Protein Sequence Analysis of The Zika Virus and The Dengue Virus Using Smith Waterman Algorithm

Mohammad Syaiful Pradana[1,a)] and Siti Amiroch[1,b)]

[1]*Departemen of Mathematics Faculty of Mathematics and Natural Science Universitas Islam Darul Ulum Lamongan, Indonesia.*

[a)]Corresponding author: syaifulp@unisda.ac.id
[b)]Siti.Amiroch@unisda.ac.id

Abstract. Zika virus that became a hot issue in the media recent years behind is very interesting to study, especially the Zika virus has similar to infection to Dengue fever. The virus that is infect to humans through mosquito bites by Aedes genus, Especially Aedes aegypti is identical mosquito that spread Dengue, yellow fever and chikungunya. In Brazil, an increase syndrome Zika virus infection in the community have observed by local health authorities, and also in northeast Brazil an increase of microcephaly condition in babies born. In addition, the Zika virus sporadic infection has reported by approximately 13 Americas countries that show very rapid expansion, and also Indonesia, the euphoria discussed increasingly after the positive patient infected by Zika virus has discovered in Jambi on January 2016. Moving on from this, the authors wanted to know how the sequences protein Zika virus when compared with the Dengue virus, the percentage of identical sequence and also the calculation of local alignment its using the Smith Waterman algorithm. In addition it will also be known genetic mutations that occur in Zika virus from its origin until Zika virus into Indonesia and phylogenetic tree spread of the virus to get to Jambi. So the researcher construct sequence alignment tools by using graphical user interface in MATLAB program based on Smith Waterman algorithm, and also a link in the system with the browser as an option for online data retrieval. From the result of sequence alignment by MATLAB and BLAST, that the identical values of sequence using MATLAB is higher than BLAST values. Likewise, the duration of computation that MATLAB computation is more effective than BLAST computation.

INTRODUCTION

The Zika virus (ZIKV) was identified at 1947 in monkeys rhesus and in 1952 human identified in Tanzania Republic and Uganda. Zika virus outbreaks are also detected in Africa, America, and Asia. The virus that is infect to humans through mosquito bites by *Aedes* genus, especially *Aedes aegypti* in the tropics is the identical mosquito that spread Dengue, yellow fever and chikungunya. The Zika virus indication are similar to Dengue, there are skin rash, fever, conjunctivitis, muscle and joint pain. Recently in Brazil, an increase of Zika virus syndrome in society have observed by local health authorities, and also an increase microcephaly condition in babies born (enlarged head) in northeastern Brazil [1].

The Zika virus outbreak was reported from Pacific at Yap 2007 and Polynesia 2013, and by 2015 from the Africa (Cabo Verde) and Americas (Colombia and Brazil). In addition, approximately 13 Americas countries have detected Zika virus sporadic infection in rapid expansion. And surprisingly this virus reached to Indonesia (Jambi) which has been reported on January 2016 ago.

There are two clusters of the spread of Zika virus, that are Asian and African clusters. Asian cluster from South Africa, Easter Island Chili, New Caledonia, Philipina, Yap Micronesia, Thailand, Cambodia, and arrived in Indonesia. While African cluster from Senegal, El Salvador, Nigeria, Central African Republic, Uganda, Malaysia and arrived in French Polinesia [2].

Based on the background of Zika virus euphoria and similarity of symptoms with Dengue virus infection. It is very interesting topic to study by research to find out Zika virus protein sequence compared with Dengue virus, identical percentage, local alignment calculation and genetic mutation using Smith Waterman algorithm. The one of local alignment algorithm is Smith Waterman algorithm. It looks simple for the development of dynamic program-based algorithms with appropriate local alignment, this algorithm is instrumental in bioinformatics [3].

Proceedings of the Symposium on BioMathematics (SYMOMATH) 2018
AIP Conf. Proc. 2084, 020011-1–020011-5; https://doi.org/10.1063/1.5094275
Published by AIP Publishing. 978-0-7354-1814-1/$30.00

Sequence Alignment (SA) is a procedure to aligning DNA or proteins sequences to find resemblance among the sequences or to prove that the two sequences are compared from the same sequence. Sequence alignment has two methods, there are global and local alignment. In global alignment, it is performed for the whole sequence using as many nucleotide as possible in the DNA. Menwhile, local alignment is parallel to some of the sequences usually parts that have a high enough level of similarity.

As bioinformatics develops as a science that applies computational techniques to analyzing biological information, bioinformatics also includes the application of mathematical, statistical and informatical methods to solve biological problems, primarily by using biological information in DNA or protein sequences. Furthermore, biological information is analyzed to determine the similarities between sequences. The similarities sequences of Zika virus and Dengue virus indicates that both are from the same genus that has been mutated. However, the low sequence similarity will prove that both are not of the same genus even when viewed from the symptoms look almost identical.

DNA and Protein

Deoxyribo Nucleic Acid (DNA) is a polymers composed by nucleotides as the basic molecules that carry the properties of genes. DNA is composed by four types of nucleotides that are covalently bonded. The types of nucleotides are represented by the character A (*adenine*), C (*cytosine*), G (*guanine*), and T (*thymine*) [4].

Proteins is composed by simple molecular chains called amino acids. The final shape of a protein is depend on the proper identity, the chain sequence of amino acid and the atomic interaction between the cell medium (mostly water) and amino acids. Protein formation use a combination of 20 amino acids, there are A (*Alanine*), R (*Arginine*), N (*Asparagine*), D (*Aspartic Acid*), C (*Cyteine*), Q (*Glutamine*), E (*Glutamic Acid*), G (*Glycine*), H (*Histidine*), I (*Isoleucine*), L (*Leucine*), K (*Lysine*), M (*Methionine*), F (*Phenylalanine*), P (*Proline*), S (*Serine*), T (*Threonine*), W (*Tryptophan*), Y (*Tyrosine*), and V (*Valine*).[5].

The symbol is an unique sequence depend by gene encoding consisting of three set of nucleotides are called codon such as: GAG and GAA representing E, AAG and AAA representing K, AGC and AGA representing R.

Sequence Alignment (SA)

DNA/RNA and protein sequence are commonly determined based on biological sequence. According to Shen [4], biological sequence described using the following notation:

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

Which X, Y, Z denote sequence, and x_i, y_i, z_i are basic units of sequence at i-th 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 direcly represent 20 amino acid molecules.

Sequence alignment is analysis of position and mutation by comparing biological sequences correctly, aligning between the two sequence and determining the movement of mutation is the main idea of sequence alignment.

Smith Waterman Algorithm

Smith Watermans Algorithm applied to this paper which used a local alignment algorithm. Two important aspect of Smith Watermans Algorithm are:

- Used two-dimensional table to calculate
  Smith Watermans algorithm adds zero value when computing $s(i, j)$ so that a negative value will never happen to this algorithm.

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

- Traceback Algorithm
  Starting and ending points of the traceback method on the Smith Watermans algorithm selected with maximum value. The last point is the first zero value element on the backtrace process. A starting point with maximum value will assuring maximum value on local alignment sequences and the last point is the first zero value element ensuring that the section is not exceeded.

Genetic Mutation

Mutations is a change of genetic sequence, which is caused by the differences among organisms. This changes occur on many levels, with very different consequences [6]. Mutation in DNA sequences are classified into four types [3]:

1. Type-I : there is a nucleotide change, for example from ”a” to ”g”
2. Type-II : there is a nucleotide section that change the order of its position, for example the ”accgu” section changed the sequence to ”guacc”.
3. Type-III : there is an insertion new section nucleotide in the sequence, for example the insertion of nucleotide ”aa” in the middle of ”gguugg” section will turning to ”gguaaugg”
4. Type-IV : there is a nucleotide section elimination in the sequence, for example removing the ”ag” nucleotide from the ”acaguua” section to ”acuua”.

In the type-I and type-II, the position of all nucleotides does not change, so the mutations are called substitution mutation. While the nucleotide position movement in type III and V mutation, it is called a transfer mutation.

RESULT AND DISCUSSION

The data sequences of virus were taken online in genbank, which the world’s largest gene database belonging to the United States. The retrieve the data, we accessed the National Center for Biotechnology Information (NCBI) website [7]. The virus data are stored FASTA code in (.txt) file and the used its code to aligning sequence and find phylogenetic tree.

In this research, we used 19 sample data of Zika virus and 4 sample protein sequence data of Dengue virus type-I, type-II, type-III, and type-IV. Furthermore, a virus sequence was taken from 19 sample protein sequence (Zika virus Jambi, Indonesia) and aligned each of Dengue virus sequence, which is taken from Makasar, Indonesia. Furthermore, analizing the dissimilarity, similarity, duration process, percentages, and mutation sequence. The analysis process by comparing each Dengue virus to Zika virus. We provide link to BLAST on MATLAB and system browser. Table 1 and 2 respectively, show the protein sequence data were taken online in genbank.

TABLE 1. 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

TABLE 2. Protein sequence data of infected Zika virus patient

No	Access Code	Sequence Length	Country	Date of Sample Collection	Explanation
1	KM078936	976	Easter Island Chili	1 Maret 2014	Partial cds
2	KJ873160	893	New Caledonia	3 April 2014	Partial cds
3	KJ776791	10.807	French Polinesia	28 Nov 2013	Complete genom
4	KM851039	789	Thailand	19 Juli 2014	Partial cds
5	KF993678	10.141	Canada	19 Feb 2013	Partial cds
6	AMK 49492	383	Indonesia (Jambi)	30 Des 2014	Partial cds
7	JN860885	10.269	Cambodia	2010	Partial cds
8	EU545988	10.272	Yap Micronesia	Juni 2010	Complete cds
9	KM851038	789	Philippines	9 Mei 2012	Partial cds
10	HQ234499	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	KF268948	10.788	Central African Republic	1976	Complete cds; Host: aedes Africanus
14	KF383091	708	Senegal	2001	Partial cds
15	HQ234500	10.251	Nigeria	1968	Partial cds
16	KF383084	708	Senegal	1991	Partial cds
17	HQ234501	10.269	Senegal	1984	Partial cds
18	KF383113	708	Cote de Ivoire	1980	Partial cds Complete cds;
19	DQ859064	10.290	South Africa	-	Spondweni virus

TABLE 3. MATLAB result based Smith Waterman algorithm

Sequence			Similarity/Dissimilarity		Percentace (%)		Duration (s)
DEN-V	Type	ZIK-V	Similarity	Dissimilarity	Similarity	Dissimilarity	
AHG06327	1	AMK49492	273	108	71.4660	28.27	0.062
AHG06364	2	AMK49492	272	108	71.0183	28.20	0.094
AHG06376	3	AMK49492	274	107	71.5405	27.94	0.359
AHG06382	4	AMK49492	271	110	70.7572	28.72	0.156

Furthermore, the sequence of Zika virus from Jambi and sequence of Dengue virus proteins in the table 1 and 2 are stored in (.txt) file and then inputted for sequence simultaneous alignment process in MATLAB software.

Using the Smith Waterman algorithm, the alignment between Zika virus and Dengue virus of each type are shown in table 3.

The output of four alignments comparison using MATLAB and Basic Local Alignment Search Tool (BLAST) are shown in table 4.

From table 4, it can be concluded that the identical value of Dengue and Zika virus alignment output is more thorough by using MATLAB than the BLAST. Its proved by four numbers decimal accuracy in MATLAB while the BLAST shows only two numbers decimal and also the duration of computation time on MATLAB simulation is shorter than the computational time on BLAST.

TABLE 4. MATLAB and BLAST comparison

Sequence			Sequence length		Identical value		Duration (s)	
DEN-V	Type	ZIK-V	DEN-V	ZIKA	MATLAB	BLAST	MATLAB	BLAST
AHG06327	1	AMK49492	3392	383	71,466 %	71 %	0,062	12,16
AHG06364	2	AMK49492	3391	383	71,0183%	71 %	0,094	11,27
AHG06376	3	AMK49492	3390	383	71,5405%	71 %	0,359	6,58
AHG06382	4	AMK49492	3387	383	70,7572%	71 %	0,156	10,68

CONCLUSION

From the whole process, the result of alignment of both viruses by using Smith Waterman algorithm simulated in MATLAB more effectively in the accuracy and duration of computation time when compared with BLAST. The mutations percentage (dissimilarity) between Zika virus and Dengue virus approximately 28% and the similarity approximately 71%. Based on the simulation results, the mutations of both viruses classified in type I.

ACKNOWLEDGMENTS

This paper is an output of the research of beginner lecturers in 2017 funded by the Directorate of Research and Community Service, Directorate General of Strengthening Research and Development at the Ministry of Research, Technology and Higher Education Number: 120/SP2H/LT/DRPM/IV/2017, on April 3rd 2017.

REFERENCES

[1] "Zika Virus."[Online]. Available: http://www.who.int/mediacentre/factsheets/Zika/en/. [Accessed: 23-Feb-2016].

[2] M. S. Pradana and S. Amiroch, "Zika Virus Mutation and The Spreading to Indonesia,"Int. J. Comput. Sci. Appl. Math., vol. 4, no. 1, pp. 15–18, 2018.

[3] S. Shen, Theory and Mathematical methods in Bioinformatics. Springer Science & Business Media, 2008.

[4] "DNA."[Online]. Available: https://chem.libretexts.org/Textbook_Maps/Organic_Chemistry/Book%3A_Organic_Chemistry_with_a_Biological_Emphasis_(Soderberg)/Chapter_01%3A_Chapter_1%3A_Introduction_to_organic_structure_and_bonding_I/1.3%3A_Structures_of_some_important_biomolecules/Introduction_to_nucleic_acid_(DNA_and_RNA)_structure. [Accessed: 23-Feb-2016].

[5] N. Cristianini and M. W. Hahn, *Introduction to computational genomics: a case studies approach*. Cambridge University Press, 2006.

[6] "Genetic Mutation."[Online]. Available: http://www.nature.com/scitable/topicpage/genetic-mutation-1127. [Accessed: 10-May-2016].

[7] "National Center for Biotechnology Information (NCBI)."[Online]. Available: http:/www.ncbi.nlm.nih.gov. [Accessed: 19-Feb-2016].