International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Mathematical Analysis of DNA Curves

R. Sengothai

Ph.D. Research Scholar in Mathematics

University of Madras Chennai

ABSTRACT

Effective representation of DNA sequences is one of the important tasks in the study of genome sequences. In this paper, we propose a graphical representation of DNA sequences based on nucleotide ring structure. In the proposed representation, we convert DNA sequences into 16 dinucleotides on the surface of the hexagon so that it can preserve nucleotide's chemical property and positional information. Our approach can provide capability of efficient similarity comparison between DNA sequences and also high comparison accuracy. Furthermore, our approach satisfies uniqueness and no degeneracy of DNA sequences. In the experimental study, we use phylogeny analysis for evolutionary relationship among different species. Extensive performance study shows that the proposed method can give better performance than existing methods in comparison with the degree of similarity. In this paper, we converted DNA sequences into DNA curves without any loss of information and degeneracy.

Keywords: β-globin gene, DNA curve, hexagon, ring structure

I. INTRODUCTION

The rapid growth of biological sequences, such as of DNA, RNA, and protein, has demanded effective analysis methods for large biological sequences. Additionally, the analysis results are very helpful to biological researchers for predicting genes' structure and function, as well as similarity comparison between genes and different species.

For biological sequence analysis, two approaches have been mainly used: (i) sequence alignment method and (ii) non-sequence alignment method. The first approach obtains a degree of similarity between DNA sequences by comparing alignment scores of two sequences. This approach suffers from expensive computational cost as the length of sequences grows exponentially. The second approach analyzes DNA sequences by establishing a statistical model or a graphical representation model, or some machine learning model of DNA sequences. Recently, this approach is popularly studied due to the fact that it can give better accuracy and low computational overhead.

In the case of non-sequence alignment method, effective DNA sequence representation or feature selection from DNA sequences is essential for DNA sequence analysis, in areas such as gene prediction, similarity comparison between genes of different species, and finding gene structure and function. For this purpose, several graphical representations have been proposed according to chemical structures of 4 nucleotides, reflecting their distribution with different chemical structure and allowing numerical characterization. As for feature selection,

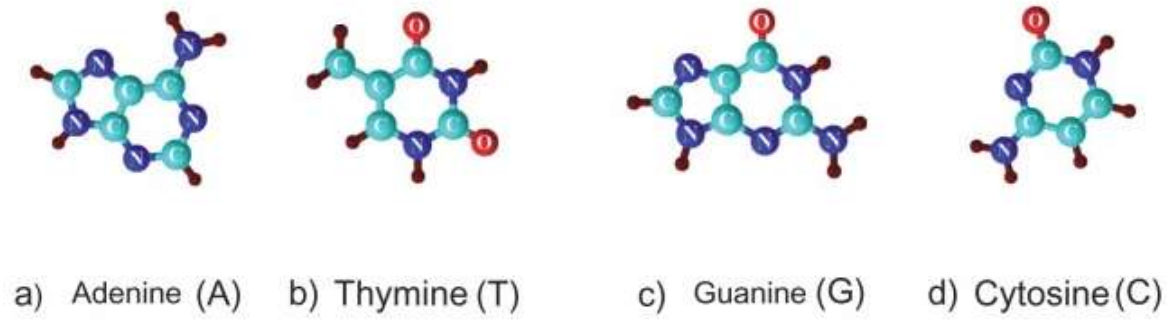
several machine learning techniques are effectively applied such as principal component analysis (PCA), neural network, and several classification models.

In this paper, we propose a graphical representation of DNA sequences based on nucleotide ring structure. In the proposed representation, we convert DNA sequences into 16 dinucleotides on the surface of the hexagon so that the nucleotide's chemical property and positional information is preserved. Our approach satisfies uniqueness and no degeneracy of DNA sequence is observed. It can also provide capability of efficient similarity comparison between DNA sequences in addition to high comparison accuracy. Extensive performance study shows that the proposed method can give better performance than existing methods in comparison with the degree of similarity.

II. DNA SEQUENCE VISUALIZATION BY HEXAGONAL STRUCTURE

2.1 Chemical structure and classification of DNA bases

As stated previously, DNA sequences are the strings of four bases, that is, A, T, C and G. The core of these bases is heterocyclic organic compound, which forms ring in their chemical structure. Of them purines (A and G) have two rings while pyrimidines (C and T) have one. The chemical ring structures of those four bases are depicted in Figure 1.



1. Heterogenic cycle of four bases

The element of these cycles are carbon and nitrogen, hydrogen and oxygen. In Figure 1, sky balls are carbon and blue balls are nitrogen. The hexagonal cycle has nitrogen in positions 1 and 3, and carbon in positions 2, 4, 5, and 6. Other than carbon and nitrogen, the bases have oxygen and hydrogen bonded with carbon and nitrogen in different number. Hence, the bases differ in molecular weight. The molecular weight of A, T, C and G are 135.13, 112.1, 111.1 and 151.13 respectively. Their ascending order in terms of molecular weight is C→T→A→G.

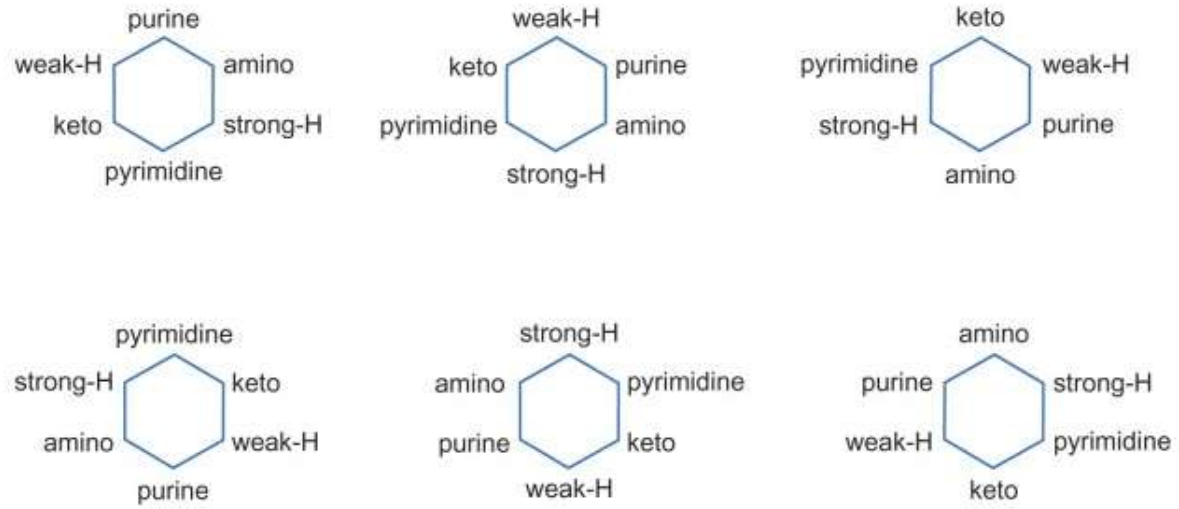
The bases also differ by heterogenic cycle, functionality, and their bonding with hydrogen. A and C fall into the amino category while G and T are in the keto group, based on their functionality. A and T are bonded by three hydrogen bonds, and hence are in strong-H group while G and C are in weak-H group as they are bonded by only two hydrogen bonds.

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

2.2 Proposed DNA encoding

The proposed encoding of dinucleotides for DNA sequence visualization is solely based on ring structure of DNA bases and their molecular weight. The bases are paired to make dinucleotides in such a way that their ascending/descending order in terms of molecular weight remains intact. The dinucleotides are placed on the 6 end of the heterocyclic hexagon as well as at the midpoint of each arm of the hexagon. The six dinucleotides which are placed on the 6 ends of the hexagon are in ascending order. The midpoint dinucleotides are positioned by descending order of molecular weight.

We place six ordered dinucleotides on opposite ends of the heterogenic cycle. The opposite ends are 1–4, 2–5, and 3–6. Any class (purine, pyrimidine; amino, keto; strong-H, weak-H) can be positioned at either end of the hexagon. Therefore, there are six possible combinations, as shown in Figure 2. The names of these combinations are Cycle 1, Cycle 2, Cycle 3, Cycle 4, Cycle 5, and Cycle 6 respectively.

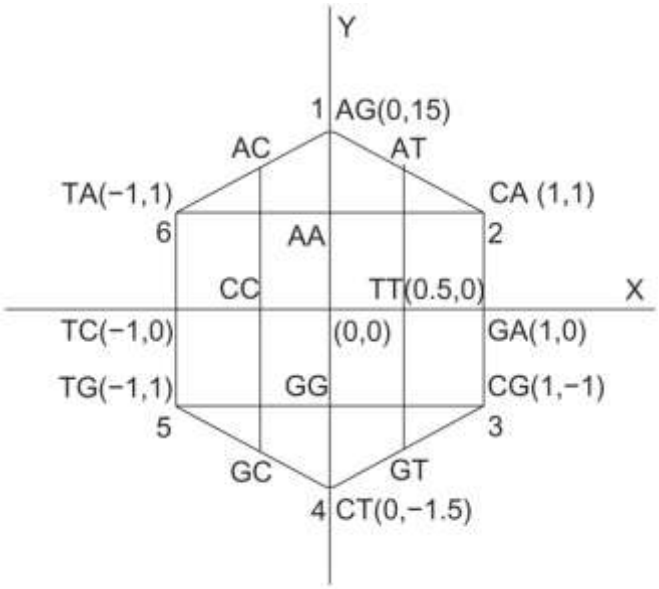


2. Six combinations of heterogenic cycle in 2D space

In Cycle 1, purines and pyrimidines are positioned at ends 1 and 4 of the hexagon respectively. A and G, the purines, form two dinucleotides: AG and GA. We keep AG on end 1 as it retains the sequence C→T→A→G. For the same reason, CT are placed on end 4, and CA (amino) and TG (keto) are placed on the 2 and 5 ends, respectively, and CG (strong-H) and TA (weak-H) are positioned on the 3 and 6 ends of the hexagon, respectively.

Conversely, midpoint of 2–3, 3–4 and 4–5 arms are determined by the following rule: take the uncommon nucleotides and form a dinucleotide with them such that descending order (G→A→T→C) of molecular weight prevails. As for example, the midpoint of 2–3 arms is GA because the commonality between CA and CG is C. So, G and A are uncommon. This rule is different for the midpoint of 5–6, 6–1 and 1–2 arms: take the common

nucleotide as well as the other which is not available on both ends. For example, the midpoint of 5–6 arms is TC because T is common between TA and TG, while C is neither in TA nor in TG. We follow these simple rules to position the 12 dinucleotides on the hexagon (six dinucleotide on six ends + six dinucleotide on midpoint of each arm of the hexagon). Based on the above discussions, Cycle 1 is drawn in the 2D Cartesian space, shown in Figure 3.



3. Cartesian coordinates of 16 dinucleotide in a hexagon

From Figure 3, we can derive the set of position coordinates of 16 dinucleotide: (0, 1.5) → AG, (0.5, 1.25) → AT, (1, 1) → CA, (1, 0) → GA, (1, −1) → CG, (0.5, −1.25) → GT, (0, −1.5) → CT, (−0.5, −1.25) → GC, (−1, −1) → TG, (−1, 0) → TC, (−1, 1) → TA, (−0.5, 1.25) → AC, (0, 1) → AA, (−0.5, 0) → CC, (−1, 0) → GG, (0.5, 0) → TT.

Let $S = \{s_1, s_2, \ldots, s_N\}$ be a DNA sequence where $s_i \in \Sigma = \{A, T, C, G\}$ and $i = 1, 2, 3, \ldots, N$. S is mapped into a series of points $P_1, P_2, \ldots, P_{N-1}$. We introduce a map function φ such that S can be formulated as $S = \varphi(s_i s_{i+1})\varphi(s_{i+1}s_{i+2}) \ldots \varphi(s_{N-1}s_N)$ where,

$$P_i = \phi(s_i \; s_{i+1}, i) = \phi(x_i, y_i, i) = \phi(x_{s_i s_{i+1}}, y_{s_i s_{i+1}}, i), \qquad i = 1,2,3,\ldots, \; N-1$$

x_{sisi+1}, y_{sisi+1} and i represent the x-coordinate, y- coordinate, and z-coordinate respectively. Thus, we connect the N-1 points from the first one and derive a 3D curve.

To locate the local and global features of the 3D curve as well as to visualize the 3D representation of this curve, we take another numerical representation. Let $X_i=\sum_{k=11}x_k, Y_i=\sum_{k=11}y_k$, we derive another mapping function for cumulative feature of the 3D curve such that

$\lambda(s_is_{i+1}) = (X_i, Y_i, i)$, where $i = 1, 2, 3, \ldots, N - 1$.

Connecting N-1 points from the first one, we get the proposed novel 3D zigzag curve.

2.3 Example of the proposed method

The following example is used with the arbitrary DNA sequence ATACGATGCAG. The length of the string is 11, hence there are 10 di-nucleotide. The 3D coordinate for all cycles of the sequence is shown in Table 1.

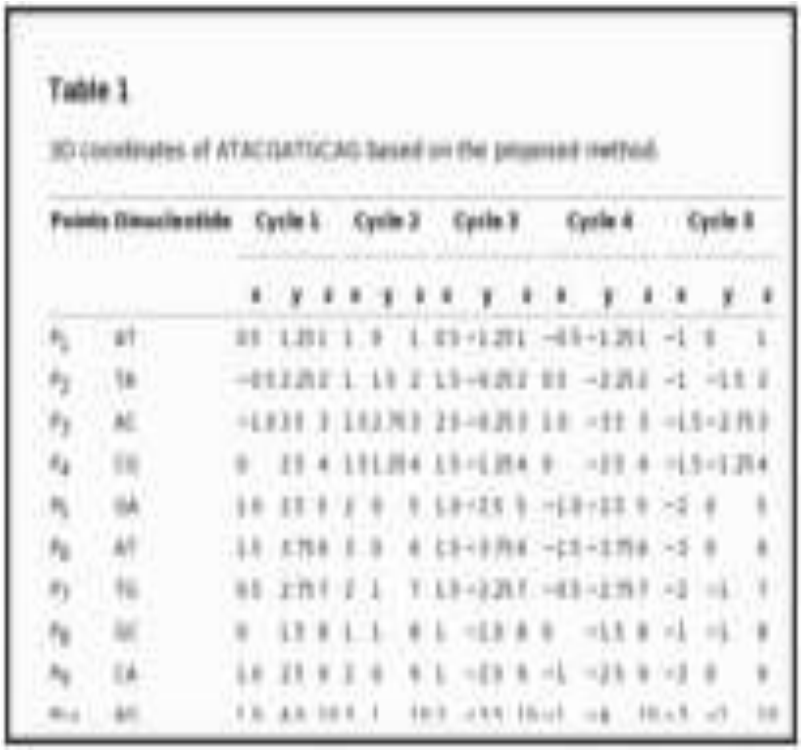


Table 1. 3D coordinates of ATACGATGCAG based on the proposed method

As for graphical representation, the 10 points P_1, P_2, …, P_{10} are plotted in 3D space for the example sequence ATACGATGCAG. The six possible DNA curves for the example sequence are shown in Figure 4.

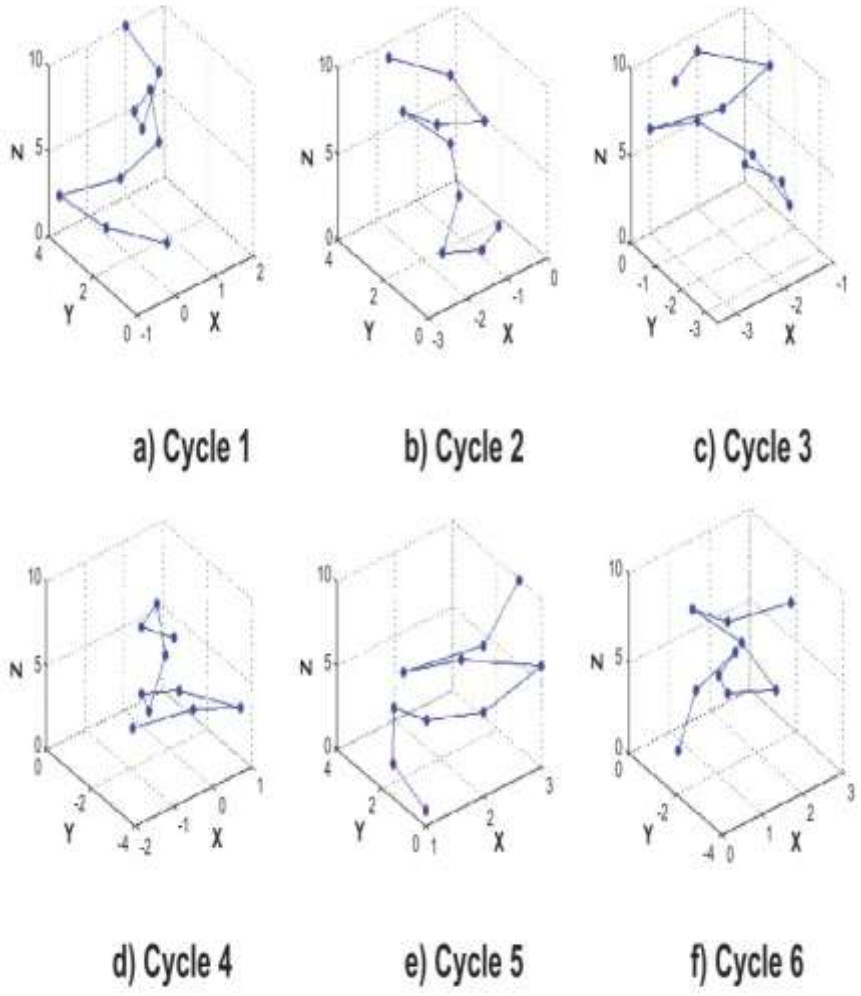


Figure 4. The graphical representation of the proposed model for the example sequence ATACGATGCAG

In this way, each DNA sequence is converted into a series of points. Then DNA curves are drawn from those points. Connecting N-1 points from the first one, we get the proposed novel 3D zigzag curve in the 3D space. The DNA curve is helpful to easily distinguish among different species.

It can easily be seen that the example graphical representation does not hold any overlapping or loop. This property will be retained for any DNA sequence because the value of "i" in the proposed method is unique in every point.

III. GRAPHICAL REPRESENTATION OF THE PROPOSED METHOD

The proposed model is useful to show the hidden properties of long DNA sequences which are not seen from the sequence. The pictorial presentation of the proposed method proves that it is very useful to understand the evolutionary similarity/dissimilarity of different species. Figure 5 shows the 3D zigzag curve based on Cycle 1 of first exon of β-globin for 11 different species. The graphical representation clearly shows that:

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

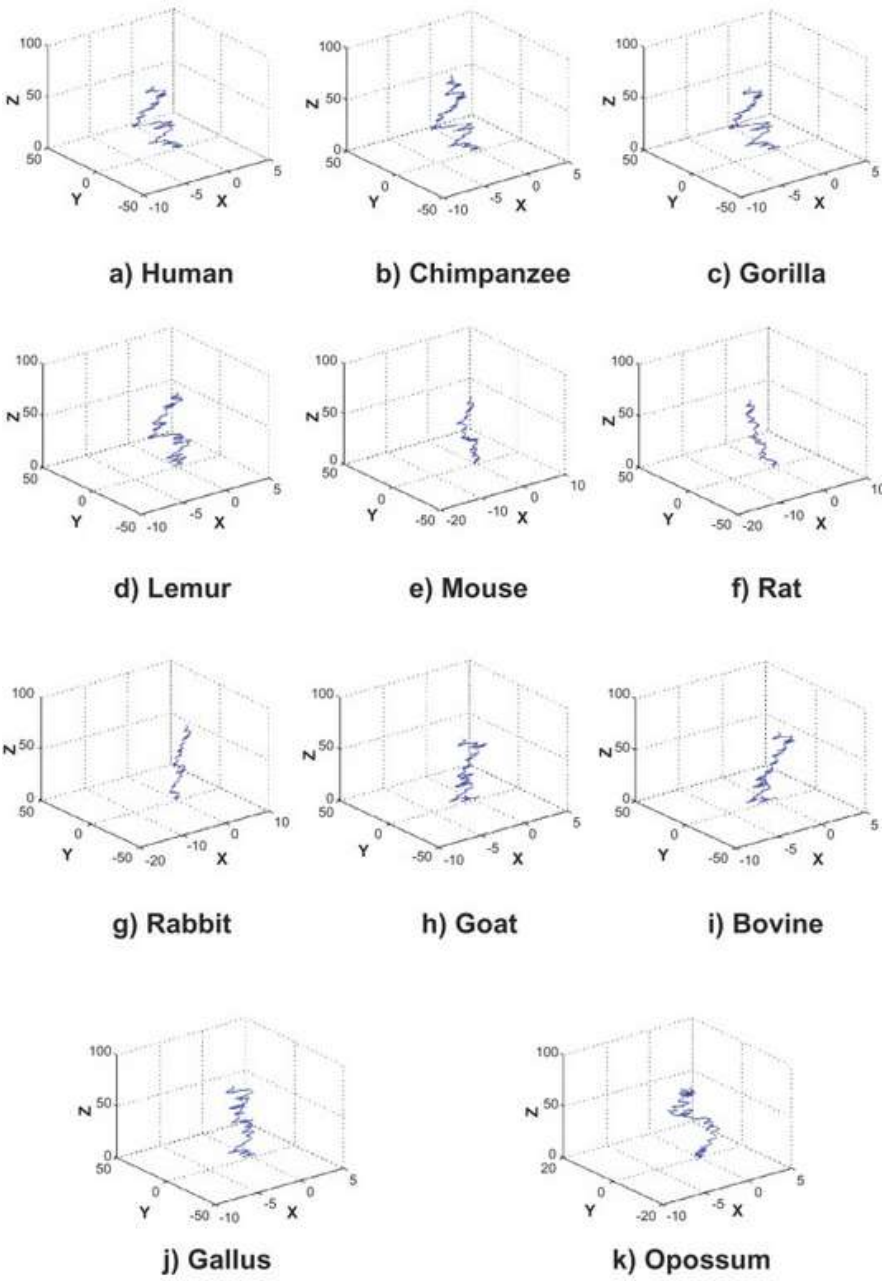


Figure 5. DNA curves of 11 different species

i.DNA curves of human, gorilla, chimpanzee and lemur are closely similar;

ii.Mouse and rat have also same DNA curves, so as rabbit's DNA curve;

iii.Goat and bovine are similar; and that

iv.Gallus and opossum seem to be outliers.

IV.EXPERIMENTAL ANALYSIS

4.1 Performance metric, dataset and experimental environment

To evaluate the performance, we illustrate the use of the proposed method with an examination of similarities/ dissimilarities among the β-globin gene of 11 different species, listed in Table 2, which were also previously studied. The table shows the different important characteristics of the dataset. First, we show the overall performance of the proposed method. To do this, two features are extracted from the DNA curves: (i) geometric center and (ii) mathematical descriptor. Each DNA sequence is finally represented by their mathematical descriptors. These descriptors form six dimensional feature vectors. After that, the Euclidian distance is calculated among feature vectors of the DNA sequences. Secondly, we draw the phylogenic tree from similarity/dissimilarity matrix using UPGMA method in PHYLIP package. Finally, we compare the proposed method with the already mentioned research works to show its superiority to others.

Table 2

The first exon of β-globin gene of 11 different species.

Species	ID/Accession	Database	Length
Human	U01317	NCBI	92
Chimpanzee	X02345	NCBI	105
Gorilla	X61109	NCBI	93
Lemur	M15734	NCBI	92
Rat	X06701	NCBI	92
Mouse	V00722	NCBI	93
Rabbit	V00882	NCBI	92
Goat	M15387	NCBI	86
Bovine	X00376	NCBI	86
Opossum	J03643	NCBI	92
Gallus	V00409	NCBI	92

Table 2. The first exon of β-globin gene of 11 different species

4.2 Numerical analysis of the proposed method

As stated earlier, features from DNA curves are extracted two ways. Firstly, the geometric centers of the curves are calculated using the following equations. Table 3 shows the geometric center of 11 DNA curves.

Table 3

Geometrical center of 11 different species.

Species	Cycle 1			Cycle 2			Cycle 3			C[illegible]
	μ_x	μ_y	μ_z	μ_x	μ_y	μ_z	μ_x	μ_y	μ_z	μ_x
Human	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Chimpanzee	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Gorilla	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Lemur	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Rat	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Mouse	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Rabbit	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Goat	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Bovine	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Opossum	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 3. Geometrical center of 11 different species.

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

$$u_x=1N\sum_{i=1N}x_i, \qquad u_y=1N\sum_{i=1N}y_i, \qquad u_z=1N\sum_{i=1N}z_i$$

The significance of the geometric center is that it shows the average value of x, y and z coordinate. Normally, if the geometric centers are plotted, then similar species fall into same cluster. So, the geometric center is an important feature of the DNA curves for the analysis of evolutionary relationship among different species.

To make our system unbiased, we take all the possible rotations (such as Cycle 1, Cycle 2, …, Cycle 6) of hexagonal ring structure and extract the geometric center for each combination.

Secondly, mathematical descriptors are obtained from the first feature, the geometric center, using the following equation. Table 4. shows the mathematical descriptor of the 11 curves.

Table 4. Mathematical descriptor of 11 different species.

We use the Euclidian distance for similarity measurement. Let two different species be i and j. The mathematical descriptor of i are p_{1i}, p_{2i}, p_{3i}, p_{4i}, p_{5i} and p_{6i}. The same descriptors for species j are p_{1j}, p_{2j}, p_{3j}, p_{4j}, p_{5j} and p_{6j}. The Euclidian distance of i and j are then calculated using the following equation:

The similarity/dissimilarity matrix found from the above Euclidian distance metric is shown in Table 5.

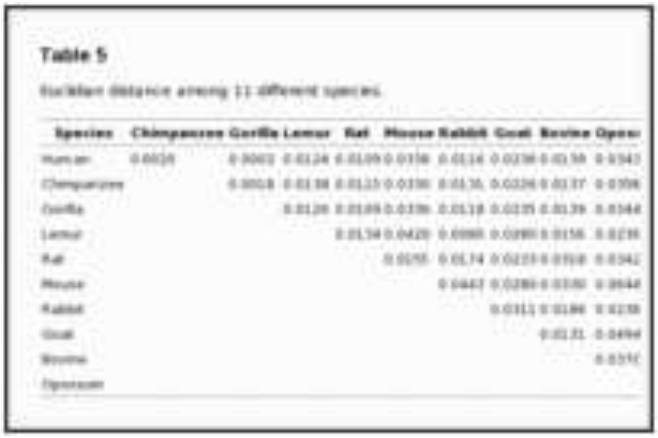


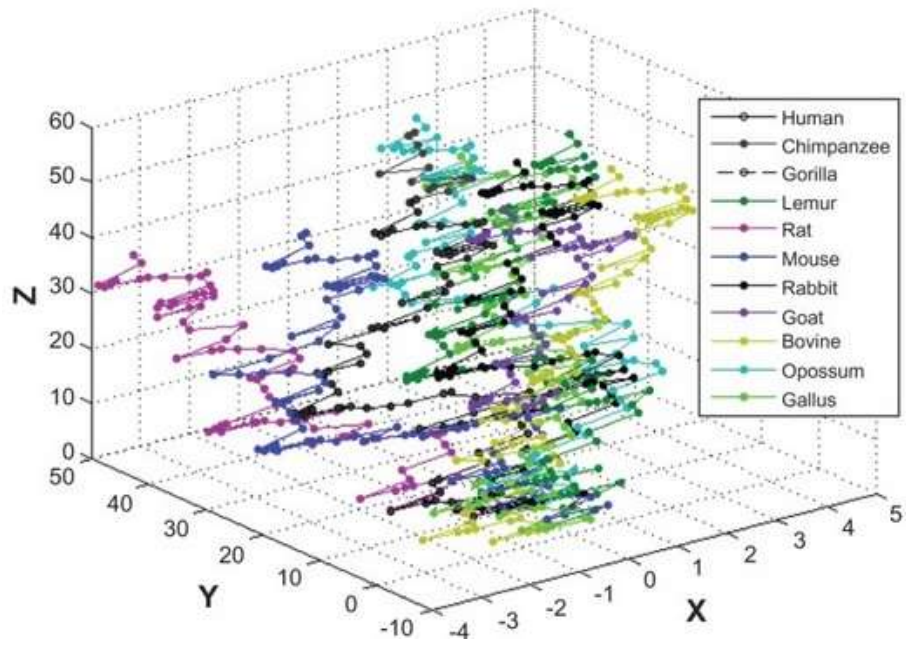
Table 5. Euclidian distance among 11 different species

Some observations are vividly depicted from Table 5 which are also consistent with the graphical representation portrayed in Section 3.4. They are as follows:

i. The smallest entry is 0.0002 for the pair (human, gorilla), showing that human and gorilla are almost same in terms of evolutionary characteristics. The same is applied for the pair (human, chimpanzee) = 0.0020. Therefore, human, chimpanzee and gorilla are similar species;
ii. The pair (goat, bovine) has the small entry 0.0131 which indicates the evolutionary similarity between goat and bovine. The biological taxonomy of bovine and goat proves that both of them are even-toed ungulates and belong to the family of "Bovidae";[16]
iii. Rat and mouse also show a small entry which indicates their evolutionary closeness;
iv. The remote mammalian opossum has the largest entry to all other mammalians.

V.CONCLUSION

A graphical method based on dinucleotides and their positional information is proposed in this research work. Graphical as well as numeric analyses of the model show that the proposed novel method is compatible with the natural consistency in terms of evolutionary relationship of 11 different species. In this paper, DNA sequences are transformed into 3D DNA curves, and features from those curves are then extracted. DNA curves are represented by their feature vector. Subsequently, Euclidian distance is applied to those feature vectors to deduce the evolutionary relationship among 11 different species. Tri-nucleotide based DNA sequence analysis using the proposed method would be one recommended future work.



DNA curves of 11 different species in one graph

REFERENCES

[1.] Hamori E, Ruskin J. H curves: a novel method of representation of nucleotide series especially suited for long DNA sequences. J Biol Chem. 1983;258(2):1318–27.

[2.] Li Y, Huang G, Liao B, Liu Z. H-L curve: a novel 2-D graphical representation of protein sequences. MATCH Commun Math Comput Chem. 2009;61(2):519–32.

[3.] Guo X, Randic M, Basak SC. A novel 2-D graphical representation of DNA sequences of low degeneracy. Chem Phys Letter. 2001;350(1–2):106–12.

[4.] Jafarzadeh N, Iranmanesh A. A novel graphical and numerical representation for analyzing DNA sequences based on codons. MATCH Commun Math Comput Chem. 2012;68:611–20.

[5.] Yu JF, Wang JH, Sun X. Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation. MATCH Commun Math Comput Chem. 2010;63:493–512.

[6.] Liao B, Zhu W, Liu Y. 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree. MATCH Commun Math Comput Chem. 2006;56(1):209–16.

[7.] Cao Z, Liao B, Li R. A group of 3D graphical representation of DNA sequences based on dual nucleotides. Internat J Quant Chem. 2008;108(9):1485–90.

[8.] Chi R, Ding K. Novel 4D numerical representation of DNA sequences. Chemical Physics Letters. 2005;407(1–3):63–7.

[9.] Liao B, Li R, Zhu W, Xiang X. On the similarity of DNA primary sequences based on 5-D representation. J Math Chem. 2007;42(1):47–57.

[10.] Liao B, Wang T. Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. J Chem Inf Comput Sci. 2004;44(5):1666–70.

[11.] Wu R, Hu Q, Li R, Yue G. A novel composition coding method of DNA sequence and its application. MATCH Commun Math Comput Chem. 2012;67:269–76.

[12.] Qi X, Wu Q, Zhang Y, Fuller E, Zhang CQ. A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinform Online. 2011;7:149–58.

[13.] Ewens J, Grant G. Statistical Methods in Bioinformatics: An Introduction. 2nd ed. New York: Springer Science; 2005.

[14.] Randić M, Vračko M, Lerš N, Plavšić D. Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett. 2003;368(1–2):1–6.

[15.] Randić M, Vračko M, Lerš N, Plavšić D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. Chem Phys Lett. 2003;371(1–2):202–7.