Short segment search method for phylogenetic analysis using nested sliding windows

A A Iskandar¹, A Bustamam¹, and H Trimarsanto²

¹ Department of Mathematics, Universitas Indonesia, Depok, Indonesia
² Bioinformatics Laboratory, Eijkman Institute for Molecular Biology, Indonesia

E-mail: {afifai,alhadi}@sci.ui.ac.id, anto@eijkman.go.id

Abstract. To analyze phylogenetics in Bioinformatics, coding DNA sequences (CDS) segment is needed for maximal accuracy. However, analysis by CDS cost a lot of time and money, so a short representative segment by CDS, which is envelope protein segment or non-structural 3 (NS3) segment is necessary. After sliding window is implemented, a better short segment than envelope protein segment and NS3 is found. This paper will discuss a mathematical method to analyze sequences using nested sliding window to find a short segment which is representative for the whole genome. The result shows that our method can find a short segment which more representative about 6.57% in topological view to CDS segment than an Envelope segment or NS3 segment.

1. Introduction

Dengue fever is a disease caused by dengue virus and transmitted by Aedes aegypti [1]. This infection is characterized by high fever, pain in the joints and red spots on the skin. This disease is common in tropical climates such as Asia, Africa, the Middle East, the Pacific, and Central and South America [1]. Based on the clinical level, dengue disease is categorized into ordinary dengue fever, dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS). DHF and DSS is a disease that can be life-threatening. Dengue virus (DENV) included in the group B.arthropod-borne viruses (Arboviruses) in the flavivirus genus (family flaviviridae) [2]. Dengue virus consists of single-stranded RNA enveloped by the nucleocapsid which diameter is 30 nm, nucleocapsid enveloped by a fatty sheath with thickness is 10 nm so that the overall diameter is 50 nm virion [2].

Research for dengue virus at the level of molecular biology will involve analysis of the gene sequences [1] [2]. Analysis of the gene sequences is used for molecular epidemiology studies, identification of strains, as well as evolution and phylogenetic studies [3]. Analysis using the entire gene sequences of dengue virus genome, particularly from all segments of the coding DNA sequence (CDS), can improve the accuracy and quality of analysis and the research itself. Since costs of the experimental sorting line to get the base sequence of the entire genome are not cheap and need a long time, many researchers only analyzed the regional envelope, assuming that the envelope protein is a protein that is the most varied and must often adapt to the immune system of a patient. However, some recent studies indicate that some of the analysis, especially phylogenetic analysis, which is done by simply using the envelope area can give different results with the overall segment analysis using ORF. This indicates that the envelope area does not
represent the entire genome of dengue virus. Therefore, if there are other areas of the dengue virus genome that can represent the whole genome of the virus to the analysis of the gene sequences, especially for phylogenetic analysis, it will increase the accuracy and quality of research at the molecular level without having to sacrifice time and lower costs.

To reduce the cost and time to perform the phylogenetic analysis, we need a short segment of the genome that can be used as representative of the whole genome sequences. The purpose of this research is to propose a method of finding a short segment of DNA sequences for phylogenetic analysis, the study of evolution and classification and identification of virus strains.

2. Material and Methods

2.1. Data Acquisition

This research uses DNA sequence data from virus DENV 4 which can be downloaded at http://www.ncbi.nlm.nih.gov. This study used 152 samples virus DENV 4 with the following criteria:

- Sequences only contain characters 'A','G','T' and 'C'
- One sample of DNA contains $8000 \leq x \leq 10000$ bp
- Data was aligned using multiple sequences alignment

2.2. Evolutionary Model

The evolutionary model is a mathematical model to describe the nucleotide mutation that took place in the gene sequences[4]. Nucleotide Mutation is defined as the process of changing a nucleotide to another nucleotide[6]. This substitution is divided into two, namely the transition and transversion[4]. The transition is a substitution that took place among purine or pyrimidine, whereas transversion was the substitution of one purine with a pyrimidine or otherwise[6]. Each nucleotide position assumed to be independent when mutations in the gene sequences occur. Therefore, mutations in the sequence can be described by a mutation which took place on only a single site.

The objective of evolutionary models is to determine the evolutionary distance between rows, the shorter the distance of its evolution, the closer kinship of the species (or sequences) are concerned. One of the simplest evolutionary models is genetic proportional distance. model. This method is said to be simple because to calculate the distance between two sequences counted for the number of base pairs (bp) are different which are further divided by the length of the sequence is assuming the sequence is already aligned. To write the distances between pairs of sequences in the database, used a matrix which entry of the matrix represents the distance between each pair of species.

To build a genetic proportional distance, we use the following algorithm:

2.3. Neis Minimum Genetic Distance

Nei’s minimum genetic distance is a one of popular model for calculate the genetic distance [5] [6]. In this study, the method used to calculate how informative segments are taken for performing phylogenetic analysis.

Consider $X$ and $Y$ are populations, we can compute their genetic distance as :

$$D_{mn} = \frac{J_x + J_y}{2} - J_{xy}$$

In this research, we use this model to compute a distance between a CDS and a short segment.
Algorithm 1: Build a Distance Matrix

1: \textbf{procedure} P\_DISTANCE(mseq)
2: \hspace{1em} matrix \leftarrow \text{zero\_matrix(length(mseq))}
3: \hspace{1em} i \leftarrow 1
4: \hspace{1em} \textbf{while} i \leq \text{length(mseq)} - 1 \textbf{do}
5: \hspace{2em} j \leftarrow i + 1
6: \hspace{1em} \textbf{while} j \leq \text{length(mseq)} \textbf{do}
7: \hspace{2em} distance \leftarrow 0
8: \hspace{2em} k \leftarrow 1
9: \hspace{2em} \textbf{while} k \leq \text{length(mseq[0])} \textbf{do}
10: \hspace{3em} \text{if} mseq[i] \neq mseq[j] \text{ then}
11: \hspace{4em} distance \leftarrow distance + 1
12: \hspace{3em} \text{end if}
13: \hspace{2em} k \leftarrow k + 1
14: \hspace{1em} \textbf{end while}
15: \hspace{1em} distance \leftarrow \frac{distance}{mseq[0]}
16: \hspace{1em} matrix[i][j] \leftarrow distance
17: \hspace{1em} matrix[j][i] \leftarrow distance
18: \hspace{1em} j \leftarrow j + 1
19: \hspace{1em} \textbf{end while}
20: \hspace{1em} i \leftarrow i + 1
21: \hspace{1em} \textbf{end while}
22: \hspace{1em} \textbf{return} matrix
23: \textbf{end procedure}

2.4. General Time Reversible Evolutionary Model

The general time reversible (GTR) is the evolutionary model that used for analysis of DENV sequences. Suppose \( Q \) is a matrix form of GTR model, so we can write:

\[
Q = \begin{bmatrix}
-\mu x & a \mu \pi_C & b \mu \pi_G & c \mu \pi_T \\
 g \mu \pi_A & -\mu y & d \mu \pi_G & e \mu \pi_T \\
 h \mu \pi_A & i \mu \pi_C & -\mu z & f \mu \pi_T \\
 j \mu \pi_A & k \mu \pi_C & l \mu \pi_G & -\mu w
\end{bmatrix}
\]

Where:
- \( x = a \mu \pi_C + b \mu \pi_G + c \mu \pi_T \)
- \( y = g \mu \pi_A + d \mu \pi_G + e \mu \pi_T \)
- \( z = h \mu \pi_A + i \mu \pi_C + f \mu \pi_T \)
- \( w = j \mu \pi_A + k \mu \pi_C + l \mu \pi_G \)

GTR model is a complex models because this model is a generalization of the evolutionary model based on Markov Chain [4].

2.5. Parsimony Informative Site

Parsimony method of tree construction work by finding all the possible topology of the tree and construct the line of ancestors that requires a minimum number of evolutionary ranks obtained today [7].

To save computing time, only a small number of sites with the highest phylogenetic information that is used to construct the tree with this method. This site is called parsimony informative sites [6].
To find a parsimony informative site, we built an algorithm as follows:

**Algorithm 2** Parsimony Informative Site

1. **procedure** PIS(multisequence)
2. \( PIS\_location \leftarrow \emptyset \)
3. \( j \leftarrow 1 \)
4. **while** \( j \leq \text{length(multisequence[0])} \) **do**
5. \( \text{if contains at least two types of nucleotides then} \)
6. \( PIS\_location \leftarrow \{j\} \)
7. **end if**
8. \( j \leftarrow j + 1 \)
9. **end while**
10. **return** \( PIS\_location \)
11. **end procedure**

2.6. **Sliding Window Analysis**

The segment of DNA sequence is depends on the position, we can not take a random segment of a sequence to find a short representative segment. So, we need a sliding window analysis to find a segment without changing a position of that segment on the whole sequence. Sliding window analysis is a method to analyze the data by analyzing a subset of the data using the technique of 'shift' until all subsets of the data are analyzed.

2.7. **Short Segment Search using Nested Sliding Window**

Unlike the sliding window analysis that we discussed earlier, before sliding window which is more focused on the formation of scatter map that will be analyzed, nested sliding window is a sliding window that is focused on short segments of the CDS. However, to determine that the segment is candidates for the segment, nested sliding window is very dependent on the sliding window analysis that previously discussed. Illustration of nested sliding window can be seen from the following figure:

![Figure 1: Nested sliding window illustration](attachment:image.png)

2.8. **Tree Comparison using Disagree Method**

The disagree method is a method for comparing two trees using the novel algorithm and return taxa that make the disagreement between those trees [8]. To perform this process, we use TopD software. If this algorithm returns small number of taxa branch, we can say that those trees are similar.
3. Experiment

This study uses the following framework:

![Short segment search method flow](image)

In this research, we found 30 segments which is more representative than envelope segment and NS3 segment. The results for each step explained as follow:

3.1. Sliding Window Plot

A sliding window computation for CDS is used as a population which needed for computing a distance between a short segment and CDS. A plot of CDS can be seen in figure below.

![Sliding window result on CDS](image)

Fig.3 show a variance point of parsimony informative site and proportional genetic distance from CDS. We can see that many points accumulates in one coordinate. The aim of this analysis method is to minimize accumulate points without change phylogenetic information, so we need a short segment that represents that points. We did the same thing for an envelope segment and NS3 segment, so we have other pairs of parsimony informative sites and proportional genetic distance points.

TABLE 1 shows the results of minimum Nei’s genetic distance using some weighting function.

After the computation using sliding window, we choose two function $f(x) = \log(5x)$ and $f(x) = \log(x)$ for short segment filtration using nested sliding window.

3.2. Finding short segment

At this stage, we found a representative segment using nested sliding window with a large window size is 1000 bp with three codons shifts step (9bp), and a small window size is 100 bp with 1bp shift step. A length of aligned sequences of CDS are about 10164 bp, with sliding window analysis that we defined before, we can find about 1019 short segments. However, from our methods, it reduced to 200 segments.
Table 1: Sliding window computation result

| Functions | Protein | Mean       | Standard Deviation |
|-----------|---------|------------|-------------------|
| Without Weighted Function | Envelope 0.172274829359 | 0.117744229463 |
| | NS3 0.188885006042 | 0.133148237803 |
| $f(x) = \log(x)$ | Envelope 0.172274829359 | 0.117744229463 |
| | NS3 0.188885006042 | 0.133148237803 |
| $f(x) = \log(5x)$ | Envelope -0.145138999776 | 0.106608939153 |
| | NS3 -0.1470009567 | 0.102611490732 |
| $f(x) = \log(5y)$ | Envelope 0.802650015027 | 0.0556421676208 |
| | NS3 0.0891129912456 | 0.0645141471741 |

Table 2: Nested sliding window result

| Functions | Total Segments |
|-----------|----------------|
| log x     | 44 segments    |
| log 5x    | 198 segments   |

Table 3: Tree distance based on disagreement

| Segment       | Tree Distance |
|---------------|---------------|
| Envelope      | 0.361842105   |
| NS3           | 0.361842105   |
| Best Segment 1| 0.296052632   |
| Best Segment 2| 0.296052632   |
| Best Segment 3| 0.309210526   |

Next, we constructed a phylogenetic tree from all segments that we have got from sliding window analysis with PhyML software, and then we compare those trees with CDS tree. From TABLE III, we can see that our method can find a better tree distance with disagreement method, there are 30 segments which have a smaller disagreement value than NS3 or envelope. Tree that have a small disagreement is considered as a tree that has a close resemblance to the CDS, that are considered to represent the entire genome.

4. Conclusion and Future Works
Phylogenetic analysis using CDS (Coding DNA Sequence) or a whole genome are not cheap and take a long time for the computational side. In this research, we proposed a method for searching a short segment for phylogenetic analysis which better than the envelope segment and the NS3 segment on Dengue Virus sample. So, the result from this methods can be used for further phylogenetic analysis more efficient than using a whole genome. To make computational process faster we plan to apply parallel computing for sliding window process [9].

Acknowledgement
This research was supported by research grant HIBAH PUPT 2016 and Bioinformatics laboratory at Eijkman Institute for Molecular Biology

References
[1] Lozano-Fuentes S, Hayden M H, Welsh-Rodriguez C et al. 2012 The Dengue Virus Mosquito Vector Aedes aegypti at High Elevation in Mexico. In: *the American Journal of Tropical Medicine and Hygiene* 7(5) pp
902–909 http://doi.org/10.4269/ajtmh.2012.12-0244

[2] Zhang H et al. 2014 Spatiotemporal characterization of dengue virus in mainland China: insight into the whole genome from 1978 to 2011. In: PLoS ONE 9 pp 1–10

[3] Hochreiter S 2008 Bioinformatics I: Sequence Analysis and Phylogenetics http://www.bioinf.jku.at[On-line]

[4] Isaev A 2004 Introduction to mathematical method in bioinformatics Berlin: Springer

[5] Nei M 1987 Molecular genetics Columbia University New York Press

[6] Lemey P, Salemi M, Vandamme A 2010 The phylogenetic handbook New York: Cambridge

[7] Xiong J 2006 Essential bioinformatics New York: Cambridge

[8] Puigbo et al. 2007 TOPD/FMTS: a new software to compare phylogenetic trees human diseases: principles and evolution approaches In: Oxford Journal 23(12) pp 1556–1558

[9] Bustamam A, Ardaneswari G, Lestari D 2014 Performance Evaluation of Fast Smith-Waterman Algorithm for Sequence Database Searches using CUDA GPU-Based Parallel Computing In: Journal of Next Generation Information Technology (JNIT) 5(2) pp 38–46