Determining Geographical Spread Pattern of MERS-CoV by Distance Method using Kimura Model

Siti Amiroch[1,a)] and Arif Rohmatullah[1,b)]

[1]*Department of Mathematics*
Faculty of Mathematics and Natural Sciences Universitas
Islam Darul 'Ulum Lamongan, Indonesia

[a)]Corresponding author: amirast_117@yahoo.com
[b)]arifrohmatullahkc@gmail.com

Abstract. MERS-CoV or generally called as Middle East Respiratory Syndrome Coronavirus, a respiratory disease syndrome caused by a corona virus that attacks the respiratory tract ranging from mild to severe acute indication of fever, cough and shortness of breath. The cases happened relate to the countries in the Arabian Peninsula (Middle East) and there were 356 deaths have been reported due to the spread of the epidemic MERS.The data used in the case of MERS are the data DNA sequences taken from Genbank, the online database of the United States that stores the results of molecular biological experiments from all over the world (http://www.ncbi.nlm.nih.gov). In this case, bioinformatics plays an important role of reading sequences of DNA and genetic information by using the main device in the form of software that is supported by the availability of the Internet, while the analysis there in made and proven with mathematical methods.In similar research conducted by molecular biologists and physicians, the process of DNA sequencing is done with software that is already available like BLAST. In order to determine the MERS geographical distribution patterns in the Arabian Peninsula is done with program Clustal W, Bayesian, Phylip, etc. In this study, the writer use the Matlab simulation for all processes starting sequence alignment, counting the number of transitions and transversion substitutions for each sequence and its location up to the process of forming a phylogenetic tree that figures out the pattern of spread of the epidemic MERS. Mathematical analysis performed on a decline in the formula is to find Kimura evolutionary models and the process of forming a phylogenetic tree (the pattern of the epidemic MERS distribution) with neighbor joining algorithm.Finally it was obtained the pattern of geographical spread with 6 groups epidemic of MERS which ultimately turns out that all the MERS viruses that were spread in the Arabian Peninsula everything are almost the same as the virus sequence found in al-Hasa.

INTRODUCTION

MERS which is another name of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) attacks the respiratory system. MERS is usually acknowledged by some indications such as: fever, cough and shortness of breath. It is called Middle East Respiratory Syndrome Coronavirus because almost all the cases relating to the countries in the Arabian Peninsula (Middle East), and most infected people are living in the Middle East or after contact with an infected person who has just traveled from the Middle East. In 2012, the first case of MERS was acknowledged after an investigation on genome sequencing of an isolated virus from patient's sputum samples that got sick in 2012 because of the new influenza epidemic. There were about 22 countries, such as Saudi Arabia, Malaysia, Jordan, Qatar, Egypt, United Arab Emirates, Kuwait, Turkey, Oman, Algeria, Bangladesh, Austria, the United Kingdom, and the United States that informed about MERS's cases in Juni 2014.

In Spite of MERS is a case that has been a long time coming, but until now they have been reported cases of death caused by MERS transmission. According to WHO data per February 5, 2015, there were 356 deaths have been reported due to the spread of this MERS epidemic.

An interesting thing about MERS-CoV transmission cases is described in a pattern derived from a mathematical calculation. In the beginning, there was some confusions about the beginning of the emergence of MERS case.

Symposium on Biomathematics (SYMOMATH 2016)
AIP Conf. Proc. 1825, 020001-1–020001-8; doi: 10.1063/1.4978970
Published by AIP Publishing. 978-0-7354-1493-8/$30.00

Several sources mentioned that MERS case originated in Jordan, Saudi Arabia, Oman, even Qatar. But, so far, the MERS case is associated with countries in the Middle East and there is a possibility that the case will spread to other countries through tourists, travelers, migrant workers, or pilgrims who get infected after get in touch with animals (for example when visiting farms or markets). Then, it can be described the pattern of the geographical spread of the virus from that. By aligning and analyzing the DNA sequence of the MERS virus, it will be obtained MERS genetic distance of each species and represented in a matrix called distance matrix. Because basically evolution is a transition and transversion mutation at nucleotide, so the distance matrix formed is converted into evolutionary distance matrix using Kimura evolutionary models.

Finally, from that evolutionary distance matrix, it can be formed a phylogenetic tree which is a description or geographic distribution patterns of MERS-CoV in the world.

Sequence and Sequence Alignment

DNA sequence, RNA sequence, and protein sequence are commonly determined based on biological sequence. At stated in Shen [7], biological sequences described with the following notation:

$$X = (x_1, x_2, \dots . , x_{n_a}), \quad Y = (y_1, y_2, \dots . y_{n_b}), \quad Z = (z_1, z_2, \dots . z_{n_b})$$

In which X, Y, Z, express *sequence*, while x_i, y_i, z_i, are basic units of sequence in $_i$ position, in which those elements are obtained from the set $V_q = \{0, 1, \dots , q-1\}$. The length of *X, Y* dan *Z* is expressed by n_x, n_y, n_z. If *X, Y, Z* are DNA/RNA *sequence*, so $V_4 = \{a, c, g, t\}$ or $\{a, c, g, u\}$, whereas if the protein sequence, then q = 20 represents the 20 amino of acid molecule.

Whereas sequence alignment in Shen [7], is a method of the position analysis and the type of mutation that are important in biological sequence that enables the comparison properly. *Alignment* between the two sequences is called *Sequence Alignment*, while alignment involving multiple sequence alignment called the Multiple Sequence Alignment. Determining the movement of mutation is the core idea of sequence alignment.

Algorithm Needleman Wunchs

This research applies Needleman Wunch algorithm which is based on a global alignment algorithm. In Shen [7] the steps of the algorithm as follows:

1. Formulate two sequences in the two-dimensional table
   If given a sequence $A = (a_1, a_2, \dots , a_n)$ and $B = (b_1, b_2, \dots , b_m)$ so, to align the two sequences are formed in a two-dimensional table as follows:

TABLE 1. Two dimensional table of sequence *A,B*

		a_1	a_2		a_n
	$s(0,0)$	$s(1,0)$	$s(2,0)$	....	$s(n,0)$
b_1	$s(0,1)$	$s(1,1)$	$s(2,1)$	....	$s(n,1)$
b_2	$s(0,2)$	$s(1,2)$	$s(2,2)$	....	$s(n,2)$
....	....	....	....	....	....
b_m	$s(0,m)$	$s(1,m)$	$s(2,m)$	....	$s(n,m)$

2. Calculate element $S(i,j)$ from two-dimensional table
   Each element $S(i,j)$ that is contained in a two-dimensional table is established by three elements, namely:

   - $S(i-1, j-1)$ in the top left corner
   - $S(i-1, j)$ on the left side
   - $S(i, j-1)$ that is on top

   The initial step is to determines the score $S(i,0)$, and the score $S(0,j)$. In other words, assume that scores a penalty on a virtual circuit symbol is *d*. So $S(0,j) = -jxd$ and $S(i,0) = -ixd$ and $S(0,0) = 0$. While the element $S(i,j)$ can be calculated using the following formula:

$$s(i,j) = max\{s(i-1, j-1) + s(a_i, b_j), s(i-1, j) - d, s(i, j-1) - d\}$$

3. Traceback algorithm
   The final value $s(n,m)$ is the maximum value of sequence alignment (A, B), in which $s(n,m)$ is the starting point for the method backward (backward flow). For any $s(i,j) = (i-1, j-1) + s(a_i, b_j)$, then the flow of backward is from $(i,j) \rightarrow (i-1, j-1)$, then from $s(n,m)$ to the last $s(0,0)$. This is what backward method. Then it will be obtained alignment of these sequences in the following way:
   - Denote the pair of nucleic acids as a_i, b_i if the backward flow is started from a_i, b_i to the top left corner.
   - Insert a symbol on a virtual vertical sequences and denote as $(a_i, -)$ if the backward flow is horizontal.
   - Insert a virtual symbol on horizontal sequences and denote it as $(-, b_i)$ if the backward flow is vertical.
4. Last, an optimal alignment of the two sequences is obtained.
   In applying the backward method, it is often obtained non-single solutions, so it is possible to gain some optimal alignment with the same optimum score.

Distance Matrix

In Isaev [4] a distance matrix is formed by a function of distance (distance function) defined as follows:

Definition 1. Suppose M is a set and $d: MxM \rightarrow R$ is a function, d said to be a distance function or a function of distance on M if

(i). $d(u,v) > 0$ for every $u, v \in M, u \neq v$,
(ii). $d(u,u) = 0$ for every $u \in M$,
(iii). $d(u,v) = d(v,u)$ for every $u, v \in M$,
(iv). Meet triangle inequality $d(u,v) \leq d(u,w) + d(w,v)$ for every $u, v, w \in M$

If d is the distance function of M, then $u, v \in M$, number $d(u,v)$ referred to as the distance between u and v. The set will be used here is a finite set $M = \{x1, x2, \ldots, xN\}$ which is the set of sequences (OTU) which will be formed its phylogenetic tree. It is assumed that distance function d is defined in M and d biologically is relevant, the intention is d in accordance with the genetic information contained in sequences (OTU) in M. For example $d(x1, x2) > d(x3, x4)$ means OTU $x1$ with $x2$ more distance evolutionary relationship or kinship than OTU $x3$ with $x4$. To simplify the writing, $d(xi, xj)$ is written as d_{ij} with $i, j \in \{1,2, \ldots, N\}$. Based on that distance function can be obtained distance matrix, $M_d = (d_{ij})$ with the following formal definition.

Definition 2. Suppose d is a distance function, Md called distance matrix defined by

$$Md = \begin{pmatrix} d_{11} & d_{12} & \ldots & d_{1j} & \ldots & d_{1N} \\ d_{21} & d_{22} & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ d_{i1} & \ddots & \ddots & d_{ij} & \ddots & d_{iN} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{Nj} & \cdots & d_{NN} \end{pmatrix}$$

with $i,j = 1,2,3,\ldots..N$ and N is the total OTU involved.

Kimura Evolutionary Model

In Christianini [1], Motoo Kimura proposed a model based on more than one parameter α showing the general possibility of replacement (substitution). In widespread use on models with two parameters, it will distinguish between possible transitions and transversions.

Having known many good substitution transition and transversion in each alignment, then look for the value of P and Q, where P is the number of transitions divided substitution length alignment, while the value of Q is the

number of substitution transversion divided the length of alignment. The results obtained from the evolutionary distance value through a formula Kimura [5]:

$$K = -\frac{1}{2}\ln\{(1-2P-Q)\sqrt{1-2Q}\}$$

where

K = distance value of evolutionary model of Kimura
P = probability of transition substitution (substitution Type I)
Q = probability transversion substitution (substitution Type II)

Distance Method

As stated by Saitou N and Nei M [6], neighbor-joining method is a distance-based method used to construct phylogenetic trees. Phylogenetic tree formed based on DNA or protein sequences. Each OTU (Operational taxonomic units) represent the sequence of MERS-CoV virus taken from samples that have been infected from some regions and countries that have reported sample of the data to GenBank. The input of this algorithm is a distance matrix, which is the elements are obtained by finding the difference of the nucleotide from the aligned sequence. And the steps of this neighbor joining algorithm is described once in the results and discussion.

RESULT AND DISCUSSION

This research used 65 samples of DNA sequence data of patients that was infected MERS virus accessed from NCBI (www.ncbi.nlm.nih.gov). The overall sample represents 37% of the 178 MERS cases noted worldwide as in Cotten et al [2]. From the entire whole sample used, each sequence is aligned. Then, the total of substitution type I (transitions) and the number of substitutions type II (transversion) are obtained in order to obtain evolutionary distance of kimura models for each result of those sequences alignment. From the evolutionary distance matrix, then phylogenetic tree is formed using neighbor joining algorithm based distance in which the whole process is simulated in matlab.

Most of the data sequence of MERS-infected patients were used in this study complete genomes, in the form of complete genome data with the number of base pairs (bp) were great, approximately 30.119 bp. Due to the limitation of matlab in accommodating the matrix (30,000 x 30,000), before the alignment process, limitation of the data were done by the following classifications:

- For a sequence with a length of > 10,000 bp, genome is taken by 75% of the overall length of sequence
- For a sequence with a length of <10,000 bp, genome fully rendered

After determining the sequence that will be used, the first step for data retrieval is to open a website www.ncbi.nlm.nih.gov. Then click All Database menu, Move the pointer to the nucleotide for nucleotide searching. In the search menu, type the nucleotide sequences code requested. After performing the required nucleotide sequence, show in FASTA code then save the code in a txt file. Then, FASTA code in this txt file is inserted to process sequence alignment which is simulated in Matlab.

Because the data used in this study were 65 sequences, thus the process of branching formation to form tree occurred as many as 64 cycles.

Some steps to determine the phylogenetic tree branching (neighbor joining algorithm), the authors demonstrated in cycle 58 as in Irawan and Amiroch [3], namely:

a. Published input from the new evolutionary distance matrix
b. Step 1: Calculate S_i for each OTU by following the formula

$$S_i = \frac{1}{N-2}\sum_{k=1}^{N} D_{ik}$$

In which S is the sum of the distance (D) between an OTU to another OTU, divided by (N-2), and N is the total sum of OTU, D_{ik} is the distance from i to k in the matrix's evolution. While N in this cycle is 8.

c. Step 2: Find the minimum value for each pair of sequences:

$$M_{ij} = D_{ij} - S_i - S_j$$

d. Step 3: Define a new OTU namely U_1 Which replaces the smallest pair (eg. C and D). Furthermore, these taxa are combined as U_1 follows the formula:

$$S_{CU_1} = 0.5(D_{CD} + S_C - S_D)$$
$$S_{DU_1} = 0.5(D_{CD} + S_D - S_C)$$

e. Step 4:
Connect taxa U_1 with C and U_1 with D, each of them following the length of edge as the result of calculation in step 3. At the tree obtained, The length of the branch represents the distance of the evolution.

f. Step 5:
Combine the new distance of all taxa to U_1

$$D_{AU_1} = 0.5(D_{CA} + D_{DA} - D_{CD})$$
$$D_{BU_1} = 0.5(D_{CB} + D_{DB} - D_{CD}), \text{ so on.}$$

Results from a new distance of U_1 next to all taxa then included in the new evolutionary of distance matrix.

The next calculation step is the same as in cycle 58, the value is different because N is different, the smallest Mij different, that finally the new distance to each taxa is also different. And this phase continues until all branches are formed.

Cycle 58:
Obtained new evolutionary distance matrix:

d_new = (1.0e-04*)

0	0.0210	0.0302	0.4349	0.0210	0.0198	0.0199	0.0199
0.0210	0	0	0.4157	0	0	0	0
0.0302	0	0	0.4156	0	0	0	0
0.4349	0.4157	0.4156	0	0.4156	0.4151	0.4151	0.4151
0.0210	0	0	0.4156	0	0	0	0
0.0198	0	0	0.4151	0	0	0	0
0.0199	0	0	0.4151	0	0	0	0
0.0199	0	0	0.4151	0	0	0	0

Obtained the value of S as follows: (1.0e-04 *)

S =	0.0945	0.0728	0.0743	0.4878	0.0728	0.0725	0.0725	0.0725

Obtained the value of M_{ij} as follows: (1.0e-04 *)

M =

0	-0.1462	-0.1385	-0.1474	-0.1463	-0.1471	-0.1470	-0.1470
-0.1462	0	-0.1471	-0.1449	-0.1456	-0.1453	-0.1453	-0.1453
-0.1385	-0.1471	0	-0.1466	-0.1471	-0.1468	-0.1468	-0.1468
-0.1474	-0.1449	-0.1466	0	-0.1450	-0.1452	-0.1453	-0.1453
-0.1463	-0.1456	-0.1471	-0.1450	0	-0.1452	-0.1453	-0.1453
-0.1471	-0.1453	-0.1468	-0.1452	-0.1452	0	-0.1450	-0.1450
-0.1470	-0.1453	-0.1468	-0.1453	-0.1453	-0.1450	0	-0.1450
-0.1470	-0.1453	-0.1468	-0.1453	-0.1453	-0.1450	-0.1450	0

Cycle 59 :
Obtained new evolutionary distance matrix:

d_new = 1.0e-06 *

0	0.0937	0.5458	0.0865	0.0010	0.0072	0.0072
0.0937	0	0	0	0	0	0
0.5458	0	0	0	0	0	0
0.0865	0	0	0	0	0	0
0.0010	0	0	0	0	0	0
0.0072	0	0	0	0	0	0
0.0072	0	0	0	0	0	0

Obtained the value of S as follows: (1.0e-06 *)

S = 0.1483 0.0187 0.1092 0.0173 0.0002 0.0014 0.0014

Obtained matrix M_{ij} as follows: (1.0e-06 *)

M =

0	-0.0733	0.2884	-0.0791	-0.1475	-0.1425	-0.1425
-0.0733	0	-0.1279	-0.0360	-0.0189	-0.0202	-0.0202
0.2884	-0.1279	0	-0.1264	-0.1094	-0.1106	-0.1106
-0.0791	-0.0360	-0.1264	0	-0.0175	-0.0187	-0.0187
-0.1475	-0.0189	-0.1094	-0.0175	0	-0.0016	-0.0016
-0.1425	-0.0202	-0.1106	-0.0187	-0.0016	0	-0.0029
-0.1425	-0.0202	-0.1106	-0.0187	-0.0016	-0.0029	0

In the same way, cycle continues until cycle 64.

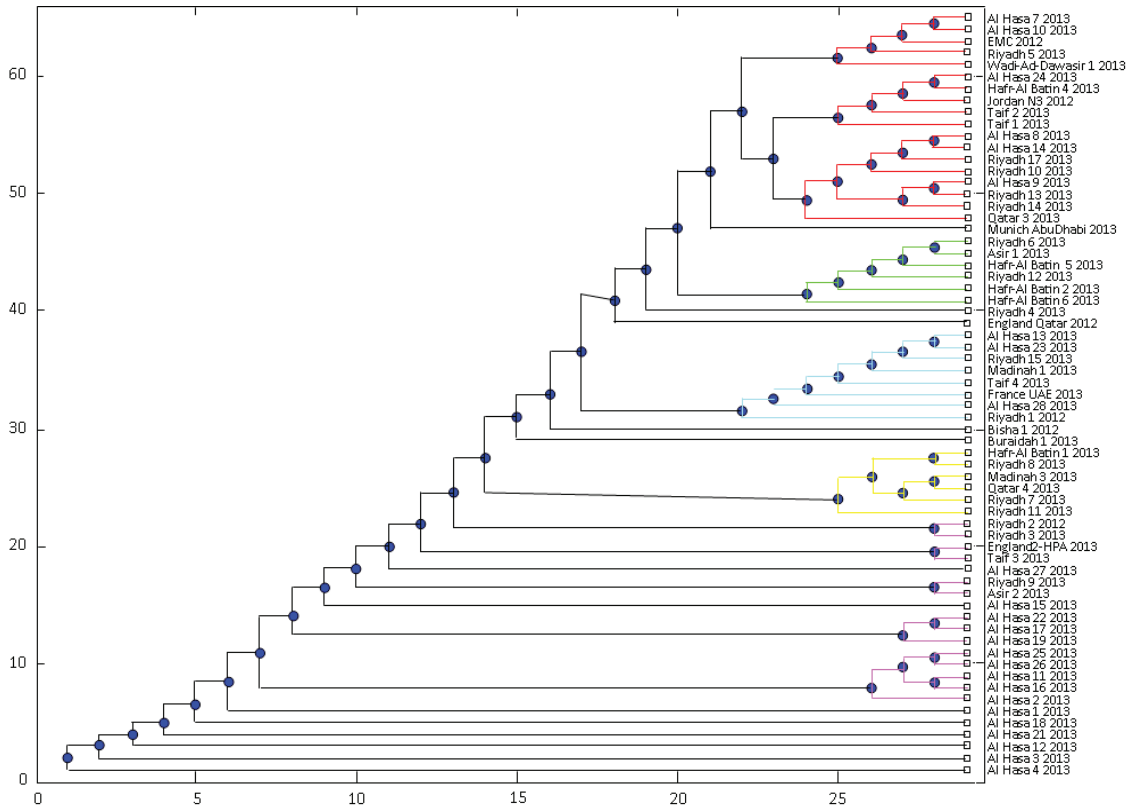

FIGURE 1. Phylogenetic tree of MERS deployment pattern resulted from Matlab simulation

Evolutionary distance obtained from each cycle will ultimately determine the length of the evolutionary distance from one individual to another, which indicates the level of genetic closeness between species. And from the formation steps of MERS deployment pattern, resulting phylogenetic tree visualization Matlab output above.

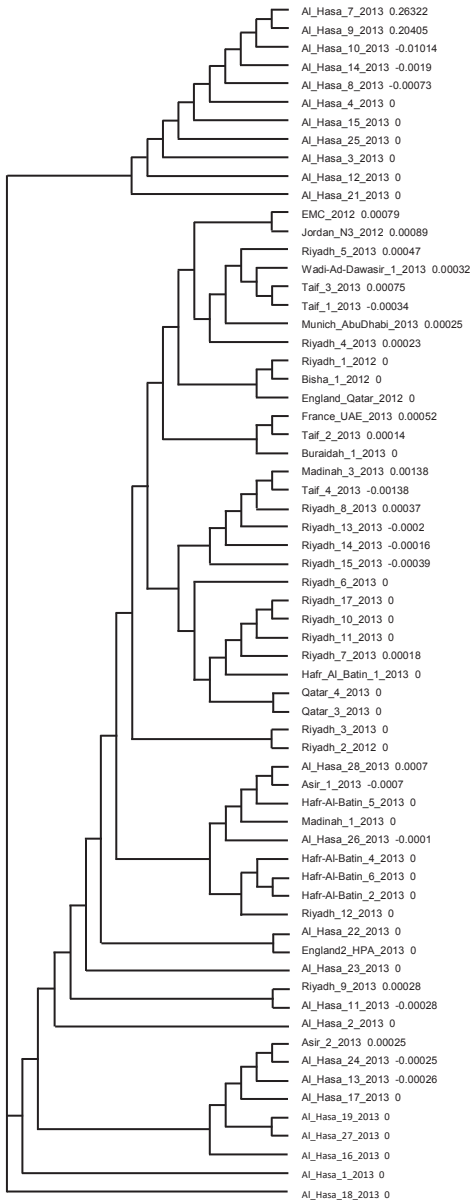

FIGURE 2. Phylogenetic tree pattern of geographical spread MERS Co-V resulted form MUSCLE

Based on the pattern of spread resulted from Matlab formations as shown above, it can be seen that the pattern of spread of MERS Co-V in the Arabian Peninsula was divided into 6 groups. The first group is the **Al-Hasa**. The closeness of the result of genetic and phylogenetic tree known that virus originated from Al-Hasa. In this group, al-Hasa Virus survived from April 22nd until May 30th. The second group is **Riyadh**. Continuing virus of al-Hasa, but in this second group, it can be seen that the virus survived 10 months from October 30th, 2012 until the beginning of August 2013. The third group is **Hafr al-Batin1**. Hafr Al-Batin lineage virus started from the June 4th, 2013 until October 1st, 2013, lasted in 4 months. Group 4 is a lineage of **Buraidah1**. In this group, the virus survived from October 23rd, 2012 to September 1st, 2013, long enough deployment which was up to 11 months. Group 5 is a

lineage **Asir1**. In this group, the virus survived on July 2nd, 2013 until August 28th, 2013, adequately short just in two months. And a group of 6 is **Jordan** lineage virus began on April 15th, 2012 until October 1st, 2013, for the last 6 months. Of those six groups, described from the figure the closeness distance of the virus genetic and early deployment was **al-Hasa**.

In the figure 2, at glance, the geographic distribution patterns of MERS that occur were not entirely the same, but when they were viewed from the closeness and the beginning of the virus originated, it seems clear that the initial spread of the virus is from Al-Hasa.

The process of computing with matlab takes 3.66 hours commencing from the running program, began to see each sequence, sequence alignment, kimura calculation, distance matrix calculation until the formation of the pattern. While on MUSCLE, the process of data entry is done manually one by one for 65 minutes and running process lasts up to 2 hours.

CONCLUSION

From the whole process of the formation of the geographical pattern of MERS-CoV, it can be concluded that distance method using Kimura Model successfully applied to determine the geographic distribution patterns MERS Co-V. The beginning of the spread of the virus is from Al-Hasa. It can be shown on the pattern of spread of MERS either the output Matlab or result form MUSCLE. Overall the phylogenetic tree Matlab output results more accurate because even though both use the distance method (with a neighbor joining algorithm), but output Matlab models take into account the changes that Kimura evolutionary transition and transversion at each nucleotide, while the results of MUSCLE ignore it. And because the process in Matlab was more detail, it affected the processing of running time program. To produce the pattern of spread MERS Co-V with Matlab simulation, it took 4 hours, while the MUSCLE took 2 hours.

ACKNOWLEDGMENTS

This research is the outcome of a research lecturer starters in 2016 were financed by Daftar Isian Pelaksanaan Anggaran (DIPA), Direktorat Jenderal Penguatan Riset dan Pengembangan (Kemristekdikti) number SP/DIPA-023.04.1.673.453/2016, Revision 01 by March 3rd, 2016.

REFERENCES

1. Christianini, N., Hahn, M.W, *Introduction to Computational Genomics A Case studies Approach,* Cambridge University Press, New York, 2006.
2. Cotten et all, *Spread, Circulation, and Evolution of the Middle East Respiratory Syndrome Coronavirus*, (http://mbio.asm.org vol 5 issue 1e01062-13, 2014), accessed on 12th April 2015.
3. Irawan I, Amiroch S, *Construction Of Phylogenetic Tree Using Neighbor Joining Algorithms To Identify The Host And The Spreading Of SARS Epidemic*, Journal of Theoretical and Applied Information Technology Vol.71 No.3 pg. 424-429 (2015).
4. Isaev, A., *Introduction to Mathematical Methods in Bioinformatics*, Springer (2006).
5. Kimura, Motoo, *A Simple Method for Estimating EvolutionaryRates of Base Substitutions Through Comparative Studies of Nucleotide Sequence,* Journal of Molecular Evolution vol 16 pg 111-120 (1980).
6. Saitou N, Nei M, *The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.* Mol. Biol. Evol. 4(4): 406-425 (1987).
7. Shen, S.N., *Theory and Mathematical Methods for Bioinformatics,* Biological and Medical Physics, biomedical Engineering, Springer, 2007.