Abstract-- Introduction of graphic representation for nucleotide or protein sequences can provide intuitive overall pictures as well as useful insights for performing large-scale similarity analysis. In this paper, we are analyzing the similarity/dissimilarity of the mitochondrial genome sequences from twenty-four mammal species. The analysis is important in finding the relatedness among the species and eventually finding the evolutionary relationship. The evolutionary tree or phylogenetic tree is constructed by Unweighted Pair Group Method with Arithmetic Mean (UPGMA). The graphical representation of DNA sequence using Adjacent Nucleotide Pair has been constructed. The distance matrix required for the construction of phylogenetic tree has been built by applying the biological geometry method.

Keywords—Graphical representation; similarity analysis; sequence comparison

I. INTRODUCTION

In bioinformatics one of the primary challenges is dealing with large volume of genomic sequences. A lot of computational and statistical techniques have been proposed by researchers to compare the nucleotide or protein sequences. Existing methods for analyzing similarity/dissimilarity of biological such sequences can be broadly classified into two categories: (i) Alignment methods and (ii) Alignment free methods. Multiple sequence alignment requires the alignment of more than two sequences which is computationally difficult. Therefore an alignment free approach is came into existence in 1983 by Hamori and Ruskin [1]. The graphical representation of a DNA sequence facilitates viewing, sorting and comparing various gene sequences with intuitive pictures and pattern. Many other graphical representations have been proposed subsequently [2-5].

The graphical representation methods to display sequences have the advantage of visual indications of trends and inherent features. Hamori and Ruskin [1] first proposes a 3D graphical representation (H-curve) for DNA sequences which can be displayed and manipulated conveniently. Some different graphical approaches representing DNA sequences have been reported by quite a lot of authors. Gates [3] presented a canonical choice of vectors which has certain advantages, both for exploring structure and for comparing sequences. Thereafter Leong and Morgenthaler [4] designed random walk plot of DNA sequence to show the base composition in a compact form, whereas the gap plot visualizes positional correlations. The familiar dot-matrix types of graphs have been widely used to determine systematic in DNA and protein sequences. The idea is to read a DNA sequence base by base and plot succeeding points on the graph with four orthogonal unit vectors representing four kinds of bases. The different graphical representation needs numerical characterization to identify the regions of biological interest. Suitable mathematical descriptor helps in finding the numerical (quantitative) representation of similarity/dissimilarity between the sequences, and gives distance matrix for the set of sequences. The elements of similarity/dissimilarity matrix are used to construct phylogenetic tree. Some means for comparing phylogenies are desirable in order to assess the quality of phylogenetic trees from different construction methods and that are capable of enlightening where two trees agree or differ.

This paper finds the similarity/dissimilarity of the mitochondrial genome sequences from twenty-four mammal species that uses graphical representations and numerical characterization for measuring similarity between sequences. The graphical representation of DNA sequence has been constructed using Adjacent Nucleotide Pair (ANP) [6] and the numerical characterization is performed by moment vectors [7, 8].

II. MATERIAL AND METHOD

A. Graphical representation of DNA sequence
The representation of the sequence of bases in a DNA or RNA using graphical methods was initiated several years ago with a three dimensional model. We have taken the alignment free technique proposed by Gupta et al., [6] for obtaining the graphical representation of DNA sequence. The vectors corresponding to each of the sixteen ANPs have been created and the coordinates have been assigned to them based on chemical properties of nucleic bases, Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). The vectors corresponding to 16 dual nucleotide pairs are as follows: TC\( \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \), CT\( \left( \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right) \), GT\( \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \), TG\( \left( \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right) \), GC\( \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \), CG\( \left( \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right) \), CC\( \left( 1, 0 \right) \), TT\( \left( 1, 0 \right) \), GG\( \left( 1, 0 \right) \), AA\( \left( 1, 0 \right) \), TA\( \left( 1, 0 \right) \), AT\( \left( 1, 0 \right) \), CA\( \left( 1, 0 \right) \), AC\( \left( 1, 0 \right) \), AG\( \left( 1, 0 \right) \), GA\( \left( 1, 0 \right) \). The value of x-coordinate is 1 for all ANPs, y-coordinate value has been assigned in first and fourth quadrant according to the chemical classification of nucleotides. The ANP method in [9] have shown the representation is more sensitive and informative as compared to other dual nucleotide representations [5, 10-12].

### B. Numerical characterization

The graphical representation and numerical characterization gave a rich view of the complexities of nucleotide sequences. Recently, the majority of work on quantitative measurement of similarity among DNA sequences has been proposed. A review of various numerical characterization majorly used by researchers have been remarkably summarized by Vinga and Almeida [13]. Different graphical representations of DNA, both 2D and 3D, lead us to get different numerical characterizations. We have used the concept of moment vectors [6-8, 14] for numerical characterization of given graphical representation. The construction of moment vector by graphical representation was originally proposed by Yau et al. [7] for proteins map. Later Yu et al. [8] has used it for DNA sequence.

An m-dimensional moment vector, where \( m = n-1 \), and \( n \) is the length of DNA sequence, is constructed for the proposed m-ANP graphical representation. The number of dimension is one less than the number of nucleotide bases because making pairs of \( n \) nucleotides will produce \( n-1 \) pairs. Moments are determined to use the global information in the sequence \( (1, y_1), (2, y_2), \ldots, (n-y_{n-1}) \) from \( n \) length genome sequence. The moments for DNA sequence is defined in [6] for ANPs as:

\[
M_i = \sum_{j=1}^{n} x_j y_j^i, \quad j = 1, 2, \ldots, n-1
\]

where \((x_i, y_i)\) represents the coordinate position of \(i^{th}\) ANP in the DNA sequence.

Using these moment vectors of DNA sequences, a genome space is constructed as a subspace in Euclidean space. Each genome sequence can be represented as a point in this space. Therefore, this genome space can be used to make comparative analysis to study the clustering and phylogenetic relationship among genomes. The biological (evolutionary) distance between two genomes can be obtained by computing the Euclidean distance among the corresponding points in the genome space. After obtaining the distance matrix for every pair of DNA sequence, we construct the phylogenetic tree by UPGMA method of MEGA 5 package [15].

### C. Data Set

The mitochondrial sequence data that we are using to analyze similarity is shown in Table 1. The detail description of twenty four complete mitochondrial

<table>
<thead>
<tr>
<th>Accession No</th>
<th>Length (nt)</th>
<th>Description</th>
<th>Accession No</th>
<th>Length (nt)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V00662</td>
<td>16569</td>
<td>H.sapiens</td>
<td>AY488491</td>
<td>16355</td>
<td>Buffalo</td>
</tr>
<tr>
<td>D38116</td>
<td>16563</td>
<td>pygmy chimpanzee</td>
<td>EU442884</td>
<td>16774</td>
<td>Wolf</td>
</tr>
<tr>
<td>D38113</td>
<td>16554</td>
<td>Chimpanzee</td>
<td>EF511003</td>
<td>16990</td>
<td>Tiger</td>
</tr>
<tr>
<td>D38114</td>
<td>16364</td>
<td>Gorilla</td>
<td>EF511002</td>
<td>16954</td>
<td>Leopard</td>
</tr>
<tr>
<td>X99256</td>
<td>16472</td>
<td>Common gibbon</td>
<td>X97336</td>
<td>16829</td>
<td>Indian rhinoceros</td>
</tr>
<tr>
<td>Y18001</td>
<td>16521</td>
<td>baboon</td>
<td>Y07726</td>
<td>16832</td>
<td>White rhinoceros</td>
</tr>
<tr>
<td>AY863426</td>
<td>16389</td>
<td>Vervet monkey</td>
<td>AJ224821</td>
<td>16866</td>
<td>African elephant</td>
</tr>
<tr>
<td>D38115</td>
<td>16389</td>
<td>Bornean orangutan</td>
<td>DQ316068</td>
<td>16902</td>
<td>Asiatic elephant</td>
</tr>
</tbody>
</table>
genome sequences along with the accession number is shown in Table 1, each of which has length of more than 16000 nucleotides. All these DNA sequences have been downloaded from the GenBank (http://ncbi.nlm.nih.gov/genbank).

III. Result

In this section, we report the results obtained in the form of evolutionary relationship (phylogenetic tree) for the sequence data shown in preceding section. The phylogenetic tree has been constructed from Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [16] method which uses the distance matrix as input. The distance matrix has been constructed from computing the Euclidian distance between the moment vectors. We have constructed 10D moment vectors for each of the DNA sequence. We plot these 10D moment vectors in genome space and shown by a point. The distance between these points in genome space of biological geometry is calculated as Euclidian distance. We observe in the phylogenetic tree shown in Fig. 1, the last nine mammals are grouped into a cluster because they are primates. This evolutionary relationship coincides with the phylogenetic relationship among them found by Raina et al. [17]. Similarly hedgehog is least similar to all other species. All computation have been performed on our Intel(R) Core(TM)2 Duo CPU E7500 @2.93GHz, with 2.99 GB of RAM machine in Windows environment. The code for obtaining the graphical curve and distance matrix is executed in MATLAB.
IV. DISCUSSION

The availability of large amounts of genetic data allows us to study and compare the results from different phylogenetic reconstruction methods. The resultant phylogenetic tree in Fig. 1 shows the evolutionary relationship among the mitochondrial genome sequences. The result is in agreement with the previous studies as found by Raina et al. [17] and Yu et al. [8]. The last nine mammals are grouped in a cluster because they are primates. Computationally alignment free approaches perform better than alignment method for biological sequences. a large number of existing methods for phylogenetic inference requires multiple alignment of sequences. The alignment methods assume some sort of an evolutionary model [18]. The choice of evolutionary model is totally depends on the researchers. Therefore, the results obtained from different models are different. In other word, since these results require human intervention therefore the results might be controversial. The proposed alignment-free method does not use any evolutionary model. The results are automatically generated and do not need human intervention. The proposed graphical representation and numerical characterization approach can also handle large volumes of DNA sequences more quickly and more easily than multiple sequence alignment methods.

V. CONCLUSION

We have performed the similarity analysis of mitochondrial genome sequences by an alignment
free method. Mitochondrial DNA is not highly conserved in nature and has a rapid mutation rate, therefore it is very useful for studying the evolutionary relationships of organisms. The graphical representation has been constructed by assigning coordinates to adjacent nucleotide pairs. Then moment vectors have been constructed for this graphical curve corresponding to each nucleotide sequence. The distance matrix is similarity matrix which represents the degree of similarity between each pair of nucleotide sequence. The distance matrix is computed by calculating the Euclidian distance between the points represented by moment vectors in genome space. Finally the evolutionary relationship among the species is represented by phylogenetic tree. The UPGMA method used for the construction of phylogenetic tree uses the distance matrix as input. The result shown by us is in agreement with the studies in past. Therefore, we can conclude that alignment free method which is computationally very less expensive than alignment methods produces similar results.

REFERENCES


