 X1659 X578 X128 X1506 X655 X1661 X1024 X1057 X1033 X1335 X91 X696 X194 X1317 X1114 X285 X474 X519 X498 X533 X561

Fig. 6 Circular cladogram of 2305 SARS-CoV-2 DNA data obtained by employing the average linkage method using MSA

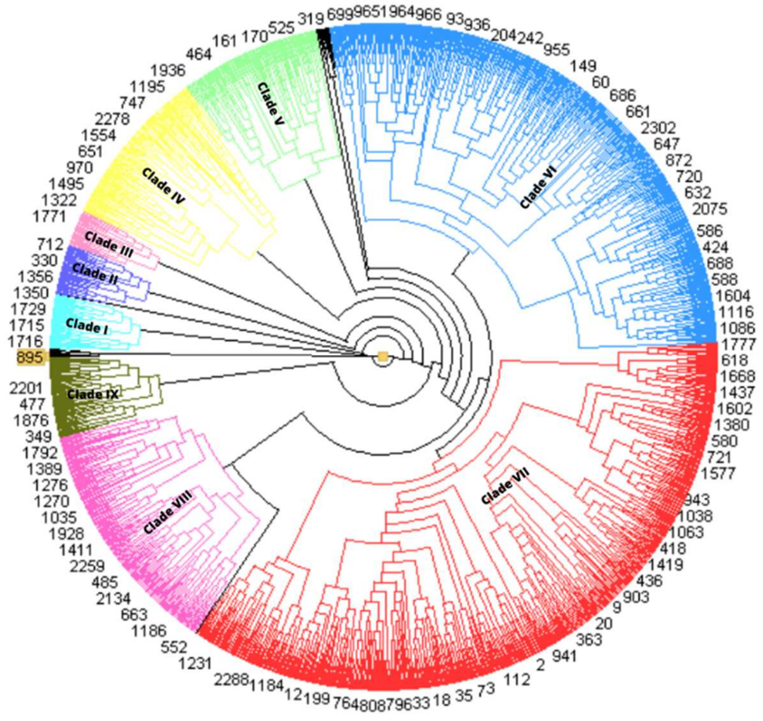


Fig. 7 The circular cladogram of 2305 SARS-CoV-2 DNA data obtained by employing the complete linkage method using MEV

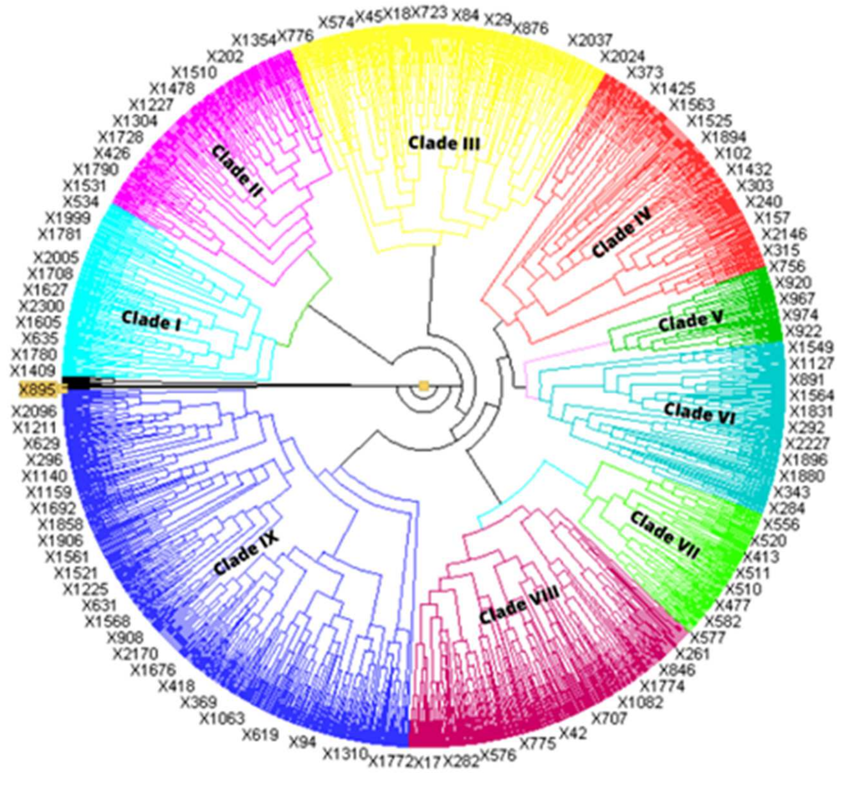

Clade I
Clade II
Clade III
Clade IV
Clade V
Clade VI
Clade VII
Clade VIII
Clade IX
X895
X574 X45 X18 X723 X84 X29 X876 X2037 X2024 X373 X1425 X1563 X1525 X1894 X102 X1432 X303 X240 X157 X2146 X315 X756 X920 X967 X974 X922 X1549 X1127 X891 X1564 X1831 X292 X2227 X1896 X1880 X343 X284 X556 X520 X413 X511 X510 X477 X582 X577 X261 X846 X1774 X1082 X707 X42 X775 X576 X282 X17 X1772 X1310 X94 X619 X1063 X369 X418 X1676 X2170 X908 X1568 X631 X1225 X1521 X1561 X1906 X1858 X1692 X1159 X1140 X296 X629 X1211 X2096 X1409 X1780 X635 X1605 X2300 X1627 X1708 X2005 X1781 X1999 X534 X1531 X1790 X426 X1728 X1304 X1227 X1478 X1510 X202 X1354 X776

Fig. 8 Circular cladogram of 2305 SARS-CoV-2 DNA data obtained by employing the complete linkage method using MSA

III. Results and Discussion

These are the 18-dimensional MEV and distance matrix data from 2305 DNA SARS-COV-2 sequences (TABLE III). This study compared the results of the grouping using the MSA. The SARS-CoV-2 virus circular cladogram is presented below. Figure 5 presents the circular cladogram of 2305 SARS-CoV-2 DNA data obtained using the average linkage method using the MEV. Figure 6 presents the circular cladogram of 2305 SARS-CoV-2 DNA data obtained using the MSA's average linkage method.

TABLE III
THE DATA OF 18-DIMENSIONAL MULTIPLE ENCODING VECTOR FROM 2305 DNA SARS-COV-2 SEQUENCES

	X1	X2	X3	X4	X5	…	X2303	X2304	X2305
n_R	14777	14810	14837	14817	14810	…	14744	14711	14786
μ_R	30168.16	30224.05	30175.1	30175.45	30173.56	…	30104.19	30210.21	30167.97
D_2^R	20112.44	20149.7	20117.07	20117.3	20116.04	…	20069.79	20140.48	20112.32
n_Y	15082	15110	15086	15086	15085	…	15045	15019	15078
μ_Y	29558.07	29623.97	29677.05	29637.39	29623.5	…	29501.9	29590.68	29583.74
D_2^Y	19705.71	19749.64	19785.03	19758.59	19749.33	…	19668.26	19727.45	19722.82
n_M	14406	14440	14466	14445	14442	…	14365	14334	14409
μ_M	30945.08	30998.49	30948.98	30952.55	30942.42	…	30898.44	31004.77	30957.29
D_2^M	20630.4	20666	20633	20635.38	20628.63	…	20599.31	20670.19	20638.54
n_K	15453	15480	15457	15458	15453	…	15424	15396	15455
μ_K	28848.44	28915.9	28964.74	28924.16	28918.04	…	28776.98	28866.09	28862.09
D_2^K	10302.17	10320.17	10304.84	10305.51	10302.17	…	10282.84	10264.17	10303.51
n_S	11351	11372	11353	11354	11352	…	11315	11283	11343
μ_S	39273.62	39361.43	39435.21	39379.04	39364.91	…	39227.23	39388.67	39325.02
D_2^S	7567.46	7581.46	7568.793	7569.46	7568.127	…	7543.46	7522.126	7562.127
n_W	18508	18548	18570	18549	18543	…	18474	18447	18521
μ_W	24086.6	24132.96	24109.2	24104.25	24099.15	…	24025.99	24091.85	24084.21
D_2^W	16058.01	16088.91	16073.07	16069.77	16066.37	…	16017.59	16061.5	16056.41

TABLE IV
DISTANCE MATRIX OF 2305 DNA SARS-COV-2 SEQUENCES

	X1	X2	X3	X4	X5	…	X2301	X2302	X2303	X2304	X2305
X1	0.00	191.28	268.33	177.23	152.52	…	168.12	185.90	185.90	226.77	66.15
X2	191.28	0.00	153.18	100.53	110.34	…	353.64	373.46	373.46	262.20	149.64
X3	268.33	153.18	0.00	91.72	116.34	…	402.86	420.93	420.93	307.82	211.65
X4	177.23	100.53	91.72	0.00	29.05	…	319.89	338.43	338.43	252.60	122.12
X5	152.52	110.34	116.34	29.05	0.00	…	294.01	312.29	312.29	243.02	100.06
…	…	…	…	…	…	…	…	…	…	…	…
X2301	168.12	353.64	402.86	319.89	294.01	…	0.00	36.46	36.46	268.93	208.39
X2302	185.90	373.46	420.93	338.43	312.29	…	36.46	0.00	0.00	300.48	227.54
X2303	185.90	373.46	420.93	338.43	312.29	…	36.46	0.00	0.00	300.48	227.54
X2304	226.77	262.20	307.82	252.60	243.02	…	268.93	300.48	300.48	0.00	201.22
X2305	66.15	149.64	211.65	122.12	100.06	…	208.39	227.54	227.54	201.22	0.00

TABLE V
DIFFERENCES IN THE AVERAGE LINKAGE METHOD GROUPING RESULTS USING MEV AND MSA

Clade	Method	Ancestors Code	Accession Number	Country
I	MEV	1123	EPI_ISL_779593	Australia
	MSA	1022	EPI_ISL_862292	India
II	MEV	1642	EPI_ISL_882768	Italy
	MSA	1154	EPI_ISL_794721	Australia
III	MEV	741	EPI_ISL_498180	Vietnam
	MSA	1058	EPI_ISL_751255	India
IV	MEV	641	EPI_ISL_803968	Singapore
	MSA	1717	EPI_ISL_900276	France
V	MEV	1348	EPI_ISL_730044	Nigeria
	MSA	1284	EPI_ISL_860568	South Africa
VI	MEV	1064	EPI_ISL_862413	India
	MSA	443	EPI_ISL_861717	Malaysia
VII	MEV	420	EPI_ISL_728164	Malaysia
	MSA	513	EPI_ISL_812932	Thailand
VIII	MEV	1511	EPI_ISL_906689	USA
	MSA	333	EPI_ISL_833342	Philippines
IX	MEV	334	EPI_ISL_833343	Philippines
	MSA	1393	EPI_ISL_906282	Nigeria

TABLE VI
DIFFERENCES IN THE COMPLETE LINKAGE METHOD GROUPING RESULTS USING MEV AND MSA

Clade	Method	Ancestors Code	Accession Number	Country
I	MEV	333	*EPI_ISL_833342*	Philippines
	MSA	1393	EPI_ISL_906282	Nigeria
II	MEV	28	EPI_ISL_529719	Indonesia
	MSA	1536	EPI_ISL_804835	Brazil
III	MEV	1754	EPI_ISL_900425	France
	MSA	1058	EPI_ISL_751255	India
IV	MEV	2168	EPI_ISL_904651	Spanish
	MSA	443	EPI_ISL_861717	Malaysia
V	MEV	1660	EPI_ISL_900577	Italy
	MSA	955	EPI_ISL_902629	Japan
VI	MEV	2212	EPI_ISL_735306	Turkey
	MSA	1424	EPI_ISL_936484	USA
VII	MEV	323	EPI_ISL_434554	Philippines
	MSA	536	EPI_ISL_882774	Thailand
VIII	MEV	455	EPI_ISL_877236	Malaysia
	MSA	2233	EPI_ISL_735354	Turkey
IX	MEV	2161	EPI_ISL_902744	Spanish
	MSA	2	EPI_ISL_435282	Indonesia

After forming a circular cladogram, the diagram is cut with a large clade and forms a clustering. The circular cladogram above is divided into clusters that have many nodes. The conclusion of the circular cladogram in Fig. 5 and Fig. 6 are as follows:

- The grouping results obtained by employing the average linkage method using the MEV and MSA produced nine clusters.
- From each clade, the ancestors of the two methods are determined. TABLE V shows the analysis results of each clade.
- We conclude that the ancestor of SARS-CoV-2 came from code 895 based on the results presented in the circular cladogram obtained by employing the average linkage method using MEV and MSA. Code 895 is a virus with the accession number EPI_ISL_610156 from China.
- The result presented in the circular cladogram obtained using MEV Clade VI is the DNA code group with the most number of members. For example, the distribution of COVID-19 codes in Indonesia, namely code 1 to code 309, is mostly in Clade VI. The COVID-19 ancestor most closely related to Indonesia is code 1064 from India. Meanwhile, if we use MSA Clade II, it has the most number of members.
- The DNA sequences of SARS-CoV-2 from the same country do not necessarily belong to the same clade based on their geographic location. According to the MEV grouping results, the DNA sequence at code 330 from the Philippines is in the same clade as code 2021 from Germany. Similarly, when using MSA to group DNA sequences, the DNA sequence at code 1147 from Australia is in the same clade as code 1767 from France and 2097 from Germany.

After analyzing the cladogram using the average linkage method, it was simulated using the complete linkage method. Figure 7 presents the circular cladogram of 2305 SARS-CoV-2 DNA data obtained using the complete linkage method using MEV. Figure 8 presents the circular cladogram of 2305 SARS-CoV-2 DNA data obtained using the MSA's complete linkage method.

TABLE VII
THE DIVISION OF CLUSTERS AND SARS-COV-2 USES THE K-MEANS CLUSTERING METHOD

K-Means Clustering	**Code**
Cluster I	88; 137; 333; 348; 350; 827; 840; 1075; 1386; 1387; 1391; 1403; 1415; 1709-1724; 2007; 2008
Cluster II	87; 134; 164; 165; 317; 318; 320; 321; 329; 335; 338; 342; 345; 349; 1150; 1179; 1198; 1322; 1336; 1344; 1361; 1409; 1446; 1453; 1511; 1536; 1537; 1567; 1647; 1655; 1770-1773; 1815-1820; 1834-1837; 1840; 1876; 1878; 1887; 1903; 1904; 2023-2026; 2155; 2156; 2158; 2161; 2163; 2164
Cluster III	136; 339; 341; 1120; 1154; 1155; 1323; 1329; 1358; 1379; 1393; 1725; 1726-1728; 2009-2011; 2246
Cluster IV	730; 732-737; 739-753; 994; 1003; 1016; 1137- 1146; 1164-1171; 1194-1196; 1199-1206
Cluster V	316; 330-332; 336; 337; 344; 706; 712; 1127; 1134; 1153; 1346; 1349; 1352; 1357; 1359; 1375; 1445; 1749-1769; 2021; 2022; 2253
Cluster VI	1-7; 9-26; 29-83; 85; 86; 89-95; 97-116; 119-132; 138-158; 160-163; 167-173; 175-250; 252-291; 294-314; 323-325; 327; 340; 351-487; 489-505; 507-517; 519-531; 534; 539-701; 703; 704; 707; 709-711; 713; 714; 718-729; 731; 738; 754-839; 841-882; 884-894; 896-993; 995-1002; 1004-1015; 1020-1057; 1059-1066; 1068-1074; 1076-1119; 1121; 1124; 1129; 1130; 1132; 1133; 1135; 1136; 1148; 1157-1163; 1172-1178; 1180-1188; 1191-1193; 1207-1214; 1225-1321; 1324-1328; 1330-1335; 1337-1339; 1341-1343; 1345; 1347; 1348; 1362-1365-1374; 1376-1378; 1380-1385; 1388-1390; 1392; 1394-1402; 1407; 1408; 1410-1414; 1416-1426; 1428; 1429; 1431-1440; 1442; 1443; 1447-1452; 1454; 1459-1480; 1482-1510; 1512-1530; 1538-1566; 1568; 1569; 1570; 1571; 1572; 1573; 1574; 1575; 1576; 1577; 1578; 1579; 1580-1641; 1644-1646; 1648-1654; 1656; 1657; 166-1708; 1774; 1776-1807; 1812-1814; 1822-1831; 1838; 1841-1850; 1852; 1853; 1856; 1857; 1859-1867; 1869-1875; 1877; 1879-1882; 1884; 1885; 1888; 1891-1893; 1895-1901; 190-1921; 1923-1926; 1928-1935; 1937-1939; 1941-1996; 1999; 2000; 2002-2006; 2027-2148; 2167-2183-2200; 2202; 2203; 2206-2241; 2243-2252; 2254-2305
Cluster VII	27; 84; 96; 117; 118; 133; 135; 159; 166; 174; 251; 292; 293; 315; 319; 322; 326; 328; 343; 347; 464; 477; 478; 488; 506; 518; 532; 533; 535-538; 695; 702; 705; 708; 715; 1017-1019; 1122; 1125; 1126; 1131; 1147; 1149; 1151; 1152; 1189; 1190; 1197; 1215-1224; 1340; 1351; 1354; 1355; 1360; 1366; 1406; 1441; 1444; 1481; 1500-1535; 1642; 1643; 1658-1661; 1775; 1808; 1809-1811; 1821; 1832; 1833; 1839; 1851; 1854; 1855; 1858; 1868; 1883; 1886; 1889; 1890; 1894; 1902; 1919; 1922; 1927; 1933; 1936; 1940; 1997; 1998; 2001; 2149-2154; 2157; 2159; 2160; 2162; 2165; 2166; 2184; 2201; 2204; 2205; 2242
Cluster VIII	334; 883; 895; 1058; 1067; 1156; 1353; 1424; 1427; 1430
Cluster IX	8; 28; 346; 716; 717; 1123; 1128; 1350; 1356; 1404; 1405; 1729-1748; 2012-2020

After forming a circular cladogram, the diagram is cut with a large clade and forms a clustering. The circular cladogram above is divided into clusters that have many nodes.

The conclusion of the circular cladogram in Fig. 7 and Fig. 8 are as follows:

- The grouping results using the complete linkage method utilizing the MEV and MSA produce nine clusters.
- From each clade, the ancestors of the two methods are determined. TABLE VI presents the analysis results of each clade.
- The circular cladograms formed using the complete linkage method utilizing MEV and MSA indicate that the ancestor of SARS-CoV-2 originated in code 895. Code 895 is a virus with the accession number EPI_ISL_610156 from China.
- The result presented in the circular cladogram using MEV Clade VII is the DNA code group with the most

members. For example, the distribution of the COVID-19 codes in Indonesia, namely, code 1 to code 309, is mostly in Clade VII. The COVID ancestor closest to Indonesia comes from code 323 from the Philippines. Meanwhile, using MSA Clade IX, it can be found that the group has the most ancestors from code 2 from Indonesia.
- The DNA sequences of SARS-CoV-2 from the same country do not necessarily belong to the same clade based on their geographic location. The DNA sequence at code 137 from Indonesia, for example, is in the same clade as code 1724 from France, according to MEV grouping results. The same is the case of the DNA sequences originating from the Philippines, which is in code 315 to code 350, separated into several clades, Clade I and Clade VII. Similarly, when using MSA, it can be found that the DNA sequence at code 333 from the Philippines is in the same clade as code 1086 from India and 1488 from the United States.

After finding the number of clusters using the hierarchical clustering method, the number will be used as a reference for the K-Means clustering method. In this study, the SARS-CoV-2 DNA was divided into nine groups. Table VII presents the distribution of clusters and the SARS-CoV-2 DNA code.

The conclusions based on TABLE VII are as follows:

- Based on the results of the simulation using K-Means clustering, the first cluster consists of 31 SARS-CoV-2 viruses; the second cluster, 60 SARS-CoV-2 viruses; the third cluster, 19 SARS-CoV-2 viruses; the fourth cluster, 54 SARS-CoV-2 viruses; the fifth cluster, 43 SARS-CoV-2 viruses; the sixth cluster, 1919 SARS-CoV-2 viruses; the seventh cluster, 129 SARS-CoV-2 viruses; the eighth cluster, 10 viruses; SARS-CoV-2 and the ninth cluster, 40 SARS-CoV-2 viruses.
- The DNA sequences of SARS-CoV-2 from the same country do not necessarily belong to the same clade based on their geographic location. For example, the DNA sequence in code 88 from Indonesia is in cluster I, the same as the DNA sequence in code 827 from China and in code 1724 from France.
- Viruses with a greater number of base pairs have properties (proteins) that allow them to infect their hosts with more viruses. In terms of virus characteristics based on the number of base pairs, the viruses in the sixth cluster have more base pairs and can generate more protein than the SARS-CoV-2 virus found in other classes.

IV. Conclusion

The findings of this study can be summarised as follows. According to the 2305 data of DNA sequences of SARS-CoV-2 from various countries in 2020 and 2021, the SARS-CoV-2 DNA originated in code 895, which came from China, specifically in Yunnan province. This is relevant with previous research by [23], who found that the DNA and RNA from coronavirus were 96.6% identical or similar to those of the virus in bats (BatCoV RaTG13) from Yunnan province. This confirms that coronavirus (SARS-CoV-2) originates from bats. The ancestor of SARS-CoV-2, which is most closely related to Indonesia, is also from India. Second, using the complete linkage method, it can be found that the DNA sequence originating from Indonesia, code 28, is the ancestor of each DNA sequence from countries in Clade II, such as Timor Leste, Nigeria, France, and Germany. Third, the K-Means clustering method is used to generate nine clusters. The sixth cluster contains a virus with several base pairs capable of producing more protein than the SARS-CoV-2 virus. Fourth, when using hierarchical clustering and K-Means clustering, the DNA sequences of SARS-CoV-2 from the same country are not always in the same cluster as the DNA sequences of SARS-CoV-2 from Indonesia and France, which are in the same cluster. The limitation of this study is the unavailability of protein interaction data, which could have been helpful in the discovery of the COVID-19 vaccine. In the future, this method can be compared with the other methods, and in the analysis, we can also test its computational speed and complexity.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgment

This study was funded by PUTI KI2Q2 2020 grants NKB-778/UN2.RST/HKP.05.00/2020. We would also like to thank BRIN/DIKTI for their significant thoughts and experience, which helped to strengthen our research in numerous ways.

References

[1] S. Namasudra, “Data Access Control in the Cloud Computing Environment for Bioinformatics,” *International Journal of Applied Research in Bioinformatics*, vol. 11, no. 1, pp. 40–50, Jan. 2021, doi: 10.4018/ijarb.2021010105.

[2] S. R. Manisekhar, G. M. Siddesh, and S. S. Manvi, “Introduction to Bioinformatics,” in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, Springer Singapore, 2020, pp. 3–9. doi: 10.1007/978-981-15-2445-5_1.

[3] E. Banjarnahor, A. Bustamam, T. Siswantining, and W. Mangunwardoyo, “K-Means Clustering and Analyze of SARS-CoV 2 DNA based on Multiple Encoding Vector and K-Mer Method,” *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, pp. 18647–18658, 2021.

[4] M. Crochemore, G. Fici, R. Mercacs, and S. P. Pissis, “Linear-Time Sequence Comparison Using Minimal Absent Words & Applications,” in *2016: Theoretical Informatics*, Springer Berlin Heidelberg, 2016, pp. 334–346. doi: 10.1007/978-3-662-49529-2_25.

[5] A. Bustamam, H. Tasman, N. Yuniarti, Frisca, and I. Mursidah, “Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV),” 2017. doi: 10.1063/1.4991238.

[6] A. Bustamam, T. Siswantining, N. L. Febriyani, I. D. Novitasari, and R. D. Cahyaningrum, “Protein sequences clustering of herpes virus by using Tribe Markov clustering (Tribe-MCL),” 2017. doi: 10.1063/1.4991254.

[7] A. Bustamam, E. D. Ulul, H. F. A. Hura, and T. Siswantining, “Implementation of hierarchical clustering using k-mer sparse matrix to analyze MERS CoV genetic relationship,” 2017. doi: 10.1063/1.4991246.

[8] Y. Li, L. He, R. L. He, and S. S.-T. Yau, “A novel fast vector method for genetic sequence comparison,” *Scientific Reports*, vol. 7, no. 1, Sep. 2017, doi: 10.1038/s41598-017-12493-2.

[9] K. Qian and Y. Luan, “Phylogenetic analysis of DNA sequences based on fractional Fourier transform,” *Physica A: Statistical Mechanics and its Applications*, vol. 509, pp. 795–808, Nov. 2018, doi: 10.1016/j.physa.2018.06.044.

[10] H.-H. Huang and S. B. Girimurugan, “A Novel Real-Time Genome Comparison Method Using Discrete Wavelet Transform,” *Journal of Computational Biology*, vol. 25, no. 4, pp. 405–416, Apr. 2018, doi: 10.1089/cmb.2017.0115.

[11] A. Criscuolo, "A fast alignment-free bioinformatics procedure to infer accurate distance-based phylogenetic trees from genome assemblies," *Research Ideas and Outcomes*, vol. 5, Jun. 2019, doi: 10.3897/rio.5.e36178.

[12] Y. Zhang, J. Wen, and S. S.-T. Yau, "Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method," *Genomics*, vol. 111, no. 6, pp. 1298–1305, Dec. 2019, doi: 10.1016/j.ygeno.2018.08.010.

[13] L. Muflikhah, Widodo, W. F. Mahmudy, and Solimun, "DNA Sequence of Hepatitis B Virus Clustering Using Hierarchical k-Means Algorithm," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2019. doi: 10.1109/icetas48360.2019.9117565.

[14] Y. Gao, T. Li, and L. Luo, "Phylogenetic study of 2019-nCoV by using alignment-free method." 2020.

[15] Y. Ma, Z. Yu, R. Tang, X. Xie, G. Han, and V. V Anh, "Phylogenetic Analysis of HIV-1 Genomes Based on the Position-Weighted K-mers Method," *Entropy*, vol. 22, no. 2, p. 255, Feb. 2020, doi: 10.3390/e22020255.

[16] G. Gamage *et al.*, "Phylogenetic Tree Construction Using K-Mer Forest- Based Distance Calculation," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 07, p. 4, Jun. 2020, doi: 10.3991/ijoe.v16i07.13807.

[17] S. Das, A. Das, B. Mondal, N. Dey, D. K. Bhattacharya, and D. N. Tibarewala, "Genome sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides," *Gene*, vol. 730, p. 144257, Mar. 2020, doi: 10.1016/j.gene.2019.144257.

[18] L. He, R. Dong, R. L. He, and S. S.-T. Yau, "Positional Correlation Natural Vector: A Novel Method for Genome Comparison," *International Journal of Molecular Sciences*, vol. 21, no. 11, p. 3859, May 2020, doi: 10.3390/ijms21113859.

[19] N. De Maio, "The Cumulative Indel Model: Fast and Accurate Statistical Evolutionary Alignment," *Systematic Biology*, Jul. 2020, doi: 10.1093/sysbio/syaa050.

[20] S. Amiroch, M. I. Irawan, I. Mukhlash, A. Nur, M. Ansori, and A. Nidom, "Identification of the Spread of the Influenza Virus Type A / H9N2 in Indonesia Using the Neighbor-Joining Algorithm with Felsenstein Models" 2021.

[21] J. K. Das, A. Sengupta, P. P. Choudhury, and S. Roy, "Mapping sequence to feature vector using numerical representation of codons targeted to amino acids for alignment-free sequence analysis," *Gene*, vol. 766, p. 145096, Jan. 2021, doi: 10.1016/j.gene.2020.145096.

[22] E. Banjarnahor, A. Bustamam, W. Mangunwardoyo, and D. Sarwinda, "Implementation of Hierarchical Clustering Method in Analyzing Genetic Relationship on DNA SARS-CoV-2 Sequences," *Journal of Physics: Conference Series*, vol. 1811, no. 1, p. 12074, Mar. 2021, doi: 10.1088/1742-6596/1811/1/012074.

[23] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, Feb. 2020, doi: 10.1038/s41586-020-2012-7.

[24] Y. Gao *et al.*, "Structure of the RNA-dependent RNA polymerase from COVID-19 virus," *Science*, vol. 368, no. 6492, pp. 779–782, Apr. 2020, doi: 10.1126/science.abb7498.

[25] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, and P.-R. Hsueh, "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges," *International Journal of Antimicrobial Agents*, vol. 55, no. 3, p. 105924, Mar. 2020, doi: 10.1016/j.ijantimicag.2020.105924.

[26] T. Li, A. Rezaeipanah, and E. M. Tag El Din, "An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3828–3842, Jun. 2022, doi: 10.1016/j.jksuci.2022.04.010.

[27] S. Pasupathi, V. Shanmuganathan, K. Madasamy, H. R. Yesudhas, and M. Kim, "Trend analysis using agglomerative hierarchical clustering approach for time series big data," *J Supercomput*, vol. 77, no. 7, pp. 6505–6524, Jul. 2021, doi: 10.1007/s11227-020-03580-9.

[28] M. T. Lwin, M. M. Aye, and others, "A modified hierarchical agglomerative approach for efficient document clustering system," *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, vol. 29, no. 1, pp. 228–238, 2017.

[29] P. Yildirim and D. Birant, "K-Linkage: A New Agglomerative Approach for Hierarchical Clustering," *Advances in Electrical and Computer Engineering*, vol. 17, no. 4, pp. 77–88, 2017, doi: 10.4316/aece.2017.04010.

[30] A. H. E. Chan, K. Chaisiri, S. Saralamba, S. Morand, and U. Thaenkham, "Assessing the suitability of mitochondrial and nuclear DNA genetic markers for molecular systematics and species identification of helminths," *Parasit Vectors*, vol. 14, no. 1, p. 233, Dec. 2021, doi: 10.1186/s13071-021-04737-y.

[31] I. Annaki *et al.*, "Clustering analysis of human navigation trajectories in a visuospatial memory locomotor task using K-Means and hierarchical agglomerative clustering," *E3S Web of Conferences*, vol. 351, p. 01042, May 2022, doi: 10.1051/e3sconf/202235101042.

[32] C. Wu, Q. Peng, J. Lee, K. Leibnitz, and Y. Xia, "Effective hierarchical clustering based on structural similarities in nearest neighbor graphs," *Knowl Based Syst*, vol. 228, p. 107295, Sep. 2021, doi: 10.1016/j.knosys.2021.107295.