Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method

By Siti Amiroch

WORD COUNT 2699 TIME SUBMITTED 14-JAN-2019 05:15PM
PAPER ID 43525011

Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method

S Amiroch[1], M S Pradana[1], M I Irawan[2] and I Mukhlash[2]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Islam Darul 'Ulum Lamongan, Indonesia
[2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail:siti.amiroch@unisda.ac.id

Abstract. Multiple Alignment (MA) is a particularly important tool for studying the viral genome and determine the evolutionary process of the specific virus. Application of MA in the case of the spread of the Severe acute respiratory syndrome (SARS) epidemic is an interesting thing because this virus epidemic a few years ago spread so quickly that medical attention in many countries. Although there has been a lot of software to process multiple sequences, but the use of pairwise alignment to process MA is very important to consider. In previous research, the alignment between the sequences to process MA algorithm, Super Pairwise Alignment, but in this study used a dynamic programming algorithm Needleman wunchs simulated in Matlab. From the analysis of MA obtained and stable region and unstable which indicates the position where the mutation occurs, the system network topology that produced the phylogenetic tree of the SARS epidemic distance method, and system area networks mutation.

1. Introduction

Multiple alignment (MA) is the alignment of multiple (more than two) sequences simultaneously. One of the main purposes of a MA is to form a phylogenetic tree, a topology tree that describe the kinship among multiple sequences. In this study, MA is developed from sequence alignments by using the algorithm of Needleman Wunchs, one of the main dynamic programming algorithms that applies the principle of global alignment for pairwise alignment.

There are several kinds of analysis that can be applied in MA, but in general, all analyses are used to determine the network mutation from MA to network decomposition of orthogonal alignment itself.

Although SARS epidemic has long appeared, the spread pattern is still interesting to be studied, especially from mathematical side which is supported by matlab software with its simulation capabilities to allign MA. The pattern of the spread of the SARS virus during the winter and spring of 2003 can be described as a tree forming a network connectivity that branches out as an epidemic is transmitted from one individual to another. Along with technological advances in DNA sequencing, it is concluded that phylogenetic tree diagram is closely related to DNA analysis which is based on whole genome comparisons between different SARS viruses and it will show mutation traces in these individuals [2]. This study used distance method in the process of phylogenetic tree formation.

2. Methodology

2.1 Data Preparation

The data used in this study were 6 DNA sequences of patients infected by SARS virus and one DNA sequence of Palm Civet (in 7th sequence) known as the host of SARS epidemic [3]. DNA sequences were retrieved from GenBank (www.ncbi.nlm.nih.gov). Fortunatelly the research used balance data, so it is in the preprocessing not necessary to integrate data and machine learning[5].

2.2 Analysis of Multiple Alignment Procedures

Sequence Alignment is a comparison between two biology sequences. It is an important method in analysis of position and type of mutation. The most important thing in sequence alignment is to determine mutation displacement [6]. Given two sequences A and A_2 which are defined as:

$$A = (a, a_{12}, \dots, a_{n_a}) \text{ and } A_2 = (a_{21}, a_{22}, \dots, a_{2n_a}) \quad (1)$$

The elements of the sequence A and A_2 representing the DNA sequence have a range $V_5 = \{0,1,2,3,4\}$ or $\{a, c, g, t, -\}$. While Multiple Sequence is a collection of sequences expressed as:

$$\mathcal{A} = \{A_1, A_2, \dots, A_m\} \quad (2)$$

For every A_s, A_s is a separated sequence defined on V_q, and expressed as :

$$A_s = (a_{s,1}, a_{s,2}, \dots, a_{s,n_s}), \qquad s = 1,2, \dots, m \quad (3)$$

where n_s is the length of sequence A_s and m is the number of sequences in each group.

Multiple Alignment (MA) is the alignment performed on multiple sequences. Suppose A is a multiple sequence consisting of 7 sample data and A' is the result of MA. According to [6], the procedures for the analysis of MA output are as follows: First, penalty matrix $W = (w_{s,t})$ is calculated based on A', and then the minimum distance of G_1 tree is constructed. After that *k-order* graph G_k and *k-order* network mutation $G_k(W)$are constructed. Finally, based on the minimum distance of G tree and mutation region of matrix Δ, network is orthogonalized and the appropriate graph for the network of orthogonal decomposition is formed.

2.3 Penalty Matrix

Penalty Matrix is defined by [6], e.g. C is an alignment matrix induced by multiple sequence A. If penalty function $W=w(a,b)$ is defined on a given V_5, for any $s, t \in M$, there are two expansions $C_{s,t}$ and $C_{t,s}$ based on the sequence A_s and A_t. Penalty score for pair $C_{s,t}$ and $C_{t,s}$ is defined by:

$$w_{s,t}(\bar{C}) = w(C_{s,t}, C_{t,s}) = \sum_{j=1}^{n_{s,t}} w(C_{s,t;j}, C_{t,s;j}) \quad (4)$$

While the matrix

$$\bar{W}(\bar{C}) = [W_{s,t}(\bar{C})]_{s,t=1,2,\dots,m} \quad (5)$$

is a penalty matrix induced by pairwise alignment of multiple sequences A and simplified as a penalty matrix.

2.4 Distance Method

Distance method is one of methods for tree formation from a set of distance between each pair of sequences that has been aligned. The set of distance is written in matrix form so-called distance matrix

[1]. Formally, a distance matrix is formed based on distance function defined as follows [4]:

Suppose M is a set and $d: MxM \rightarrow R$ is a function, d is said to be a distance function 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) Meets the 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 on M, then for $u,v \in M$, numbers $d(u,v)$ is referred to as the distance between u and v. Distance matrix is obtained based on distance function, $M_d = (d_{ij})$ with $i, j = 1,2,3,$ ..... N and N is the number of sequences involved. From the definition, if a penalty matrix meets a function of distance, then the penalty matrix is also referred to as a distance matrix.

2.5 Graph Theory

Graph theory, a part of mathematics studying the graph, is used to model the relationship of ordered pairs of a particular set of objects. Graf (G) is written with the notation $G = (V,E)$ where (V) is not an empty set of points and (E) is the set of edges connected pair of points [7]. In an undirected graph, there is no difference in the direction at two points connecting each side. While in a directed graph, its side points towards one of the main points. Graf with weight is called a weighted graph, in which each connected pair represents a specific numerical value (weight). A directed graph with weighted side is called a *Network*.

3. Results and Discussion

3.1 Analysis of Topological Network System

Topology network system generated by MA output is $G(W) = \{M, V, W\}$ where W is a penalty function of the output MA. By following the alignment of multiple analysis procedures and implementing algorithms Wunch Needleman alignment for either pairwise alignment or the MA simulated in Matlab, the penalty matrix obtained is as follows:

$$\bar{W}(\bar{C}) = \begin{array}{c} \begin{matrix} A & B & C & D & E & F & G \end{matrix} \\ \begin{bmatrix} 0 & 8 & 9 & 12 & 12 & 11 & 3 \\ 8 & 0 & 1 & 4 & 4 & 3 & 7 \\ 9 & 1 & 0 & 3 & 3 & 2 & 8 \\ 12 & 4 & 3 & 0 & 2 & 1 & 11 \\ 12 & 4 & 3 & 2 & 0 & 1 & 11 \\ 11 & 3 & 2 & 1 & 1 & 0 & 10 \\ 3 & 7 & 8 & 11 & 11 & 10 & 0 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \end{matrix} \end{array}$$

where A represents the sequence of Guangzhou 16/12/02, B represent the sequence of Guangzhou Hospital, C represent the sequence of Metropole 02/21/03, D represents the sequence of Hanoi 02/26/03, E represents the sequence of Toronto 27/02/03, F represents the sequences of Singapore 01/0/03, and G represent sequences of Palm Civet. Palm Civet is a ferret alleged as host of SARS epidemic [3]. Sequence of palm civet is shown to know which DNA sequences that is closer to the host. A sequence closest to the host means the beginning point of the spread of SARS epidemic. The analysis of topological network system also results in a stable area showing the same nucleotide position in the multiple alignment and unstable regions showing different nucleotide positions. In this unstable region. mutation between sequences is located. Below is a table of unstable regions and position of each nucleotide in the MA of SARS epidemic:

Table 1. The number of nucleotides in the unstable region and their position.

No.	Position of nucleotide	Amount of nucleotide				
		A	C	G	T	-
1	230	4	0	3	0	0
2	654	0	2	0	5	0
3	716	0	5	0	2	0
4	731	0	4	0	3	0
5	931	2	0	5	0	0
6	1026	4	0	3	0	0
7	1031	5	0	2	0	0
8	1502	1	0	0	6	0
9	1729	0	0	1	6	0
10	2332	0	0	2	5	0
11	2380	0	6	0	1	0
12	3075	0	2	0	5	0
13	3381	0	1	0	6	0
14	3487	6	0	1	0	0

Table 1 shows that only 14 positions away from 3768bp sequence which is an area unstable (0,4%), changes in nucleotide mutation causes. At each position, the number of changed nucleotide is not the same. For example, the number of nucleotide A (adenine) of position 230 is 4 and nucleotide G (Guanine) is 3. However, there is a different number in different. position of unstable areas.

3.2 Analysis of mutation region network system

The second analysis is the network system of a mutation region on MA of SARS epidemic. This section discusses how to build a graph and a tree generated by the SARS epidemic. The graph is in the form of a phylogenetic tree, a family tree reflecting the evolutionary relationships between sequences. In this case, the phylogenetic tree illustrates evolutionary relationship showing the epidemic spread from one sequence to another in which each sequence is taken from different areas. Therefore the output of the phylogenetic tree will indicate the spread of SARS epidemic among regions/countries. In this study, the input in the formation of a phylogenetic tree is distance matrix.

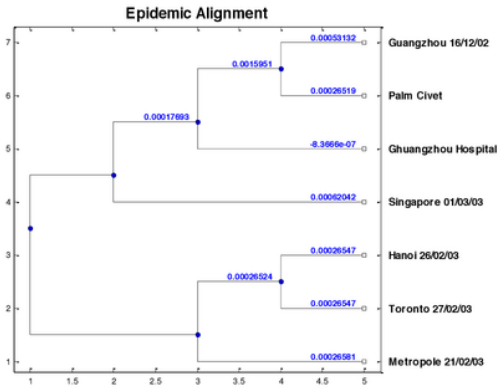

Figure 1. Phylogenetic tree epidemic of spread of SARS.

Figure 1 shows that mutation region or the spread of SARS epidemic is originated from GuangZhou 12/16/02 because DNA sequence of the patients from 12/16/02 GuangZhou genetic has the closest

distance to Palm Civet as host of the SARS virus [3]. Then it spread to Guangzhou Hospital continues to Singapore. After that then spread to Metropole, Hanoi and Toronto simultaneously.

3.3 Analysis of Network Systems of Mutation Mode

Before performing a network analysis of mutation mode, non-directional graph showing relationship between sequence mutations is visualized from penalty matrix.

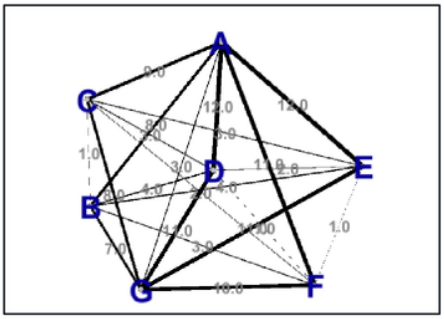

Figure 2. Decomposition of network mutation of SARS epidemic.

Notation on node indicates the name of sequence encoded as A, B, ... G with each code representing a sequence name as stated in previous discussions. In Figure 2, the label on the side shows the number of mutations. The thicker lines show that more mutations occur between those nodes. As shown in previous discussion, there are 14 different mutations in unstable regions on 7 DNA sequences of this SARS epidemic. Figure 2 shows that some mutations occur only in the arc orthogonal of 1st order, for example in $\Delta\, ABE,\ \Delta\, ABF,\ \Delta\, ABD,\ \Delta\, AFD,\ \Delta\, BFD$. At $\Delta\, ABE$: |AE| = |AB| + |BE| and the structure of modulus H_{AE}, H_{AB}, H_{BE} is mutually orthogonal. The following table represents multiple mutations between two sequences representing mutation mode H_{AE} (a mutation in sequence of Guangzhou 16/12/02 to sequence of Toronto 27/03/03), mode mutation H_{AB} (a mutation in sequence of Guangzhou 16/12/02 to sequence of Guangzhou Hospital), and mode mutation H_{BE} (a mutation in sequence of Guangzhou Hospital to sequence of Toronto 27/03/03).

Table 2. Movements between two sequences on mutation mode H_{AB}, H_{AE}, H_{BE}.

Sequence 1	Sequence 2	Nucleotide		Position	Mutation	Percentage	
		Diff.	Similar			Diff.	Similar
A. Guangzhou 16/12/02	B.Guangzhou Hospital	8	3760	654	C to T	0.2%	99.8%
				716	T to C		
				931	A to C		
				1031	G to A		
				1502	A to T		
				2332	G to T		
				2380	T to C		
				3075	C to T		
A.Guangzhou 16/12/02	E. Toronto 02/27/03	12	3756	230	A to G	0.3%	99.7%
				654	C to T		
				716	T to C		

				731	C to T		
				931	A to C		
				1026	G to A		
				1031	G to A		
				1502	A to T		
				1729	T to G		
				2332	G to T		
				2380	T to C		
				3075	C to T		
B. GuangZhou Hospital	E. Toronto 27/02/03	4	3764	230	A to G	0.10 %	99.90%
				731	T to C		
				1026	G to A		
				3381	T to C		

4. Conclusion

The result of the analysis of multiple allignment shows that there are 14 unstable regions of 3768bp where mutations occur with a different nucleotide number. The areas are located at position 230, 654, 716, 731, 931, 1026, 1031, 1502, 1729, 2332, 2380, 3075, 3381, and 3487. The result of the analysis of mutation region network shows that the spread of the SARS epidemic stems from Guangzhou 16/12/02, then spread to the Guangzhou Hospital continues to Singapore. After that then spread to Metropole, Hanoi and Toronto simultaneously. While the analysis of mutations mode network system, it is known that there are only a few mutations are orthogonal to the order-1.

Acknowledgments

This research is the outcome of a cooperation research between universities in 2017 were financed by Direktorat Riset dan Pengabdian Masyarakat, Direkorat Jenderal Penguatan Riset dan Pengembangan, Kementerian Riset, Teknologi dan Pendidikan Tinggi (Kemenristekdikti) number 101/SP2H/PPM/DRPM/IV/2017, April 3rd 2017.

References

[1] Amiroch, S., Rohmatullah, A., 2017, *Determining Geographical Spread Pattern of Mers-CoV by distance methods using Kimura Model,* AIP Conference Proceedings 1825 (1), 020001.
[2] Christianini, N. , Hahn, MW, 2006, *Introduction to Computational Genomics studies A Case Approach,* Cambridge University Press, New York.
[3] Guan, et al., 2003, *Isolation and characterization of Viruses Related to the SARS coronavirus from Shouthern Animals in China.* Science 302, 276 (2003), The American Association for the Advancement of science, Washington,(www.sciencemag.org), accessed on 29 January 2014.
[4] Isaev, A., 2006, *Introduction to Mathematical Methods in* Bioinformatics, Springer.
[5] Mahdiyah, U, M. Isa Irawan, Elly M. Imah, 2016, *Integrating Data Selection and Extreme Learning Machine for Imbalance Data,* Journal Procedia Computer Science Vol. 59 pg. 221-229
[6] Shen, SN (2007), *Theory and Mathematical Methods for Bioinformatics,* Biological and Medical Physics, Biomedical Engineering, Springer.
[7] Yallen, Jay, (2012), *Graph Theory and its Application Second Edition.* New York, Chapman & Hall.

Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method

ORIGINALITY REPORT

21%

SIMILARITY INDEX

PRIMARY SOURCES

1	Siti Amiroch, Arif Rohmatullah. "Determining geographical spread pattern of MERS-CoV by distance method using Kimura model", AIP Publishing, 2017 Crossref	87 words — 3%
2	S Sa'adah, F S Tapilouw, T Hidayat. "Undergraduate Students' Difficulties in Reading and Constructing Phylogenetic Tree", Journal of Physics: Conference Series, 2017 Crossref	80 words — 3%
3	earchive.tpu.ru Internet	50 words — 2%
4	"Network Structures of Multiple Sequences Induced by Mutation", Biological and Medical Physics Biomedical Engineering, 2008 Crossref	39 words — 2%
5	"Multiple Sequence Alignment", Biological and Medical Physics Biomedical Engineering, 2008 Crossref	39 words — 2%
6	www.lppm.its.ac.id Internet	38 words — 1%
7	shura.shu.ac.uk Internet	32 words — 1%
8	I Aisah, M Suyudi, E Carnia, Suhendi, A K Supriatna. "Representation mutations from standard genetic	24 words — 1%

codes", IOP Conference Series: Materials Science and Engineering, 2018
Crossref

9	Wim Gaastra, Frits R. Mooi, Antoine R. Stuitje, Frits K. Graaf. " The nucleotide sequence of the gene encoding the K88ab protein subunit of procine enterotoxigenic ", FEMS Microbiology Letters, 1981 Crossref	22 words — 1%
10	stats.ma.ic.ac.uk Internet	19 words — 1%
11	toc.proceedings.com Internet	18 words — 1%
12	M Basyuni, R Wati. "Bioinformatics analysis of the oxidosqualene cyclase gene and the amino acid sequence in mangrove plants", Journal of Physics: Conference Series, 2017 Crossref	17 words — 1%
13	sci.nic.in Internet	17 words — 1%
14	N.W. Rahayu, S.N. Huda. "The Use of ICT to Support Perpetual Undergraduate Students", IOP Conference Series: Materials Science and Engineering, 2017 Crossref	17 words — 1%
15	repository.unikama.ac.id Internet	14 words — 1%
16	airccj.org Internet	12 words — < 1%
17	www.bionewsonline.com Internet	12 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	ON		