Abstract. This research compares DNA sequences of H5N1 virus to analyze them with a tree diagram method. We use tree diagram method to analyze similarity level of nucleotides. A tree diagram method is one of several methods to align a pair of DNA sequences at entirely length. This method uses concept data structure in general tree with post-order traversal. Furthermore, we obtain mutation level of nucleotides. The scoring equation and parameters are determined.

Keywords and Phrase: Global Alignment, Pair of DNA Sequences, Tree Diagram Method.

1. INTRODUCTION

The comparison of the existing sequences is a modern method for studying the evolutionary interaction between genes. It is based on the alignment-the process of arranging two or more sequences to achieve the maximum level of identity (for evaluation purposes), the degree of similarity and eventual homology. Sequence alignment is an important method in DNA and protein analysis [6,3]. Increasing of new biological sequences is basic of any sequence analysis [4]. Bioinformatics is a collection of mathematical, statistical and computational methods for analyzing biological sequences, that is, DNA, RNA and amino acid (protein) sequences.

We will compare the DNA sequences of H5N1 to analyze their similarity level. H5N1 is Influenza-A virus that has a segmented single strain negative RNA linear genome. Inside the host cell, virus’ RNA will be a reverse transcription into RNA-DNA hybrid and eventually forms the DNA. Furthermore, the virus’ DNA will enter into the nucleus’ host cell. DNA virus will damage the DNA host’s and to forms mRNA (messenger RNA). mRNA will translate to produce a viral envelope protein to form new viruses. The virus has very high mutation rate, so it can create different variations of the viruses [5].

In the previous research, the similarity level in segment HA and NA protein was determined by [1] using tools EMBOSS, this tools applied “needle” algorithm. This research use tree diagram method to align a pair of DNA sequences [7]. This method consists of three parts: (i) simple alignment algorithm, (ii) extension algorithm, (iii) Graphical Simple Alignment Tree (GSA tree). These theories will be explained as follows.

2011 Mathematics Subject Classification: Computer Science or Biology and Other Natural Sciences
1.1 Sequences Alignment of DNA

In this section we give an introduction of DNA sequences and Sequence Alignment that used in the next discussion.

1.1.1 DNA Sequences

DNA (deoxyribonucleic acid) sequences are associated with the four-letter DNA alphabet \{A, C, G, T\}, where A, C, G and T stand for the nucleic acids or nucleotides Adenine, Cytosine, Guanine and Thymine respectively. Most DNA sequences currently is being studied come from DNA molecules that found in chromosomes. They that are located in the nuclei of the cells of living organisms. More information about DNA can be found in [4].

DNA sequences are string of letters from a four-letter alphabet called nucleotides (A, C, G, T). The length of sequence is a variable and not all sequences are of the same length. Generally, we use the following description of DNA sequence:

\[ A = (a_1 \ldots a_m) \quad B = (b_1 \ldots b_n) \]  

Where the capital letters A, B represent the sequence and \(a_i\), \(b_i\) represent the basic units of the sequence at position i, whose elements are obtained from the set \{A, C, G, T\}. For instance, DNA sequence can have substitutions (change of one nucleotide for another), insertions and deletions (gain or loss of one or more nucleotides) and therefore algorithms should include the possibility of gaps [9]. There are some biology assumptions about beginning and the end of a sequence that are useful for algorithm development [2].

1.1.2 Sequence Alignment

An alignment between two sequences is simply a pairwise match between the characters of each sequence. A true alignment of DNA sequences is one that reflects the evolutionary relationship between two or more homologs (sequences that share a common ancestor).

Given two sequences in first equation with lengths is \(m, n\) respectively, let \(c\) be the total length of the alignment we have, i.e.,

\[ \max(n, m) \leq c \leq n + m \]

The alignment is represented by a \(2 \times c\) matrix, \(X(A,B)\), where 2 is row and \(c\) is column. The row are the sequences and the columns are matches, mismatch, insertion and deletion (called “indels”) [2].

Example: Given two sequences, \(X = GAATTAGTTA\) and \(Y = GGATCGA\). Where length is \(m = 10, n = 7\) respectively. One possible arrangement can be defined as:

\[ A(X,Y) = \begin{array}{cccccc}
G & A & A & T & T & A \\
G & G & A & T & C & G & A
\end{array} \]

Since there are different ways of arranging the sequences and find mismatch and indels, so we are interested in the best possible arrangement. Furthermore, the “best” alignment depends on how the scoring of matches, mismatches and gaps. Since we are interested in the similarity of two sequences, we would reward a match and penalize a mismatch/gap. Thus, the first step is to define an appropriate scoring equation in order to quantify the sequence alignments.

A scoring equation can be designed to quantify the edit distance (mutations, insertions and deletions):

Alignment score: \[ S = \alpha p - \beta q - r(o + ke), \]

Where \(\alpha\) is score of each match, \(\beta\) is score of each mismatch and \((o + ke)\) is
score of each gaps. The parameters of p, q and r denote the total amount of matches, the total amount of mismatches and the total amount of gaps, respectively [7].

For nucleotide sequences, Sequence similarity and sequence identity are synonymous [10]. So, similarity or identity level for nucleotide sequence can be counted by using:

$$S = \left[ \frac{L_s \times 2}{L_a + L_b} \right] \times 100$$

(3)

Where, S is the percentage sequence similarity, $L_s$ is the number of aligned residues with similar characteristics, $L_a$ and $L_b$ are the total lengths of each individual sequence in alignment.

1.2 Tree Diagram Methods

In this section we discuss about Tree Diagram Methods and their theory related to the problem discussed in this research. Tree diagram method is one method to align a pair of DNA sequence in entire length (global alignment). This method consists three parts: (1) improved simple alignment algorithm, (2) extension algorithm, and (3) GSA Tree [7].

1.2.1 Improved Simple Alignment Algorithm

Given two sequences $X=\{x_1,x_2,\ldots,x_m\}$ and $Y=\{y_1,y_2,\ldots,y_n\}$, where $m$ and $n$ is denote the length of $X$ and $Y$, respectively. So improved simple alignment algorithms can be defined as step for align $X$ and $Y$ sequence, with initial position is first base ($x_1$) of $X$ overlaps with the first base ($y_1$) of $Y$. Then the sliding process is done along the left and right, respectively. These steps shown by [7] as follows:

(1) Initial position

\[
\begin{array}{ccccccc}
  x_1 & x_2 & x_3 & \ldots & x_m \\
  y_1 & y_2 & y_3 & \ldots & y_m & \ldots & y_n
\end{array}
\]

(2) Every times X moves one base position along the right direction

\[
\begin{array}{ccccccc}
  x_1 & x_2 & x_3 & \ldots & x_m \\
  y_1 & y_2 & y_3 & \ldots & y_m & \ldots & y_n
\end{array}
\]

(3) Every times X moves one base position along the left direction

\[
\begin{array}{ccccccc}
  x_1 & x_2 & x_3 & \ldots & x_m \\
  y_1 & y_2 & y_3 & \ldots & y_m & \ldots & y_n
\end{array}
\]

Every alignment has score $S$ according to the scoring equation on equation 2nd. Then at the step (2) and (3) still define another score $S'_i$ and $S'_j$, respectively is $S'_i = \alpha p + \alpha q - r(o + ke)$ and $S'_j = \alpha p + \beta q - r(o + ke)$ as stopping condition in step (2) and (3). Here, we choose typical values of parameters, such as $\alpha = 5, \beta = 4, o = 10$ and $e = 0.5$. We choose this
parameters from element of DNAFULL matrix (i.e value of $\alpha, \beta$ ) and from one standard combination in tools EMBOSS (i.e value of o and e). For the more complete explanation of this algorithm can be found in [7]. From this algorithm we choose best alignment (R) denote by maximum score of alignment and longest common substring in each step.

\subsection*{1.2.2 Extension Algorithm}

This algorithm used to protect the longer common substring than C in R from being split by C.

Let C is a common substring in R, with $C = \emptyset \cup \{C_1, C_2, ..., C_m\}$. if $C = \emptyset$ then none longest common substring in C can be extended. If $C = \{C_1, C_2, ..., C_m\}$ there are m longest common substrings, where $|C_1| = |C_2| = ... = |C_m| = k$ with k denotes the number of matches within a longest common substring. Here are the steps of the extension algorithm for the longest common substring [7]:

1. Let $K$ be length of the longest common substring of X and Y.
2. When $k = K$, none of the longest common substrings $C_i$ of R can be extended into longer common substring.
3. When $k < K$, there are exist at least a longer common substring than $C_i$. there are several sub-steps to find out the longer common substrings as the following:
   a. Let $L_L$ be the number of mismatches from the right end of $C_{i-1}$ to the left end of $C_i$. When $i = 1$, $L_L$ denotes the number of mismatches from the left end of X and Y to the left end of $C_1$. Similarly, let $L_R$ denotes the number of mismatches from the right end f $C_i$ to the left of $C_{i+1}$. When $i = 1$, $L_R$ denotes the number of mismatches from the right end of $C_i$ to the right end of X and Y.
   b. When $K < L_L$, the K mismatches are extracted from the left of $C_i$. Otherwise, the $L_L$ mismatches are extracted from the left of $C_i$. Similarly, when $K < L_R$, the K mismatches are extracted from the right of $C_i$. Otherwise, the $L_R$ mismatches are extracted from the right of $C_i$. Then the sequences extracted from left of $C_i$, $C_i$ and from the right of $C_i$ are connected into two new sub-sequences $S_{i1}$ and $S_{i2}$.
   c. Apply the simple alignment algorithm to $S_{i1}$ and $S_{i2}$. If there exist a new longer common substring within $S_{i1}$ and $S_{i2}$ than $C_i$, we will face a choice: the new longer common substring or $C_i$. If there is an increment of sore when the new longer common substring comes into being, we can replace the original $C_i$ with the new substring also called as $C_i$.
4. As for every $C_i$ of R, the original $C_i$ is replaced by the new longest common substring if the new substring exists.

Output these algorithm is data C and U from R (or repaired R) which has extend if extension is occur.
1.2.3 Graphical Simple Alignment Tree

This algorithm uses to explore the appropriate gaps in $U_j^1$. Any several step for this step [7].

1. Compute the scores of all simple alignment of $U_j^1$ by the simple alignment algorithm. A good simple alignment $R_j^1$ of $U_j^1$ is generated when its score maximum.

2. If there is increment of the score due to appropriate gaps within $U_j^1$, $U_j^1$ can be further divided into the second level sub-alignment. When and how to add the gaps in the sequences? Now, let $C_i^2$ be the longest common substrings of $R_j^1$, where $C_i^2 = \emptyset \cup \{C_{i+1}, C_{i+2}, ..., C_{i+m}\}$. Then there are two sub-steps as the following:

   (a) If $C_i^2 = \emptyset$, there is no the longest common substrings in $R_j^1$. Then $U_j^1$ can not be further broken down. The good simple alignment $R_j^1$ of $U_j^1$ becomes a leaf node in GSA tree. The two sequences of $R_j^1$ might be entirely overlapping, or partially overlapping, or one sequence might be aligned entirely internally to the other. When the two sequences of $R_j^1$ are entirely overlapping, there are no gaps within $R_j^1$. Otherwise, the hanging ends of the overlap come into being the gaps of $R_j^1$. The relative position of these gaps is fixed, and becomes the gaps within the final global alignment.

   (b) If $C_i^2 \neq \emptyset$, there are m longest common substrings.

   Then $U_j^1$ can be further divided into the second level sub-alignment by $C_i^2$. Let $U_j^2$ be the substrings spaced by $C_i^2$, where $U_j^2 = \emptyset \cup \{U_{j+1}^2, U_{j+2}^2, ..., U_{j+n}^2\}$. Then every substring $C_{i+k}^2$ (k=1,2,...,m) of $C_i^2$ becomes a leaf node in GSA tree. There are no gaps within them. As for every substring $U_{j+k}^2$ (k=1,2,...,n) of $U_j^2$, the operation flow goes back to the step (1). And the level of sub-alignment enters the next.

These above steps repeatedly do, until all U the last level sub-alignment cannot be further decomposed by the improved simple alignment algorithm and the extension algorithm. Then we can obtain a Graphical Simple Alignment Tree (GSA Tree) for string X and Y, consisting of a series of substrings.

1.2.4 General Tree Using post-order Traversal

Global alignment in this method is formed by GSA tree. In this case, GSA tree is a general tree or tree structure that has any children. We can construct any general tree in figure 1.
Figure 1. The General tree

General tree in Figure 1 is constructed from any sequence X and Y. In the position of root is consist best alignment R, from improved simple alignment algorithm and extension algorithm. And from R, we can divide two kinds of child in each level, i.e: C and U. C is a longest common substring from R and U is substrings spaced by C. There are consists two types of nodes: inner and leaf nodes. The global alignment of string X and Y is formed by all leaf nodes. To obtain global alignment, GSA tree is traversed by post-order traversal of tree. A post-order traversal of a general tree performs a post-order traversal of the root’s sub-trees from left to right, then visits the root [8]. Then all inner nodes are deleted from the result of post-order traversal [7].

2. GLOBAL ALIGNMENT DNA VIRUS H5N1 AND ITS ANALYSIS

In this section, we discuss about the analysis of sequences DNA virus H5N1 using tree diagram methods. This method is described in figure 2.

Figure 2. Construction of global alignment using tree diagram method

In figure 2. After we input a pair of sequences DNA virus H5N1, we can process alignment of two sequences in three part, i.e improved simple alignment, extension algorithm and GSA Tree. The scoring equation is $S = \alpha p - \beta q - r(o + ke)$ and we choose the parameters $\alpha = 5, \beta = 4$ (from DNAFULL matrix) and $o = 10, e = 0.5$ (default gap open and gap extension in tools EMBOSS). Then we obtain percentage of similarity and gaps to analysis mutation level in nucleotides.
We obtain DNA sequences in this research from GenBank database. This data is 2 sequences from host human and 4 sequences from host avian. The result of alignment using this method can explain in Table 1.

Table 1. The result alignment sequence DNA H5N1 virus on HA segment using tree diagram methods

<table>
<thead>
<tr>
<th>No</th>
<th>A pairs of sequences</th>
<th>Length alignment</th>
<th>similarity</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CY088769 and HQ200596</td>
<td>1774</td>
<td>1583</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(89.2%)</td>
<td>(3.8%)</td>
</tr>
<tr>
<td>2</td>
<td>CY088769 and CY091956</td>
<td>1746</td>
<td>1590</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(91.1%)</td>
<td>(2.2%)</td>
</tr>
<tr>
<td>3</td>
<td>CY088769 and HM172081</td>
<td>1709</td>
<td>1554</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(90.9%)</td>
<td>(0.4%)</td>
</tr>
<tr>
<td>4</td>
<td>CY088769 and AB569353</td>
<td>1745</td>
<td>1560</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(89.4%)</td>
<td>(2.3%)</td>
</tr>
<tr>
<td>5</td>
<td>CY088769 and AB629698</td>
<td>1749</td>
<td>1562</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(89.3%)</td>
<td>(2.8%)</td>
</tr>
<tr>
<td>6</td>
<td>HQ200596 and CY091956</td>
<td>1775</td>
<td>1631</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(91.9%)</td>
<td>(1.7%)</td>
</tr>
<tr>
<td>7</td>
<td>HQ200596 and HM172081</td>
<td>1773</td>
<td>1550</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(87.4%)</td>
<td>(3.9%)</td>
</tr>
<tr>
<td>8</td>
<td>HQ200596 and AB569353</td>
<td>1774</td>
<td>1606</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(90.5%)</td>
<td>(1.9%)</td>
</tr>
<tr>
<td>9</td>
<td>HQ200596 and AB629698</td>
<td>1774</td>
<td>1601</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(90.2%)</td>
<td>(1.9%)</td>
</tr>
<tr>
<td>10</td>
<td>CY091956 and HM172081</td>
<td>1747</td>
<td>1566</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(89.6%)</td>
<td>(2.5%)</td>
</tr>
<tr>
<td>11</td>
<td>CY091956 and AB569353</td>
<td>1747</td>
<td>1601</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(91.6%)</td>
<td>(0.3%)</td>
</tr>
<tr>
<td>12</td>
<td>CY091956 and AB629698</td>
<td>1747</td>
<td>1600</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(91.6%)</td>
<td>(0.3%)</td>
</tr>
<tr>
<td>13</td>
<td>HM172081 and AB569353</td>
<td>1746</td>
<td>1542</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(88.3%)</td>
<td>(2.6%)</td>
</tr>
<tr>
<td>14</td>
<td>HM172081 and AB629698</td>
<td>1745</td>
<td>1541</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(88.3%)</td>
<td>(2.5%)</td>
</tr>
<tr>
<td>15</td>
<td>AB569353 and AB629698</td>
<td>1742</td>
<td>1714</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(98.4%)</td>
<td>(0.0%)</td>
</tr>
</tbody>
</table>

Using Table 1, we can found similarity level in internal host (human-human and avian-avian) and external host (human-avian). The similarity level in human-human is 89.2%, avian-avian is 91.3%, and human-avian is 90.1%. Furthermore, we count mutation level of nucleotides using information of similarity and gap in the Table 1. We obtain the mutation in human-human is 7%, avian-avian is 7.3% and human-avian is 7.8%. All these result was same with the result alignment from EMBOSS tools (a tools for pairwise alignment using “needle”-algorithm).
3. CONCLUDING REMARK

According to this result, we can conclude that tree diagram method is sufficient to align the sequences by applying the concept of a tree data structure that contains simple alignment algorithms, Extension algorithms and GSA tree. This method can improved the alignment of two DNA sequences by exploring the appropriate gaps, gradually based on simple alignment algorithms and extension algorithms. Based on validation results using tools EMBOSS, shows that the optimal alignment generated from the parameters match, mismatches, penalty gap open and penalty gap extend respectively \( \alpha = 5, \beta = 4 \) and \( \gamma = 10, e = 0.5 \).

Furthermore, based from result similarity level sequence DNA H5N1 virus on segment HA, in internal and external host we conclude that in biological this is an indication that they are different species. And based from result mutation level shown that mutation level on this virus is high, and the highest is mutation level between host human-avian 7.8%.

Acknowledgement. I wish to give my gratitude to Ministry of Religion of Republic Indonesia that has given me scholarship to study magister program in Institut Teknologi Sepuluh Nopember (ITS) Surabaya.

References


SITI FAUZIYAH: Graduate Student of Mathematics Department at Institut Teknologi Sepuluh Nopember (ITS) Surabaya. 
E-mails: s_fauziyah@ymail.com

M. ISA IRAWAN: supervisor, lecturer of Mathematics Department at Institut Teknologi Sepuluh Nopember (ITS) Surabaya. 
E-mails: mii@its.ac.id

MAYA SHOVITRI: co. supervisor, lecturer of Biology Department at Institut Teknologi Sepuluh Nopember (ITS) Surabaya. 
E-mails: maya@bio.its.ac.id