The Similarity Analysis of DNA Sequence Model Based on Graph Theory and Blast Program

Y. A. Lesnussa, S. Kappuw, B. P. Tomasouw & E. R. Persulessy
Mathematic Department, Faculty of Mathematics and Natural Sciences, University of Pattimura
Jl. Ir M. Putuhena, Kampus Unpatti Poka-Ambon, POS Code 97233, Indonesia
e-mail: yopi_a_lesnussa@yahoo.com; sammykappuw@ymail.com

Abstract

DNA is a nucleotide acid in form of double helix which contains genetic instruction to determine biology development of all forms of cell's life also it relates with genetic characteristic inheritance. In this research, we will see the similarity of two DNA sequences. DNA sequences that we used are human, orangutan, and gorilla. The method that we used to analyze the similarity of DNA sequences is Graph Theory. This method started by modeling each DNA sequence into a graph, making its adjacency matrix and builds a matrix vector for each graph. From these vectors we will determine similarity of two DNA sequences. The similarity of DNA sequences is determined by the similarity level using Cosine, Correlation, and Euclid. Where, the results are shown by the smaller distance, and then showing the similarity of two DNA sequences. And then compare the result from Graph Theory with the results of Basic Local Alignment Search Tools (BLAST) program. Finally, the result of research shows that Human and Gorila have close similarity of their DNA sequences.

2010 Mathematics Subject Classification: 00A69.
Keywords DNA Sequence, Graph, Cosine, Correlation, Euclid

INTRODUCTION

Deoxyribose Nucleic Acid (DNA) is a nucleotide acid, usually in the form of a double helix that contains the genetic instructions that determine the biological development of all cellular life forms. DNA is responsible for the genetic inherited of most traits. In humans, these traits, for example from hair color to susceptibility to the disease. During cell division, DNA is replicated and can be passed to ancestry during reproduction. At present, thousands of new species are discovered each year, the DNA line and protein described every day from species not previously examined. At the current rates today, which increases exponentially, nearly 38 million new base row is read every day (Xingqin Qi, 2011). Therefore, the usefulness of seeking similarities in DNA sequences is very important to know where the species originate. The greater number of DNA sequences in a DNA database because the DNA sequence rearrangements occur during evolution over time. It is one of the challenges for bio-scientists to analyze the similarity of DNA sequences. In this case, the similarity in DNA sequences will show by its similarity level that was approached by the mathematical sciences. The similarity level will be measured using the formula distance measurement, in which the smaller distance, the more similar the two DNA sequences.

The method used to analyze the similarity in DNA sequences was by using graph. This method is a numerical scheme to characterize and analyze the similarity of DNA sequences. This methodology begins with representation of graph from DNA sequences, and then graph of DNA representing adjacency matrix, and then forms a vector comprising a matrix invariants are used to compare the DNA sequences. This research will introduce three distance measurements will be used to analyze the similarity or dissimilarity of DNA sequence, such as: Cosine, Correlation and Euclid. The aim of this research was modeling the DNA sequence in the graph, and to know the similarity or dissimilarity of the DNA sequence of several different species by using Cosine, Correlation and Euclid and comparing the results with the analysis of the BLAST program.

RESEARCH METHODOLOGY

This type of research is a literature study about how to apply the mathematical method to solve the problem in Genetic Biology. The material used in this research is DNA sequences that were obtained from Gen Bank of National Centre for Biotechnology Information (NCBI). We calculate the similarity level of DNA from several species by using Graph Theory and we then represent this calculation by using Graphic User Interface (GUI) Matlab and compared it to BLAST program. The stages of this research are shown in Figure 1 as follows:

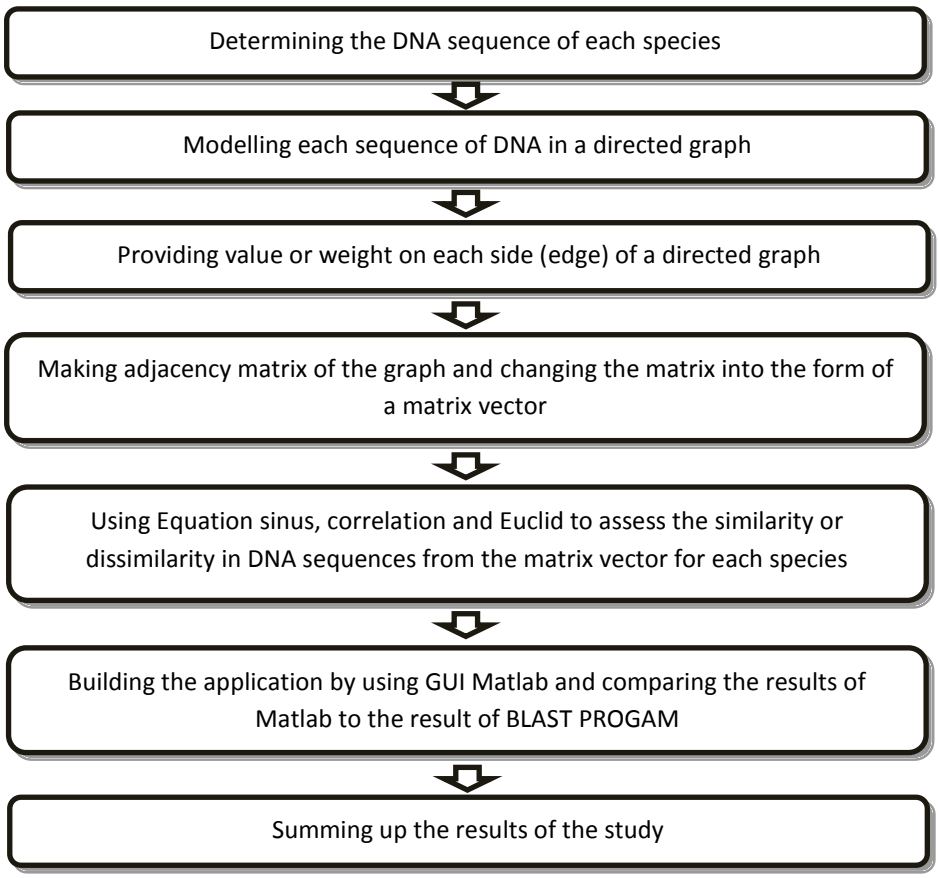

Figure 1 Research steps

RESULTS AND DISCUSSION

1. Vector Representation of DNA sequences

Representative letters to the DNA sequences are A, C, G and T. Assume $S = s_1, s_2, \ldots, s_n$ is the DNA sequence of length n, where $s_i \in \{A, C, G, T\}$. The following will discuss on how to form a double-directed and weighted graph to line up S.

Double Directional Graph, G_m

Definition 1.
Suppose $G_m = (V(G_m), E(G_m))$ is a double-directed and weighted graph of an S DNA sequence. $V(G_m)$ is a set of points with elements A, C, G and T. $E(G_m)$ is a set of edges directionally placed in

each pair of nucleotide $s_i$ands_j in S with $i < j$ so assigning weights to the side (arc) can be defined as $\frac{1}{(j-1)^{\alpha}}$, where $\alpha > 0$. The function $\frac{1}{(j-1)^{\alpha}}$ is a function down from $(j-1)$ which will reflect the fact that the two nucleotides with a smaller distance will have a stronger relationship than the two nucleotides that have a greater distance.

Theorem 1.

Suppose S is a DNA sequence and G_m is a double directed and weighted graph, then there is a one-to-one correspondence between S and G_m.

Proof:

Enough shown that the graph G_m can only form one row of S, and vice versa. Given $n_w$representing the number of nucleotides, with $W \in \{A, C, G, T\}$ appearing in the DNA sequence and x_w represents the number of loops at each point W in G_m. Obviously $x_w = (n_w - 1) \times \frac{n_w}{2}$ for each $W \in \{A, C, G, T\}$, so it can be obtained every n_w of x_w. The length of DNA sequences can be obtained from $n = n_A + n_C + n_G + n_T$. Note that there is only one arc (W', W'') with a weight of $\frac{1}{(n-1)^{\alpha}}$ in G_m. Thus, the first nucleotide in the sequence S is $W'$and second nucleotide in row S is W'', where $W''$is a j*th*-nucleotide in S of arc (W', W'') with a weight of $\frac{1}{(j-1)^{\alpha}}$.

Double Directional and Weighted Simplified Graph G_s

Suppose G_m is double directed graph. There is a possibility that there are several parallel arcs that connect from one point to another. Thus, G_m will simplify to G_s by combining multiple parallel arc/arcs into a single arc. Given a set point $V(G_s) = V(G_m)$. Denoted $A_{u,v}^m$ as the set of all arcs from point **u** to **v** in G_m. For every pair of points u and v, $A_{u,v}^m \neq \emptyset$, place the arc (u, v) from u to v in G_s, then the weight of the arc (u, v) in G_s can be formulated as:

$$W_s(u,v) = \sum_{(u,v) \in A_{u,v}^m} W_m(u,v), A_{u,v}^m \neq \emptyset$$

Establishment Vector

By the previous explanation, we can obtain the directed and weighted graph that are related to DNA sequences. The directed and simplified weighted graph G_s corresponding to matrix adjacency $M_{(4\times4)}$, which is defined as follows:

$$M = \begin{pmatrix} W_s(A,A) & W_s(A,C) & W_s(A,G) & W_s(A,T) \\ W_s(C,A) & W_s(C,C) & W_s(C,G) & W_s(C,T) \\ W_s(G,A) & W_s(G,C) & W_s(G,G) & W_s(G,T) \\ W_s(T,A) & W_s(T,C) & W_s(T,G) & W_s(T,T) \end{pmatrix}$$

Then with the command line, the matrix will be written in 16-dimensional vector, as follows:

$$\overrightarrow{R_s} = [W_s(A,A), \dots, W_s(A,T), \dots, W_s(C,A), \dots, W_s(C,T), \dots, W_s(T,A), \dots, W_s(T,T)]$$

16-dimensional vector is a vector representation of the DNA sequence. The weighting function is $f(l) = \frac{1}{\sqrt{l}}$, where *l* = (j-1).

2. Calculation of Similarity DNA Sequence

Vector can represent the DNA sequence. This vector will be used to determine the degree of similarity between the DNA sequences with other DNA sequences. The level of similarity will be calculated using three measurements of distance, namely:

i. Measurement of the first distance is defined as Cosine, which is based on assumption that two DNA sequences are similar if the vector of 16 dimensions appropriate in space of 16 dimensions and have the same direction, such as:

$$d_1(s,h) = 1 - \cos(\overrightarrow{R_s}, \overrightarrow{R_h}) = 1 - \frac{\sum_{i=1}^{16} \overrightarrow{R_s}(i) \cdot \overrightarrow{R_h}(i)}{\sqrt{\sum_{i=1}^{16} (\overrightarrow{R_s}(i))^2 \cdot \sum_{i=1}^{16} (\overrightarrow{R_h}(i))^2}}$$

ii. The second distance measurement is based on the calculation of Correlation coefficients, as the following:

$$r(s,h) = \frac{K \sum_{i=1}^{K} [\overrightarrow{R_s}(i) \cdot \overrightarrow{R_h}(i)] - \sum_{i=1}^{K} \overrightarrow{R_s}(i) \cdot \sum_{i=1}^{K} \overrightarrow{R_h}(i)}{\sqrt{K \sum_{i=1}^{K} (\overrightarrow{R_s}(i))^2 - [\sum_{i=1}^{K} \overrightarrow{R_s}(i)]^2} \times \sqrt{K \sum_{i=1}^{K} (\overrightarrow{R_h}(i))^2 - [\sum_{i=1}^{K} \overrightarrow{R_h}(i)]^2}}$$

where K is the dimension of $\vec{R}_s$ or $\vec{R}_h$ (K = 16). Thus the second distance measurement is defined as follows:

$$d_2(s,h) = 1 - r(s,h)$$

$$= 1 - \frac{K \sum_{i=1}^{K} [\overrightarrow{R_s}(i) \cdot \overrightarrow{R_h}(i)] - \sum_{i=1}^{K} \overrightarrow{R_s}(i) \cdot \sum_{i=1}^{K} \overrightarrow{R_h}(i)}{\sqrt{K \sum_{i=1}^{K} (\overrightarrow{R_s}(i))^2 - [\sum_{i=1}^{K} \overrightarrow{R_s}(i)]^2} \times \sqrt{K \sum_{i=1}^{K} (\overrightarrow{R_h}(i))^2 - [\sum_{i=1}^{K} \overrightarrow{R_h}(i)]^2}}$$

iii. The third distance measurement is defined as Euclidean distance based on the assumption that the same of two DNA sequence if the corresponding 16-dimensional vector have the same size, namely:

$$d_3(s,h) = \sqrt{\sum_{i=1}^{16} (\overrightarrow{R_s}(i) - \overrightarrow{R_h}(i))^2}$$

3. The Results of Similarity Analysis of DNA Sequence

In this research, we use the DNA sequences from Human, Orangutan and Gorilla obtained from Gen Bank database. The following shows the results of DNA sequence similarity analysis of three different species (Human, Gorilla and Orangutan) with take length of DNA sequence fragments = 8, α = 1.

i. The DNA sequence of Human (Homo sapiens), as derived from Gen Bank is as follow:

```
ORIGIN:
AAGCTTCACC GGCGCAGTCA TTCTCATAAT CGCCCACGGA CTTACATCCT CATTACTATT
CTGCCTAGCA AACTCAAACT ACGAACGCAC TCACAGTCGC ATCATAATCC TCTCTCAAGG
ACTTCAAACT CTACTCCCAC TAATAGCTTT TTGATGACTT CTAGCAAGCC TCGCTAACCT
CGCCTTACCC CCCACTATTA ACCTACTGGG AGAACTCTCT GTGCTAGTAA CCACGTTCTC
CTGATCAAAT ATCACTCTCC TACTTACAGG ACTCAACATA CTAGTCACAG CCCTATACTC
CCTCTACATA TTTACCACAA CACAATGGGG CTCACTCACC CACCACATTA ACAACATAAA
ACCCTCATTC ACACGAGAAA ACACCCTCAT GTTCATACAC CTATCCCCCA TTCTCCTCCT
ATCCCTCAAC CCCGACATCA TTACCGGGTT TTCCTCTTGT AAATATAGTT TAACCAAAAC
ATCAGATTGT GAATCTGACA ACAGAGGCTT ACGACCCCTT ATTTACCGAG AAAGCTCACA
AGAACTGCTA ACTCATGCCC CCATGTCTGA CAACATGGCT TTCTCAACTT TTAAAGGATA
ACAGCTATCC ATTGGTCTTA GGCCCCAAAA ATTTTGGTGC AACTCCAAAT AAAAGTAATA
ACCATGCACA CTACTATAAC CACCCTAACC CTGACTTCCC TAATTCCCCC CATCCTTACC
ACCCTCGTTA ACCCTAACAA AAAAAACTCA TACCCCCATT ATGTAAAATC CATTGTCGCA
TCCACCTTTA TTATCAGTCT CTTCCCCACA ACAATATTCA TGTGCCTAGA CCAAGAAGTT
ATTATCTCGA ACTGACACTG AGCCACAACC CAAACAACCC AGCTCTCCCT AAGCT
```

From the DNA sequence, take 8-Fragment for example: ACTGACAC. Then, build the directed and weighted graph and provide the value of the each direction. The next stage is to take an adjacency matrix of the graph and then to change the matrix into the form of a matrix vector.

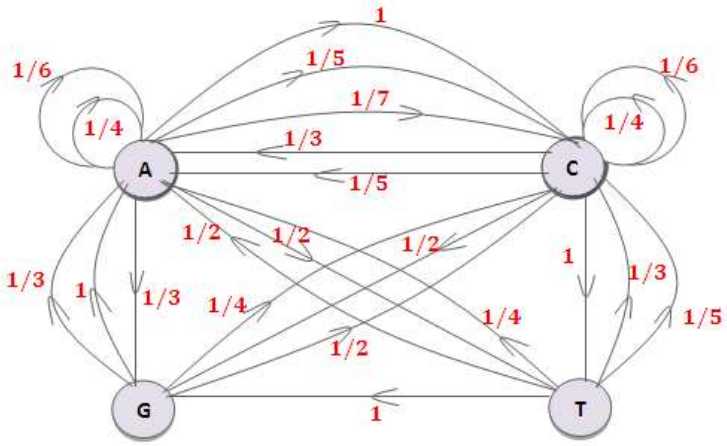

Figure 2 Graf G_m of the human DNA sequence fragments

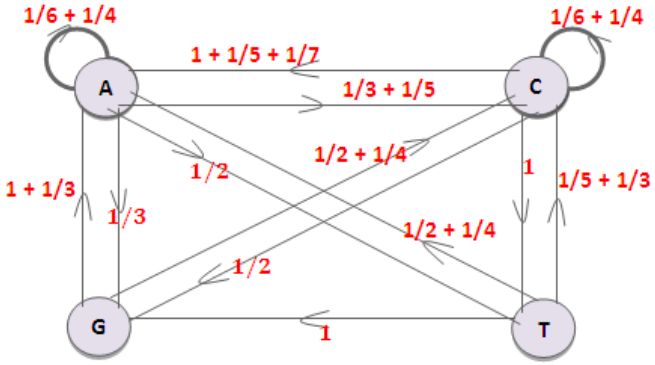

Figure 3 Graph G_S of the human DNA sequence 8-fragment ACTGACAC

$$M = \begin{pmatrix} 0.4167 & 0.5333 & 0.3333 & 0.5000 \\ 1.3429 & 0.4167 & 0.5000 & 1.0000 \\ 1.3333 & 0.7500 & 0.0000 & 0.0000 \\ 0.7500 & 0.5333 & 1.0000 & 0.0000 \end{pmatrix}$$

Thus, find the matrix vector as the following:

$$\overrightarrow{R_m} = [0.4167, 0.5333, 0.3333, 0.5000, 1.3429, 0.4167, 0.5000, 1.0000, 1.3333, 0.7500, 0.0000, 0.0000, 0.7500, 0.5333, 1.0000, 0.0000]$$

ii. The DNA sequence of Sumatran Orangutan (Pongo abelii), for example:

```
ORIGIN:
AAGCTTCACC GGCGCAACCA CCCTCATGAT TGCCCATGGA CTCACATCCT CCCTACTGTT
CTGCCTAGCA AACTCAAACT ACGAACGAAC CCACAGCCGC ATCATAATCC TCTCTCAAGG
CCTTCAAACT CTACTCCCCC TAATAGCCCT CTGATGACTT CTAGCAAGCC TCACTAACCT
TGCCCTACCA CCCACCATCA ACCTTCTAGG AGAACTCTCC GTACTAATAG CCATATTCTC
TTGATCTAAC ATCACCATCC TACTAACAGG ACTCAACATA CTAATCACAA CCCTATACTC
TCTCTATATA TTCACCACAA CACAACGAGG TACACCCACA CACCACATCA ACAACATAAA
ACCTTCTTTC ACACGCGAAA ATACCCTCAT GCTCATACAC CTATCCCCCA TCCTCCTCTT
```

```
ATCCCTCAAC CCCAGCATCA TCGCTGGGTT CGCCTACTGT AAATATAGTT TAACCAAAAC
ATTAGATTGT GAATCTAATA ATAGGGCCCC ACAACCCCTT ATTTACCGAG AAAGCTCACA
AGAACTGCTA ACTCTCACTC CATGTGTGAC AACATGGCTT TCTCAGCTTT TAAAGGATAA
CAGCTATCCC TTGGTCTTAG GATCCAAAAA TTTTGGTGCA ACTCCAAATA AAAGTAACAG
CCATGTTTAC CACCATAACT GCCCTCACCT TAACTTCCCT AATCCCCCCC ATTACCGCTA
CCCTCATTAA CCCCAACAAA AAAAACCCAT ACCCCCACTA TGTAAAAACG GCCATCGCAT
CCGCCTTTAC TATCAGCCTT ATCCCAACAA CAATATTTAT CTGCCTAGGA CAAGAAACCA
TCGTCACAAA CTGATGCTGA ACAACCACCC AGACACTACA ACTCTCACTA AGCTT
```

From the sequence of Orangutan DNA sequence above, take 8-fragments, such as: CTGATGCT. Then, build a directed and weighted graph of this sequence. The steps to build a graph are as follows:

1. Determine the vertex of graph by using the letter from DNA sequence (A, C, G, T)
2. Determine the weight of each edge by counting the step from a letter to another letter in a fragment of DNA sequence but with only one direction (Figure 4).
3. Sum all the weights in every edge which has the similar direction from a letter to another letter (Figure 5).

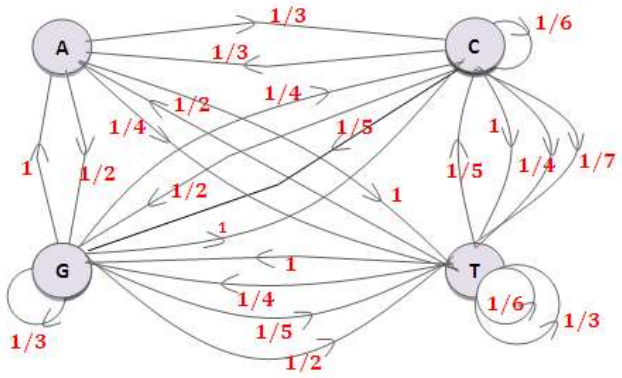

Figure 4 Graph G_m of orangutan DNA sequence 8-fragment CTGATGCT

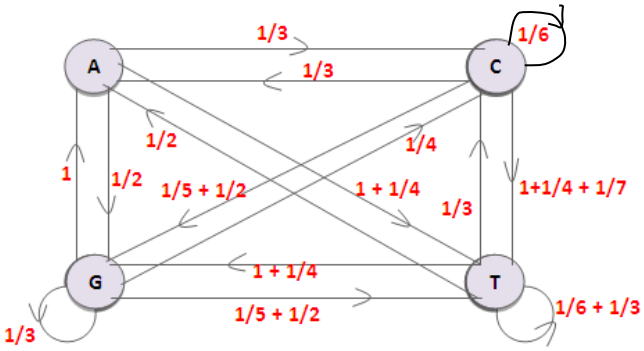

Figure 5 Graph G_s of Orangutan DNA sequence 8-fragment CTGATGCT

$$M = \begin{pmatrix} 0.0000 & 0.3333 & 0.5000 & 1.2500 \\ 0.3333 & 0.1667 & 0.7000 & 1.3929 \\ 1.0000 & 1.2500 & 0.3333 & 0.7000 \\ 0.5000 & 0.3000 & 1.2500 & 0.5000 \end{pmatrix}$$

$$\overrightarrow{R_0} = [0.0000, 0.3333, 0.5000, 1.2500, 0.3333, 0.1667, 0.7000, 1.3929, 1.0000, 1.2500, 0.3333, 0.7000, 0.5000, 0.3000, 1.2500, 0.5000].$$

iii. The DNA sequence of Western Gorilla (**Gorilla gorilla**), for example

```
ORIGIN:
AAGCTTCACC GGCGCAGTTG TTCTTATAAT TGCCCACGGA CTTACATCAT CATTATTATT
CTGCCTAGCA AACTCAAACT ACGAACGAAC CCACAGCCGC ATCATAATTC TCTCTCAAGG
ACTCCAAACC CTACTCCCAC TAATAGCCCT TTGATGACTT CTGGCAAGCC TCGCCAACCT
CGCCTTACCC CCCACCATTA ACCTACTAGG AGAGCTCTCC GTACTAGTAA CCACATTCTC
```

```
CTGATCAAAT ACCACCCTTT TACTTACAGG ATCTAACATA CTAATCACAG CCCTGTACTC
CCTTTATATA TTTACCACAA CACAATGAGG CCCACTCACA CACCACATCA CCAACATAAA
ACCCTCATTT ACACGAGAAA ACATCCTCAT ATTCATGCAC CTATCCCCCA TCCTCCTCCT
ATCCCTCAAC CCCGATATTA TCACCGGGTT CACCTCCTGT AAATATAGTT TAACCAAAAC
ATCAGATTGT GAATCTGATA ACAGAGGCTC ACAACCCCTT ATTTACCGAG AAAGCTCGTA
AGAGCTGCTA ACTCATACCC CCGTGCTTGA CAACATGGCT TTCTCAACTT TTAAAGGATA
ACAGCTATCC ATTGGTCTTA GGACCCAAAA ATTTTGGTGC AACTCCAAAT AAAAGTAATA
ACTATGTACG CTACCATAAC CACCTTAGCC CTAACTTCCT TAATTCCCCC TATCCTTACC
ACCTTCATCA ATCCTAACAA AAAAAGCTCA TACCCCCATT ACGTAAAATC TATCGTCGCA
TCCACCTTTA TCATCAGCCT CTTCCCCACA ACAATATTTC TATGCCTAGA CCAAGAAGCT
ATTATCTCAA GCTGACACTG AGCAACAACC CAAACAATTC AACTCTCCCT AAGCTT
```

From Gorilla DNA sequence, take 8-fragments, such as: GCTGACAC, and then construct the directed and weighted graph to find matrix vector.

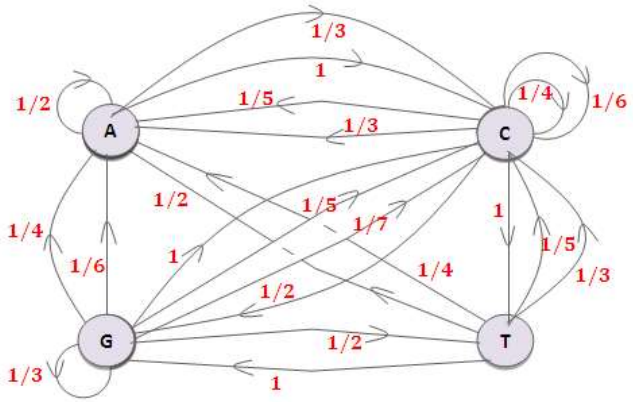

Figure 6 Graph G_m of Gorilla DNA sequence 8-fragment GCTGACAC

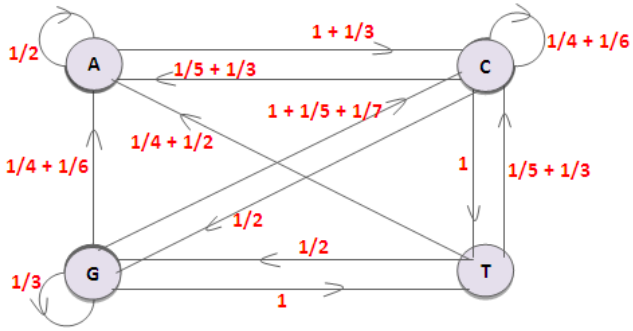

Figure 7 Graph G_s of gorilla DNA sequence 8-fragment GCTGACAC

$$M = \begin{pmatrix} 0.5000 & 2.3333 & 0.0000 & 0.0000 \\ 1.5333 & 0.9167 & 0.5000 & 1.0000 \\ 1.7500 & 2.0929 & 0.3333 & 0.5000 \\ 0.7500 & 0.5333 & 1.0000 & 0.0000 \end{pmatrix}$$

Therefore, the matrix vector of the sequence, is:

$$\overrightarrow{R_g} = [0.5000, 2.3333, 0.0000, 0.0000, 1.5333, 0.9167, 0.5000, 1.0000, 1.7500, 2.0929, 0.3333, 0.5000, 0.7500, 0.5333, 1.0000, 0.0000].$$

In the previous section of modelled fragments of DNA sequence from human, Orangutan and gorilla in the graph and adjacency matrix known as well as 16-dimensional vector $(\vec{R}_m, \vec{R}_o, \vec{R}_g)$ of each of the DNA sequences. Therefore, by using Cosine, Correlation and Euclid, the similarity level or degree of each

species. The level of similarity judged from the smaller distance, then the more similar the two DNA sequences are. In the following, will be calculated degree of similarity of some of the DNA sequences using GUI Matlab. Then the results

The Results of Calculation Using GUI Matlab

By using Matlab version 7.1, the results can be presented by using Graphic User Interface (GUI) Matlab to help user easier to calculate the value of Cosine, Correlation and Euclid. It can be seen in following figure:

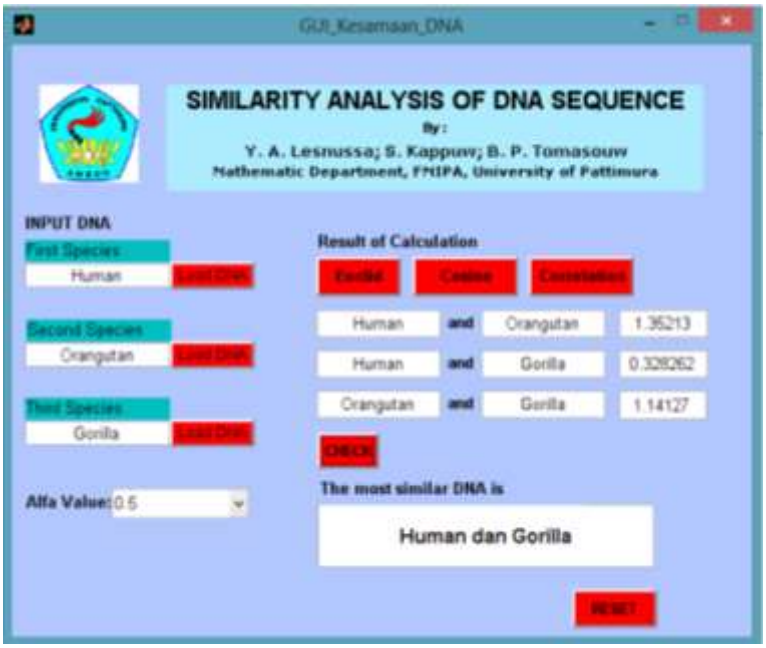

Figure 8 Visualisation of GUI Matlab

For example, take the length of DNA sequence is 895 for each species and α is 3, and the result of calculation can be seen in the following table:

Table 1 Calculation Results Using MATLAB

Species	**Cosine**	**Correlation**	**Euclid**
(Human and Orangutan)	0.0025	0.0084	22.4005
(Human and Gorilla)	0.0018	0.0084	18.1558
(Orangutan and Gorilla)	0.0046	0.0173	29.8351

It can be concluded that the human DNA sequence and the Gorilla DNA sequence are very similar because they have the smallest value of Cosine, Correlation and Euclid (Table 1). It can be seen from each results of calculation = 0.0018, 0.0084, 18.1558, is the smallest distance.

The Results Analysis Using BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identifying members of gene families. One of BLAST tools; Global Alignment,

is used to compare two sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970. The **Needleman–Wunsch algorithm** is an algorithm used in bioinformatics to align protein or nucleotide sequences. It is one of the first applications of dynamic programming to compare biological sequences.

The results of BLAST Global Alignment Nucleotide Sequences, as the following figure:

a. The Comparison of Human with Orangutan

Figure 9 Output of BLAST to Compare Human with Orangutan

b. The Comparison of Human with Gorilla

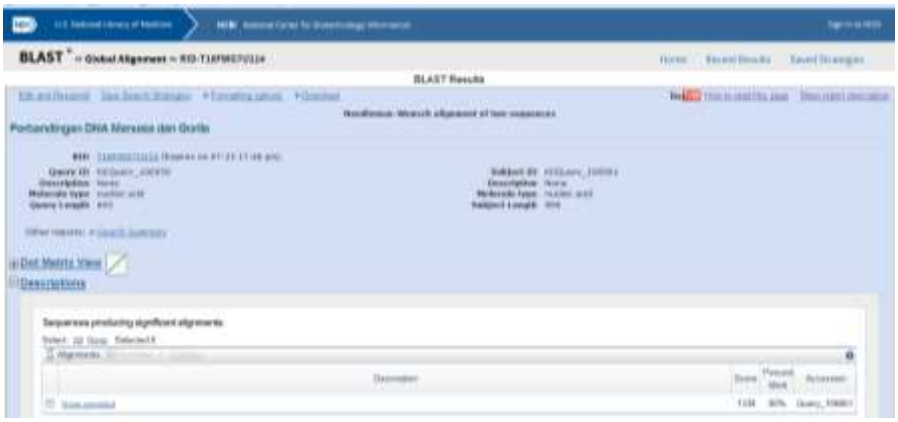

Figure 10 Output of BLAST to Compare Human with Gorilla

c. The Comparison of Orangutan with Gorilla

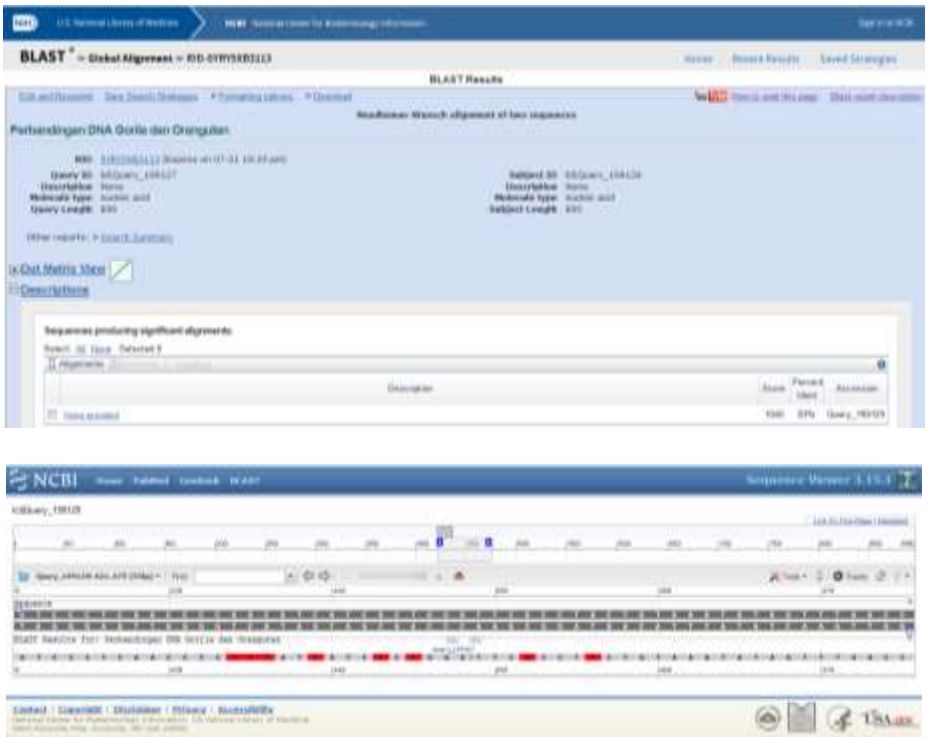

Figure 11 Output of BLAST to Compare Human with Gorilla

The results of the analysis with the BLAST program found that the percentage of similarity of each pair of DNA sequence showed that the largest percentage of DNA is found on the pair Human and Gorilla (90%), followed by Human and Orangutan (84%) and lastly, Orangutan and Gorilla (83%). Therefore, the most similar DNA sequence is Human and Gorilla.

CONCLUSION

The DNA sequence of a species can be modelled into a double directed and weighted graph and the results of the calculations using Cosine, Correlation and Euclid, has obtained that the DNA of Human and Gorilla has a similar genetic relationship compared to Human and Orangutan DNA or Orangutan and Gorilla DNA.

REFERENCES

[1] Chartrand G., and Lesniak L. (1986). *Graph and Digraph 2nd Edition*. California: Wadsworth. Inc.
[2] Hasan I. (2004). *Analisis Data Penelitian dengan Statistik*. Jakarta: Penerbit Bumi Aksara.
[3] Howard A. (2004). *Aljabar Linier Elementer*. Jakarta: Penerbit Erlangga.
[4] Xingqin Q., Qin W., Yusen Z., Eddie F., & Cun Q. Z. (2011). *A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory*, Evolutionary Bioinformatics, Libertas Academica.
[5] Tooze J. and Watson J. D. (1988). *DNA Rekombinan*. Jakarta: Penerbit Erlangga.
[6] Wibisono, S. (2008). *Matematika Diskrit.* Yogyakarta: Penerbit Graha Ilmu.
[7] Wilson R. J., and Watkins J. J. (1990). *Graph An Introductory Approach: A First Course in Discrete Mathematic*. New York: John Wiley & Sons, Inc.
[8] Zhang Y, Liao B, Ding K. (2006). *On 3D D-curves of DNA sequences*. *Mol Simul*.32:29-34.
[9] https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome