Zika Virus Mutation and The Spreading to Indonesia

Mohammad Syaiful Pradana and Siti Amiroch

***Abstract*—More than 13 countries in the Americas have reported sporadic Zika virus infection that show very rapid geographic expansion. While in Indonesia, the euphoria is also increasingly prevalent virus discussed especially after the discovery of Jambi positive patients infected with the virus Zika on January 26, 2016 last.Viruses transmitted by mosquito bites to humans are the same mosquito transmits dengue fever, chikungunya and yellow fever with symptoms that have similar resemblance to the zika virus. Based on similarity of symptoms of infection with dengue virus, can be analyzed using sequence alignment to get identical percentage, local alignment calculation, genetic mutation and the spread of the zika virus to Indonesia. From the whole process, Smith Waterman algorithm can be used to align dengue virus type 1, type 2, type 3, and type 4 with zika virus (jambi). Mutations between dengue virus and zika virus on average 28% of dissimilarity sequences in the same position between the sequence. Overall, the dengue virus type 1 mutation to the zika virus was 28.2723%, the dengue virus type 2 mutation to the zika virus was 28.1984%, the dengue virus type 3 mutation to the zika virus 27.9373% and the mutation of dengue virus type 4 to the zika virus of 28.7206%. By looking at all mutations, from the simulation results note that the mutations of both viruses include mutation type I. The phylogenetic tree showed the spread of the Zika virus to Indonesia, originally from South Africa, the islands of Chile, Caledonia, the Philippines, Yap Micronesia, Thailand, Cambodia, and finally reached Indonesia. Zika virus jambi suspected of dengue virus mutation because a few years earlier there was a dengue virus outbreaks in a long time in jambi, it turns out the virus zika jambi is not from mutation dengue virus but the virus comes from the Asian region.**

***Index Terms*—Virus mutation, virus spreading, Zika virus.**

I. Introduction

THE Zika virus was identified in 1947 in rhesus monkeys and was identified in humans in 1952 in Uganda and the Republic of Tanzania. Viruses transmitted by mosquito bites (Aedes especially Aedes aegypti) to humans are the same mosquito transmits dengue fever, chikungunya and yellow fever with symptoms that have similar resemblance to the Zika virus that is fever, skin rash, conjunctivitis, muscle and joint pain, malaise and headache. Meanwhile, the Zika virus outbreak was first reported from the Pacific in 2007 and 2013 (Yap and Polynesia), and in 2015 from the Americas (Brazil and Colombia) and Africa (Cabo Verde) until the virus reached Indonesia (Jambi) at 26 January 2016.

Based on similarity of symptoms of infection with Dengue virus, can be analyzed using sequence alignment from Zika virus sequence with Dengue virus to get identical percentage, local alignment calculation and genetic mutation that happened. Sequence Alignment which is a procedure for aligning two sequences of DNA or proteins in order to find common ground between the sequences. Sequence alignment has two methods, namely global alignment and local alignement. Global alignment is the alignment for the entire sequence, using as many characters as possible in DNA. Local alignment is partial alignment of sequences, taking part that has a high enough level of resemblance. One of the local alignment algorithms is Smith Waterman. Although it looks simple for the development of algorithms based on dynamic programming, this algorithm is very instrumental in bioinformatics. Next, constructing phylogenetic trees of zika virus to know the spread of virus zika to Indonesia using MUSCLE.

Manuscript received October 31, 2017; accepted November 29, 2017.

The authors are with the Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Islam Darul Ulum Lamongan, CO 62253, Indonesia. E-mails: syaifulp@unisda.ac.id, siti.amiroch@unisda. ac.id

This research is the outcome of research lecturer starter in 2017, which were funded by Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Research, Technology and Higher Education in accordance with the Research Contract Budgeting year 2017, Number 120/SP2H/LT/DRPM/IV/2017 on April 3, 2017.

II. Methods

A. DNA and Protein

DNA (*Deoxyribo Nucleic Acid*) is a macro molecule composed by nucleotides as the basic molecules that carry the properties of genes. DNA is formed by four types of nucleotides that are covalently bonded and represented by the letter A (*adenine*), C (*cytosine*), G (*guanine*), T (*thymine*).

While proteins are formed from simple molecular chains called *amino acids* (aa). The final shape of the protein is determined by the proper identity, the amino acid chain sequence, and is largely the atomic interaction between amino acids and the cell medium (mostly water). There are 20 amino acids used to form proteins that are A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V. [1].

The symbol is a unique sequence determined by gene encoding of the gene consisting of three sets nucleotides are called codon, such as: AAA and AAG representing K, AGA and AGC representing R, GAA and GAG representing E, etc.

B. Sequence Alignment

DNA sequence, RNA Sequence and protein sequence are commonly determined based on biological sequence. At stated in Shen [2], biological sequence described with the following notation:

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

where X,Y,Z denote sequence, and x_i, y_i, z_i are basic units of sequence at i-position. Those elements are obtained from the set $V_q = \{0,1,...,q-1\}$. The length of X,Y,Z are expressed by n_x, n_y, n_z respectively. If X,Y,Z are DNA/RNA sequence, then $V_4 = \{a,c,g,t\}$ or $\{a,c,g,u\}$, whereas if the protein sequence, then $q = 20$ and $V_q = \{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$ which directly represent 20 amino acid molecules.

Sequence alignment in Shen [2], 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 Multiple Sequence Alignment. Determining the movement of mutation is the core idea of sequence alignment.

C. Smith Waterman Algorithm

This research applies Smith Waterman's algorithm which is based on a local alignment algorithm. Two important aspects of Smith Waterman's Algorithm are:

- Calculates values in a two-dimensional table. Smith Waterman's algorithm add 0 value when computing $s(i,j)$ so that a negative score will never happen to this algorithm.

$$S(i,j) = \begin{cases} 0 \\ s(i-1,j-1) + s(x_i,y_i) \\ s(i-1,j) - d \\ s(i,j-1) - d \end{cases} \quad (1)$$

- Traceback Algorithm. Starting and ending points of the backtrace method on the Smith Waterman's algorithm selected with maximum scores. The end point is the first element with 0 value on the backtrace process. A starting point with maximum score will guarantee maximum scores on local alignment sequences and the end point is the first element with a 0 value ensuring that the section is not exceeded.

D. Genetic Mutation

Mutations are a change in genetic sequence, which is a major cause of differences among organisms. These changes occur on many levels, with very different consequences [3]. DNA sequences mutation can be classified into 4 types [2]:

1) Type I : there is a nucleotide change, for example from "a" to "g".
2) Type II : there is a nucleotide section that changes the order of its position, for example the "accgu" section changes the sequence to "guacc".
3) Type III : the insertion of a new segment into the sequence, for example the insertion of "aa" in the middle of the "gguugg" segment will change the segment to "gguaaugg".
4) Type IV : there is elimination of the nucleotide segment in the sequence, for example removes the "ag" nucleotide from the "acaguua" segment so that the segment changes to "acuua".

In the type I and type II mutations, the positions of all nucleotides do not change, these mutations are called substitution mutations. As for type III and type IV mutations that can alter the nucleotide position, it is referred to as a transfer mutation.

III. Results and Discussions

A. Data Source

The data sequences of virus taken online in *genbank*, which the world's largest gene database belonging to the United States government by accessing the *National Center for Biotechnology Information* [4]. The sequence of each virus stored in FASTA code (.txt) format for the alignment process and construction of phylogenetic tree. In this research, we used 19 sample data protein sequence of zika virus, 4 sample data protein sequence of dengue type 1, type 2, type 3, and type 4 originating from Indonesia. Furthermore, 1 virus sequence was taken from Indonesia (Jambi) and aligned with the dengue virus sequence of each type, which is also taken from Indonesia. The dissimilarity and similarity of each sequences and percentages, duration process, and mutations between sequence could be see from the data analysis. The process is statically because the aligned sequences are only 4 dengue sequences with 1 zika sequence. Overall the process is simulated in matlab. The protein sequence data taken online in genbank are shown in Tables I and II respectively.

TABLE I
PROTEIN SEQUENCE DATA OF INFECTED DENGUE VIRUS PATIENT.

No	Access Code	Type	Sequence Length	Date of Sample Collection	Explanation
1	AHG06327	1	3392	15-02-2008	Dengue virus 1 isolate Makassar-0398, complete genom
2	AHG06364	2	3391	05-04-2010	Dengue virus 2 isolate Makassar-WS80, complete genom
3	AHG06376	3	3390	22-03-2010	Dengue virus 3 isolate Makassar-WS78, complete genom
4	AHG06382	4	3387	30-04-2008	Dengue virus 4 isolate Makassar-2007, complete genom

B. Sequence Alignment Result and Analysis

The data sequence of patients infected by zika virus (Jambi) and data sequence of dengue virus proteins in the table 1 and 2 are stored in (.txt) file inputed for sequence simultaneous alignment process in matlab using the Smith Waterman algorithm, the alignment between Zika virus (Jambi) and dengue virus of each type are shown in Table III.

Based on the output four alignments, and by comparison with the Basic Local Alignment Search Tool (BLAST), a program to compare the sequence of nucleotides or proteins to

TABLE II
PROTEIN SEQUENCE DATA OF INFECTED ZIKA VIRUS PATIENT.

No	Access Code	Sequence Length	Country	Date of Sample Collection	Explanation
1	KM 078936	976	Easter Island Chili	1 Maret 2014	Partial cds
2	KJ 873160	893	New Caledonia	3 April 2014	Partial cds
3	KJ 776791	10.807	French Polinesia	28 Nov 2013	Complete genom
4	KM 851039	789	Thailand	19 Juli 2014	Partial cds
5	KF 993678	10.141	Canada	19 Feb 2013	Partial cds
6	AMK 49492	383	Indonesia (Jambi)	30 Des 2014	Partial cds
7	JN 860885	10.269	Cambodia	2010	Partial cds
8	EU 545988	10.272	Yap Micronesia	Juni 2010	Complete cds
9	KM 851038	789	Philippines	9 Mei 2012	Partial cds
10	HQ 234499	10.269	Malaysia	1966	Partial cds; host: Aedes Aegypti
11	MR766/ ABY86749	255	EI	2015	Partial cds
12	AY632535/ AAV34151.1	10.794	Uganda	1947	Complete cds; Host: sentinel monkey
13	KF 268948	10.788	Central African Republic	1976	Complete cds; Host: aedes Africanus
14	KF 383091	708	Senegal	2001	Partial cds
15	HQ 234500	10.251	Nigeria	1968	Partial cds
16	KF 383084	708	Senegal	1991	Partial cds
17	HQ 234501	10.269	Senegal	1984	Partial cds
18	KF 383113	708	Cote de Ivoire	1980	Partial cds
19	DQ 859064	10.290	South Africa	-	Complete cds; Spondweni virus

the sequence database and calculate the statistical significance of the two sequence matches are shown in Table IV.

Table IV above, showing the identical value of the matlab simulation results is more thorough than the BLAST output. Proven by level of accuracy in matlab simulation output to 4 decimal numbers, while the BLAST shows only 2 significant figures and also the duration of computation time on matlab simulation is shorter than BLAST.

Before constructing phylogenetic trees, each sequence is aligned first. Based on 19 data sequences, the data types are not the same, some use protein data and also DNA data. So most of the DNA data is converted into protein data and accessed following the sequence of protein sequences. So the whole protein sequence data Zika virus is shown in Table V below.

TABLE III
MATLAB RESULT BASED SMITH WATERMAN ALGORITHM.

Sequence			Percentage (%) Similarity/Dissimililarity		Duration (s)
DEN-V	Type	ZIK-V	Sim	Diss	
AHG 06327	1	AMK 49492	71.4660	28.27	0.062
AHG 06364	2	AMK 49492	71.0183	28.20	0.094
AHG 06376	3	AMK 49492	71.5405	27.94	0.359
AHG 06382	4	AMK 49492	70.7572	28.72	0.156

TABLE IV
MATLAB AND BLAST COMPARISON.

Sequence			Identical value		Duration (s)	
DEN-V	Type	ZIK-V	Matlab	BLAST	Matlab	BLAST
AHG 06327	1	AMK 49492	71,466 %	71 %	0,062	12,16
AHG 06364	2	AMK 49492	71,0183%	71 %	0,094	11,27
AHG 06376	3	AMK 49492	71,5405%	71 %	0,359	6,58
AHG 06382	4	AMK 49492	70,7572%	71 %	0,156	10,68

A few of identical percentage matrix in each sequence obtained by output MUSCLE (clustal w2) are shown in Fig. 1.

```
 1: ABI54480.1  100.00   76.69   74.97   77.95
 2: AHL43476.1   76.69  100.00   91.95   90.40
 3: AEN75264.1   74.97   91.95  100.00   98.48
 4: AKH87424.1   77.95   90.40   98.48  100.00
 5: ABY86749.1   71.76    -nan   96.86    -nan
 6: AHZ13508.1   75.00   90.68   98.92   99.62
 7: AHL37808.1   75.04   90.68   98.91   99.62
 8: AMK49492.2   75.03   90.68   98.69   99.62
 9: AFD30972.1   74.97   90.68   98.86   99.62
10: ACD75819.1   74.85   90.68   98.57   99.62
11: AKH87423.1   77.57   89.83   98.10   99.62
12: AJD79008.1   80.00   90.68   97.85   99.61
13: AJA40023.1   78.11   90.71   97.64   99.60
14: AHL43498.1   76.69   93.64   94.07   94.92
15: AAV34151.1   74.91   93.64   97.08   96.20
16: AHL43469.1   77.97   94.92   94.49   94.35
17: AHF49783.1   75.20   93.62   97.60   96.18
18: AEN75265.1   75.04   94.07   97.16   96.58
19: AEN75266.1   75.12   94.49   97.40   96.58
```

Fig. 1. Some elements of the identical percentage matrix.

It can be seen that the sequences has high identical value. It shows that the sequence similarity is also very high. And for the phylogenetic tree by MUSCLE shown in Fig. 2.

Based on Fig. 2, it is known that the Zika virus separated into two clusters. For more details, this information can be seen the following Table VI.

TABLE V
THE ACCESS CODE OF THE PROTEIN USED AS THE DATA.

No	Initial Code	Protein Access Code	Seq. lenght	Country	Date of Sample Collection
1	KM 078936	AJD 79008	976	Easter Island Chili	1 Maret 2014
2	KJ 873160	AJA 40023	893	New Caledonia	3 April 2014
3	KJ 776791	AHZ 13508	10.807	French Polinesia	28 Nov 2013
4	KM 851039	AKH 87423	789	Thailand	19 Juli 2014
5	KF 993678	AHL 37808	10.141	Canada	19 Feb 2013
6	AMK 49492	AMK 49492	383	Indonesia (Jambi)	30 Des 2014
7	JN 860885	AFD 30972	10.269	Cambodia	2010
8	EU 545988	ACD 75819	10.272	Yap Micronesia	Juni 2010
9	KM 851038	AKH 87424	789	Philippines	9 Mei 2012
10	HQ 234499	AEN 75264	10.269	Malaysia	1966
11	MR766 /ABY86749	ABY 86749	255	EI	2015
12	AY63253/ AAV34151.1	AAV 34151	10.794	Uganda	1947
13	KF 268948	AHF 43978	10.788	Central African Republic	1976
14	KF 383091	AHL 43476	708	Senegal	2001
15	HQ 234500	AEN 75265	10.251	Nigeria	1968
16	KF 383084	AHL 43469	708	Senegal	1991
17	HQ 234501	AEN 75266	10.269	Senegal	1984
18	KF383113	AHL 43498	708	Cote de Ivoire	1980
19	DQ 859064	ABI 54480	10.290	South Africa	-

TABLE VI
CLUSTER I ZIKA VIRUS DEPLOYMENT.

No.	Access Code	Region
1	ABI54480	South Africa
2	AJD79008	Easter Island Chili
3	AJA40023	New Caledonia
4	AKH87424	Philipina
5	ACD75819	Yap Micronesia
6	AKH87423	Thailand
7	AFD30972	Cambodia
8	AMK49492	Indonesia (Jambi)

Since most of the spread of the virus in first cluster is Asia Region, it is called the "Asian Cluster". The phylogenetic trees show of virus spread to Indonesia which first came from South Africa, the islands of Chile, Caledonia, the Philippines, Yap Micronesia, Thailand, Cambodia, and finally reached Indonesia. As for the second cluster, the distribution area can be see the following Table VII.

From second cluster in Table VII, most are in the African region, so the second cluster is called "Cluster Africa".

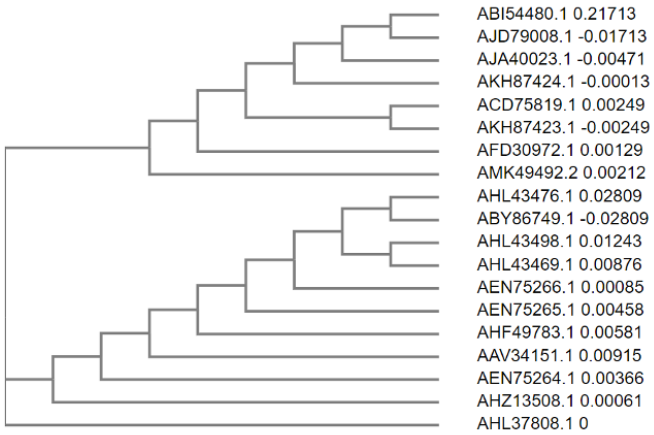

Fig. 2. Phylogenetic tree using MUSCLE.

TABLE VII
CLUSTER II ZIKA VIRUS DEPLOYMENT.

No.	Access Code	Region
1	AHL43476	Senegal
2	ABY86749	El-Salvador
3	AHL43498	Cote de Ivoire
4	AHL43469	Senegal
5	AEN75266	Senegal
6	AEN75265	Nigeria
7	AHF49783	Central African Republic
8	AAV34151	Uganda
9	AEN75264	Malaysia
10	AHZ13508	French Polinesia

IV. CONCLUSIONS

From the whole process, Smith Waterman algorithm can be used to align dengue virus type 1, type 2, type 3, and type 4 with zika virus (jambi), showed mutations between dengue virus and zika virus and the phylogenetic tree the spread of the zika virus to Indonesia.

This sequence alignment study needs to be further developed to assist practitioners in biology, medicine and pharmacy in conducting wet experiments. These simulations may be considered for input to related experiments.

REFERENCES

[1] N. Cristianini and M. Hahn, *Introduction to computational genomics: a case studies approach*. Cambridge University Press, 2006.
[2] S. Shen, *Theory and Mathematical methods in Bioinformatics*. Springer Science & Business Media, 2008.
[3] "Genetic mutation," http://www.nature.com/scitable/topicpage/genetic-mutation-1127, accessed: 2016-05-10.
[4] "National center for biotechnology information (ncbi)," http:/www.ncbi.nlm.nih.gov, accessed: 2016-02-19.