Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 4, 2021, Pages. 18647 - 18658
Revised 05 March 2021; Accepted 01 April 2021.

K-Means Clustering and Analyze of SARS-CoV 2 DNA based on Multiple Encoding Vector and K-Mer Method

[1]Evander Banjarnahor, [2]Alhadi Bustamam,[3] Titin Siswantining, [4] Wibowo Mangunwardoyo

[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Indonesia, Depok, 16424, Indonesia

[4] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Indonesia, Depok, 16424, Indonesia

Corresponding Email: evanderbanjarnahor@sci.ui.ac.id

ABSTRACT

According to WHO data, coronavirus or Severe Acute Respiratory Syndrome Coronavirus 2 (SARS CoV-2) affected more than 172.6 million people worldwide in early June 2021.This virus targets human breathing, causing lung infections and even death in humans. This virus targets human respiration, causing lung infections and even death in humans. Based on this information, it is vital to investigate the coronavirus's kinship to limit its spread. This study uses the K-Means Clustering method in grouping and uses Multiple Encoding Vector in analyzing the sequences. The sequence analysis results resulted in an 18-dimensional multiple encoding vector compared with the K-Mer method based on the translation of DNA codons into amino acids. DNA Sequences of SARS CoV-2 were collected from numerous affected countries for this investigation. The simulation results found that the DNA sequence of SARS CoV-2 consisted of two clusters and the second cluster was the group that had the most members. The results also show that this method is optimal in a grouping of data with the between ss/total ss is 81.4%.

Keywords:
Amino Acids; Bioinformatics; Covid-19; Euclidean Distance; Sequences

Introduction

Bioinformatics is an interdisciplinary science involving molecular biology, molecular chemistry, physics, mathematics, and computational science (Shen,2008). Polanski. (2007)say that bioinformatics is an emerging discipline for utilitarian purposes that introduces order into large data sets generated by new molecular biology technologies. For example, due to large-scale DNA sequencing, the need for sequence assembly tools and sequence annotation, i.e., establishing the location of protein-coding regions in DNA, developed. As a result, Bioinformatics grew to be more comprehensive, occupying a space formerly covered by other related sciences. Quantitative sciences such as Mathematical and Computational Biology, Biometry and Biostatistics, Computer Science, Genomics and Proteomics, Genetics, and Molecular and Cell Biology are included. The conclusion from this explanation is that Bioinformatics is an interdisciplinary field that is very useful in research related to macromolecular biology such as DNA, RNA, and protein that accelerates analysis using computation.

The world is now dealing with a COVID-19 viral outbreak. This virus first appeared at the end of December 2019 in Wuhan, Hubei Province, China, and has since spread worldwide. Coronavirus illness 2019 (COVID-19) is caused by the new coronavirus 2019-nCoV, commonly known as SARS Coronavirus 2 (SARS CoV-2)(Tai, Zhang, He, Jiang, & Du, 2020)**.** In their research, Lai, Shih, Ko, Tang, & Hsueh. (2020)also explained that COVID-19 is transmitted from person to person by droplets or direct contact, with an estimated incubation time of 6.4 days and a significant reproduction number of 2.24-3, 58. The common symptoms of this infection are fever, cough, flu, loss of taste, diarrhea, and many more.

Coronavirus is a severe hazard to human health, especially for people susceptible to severe acute respiratory illnesses. According to WHO estimates, more than 172.6 million people worldwide were infected with SARS CoV-2 in early June 2021. This virus affects the human respiratory tract, causing lung infections and even death in humans. More than 3.7 million individuals died as a result of coronavirus infection globally. As a result, to restrict the virus's propagation, which is becoming increasingly common, identification is required to determine a virus's kinship or grouping. The most common method for determining a virus's kinship is constructing a phylogenetic tree or clustering.

Researchers aimed to employ the DNA of SARS CoV-S sequence data as research data in this work to assist research efforts during the COVID-19 pandemic crisis internationally, particularly in Indonesia. The DNA of the SARS CoV-S sequence was grouped using the K-Means clustering technique with Multiple Encoding Vector. Wang, Gutell, & Miranker. (2008)used the biclustering approach to overcome the Multiple Sequence Alignment (MSA) problem while examining DNA sequences. Their research combined the Local MSA software, Block MSA, and biclustering. In their research, Hengused hierarchical and spatially explicit clustering in grouping DNA sequence data from Borrelia burgdorferi. According to his study, Lyme Borreliosis, a prevalent disease in North America and Europe, is caused by the tick-borne bacteria B. burgdorferi. Bustamam, Tasman, et al. (2017),in his research using the K-Means clustering method in grouping hepatitis B virus (HBV) DNA sequences. The results of the study showed that the virus in the first cluster was more likely to evolve with the HDV virus that causes hepatitis D. Bustamam et al. (2017)in his research used the Tribe Markov Clustering (Tribe-MCL) Algorithm in classifying and analyzing the protein sequences of the Herpes virus. In their research, Li, He, He, & Yau (2017)used Multiple Encoding Vectors inline alignment and DNA data analysis in mammals, bacteria, and viruses. Qian & Luan (2018)also performed a phylogenetic analysis of DNA sequences using the Fractional Fourier Transform. In their research, Xie et al. (2019) introduced the Qualitative biclustering Algorithm Version 2 as a novel biclustering algorithm (QUBIC2).In this work, QUBIC2 outperformed five other algorithms in finding biclusters in different reference data sets from E. coli, Human, and simulated data. QUBIC2 also performed effectively and consistently on the microarray, bulk RNA-Seq, and scRNA-Seq gene expression data. Swasti et al.(2019)applied the LCM-MBC Algorithm for 16215 types of interactions between HIV-1 proteins and human proteins. This study found 852 biclusters, and the biclusters that had a maximum size were four rows and 204 columns. In their research, Saadeh, Al Fayez, & Elshqeirat. (2020)analyzed gene expression microarray data using K-Means clustering, which analyzes human gene expression levels at four stages of erythropoiesis. The study's findings identified eight groups, including a cluster of 450 genes (C4) that are more active throughout the maturation period, as well as in cell division and DNA replication. Another 234-gene cluster (C7) is more active in autophagy (cell consumption/destruction), which has been linked to enucleation (expulsion of the nucleus from the cell). Banjarnahor et al.(2021)in their research used Hierarchical Clustering and Multiple Sequence Alignment in analyzing DNA sequences of SARS CoV-2. The results of this study indicate that the ancestor of SARS CoV-2 came from China. Many researchers have already analyzed the kinship of a DNA sequence of a virus either through the formation of a phylogenetic tree or by using clustering. Therefore, researchers will research the K-Means Clustering Method in DNA grouping of SARS CoV-2sequences using Multiple Encoding Vectors.

Method

Retrieval of DNA Sequences

In this research, 100 DNA of SARS CoV-S sequences will be combined for a total of 100 DNA sequences from diverse countries in 2020. Data for 100 DNA of SARS CoV-S sequences were received from GenBank using the link https://www.gisaid.org/. Clustering of 100 DNA of SARS CoV-S sequences using the K-Means Clustering technique and free source R software. In this study, each 100 DNA sequences of SARS CoV-2 from various countries were summarized in the following code:

Table 1. Table code of 100 DNA SARS CoV-2 sequences and country of origin

Code	Country of Origin	Code	Country of Origin
1 - 5	China	51 - 55	Singapore
6 – 10	USA	56 - 60	France
11 - 15	Philippines	61 - 65	Portugal
16 - 20	Malaysia	66 - 70	Pakistan
21 - 25	Thailand	71 - 75	Bangladesh
26 - 30	Vietnamese	76 - 80	Italy
31 – 32	Kenya	81 - 85	Turkey
33 - 35	Brazil	86 - 90	South Africa
36 – 40	England	91 - 95	India
41 - 45	Spain	96 - 100	Indonesia
46 – 50	Japan		

Multiple Encoding Vector

Let Ł be the set of 4 bases, i.e. {A, C, G, T} dan $Q = (S_1, S_2, \dots., S_n)$is a DNA sequence of length n, i.e. S_i ϵ L, *I* = 1,2,…,*n.* The four bases are classified into two categories based on three types of chemical and physical features. The letter R represents purines in bases A and G. Bases C and T are pyrimidines, symbolized by the letter Y. Nucleotides A and C are amino designated by M and G in another grouping. The letter T is represented by the letter K. G and C have strong H bonds and are symbolized by the letter S in the H bond. W represents the A and T bases, which have weak H bonds. This approach assigns three numerical values to each of the letters R, Y, M, K, S, and W. To define the distribution of R and Y in the sequence Q, first replace the letters A and G with R and the letters C and T with Y. The sequence then has just two sorts of letters: R and Y. For R, defined $W_R(.) : \{R, Y\} \rightarrow \{0,1\}$
so$W_R(S_i) = 1$ if$S_i = R$ and 0 otherwise. ForR, defined$n_R, \mu_R, D_2^R$to describe the sum of R, the average position of R, and the positional variation of R appearing in the order Q.

1. Let$n_R = \sum_{i=1}^{n} W_R(S_i)$indicates the number of letters R that occurs in the Q line.
2. Let $\mu_R = \sum_{i=1}^{n} i \, . \frac{W_R(S_i)}{n_R}$be the average position where the letter R appears.
3. Let $D_2^R = \sum_{i=1}^{n} \frac{(i-\mu_R)^2 W_R(S_i)}{n_R n}$be the 2nd moment on the scale from the position of the letter R.

Similarly, for Y, we define $W_Y(.) : \{R, Y\} \rightarrow \{0,1\}soW_Y(S_i) = 1$ if$S_i = Y$ and 0 otherwise. Then we get three characteristics for$Y: n_Y, \mu_Y,$ andD_2^Y. Six values are utilized to represent the distribution of the four bases about this chemical characteristic in this form of nucleotide categorization. Similarly, we define another triplet for M, K, S, and W. So, the 18-dimensional vector of the DNA sequence Q is defined by $(n_R, \mu_R, D_2^R \dots., n_W, \mu_W, D_2^W)$. This approach is known as a Multiple Encoding Vector because it encodes a DNA sequence using three letters. Figure 2 depicts this vector creation.(Li et al., 2017)

```
Pseudocode for Multiple Encoding Vector
BEGIN
Number ($n_R, \mu_R, D_2^R \dots., n_W, \mu_W, D_2^W$).
INPUT DNA sequence and represent purin A, C, G, T to R, Y, M,K,S,W
IF R in Q THEN
        $n_R = \sum_{i=1}^{n} W_R(S_i)$
        $\mu_R = \sum_{i=1}^{n} i \,. \frac{W_R(S_i)}{n_R}$
        $D_2^R = \sum_{i=1}^{n} \frac{(i - \mu_R)^2 W_R(S_i)}{n_R n}$
OUTPUT "number of letter R($n_R$)"
 OUTPUT "mean position at which the letter R appears($\mu_R$)"
 OUTPUT "scaled 2-nd moment of position of letter R ($D_2^R$)"
ITERATE Y, M, K, S, W in Q
END
```

K-Mer

K-Mer is an alternative method to calculate the similarity between DNA sequences. However, the application of this method to several long DNA sequences has the drawback of time inefficiency. K-Mer can then be used to obtain similarity values by counting the number of K-Mer that arise from several possible permutations of K-Mer used (Bustamam, Ulul, Hura, & Siswantining, 2017).

Suppose there is a DNA sequence TATGCCTAAAGGC, taking k = 4, the resulting K-Mer is: TATG ; ATGC ; TGCC ; GCCT ; CCTA ; CTAA ; TAAA ; AAAG ; AAGG ; AGGC.The K-Mer method relies heavily on determining the value of *k*. In this study, the value of *k* that will be taken as a reference is three based on DNA change into amino acids. The column for K-Mer is the number of amino acids that is 20. The row for K-Mer is *n*, which is the number of data used in the research. The value of each row in the column of that amino acid is the number of DNA codons generated in each row. Below is a table for converting DNA into amino acids (Bogard, Rouchka, & Arazi, 2008).

Table 2. Translation table to convert from a three-base codon to an amino acid.

	Amino Acid	Symbol	DNA Codon			
1	Alanine	A	GCA	GCC	GCG	GCT
2	Cystenine	C	TGC	TGT		
3	Aspartic Acid	D	GAC	GAT		
4	Glutamic Acid	E	GAA	GAG		
5	Phenylalanie	F	TTC	TTT		
6	Glycine	G	GGA	GGC	GGG	GGT
7	Histidine	H	CAC	CAT		
8	Isoleucine	I	ATA	ATC	ATT	

9	Lysine	K	AAA	AAG				
10	Leucine	L	CTA	CTC	CTG	CTT	TTA	TTA
11	Methionine	M	ATG					
12	Asparagine	N	AAC	AAT				
13	Proline	P	CCA	CCC	CCG	CCT		
14	Glutamine	Q	CCA	CAG				
15	Arginine	R	AGA	AGG	CGA	CGC	CGG	CGT
16	Serine	S	AGC	AGT	TCA	TCC	TCG	TCT
17	Threonine	T	ACA	ACC	ACG	ACT		
18	Valine	V	GTA	GTC	GTG	GTT		
19	Tryptophan	W	TGG					
20	Tyrosine	Y	TAC	TAT				

For example, there is a DNA sequence, $S = TTGCAGAAGAG$. Taking k = 3, result of K-Mer is: TTG ; TGC ; GCA ; CAG ; AGA ; GAA ; AAG; AGA ; GAG. In this case, the sum of Alanine = 1, Cystine = 1, Aspartic Acid = 0, Glutamic Acid = 2 etc.

Pseudocode for K-mer

```
BEGIN
Procedure k-mer (string seq, integer k) is L:Length, k = 3 based on amino acid codon, arr:new array of L – k +1 empty strings.
IF string 3-Mer = kodon AminoAcid THEN
        A = (string 3-mer = Alanine)
Number b1, b2, b3, b4, A
INPUT b1 = count('GCA')
INPUT b2 = count('GCC')
INPUT b3 = count('GCG')
INPUT b4 = count('GCT')
INPUT A = (b1+b2+b3+b4)
OUTPUTA “number of Alanine in 3-Mer”
ITERATE for C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y
END
```

Distance Matrix

The most popular matrix for continuous features is the Euclidean distance. The Euclidean distance equation can be calculated as follows (Bustamam, Tasman, et al., 2017).

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(X_{ij} - C_{kj})^2}$$

where

d_{ik} = the distance of object$i$to the centroid k
m = the amount of data dimension
X_{ij} = the coordinate of object$i$in the dimension j
C_{kj} = the coordinate of object $k$in the dimension$j$

K-Means Clustering

The K-Means Clustering Approach is based on fundamental theory and is similar to the EM (Expectation-Maximization) approach for estimating mixed distribution parameters. Polinski (2007) defines formalized formalized formalized formalized formalized formalized formalized formally. Regarding the development of the K-Means Clustering algorithm, it is done in the following way:

I. It is assumed that the number of K clusters is known and makes more than two assumptions.

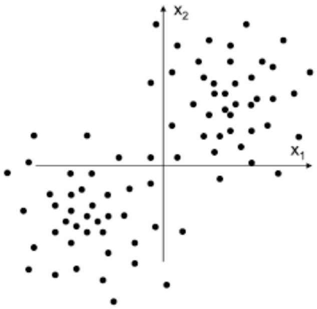

Figure1. An example of a feature vector pattern with an undefined class

II. Each cluster has a center point $x_i^C, i = 1, \dots, K$ dwith coordinates equal to the average coordinates of the data points belonging to the cluster.

III. For each$x$data point,it is decided which group does it by calculating the distance between the point and the center of all $x\ groups\ \left(x, x_i^C\right), i\ = 1, \dots, K$. The index$I$is taken from x_i^Cwhich minimizes the distance$d(x, x_i^C)$to denote the cluster containing x.

The clustering algorithm's design is pretty apparent based on assumptions (I) - (III). We will choose some initial values at random for the cluster centers $x_i^C, i = 1, \dots, K$ and then repeat the previous two procedures until convergence is achieved.

Step 1: Based on the criterion, assign each data point to a cluster (III)

Step 2: Update the value for the cluster center as established in step 1 based on the assignment (II).

Overall, the steps in this study were a DNA sequence collection of SARS CoV-2, sequence analysis using multiple encoding vectors, distance matrix, and grouping using K-Means Clustering.

The general flow of this research can be seen in Figure 2.

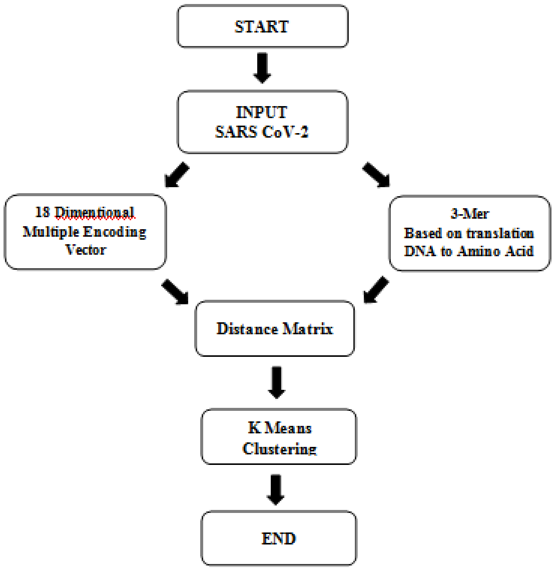

Figure 2. Flowchart of Method

Results

In the simulation using the R program, after inputting the DNA of SARS CoV-S data, the sequence alignment process was carried out using Multiple Encoding Vectors, resulting in 18-dimensional Multiple Encoding Vectors (Table 3). First, the results of the 18-dimensional distance matrix are searched using Euclidean Distance. Then the distance matrix value is simulated using the K-Means Clustering method. The results of this grouping were also compared using 3-Mer adapted to translating DNA codons into amino acids. Table 4 is the result of calculating 3-Mer based on the translation of DNA codons into amino acids.

Table 3. Dimensional Multiple Encoding Vector Data from 100 DNA of SARS CoV-S sequences

	X_1	X_2	X_3	…	X_{98}	X_{99}	X_{100}
n_R	14457	14198	14651	…	14743	14743	14743
μ_R	30788.43	31366.89	30348.17	…	30081.98	30081.98	30081.98
D_2^R	20525.97	20911.61	20232.46	…	20054.99	20054.99	20054.99
n_Y	14807	14362	14928	…	15039	15039	15039
μ_Y	30060.67	31008.71	29785.04	…	29489.9	29489.9	29489.9
D_2^Y	20040.78	20672.82	19857.03	…	19660.27	19660.27	19660.27
n_M	14111	13824	14289	…	14368	14363	14366
μ_M	31543.36	32215.5	31117.02	…	30867.11	30877.86	30871.41
D_2^M	21029.26	21477.36	20745.03	…	20578.42	20585.58	20581.28
n_K	15153	14736	15290	…	15414	15419	15416
μ_K	29374.27	30221.71	29079.86	…	28772.46	28763.13	28768.72
D_2^K	10102.17	9824.165	10193.5	…	10276.17	10279.51	10277.51
n_S	11112	10912	11240	…	11321	11316	11321
μ_S	40056.55	40812.6	39557.93	…	39174.87	39192.18	39174.87
D_2^S	7408.124	7274.789	7493.459	…	7547.46	7544.127	7547.46
n_W	18152	17648	18339	…	18461	18466	18461

μ_W	24521.17	25234.99	24245.11	…	24023.54	24017.04	24023.54
D_2^W	16347.72	16823.61	16163.68	…	16015.97	16011.63	16015.97

Table 4. Data calculation 3-Mer based on the translation of codon DNA into amino acids

	X_1	X_2	X_2	…	X_{97}	X_{98}	X_{100}
Alanine	1143	1125	1152	…	1166	1166	1166
Cystenine	1373	1345	1389	…	1404	1403	1403
Aspartic Acid	765	751	771	…	777	777	777
Glutuamic Acid	810	800	826	…	830	830	829
Phenilalanie	1509	1425	1498	…	1516	1517	1515
Glycine	1068	1058	1074	…	1089	1089	1090
Histidine	925	907	933	…	941	941	942
Isoleucine	1562	1503	1562	…	1578	1579	1579
Lysine	1434	1426	1467	…	1470	1471	1468
Leucine	3696	3571	3725	…	3763	3766	3763
Methionine	712	693	716	…	723	723	723
Asparagine	1348	1322	1364	…	1370	1370	1371
Proline	868	854	878	…	880	878	880
Glutamine	1112	1109	1135	…	1136	1135	1136
Arginine	1344	1325	1351	…	1368	1367	1367
Serine	2185	2116	2202	…	2214	2215	2215
Threonine	1973	1930	2009	…	2018	2015	2017
Valine	1943	1899	1973	…	1986	1986	1987
Tryptophan	546	535	548	…	553	553	554
Tyrosine	1208	1145	1217	…	1231	1229	1229

After finding the results of the distance matrix, the number of clusters (*k*) is determined during the simulation—determination of the number of clusters using the elbow method. The elbow method uses the total value of wss (within sum square) to determine the optimal *k*. Below are the results of the elbow simulation method in both methods. Figure 3 shows the elbow method's visualization using Multiple Encoding Vector (MEV) and using 3-Mer.

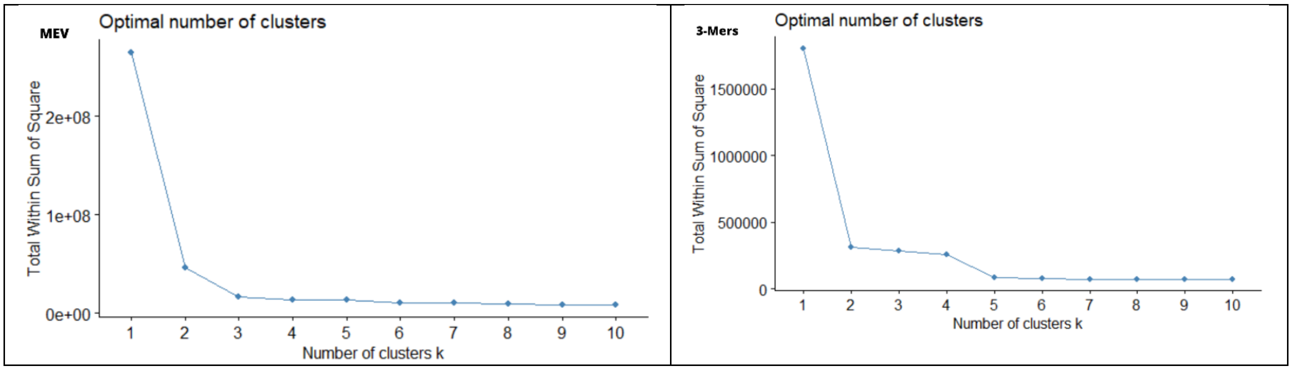

Figure 3. Visualization of *elbow* method using MEV and 3-Mer

From the two images above, we can see that the line has a fracture that forms an elbow at k = 2. Hence, by using this method, we get optimal k when it is at k = 2. After determining the value of k, the matrix value is simulated by the K method. -Means Clustering. Table 4 shows the distribution of SAR CoV-2 DNA using the K-Means clustering method with the Multiple Encoding Vector method. Table 4 shows the distribution of CoV-2 SAR DNA using the K-means clustering method with the 3-Mer method.

Table 5. The division of SARS CoV 2 DNA sequence clusters using the K-Means Clustering method using Multiple Encoding Vector

Cluster	Code
Cluster I	2,4,5,41,44
Cluster II	1,3,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,42,43,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100

Table 6. The division of SARS CoV 2 DNA sequence clusters using the K-Means Clustering method using 3-Mer

Cluster	Code
Cluster I	2,4,5,41,44,45
Cluster II	1,3,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,42,43,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100

From the simulation results of the K-means Clustering method using Multiple Encoding Vector, it is found that the value of $\frac{\text{between\_SS}}{\text{total\_SS}}$ is 81.4%. Likewise, the simulation of the K-means Clustering method using 3-Mer resulted that the value of $\frac{\text{between\_SS}}{\text{total\_SS}}$ is 82.1%. This percentage shows that the clustering that is simulated in both methods is optimal. Figure 4 is the result of visualization of K-means clustering using MEV and 3-Mer.

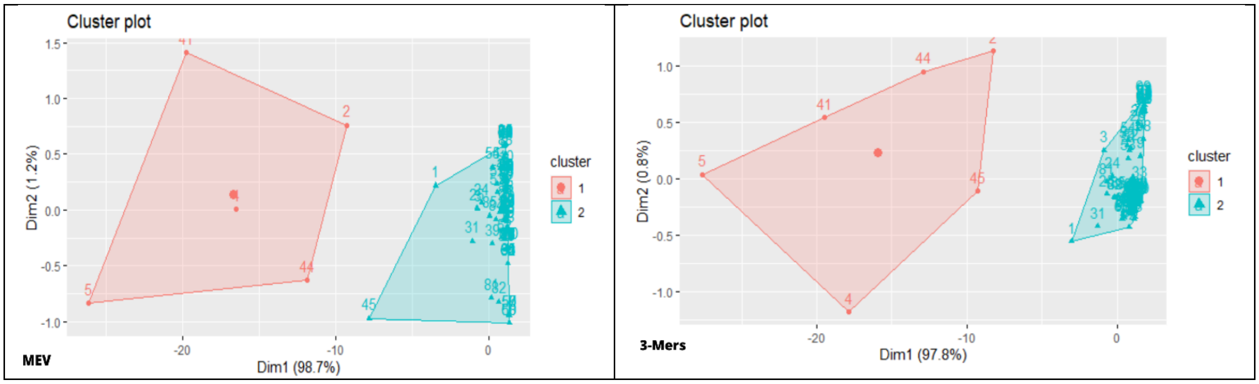

Figure 4.Visualization of K-Means Clustering using MEV dan 3-Mer

Discussion

From the simulation results using K-Means Clustering, the first cluster consists of 5 SARS CoV-2 viruses. The second cluster consists of 95SARS CoV-2 viruses. From the results of this grouping, it can be seen that cluster II is the group that has the most members, namely 95% of the entire population. Unlike the K-Means Clustering simulation results using the 3-Mer method, the first cluster consists of 6 SARS-CoV-2 viruses, and the second cluster consists of 94 SARS-CoV-2 viruses. From the results of this grouping, it can be seen that cluster II is the group that has the most members, namely 94% of the entire population. In cluster II, it can also be seen that code 96, code 97, code 98, code 99, and code 100 originate from Indonesia. In this case, it can be concluded that 100% of the DNA of SARS CoV-S sequences originating from Indonesia are in the second cluster.

The DNA sequences of SARS CoV-2 from the same nation do not necessarily belong to the same clade based on their geographical location. For example, the DNA sequence at code 2 with accession number EPI_ISL_428440 from China is in cluster II, as is code 44 with accession number EPI_ISL_565924 from Spain. Likewise, DNA sequences originating from China have codes 1-5 scattered in cluster I and cluster II.

Viruses with more base pairs have features (proteins) that allow them to infect their hosts with more viruses. Thus, in terms of virus properties based on the number of base pairs, the viruses in the third cluster have more base pairs and can make more proteins than the SARS CoV-2 virus, which is in other clusters.

Conclusion

According to the findings of this study, the results of DNA clustering SARS CoV-2using the K-Means Clustering technique were grouped into 2 clusters. The second cluster has the most members and base pairs that can create more protein than the SARS CoV-2 virus. In addition, judging from its geographical location, DNA sequences of SARS CoV-2 originating from the same country do not necessarily occupy the same clade. For example, DNA sequences originating from Indonesia are in the same cluster II as DNA sequences originating from Pakistan and India. Likewise, DNA sequences originating from China have codes 1-5 scattered in cluster I and cluster II. This study also shows that the K-means Clustering method using Multiple Encoding Vector is optimal in grouping data, which is 81.4%.

This research is still limited due to the unavailability of protein interaction data to help discover a Covid-19 vaccine. In the future, this method can also be compared with other methods in sequence analysis. Further researchers can also look for other methods to determine the initial k centroid before simulating it with K-means clustering to get more efficient clustering results.

Acknowledgment

This study was funded by PDUPT2021 grants No:NKB-183/UN2.RST/HKP.05.00/2021. 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] Banjarnahor, E., Bustamam, A., Mangunwardoyo, W., & Sarwinda, D. (2021). Implementation of Hierarchical Clustering Method in Analyzing Genetic Relationship on {DNA} {SARS}-{CoV}-2 Sequences. *Journal of Physics: Conference Series*, *1811*(1), 12074. https://doi.org/10.1088/1742-6596/1811/1/012074

[2] Bogard, C. M., Rouchka, E. C., & Arazi, B. (2008). {DNA} media storage. *Progress in Natural Science*, *18*(5), 603–609. https://doi.org/10.1016/j.pnsc.2007.12.009

[3] Bustamam, A., Siswantining, T., Febriyani, N. L., Novitasari, I. D., & Cahyaningrum, R. D. (2017). *Protein sequences clustering of herpes virus by using Tribe Markov clustering (Tribe-{MCL})*. Author(s). https://doi.org/10.1063/1.4991254

[4] Bustamam, A., Tasman, H., Yuniarti, N., Frisca, & Mursidah, I. (2017). *Application of k-means clustering algorithm in grouping the {DNA} sequences of hepatitis B virus ({HBV})*. Author(s). https://doi.org/10.1063/1.4991238

[5] Bustamam, A., Ulul, E. D., Hura, H. F. A., & Siswantining, T. (2017). *Implementation of hierarchical clustering using K-Mer sparse matrix to analyze {MERS}{\textendash}{CoV} genetic relationship*. Author(s). https://doi.org/10.1063/1.4991246

[6] Cheng, L., Connor, T. R., Siren, J., Aanensen, D. M., & Corander, J. (2013). Hierarchical and Spatially Explicit Clustering of {DNA} Sequences with {BAPS} Software. *Molecular Biology and Evolution*, *30*(5), 1224–1228. https://doi.org/10.1093/molbev/mst028

[7] Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J., & Hsueh, P.-R. (2020). 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*, *55*(3), 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924

[8] Li, Y., He, L., He, R. L., & Yau, S. S.-T. (2017). A novel fast vector method for genetic sequence comparison. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-12493-2

[9] Polanski, A. (2007). *Bioinformatics*. Berlin New York: Springer.

[10] Qian, K., & Luan, Y. (2018). Phylogenetic analysis of {DNA} sequences based on fractional Fourier transform. *Physica A: Statistical Mechanics and Its Applications*, *509*, 795–808. https://doi.org/10.1016/j.physa.2018.06.044

[11] Saadeh, H., Al Fayez, R. Q., & Elshqeirat, B. (2020). Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development. *International Journal of Machine Learning and Computing*, *10*(3).

[12] Shen, S. (2008). *Theory and mathematical methods for bioinformatics*. Berlin: Springer.

[13] Swasti, O., Bustamam, A., Lestari, D., & Mangunwardoyo, W. (2019). Biclustering protein interactions between HIV-1 proteins and humans proteins using LCM-MBC Algorithm. In B. D. Handari, H. Seno, & H. Tasman (Eds.), *Proceedings of the Symposium on BioMathematics, SYMOMATH 2018*. American Institute of Physics Inc. https://doi.org/10.1063/1.5094279

[14] Tai, W., Zhang, X., He, Y., Jiang, S., & Du, L. (2020). Identification of {SARS}-{CoV} {RBD}-targeting monoclonal antibodies with cross-reactive or neutralizing activity against {SARS}-{CoV}-2. *Antiviral Research*, *179*, 104820. https://doi.org/10.1016/j.antiviral.2020.104820

[15] Wang, S., Gutell, R., & Miranker, D. (2008). Biclustering As A Method For RNA Local

Multiple Sequence Alignment. *Bioinformatics (Oxford, England)*, *23*, 3289–3296. https://doi.org/10.1093/bioinformatics/btm485

[16] Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., … Ma, Q. (2019). {QUBIC}2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale {RNA}-Seq data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz692