Application of Graph Theory in DNA similarity analysis of Evolutionary Closed Species.

W.W.P.M.T.M.Karunasena, *Department of Mathematics, University of Kelaniya, Sri Lanka. karunasenatm@gmail.com*
G.S.Wijesiri, *Department of Mathematics, University of Kelaniya, Sri Lanka. sujeew@kln.ac.lk*

ABSTRACT

DNA is a complex molecule that consists of biological information that is passed down from generation to generation. With the evolution over time, there are different kinds of species that evolved from a common ancestor because of the occurrence of DNA sequence rearrangements. DNA sequence similarity analysis is a major challenge since the number of sequences is rapidly increasing in the DNA database. In this research, we based a mathematical method to analyze the similarity of two DNA sequences using Graph Theory. This mathematical method started by modeling a weighted directed graph for each DNA sequence, constructing its adjacency matrix, and converting it to the representative vector for each graph. From these vectors, the similarity was determined by distance measurements such as Euclidean, Cosine, and Correlation. By keeping this method as the based method, we will check whether it is applicable for any DNA fragments in considered genomes and molecular similarity coefficients can be used as distance measurements. We will obtain similarities using the graph spectrum instead of the representative vector. Then we will compare the results from the representative vector and that of the graph spectrum. The modified method is tested by using the mitochondrial DNA of Human, Gorilla, and Orangutan. It gives the same result when the number of nucleotides in DNA fragments is increased.

Keywords:
DNA, Evolutionary, Spectrum, Similarity, Weighted directed graph
Article Received: 18 October 2020, Revised: 3 November 2020, Accepted: 24 December 2020

I. Introduction

Deoxyribonucleic acid or DNA is a complex molecule that consists of the biological information that makes every species distinctive. It includes the instructions an organism needs to develop, live, and reproduce. Inside every living cell, these instructions are found and passed down during the reproductive process from a parent organism to its offspring. Deoxyribonucleic acid is created of chemical building blocks known as nucleotides. Each building block contains a phosphate group, a sugar group, and a nitrogen base. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder with the base pairs. Nitrogen bases are available in four types. They are adenine (A), thymine (T), guanine (G), and cytosine(C). The biological instructions or genetic code contained in a DNA strand is determined by the order or sequence of these bases.

DNA sequences analysis is very important in biology as the number of DNA sequences in the DNA database rapidly increases at the current rates of today. Therefore, it's essential that finding Similarities / dissimilarities in DNA sequences to know where the species originate and identify homologous sequences. But it is hard to obtain information from the DNA sequence directly because rearrangements occur during the evolution over time. Analyzing large amounts of genomic DNA sequence data is a major challenge for bio-scientists.

In mathematical biology, mathematical models are applied in biology to deal with various modeling and calculation problems. In the microscopic field of biology, DNA, RNA structures, protein sequences, and other biological networks can be represented as a graph. Thus, graph theory is established itself as a unique mathematical tool in determining various biological properties due to its ease of representing the above biological networks. With the development of the graph theory, graphs are used for different needs in different biological structures such as graphs are used to predict similarities between DNA sequences [3].

In 2011, a novel method based on graph theory was introduced for similarity calculations [4]. That method was started from a weighted directed graph for each DNA sequence, DNA representing adjacency matrix, and then comprised the matrix to a representative vector. Three distance measurements were defined to calculate the similarity between vectors. In 2017, the above method was used to calculate the similarities between Human, Gorilla, and Orangutan by using Cosine, Correlation, and Euclidean distances [2].

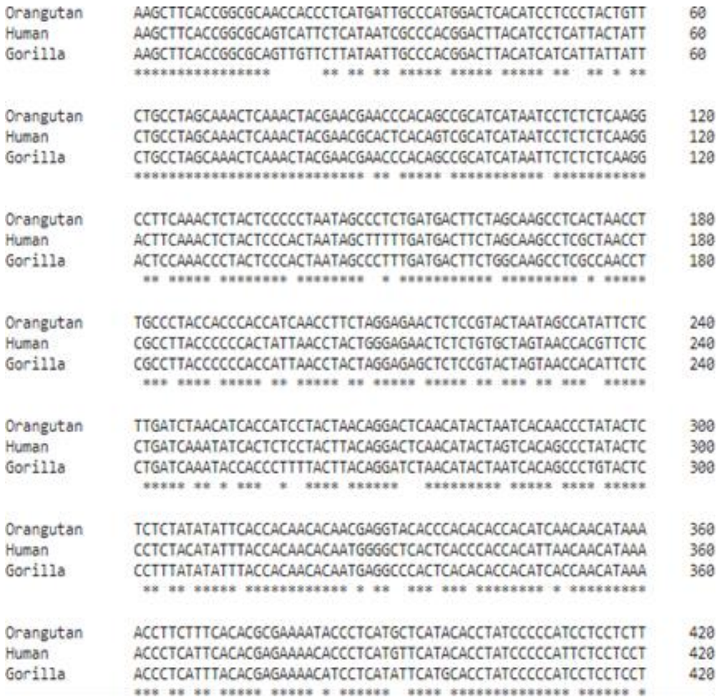
This research will modify the above mathematical model which is based on graph theory to represent DNA sequences mathematically for similarity analysis. We will check whether the above method is applicable for any regions of the genomes and apply molecular similarity measurements as distance measurements. The method was repeated by using the spectrum of the graph as a vector and compared the results of two vector representations. The modified method will be verified by using the genomes of Human, Gorilla, and Orangutan.

II. Research Methodology

The materials used in this research are mitochondrial DNA sequences of three evolutionary closed species (Human, Orangutan, and Gorilla) that were downloaded from Gen Bank of National Center for Biotechnology Information (NCBI). We use the Clustal Omega program as a multiple sequence alignment tool. The first step in this study is finding the regions of genomes that the novel method can be applied.

Genomes of the above three species were aligned by using the Clustal Omega program to detect the conserved regions and DNA variations. The regions including the conserved regions and DNA variations were used to continue the research. A sample figure of aligned genomes is shown in Figure 1.

Then we model weighted directed graphs for the randomly selected DNA regions with conserved regions and DNA variations of each DNA sequence of each species.



```
Orangutan       AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTT      60
Human           AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATT      60
Gorilla         AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATT      60
                ****************      ** ** ** ***** ***** ***** **  ** * **

Orangutan       CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATCCTCTCTCAAGG      120
Human           CTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGG      120
Gorilla         CTGCCTAGCAAACTCAAACTACGAACGAACCCACAGCCGCATCATAATTCTCTCTCAAGG      120
                *************************** ** ***** *********** ***********

Orangutan       CCTTCAAACTCTACTCCCCCTAATAGCCCTCTGATGACTTCTAGCAAGCCTCACTAACCT      180
Human           ACTTCAAACTCTACTCCCACTAATAGCTTTTTGATGACTTCTAGCAAGCCTCGCTAACCT      180
Gorilla         ACTCCAAACCCTACTCCCACTAATAGCCCTTTGATGACTTCTGGCAAGCCTCGCCAACCT      180
                 ** ***** ******** ********  * *********** ********* * *****

Orangutan       TGCCCTACCACCCACCATCAACCTTCTAGGAGAACTCTCCGTACTAATAGCCATATTCTC      240
Human           CGCCTTACCCCCCACTATTAACCTACTGGGAGAACTCTCTGTGCTAGTAACCACGTTCTC      240
Gorilla         CGCCTTACCCCCCACCATTAACCTACTAGGAGAGCTCTCCGTACTAGTAACCACATTCTC      240
                 *** **** ***** ** ***** ** ***** ***** ** *** ** ***  *****

Orangutan       TTGATCTAACATCACCATCCTACTAACAGGACTCAACATACTAATCACAACCCTATACTC      300
Human           CTGATCAAATATCACTCTCCTACTTACAGGACTCAACATACTAGTCACAGCCCTATACTC      300
Gorilla         CTGATCAAATACCACCCTTTTACTTACAGGATCTAACATACTAATCACAGCCCTGTACTC      300
                 ***** ** * ***  *  **** ******   ********* ***** **** *****

Orangutan       TCTCTATATATTCACCACAACACAACGAGGTACACCCACACACCACATCAACAACATAAA      360
Human           CCTCTACATATTTACCACAACACAATGGGGCTCACTCACCCACCACATTAACAACATAAA      360
Gorilla         CCTTTATATATTTACCACAACACAATGAGGCCCACTCACACACCACATCACCAACATAAA      360
                 ** ** ***** ************ * **  *** *** ******** * *********

Orangutan       ACCTTCTTTCACACGCGAAAATACCCTCATGCTCATACACCTATCCCCCATCCTCCTCTT      420
Human           ACCCTCATTCACACGAGAAAACACCCTCATGTTCATACACCTATCCCCCATTCTCCTCCT      420
Gorilla         ACCCTCATTTACACGAGAAAACATCCTCATATTCATGCACCTATCCCCCATCCTCCTCCT      420
                *** ** ** ***** ***** * ******  **** ************** ****** *
```

Figure 1: Sample area of aligned genomes

2.1 Weighted Directed Graph for each DNA sequence –

The order with alphabet letters A, G, C, and T create a DNA sequence and these alphabets are used to represent the four nitrogen bases. We can conduct a DNA sequence of length n as,

$X = x_1 x_2 x_3 \dots x_n$ where, $x_1, x_2, x_3, \dots x_n \in \{A, G, C, T\}$.

Suppose $G_x = (V(G_x), E(G_x))$ is the double directed weighted graph of DNA sequence X. Let $V(G_x) \in \{A, G, C, T\}$ be the vertex set and $E(G_x)$ be the edge set. Edges are established by considering each pair of nucleotides x_i and x_j in the sequence with $i < j$. The direction of the edge indicates which nucleotide comes first in the sequence and the loop in the graph signifies how many times the same nucleotide comes in the sequence. The weights of the edges are given by the following equation,

$$weight = \frac{1}{(j - i)^{\alpha}}$$

where, $i < j$ and $\alpha > 0$. Above function is a decreasing function. Since the maximum weight of an arc for any α is just 1, it should be considered that arcs with weights not less than 0.1 are relatively significant. $(j - i)$ would reflect the fact that the two nucleotides with smaller distance will have a stronger interactive relationship than the two nucleotides have a larger distance.

When assigning the weights of edges, it is important to choose α as weights not less than 0.1 to construct the representative vector. We can choose α according to the length of DNA sequence. If we take a long sequence, it is precise getting value for α like $\alpha = 1/2$ or } $\alpha = 1/3$ and for a short sequence } $\alpha = 1$ or $\alpha = 2$ are applicable. An example of constructing a weighted directed graph for a given DNA sequence is given below.

Suppose $X = ATCTGATA$ is a DNA sequence with 8 nucleotides and $\alpha = 1$ since it is a very short sequence.

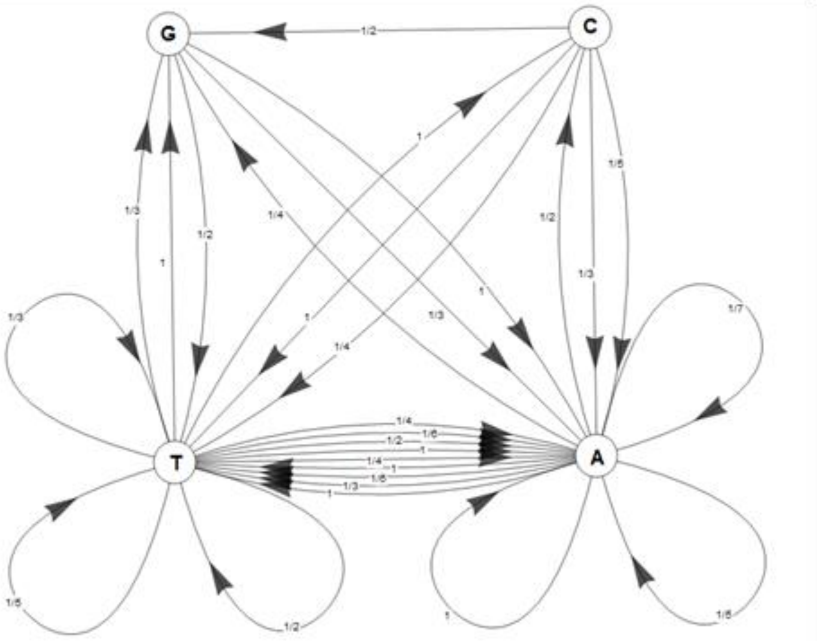


Figure 2: The weighted directed graph for sequence X

Theorem

There is a one-to-one mapping between a DNA sequence X and its corresponding weighted directed multi graph G_x .

In graph G_x there are several parallel edges that connect from one vertex to another in same direction. Thus, we can simplify the graph G_x to G_y by merging parallel edges into one edge. Since the vertex set is not changed, $V(G_x) = V(G_y)$. Suppose

$S^x_{u,v}$ is the set of all edges from vertex u to v in G_x . If $S^x_{u,v} \neq \emptyset$, an edge is assigned from u to v in G_y. The weight of that edge is given by,

$$w_y = \sum_{(u,v) \in S^x_{u,v}} w_x(u,v)\, , \quad S^x_{u,v} \neq \emptyset$$

Figure 3 shows the simplified weighted directed graph for DNA sequence X.

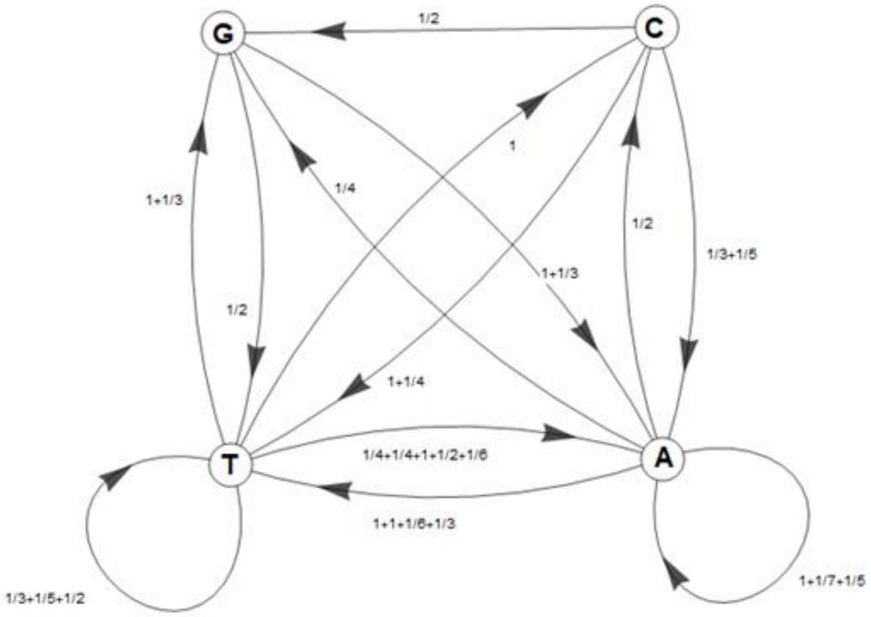


Figure 3: The simplified weighted directed graph for sequence

The next step of the research is making the corresponding adjacency matrix for each graph.

2.2. Adjacency matrix and representative vector – The adjacency matrix corresponding to the weighted directed graph G_y is defined as,

$$\begin{bmatrix} w_y(A,A) & w_y(A,C) & w_y(A,G) & w_y(A,T) \\ w_y(C,A) & w_y(C,C) & w_y(C,G) & w_y(C,T) \\ w_y(G,A) & w_y(G,C) & w_y(G,G) & w_y(G,T) \\ w_y(T,A) & w_y(T,C) & w_y(T,G) & w_y(T,T) \end{bmatrix}$$

For every DNA sequence, there is a 4×4 square matrix as the adjacency matrix. Each element of the matrix gives the interaction between two different nucleotides in the sequence.

Then we convert the 4×4 matrix to a 16-dimensional row vector $\vec{R}$. It is the representative vector for the DNA sequence.

$$\vec{R} = [w_y(A,A) \;\; \dots \;\; w_y(A,T) \;\; w_y(C,A) \;\; \dots \;\; w_y(C,T) \;\; w_y(G,A) \;\; \dots \;\; w_y(G,T) \;\; w_y(T,A) \;\; \dots \;\; w_y(T,T)]$$

The comparison of DNA sequences is converted to the comparison of 16- dimensional vectors.

This is one of the vector representations of DNA sequence that we used in this study. The representative vector is used to determine the similarities of corresponding distance measurements.

2.3. Similarity measurements–

The representative vector is used to determine the degree of similarity between two sequences.

When the distance is in the range [0,1], the common relationship between the distance (dissimilarity) and similarity is,

similarity = 1 - distance

The smaller distance reflects that corresponding DNA sequences are more similar. We can determine the degree of similarity using the following distance measurements. The first four measurements can be interpreted using the vector structure [4]. The last three measurements can be interpreted by the molecular structure of DNA complex molecule.

We take two different DNA sequences X & .Y with the same length and $\overrightarrow{R_X}$, $\overrightarrow{R_Y}$ are the corresponding 16- dimensional representative vectors.

1. Euclidean distance :

Euclidean distance gives the distance between the endpoints of two vectors. It is the shortest distance between two points along the hypotenuse.

2. Cosine distance :

Cosine distance is called Angular distance and it measures the cosine of the angle between two vectors.

3. Correlation distance :

The linear correlation similarity coefficient measures the dependence between the two vectors.

4. Manhattan distance or City-block distance :

The Manhattan distance is the distance that would be traveled from one endpoint of vector to another if a grid like a path.

In addition to above distance measurements, we can determine the similarities among DNA sequences using molecular similarity coefficients and compare with above results. Molecular similarity mainly focuses on the structural features of compounds and their representations such as shared substructures, ring systems, topologies and etc. We can apply these similarity coefficients to DNA compounds. Some coefficients that we apply here: Soergel distance, Jaccard distance, Dice distance. [1].

The spectrum of the graph was used as another vector representation of the DNA sequence.

2.4. Spectrum as a vector–

The spectrum of a graph is extensively used in graph theory to characterize the properties of a graph and gather information from its structure. Obviously, the spectrum may be changed with the small change of a graph structure. The graph spectrum is derived from a matrix representation of the graph and depends on the form of the matrix[5].

In this research, each spectrum is obtained from the corresponding 4×4 adjacency matrix of each graph. We use the spectrum as the .1×4 row vector since spectrum can be considered as the measure of graph similarity.Then the comparison between two DNA sequences was converted to the comparison between two 1×4 row vectors. Euclidean, Cosine, and Correlation are three distance measurements that use to measure the similarity using the spectrum.

III. Results and Discussion

We randomly chose regions throughout the whole genomes of three species including conserved regions and DNA variations to check whether the same similarity result can be obtained from any area of the genome. Each DNA sequence that used here consists of the length of 12 nucleotides and $\alpha = 1$ for in whole the calculations. In the following table the distance measurements are shown in pair wisely for a region that was chosen randomly.

Orangutan

ATCCCTCAACCCCAGCA**TCATCGCTGGGT**TCGCCTACTGTAAATATAGTTTAACCAAAAC 480

Human

ATCCCTCAACCCCGACA**TCATTACCGGGT**TTTCCTCTTGTAAATATAGTTTAACCAAAAC 480

Gorilla

ATCCCTCAACCCCGATA**TTATCACCGGGT**TCACCTCCTGTAAATATAGTTTAACCAAAAC 480

************* ** ** * ***** ***

****

Figure 4: Randomly chosen region with conserved and DNA variations

Table -1 Distance measurements

Species	**Distance measurements using representative vector**				**Molecular similarity measurements**			**Distance measurements using spectrum**		
	Euclidean	**Cosine**	**correlation**	**Manhattan**	**Soergel**	**Jaccard**	**Dice**	**Euclidean**	**Cosine**	**Correlation**
Human and Orangutan	**3.7362**	**0.1143**	**0.3766**	**12.338**	**3.2506**	**0.2095**	**0.1170**	**3.3873**	**0.0983**	**0.1166**
Human and Gorilla	***1.4236***	***0.0183***	***0.0640***	***3.9214***	***1.0331***	***0.0360***	***0.0183***	***1.0050***	***0.0049***	***0.0105***
Gorilla and Orangutan	**4.4499**	**0.1624**	**0.5197**	**13.9752**	**3.9553**	**0.2822**	**0.1643**	**4.1488**	**0.1355**	**0.1750**

The above three cases predict the same result that is the DNA sequences of Human and Gorilla are very similar because they have the smallest distance values for all distance measurements (Table-1). When comparing the results of the representative vector with the results of the spectrum both cases give the same result. Although most of the aligned sequences give a positive result, there are some sequences with same length are failed in this methods. Even the representative vector based method is failed in some regions when the sequence length is very short. In some DNA regions, although the representative vector-based method is passed, the spectrum-based method is failed. We can get positive results for failed regions when these regions are reused after increasing the number of nucleotides. It is difficult to build a weighted directed graph when increasing the number of nucleotides in the sequence. But gradually when the number of nucleotides in sequences is increasing, there is a high probability to give accurate results than short length DNA sequences. Sometimes the length of the sequence that we use is not enough for similarity analysis between sequences because of that the short length sequence may not be stored enough DNA variations and mutations to analyze the similarities.

IV.CONCLUSION

We applied the novel method to determine the evolutionary closeness between Human, Gorilla, and Orangutan. A unique weighted directed graph for each species is represented by using DNA fragments. The weights of arcs are known as the entries of the adjacency matrix of the weighted graph. The adjacency matrix is written as the vector form is called as a representative vector. Representative vectors of each pair of DNA sequences are used to calculate the distance measurements such as Euclidean, Cosine, Correlation, and Manhattan. The molecular similarity measurements Soergel, Jaccard and Dice are also applicable to analyze the similarity using Representative vector. Instead of the representative vector, we used the spectrum of each graph obtained from the adjacency matrix of the weighted graph since the spectrum characterizes the properties of the graph. Using the spectrum, the Euclidean, Cosine and Correlation distances were accurate with distances using representative vector. The method that used the spectrum was not more accurate for very short DNA fragments. Final calculations using both vectors become more accurate when increasing the number of nucleotides in the DNA fragments. Then this research concludes that any DNA fragments with conserved regions and DNA variations in aligned genomes of different species can be applied to detect similarity. Molecular similarity measurements are also applicable as distance measurements between DNA sequences. Spectrum of the graph is one of the suitable vector representations of DNA sequence when sequence is not very short.

REFERENCES

[1] Andrew R Leach and Valerie J Gillet. *An introduction to chemoinformatics*. Springer, 2007.

[2] YA Lesnussa, S Kappuw, BP Tomasouw, and ER Persulessy. The similarity analysis of dna sequence model based on graph theory and blast program. *EDUCATUM Journal of Science,*

[3] *Mathematics and Technology,* 4(1):41–51, 2017.

[4] Rinku Mathur and Neeru Adlakha. A graph theoretic model for prediction of reticulation events and phylogenetic networks for dna sequences. *Egyptian Journal of Basic and Applied Sciences,* 3(3):263–271, 2016.

[5] Xingqin Qi, Qin Wu, Yusen Zhang, Eddie Fuller, and Cun-Quan Zhang. A novel model for dna sequence similarity analysis based on graph theory. *Evolutionary Bioinformatics,* 7:EBO– S7364, 2011.

[6] Richard C Wilson and Ping Zhu. A study of graph spectra for comparing graphs and trees. *Pattern Recognition,* 41(9):2833–2841, 2008.