International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Clustering of Mitochondrial D-loop Sequences Using Similarity Matrix, PCA and K-means Algorithm

Can Eyupoglu*[1]

Accepted 3rd September 2016

Abstract:In this study, mitochondrial displacement-loop (D-loop) sequences isolated from different hominid species are clustered using similarity matrix, Principal Component Analysis (PCA) and K-means algorithm. Firstly, the mitochondrial D-loop sequence data are retrieved from the GenBank database and copied into MATLAB. Pairwise distances are computed using p-distance and Jukes-Cantor methods. A phylogenetic tree is created and then a similarity matrix is generated according to the pairwise distances. Furthermore, the clustering is performed using only K-means algorithm. After that PCA and K-means are used together in order to cluster mitochondrial D-loop sequences.

Keywords:*Clustering, p-distance, PCA, Jukes-Cantor, K-means algorithm, Similarity matrix.*

1. Introduction

Mitochondrial DNA (mtDNA) sequences of mammals evolve more rapidly than nuclear DNA sequences [1], [2]. This fast rate of evolution generates more change between sequences. In order for research of closely related species and populations, this rate is a benefit [1], [3]. In animal mtDNA, there are four principal kinds for sequence changes. These are sequence rearrangements, additions, deletions and nucleotide substitutions [4], [5]. Nucleotide substitutions are the most important principal for the derivation of phylogenetic relationships [4], [6]. The fastest evolving part of the mitochondrial genome is the mitochondrial control region (Displacement or D-loop) [4], [7]–[9].

Principal Component Analysis (PCA) is a classical feature extraction and data representation method [10], [11]. In addition, it can be used to reduce the dimension of similarity matrix generated according to the pairwise distances and simplify the mitochondrial D-loop sequence data structure. The main features of the mitochondrial D-loop sequences can be extracted using PCA by means of mapping high dimensional space data into low dimensional space.

In phylogenetic analysis, distance measure is a significant matter [12]. Using p-distance and Jukes-Cantor methods, pairwise distances are calculated. Jukes-Cantor method is the simplest nucleotide substitution model which estimates the evolutionary distance between two sequences. Besides, it is called one parameter model in the literature [13]. This model can be applied to nucleotide substitution in alphabet (A, C, G, T) [14].

In data mining, the field of clustering has received significant attention in recent years and has become one of the important parts of machine learning research. Clustering is the process of categorizing a finite number of objects into groups where all members have common properties. Data mining is the process of using clustering algorithms in order to analyse data for patterns and relationships. K-means clustering algorithm [15] is one of the most used and popular clustering algorithms [16]. Furthermore, K-means is a simple unsupervised learning algorithm used to solve well known clustering problems.

The rest of the paper is organized as follows. In Section 2, the methods used for clustering of mitochondrial D-loop sequences are explained. Section 3 investigates the results of the methods used in this study. Finally, conclusions being under study are summarized in Section 4.

2. Methods

In this paper, in order for clustering mitochondrial D-loop sequences isolated from different hominid species, similarity matrix, PCA and K-means algorithm are used. To calculate pairwise distances, p-distance and Jukes-Cantor methods are utilized. K-means algorithm is used alone and then it is utilized with PCA for clustering. Clustering of mitochondrial D-loop sequences using Jukes-Cantor method is shown in Figure 1. As seen in the flowchart, firstly, the mitochondrial D-loop sequence data are retrieved from the GenBank database. Secondly, retrieved sequence data are copied into MATLAB. After that pairwise distances are calculated using Jukes-Cantor method and then a phylogenetic tree is created. A similarity matrix is generated according to the pairwise distances. In addition, a dendrogram is drawn. Clustering is performed using K-means algorithm. Finally, K-means algorithm is used after applying PCA.

Clustering of mitochondrial D-loop sequences using p-distance method is shown in Figure 2. As seen in the flowchart, unlike the first method, pairwise distances are calculated using p-distance method. Following steps are performed in the same manner.

[1] *Istanbul Commerce University, Department of Computer Engineering, Istanbul, Turkey*
** Corresponding Author:Email: ceyupoglu@ticaret.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

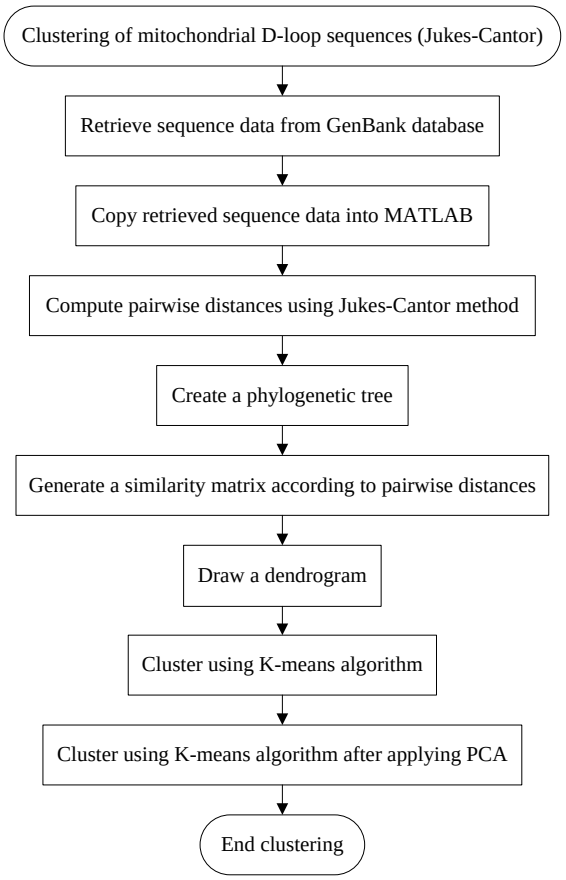

Figure 1. Flowchart of clustering of mitochondrial D-loop sequences using Jukes-Cantor method.

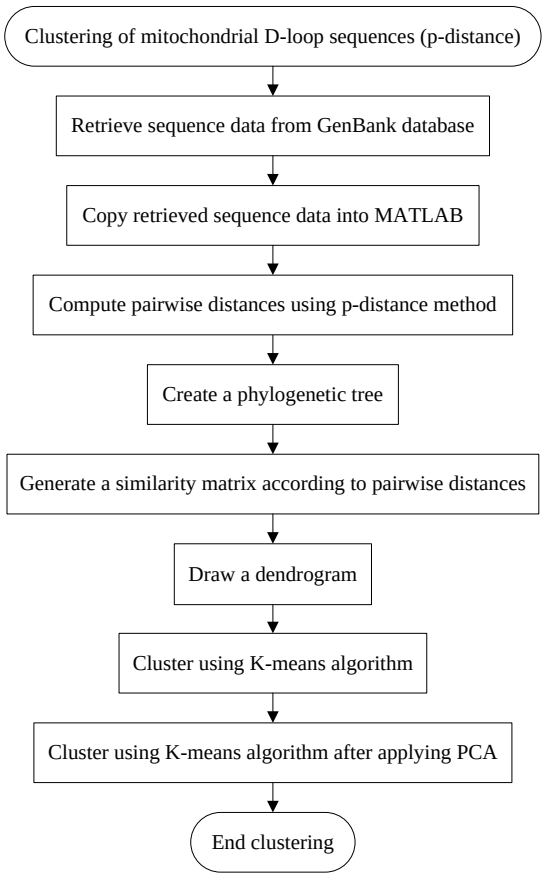

Figure 2. Flowchart of clustering of mitochondrial D-loop sequences using p-distance method.

In this work, 5 hominid species are utilized to cluster mitochondrial D-loop sequences. These are European Human, Russian Neanderthal, German Neanderthal, Chimp Troglodytes and Mountain Gorilla Rwanda. When retrieving sequence data from the GenBank database, the accession codes for the mitochondrial D-loop sequences isolated from these 5 species are used. These codes are X90314, AF254446, AF011222, AF176766 and AF089820, respectively.

3. Results and Discussion

In this study, to cluster mitochondrial D-loop sequences isolated from different hominid species, K-means algorithm is used and then it is used with PCA. Jukes-Cantor and p-distance methods are utilized in order to compute pairwise distances. The applications used for clustering of mitochondrial D-loop sequences are implemented using MATLAB R2014a. Phylogenetic tree created using Jukes-Cantor method is shown in Figure 3.

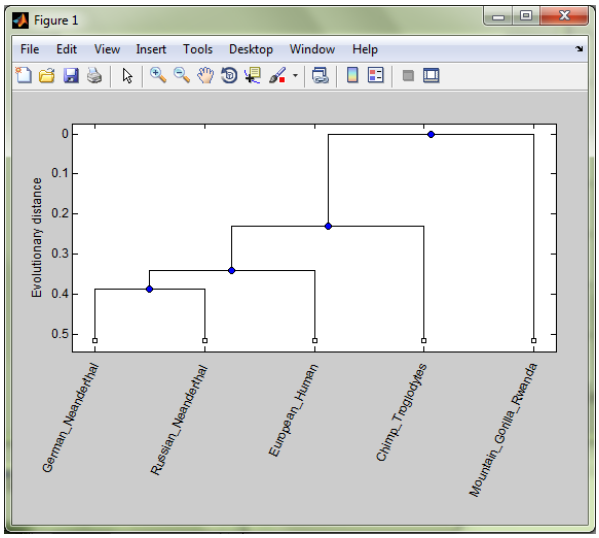

Figure 3. Phylogenetic tree (Jukes-Cantor).

Creating the similarity matrix according to the pairwise distances and drawing the dendrogram (Jukes-Cantor method) are shown in Figure 4. Converting the pairwise distances to square form (Jukes-Cantor method) is shown in Figure 5. Clustering using K-means algorithm (Jukes-Cantor method) is shown in Figure 6. Performing PCA on square form (Jukes-Cantor method) is shown in Figure 7. Clustering using K-means algorithm after applying PCA (Jukes-Cantor method) is shown in Figure 8.

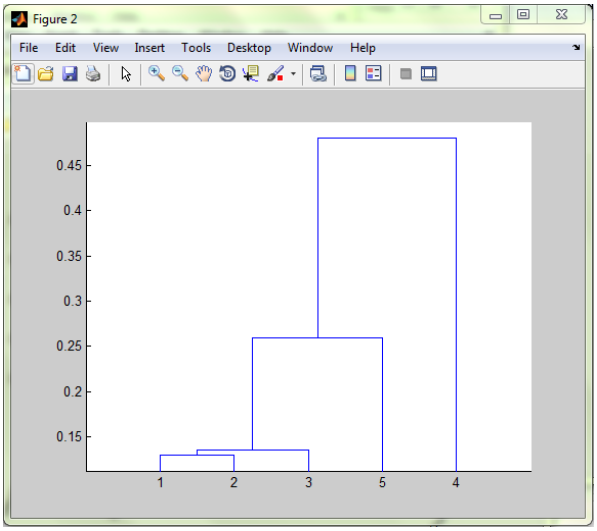

Figure 4. Creating similarity matrix and drawing dendrogram (Jukes-Cantor).

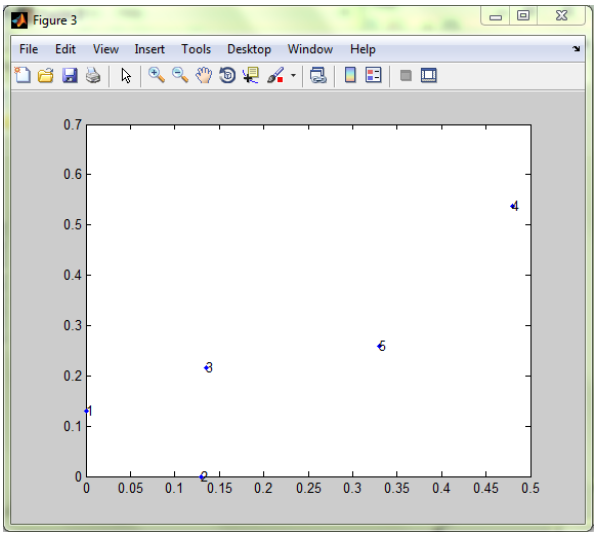

Figure 5. Converting pairwise distances to square form (Jukes-Cantor).

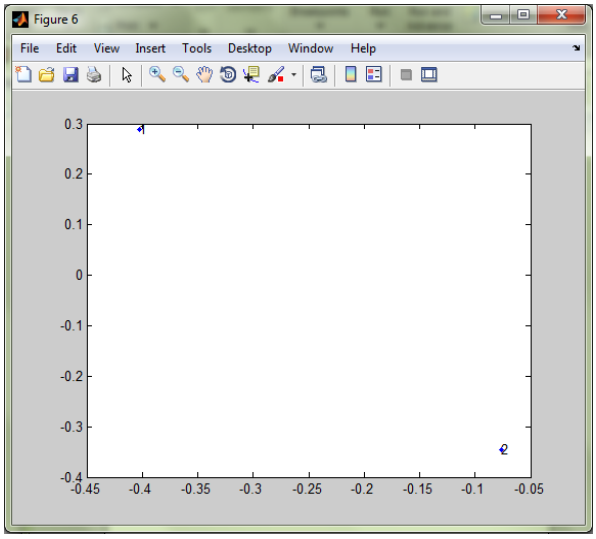

Figure 6. Clustering using K-means algorithm (Jukes-Cantor).

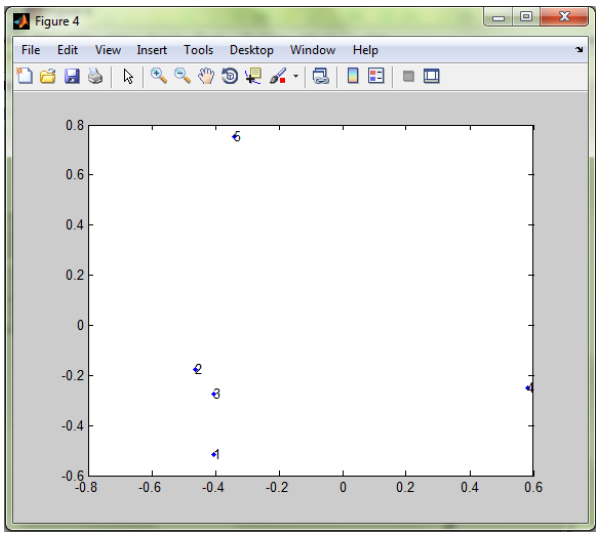

Figure 7. Performing PCA on square form (Jukes-Cantor).

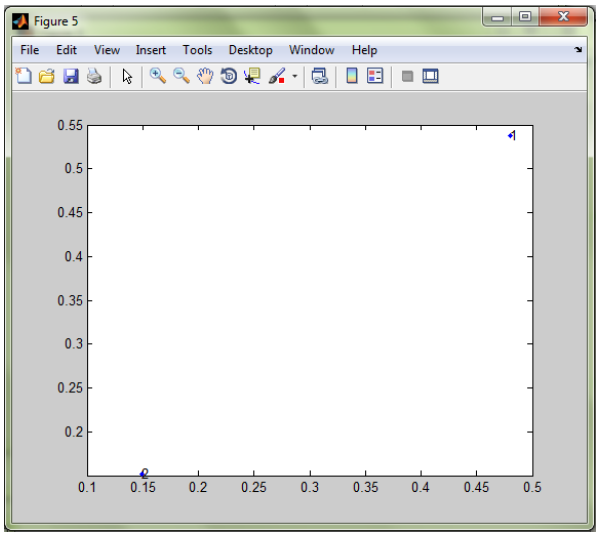

Figure 8. Clustering using K-means algorithm after applying PCA (Jukes-Cantor).

Phylogenetic tree created using p-distance method is shown in Figure 9. Creating the similarity matrix according to the pairwise distances and drawing the dendrogram (p-distance method) are shown in Figure 10. Converting the pairwise distances to square form (p-distance method) is shown in Figure 11. Clustering using K-means algorithm (p-distance method) is shown in Figure 12. Performing PCA on square form (p-distance method) is shown in Figure 13. Clustering using K-means algorithm after applying PCA (p-distance method) is shown in Figure 14.

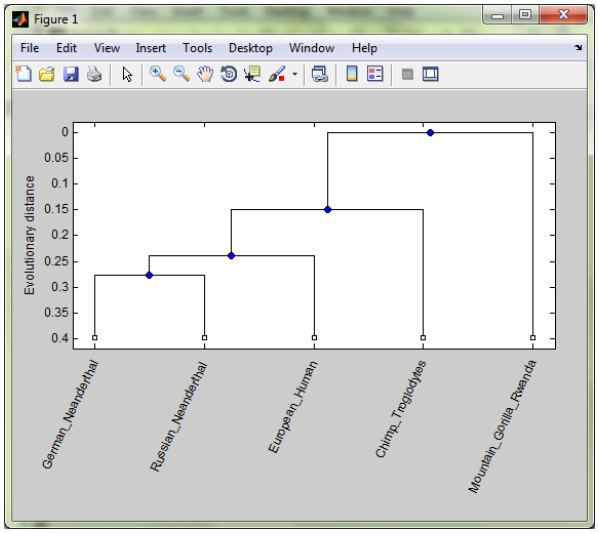

Figure 9. Phylogenetic tree (p-distance).

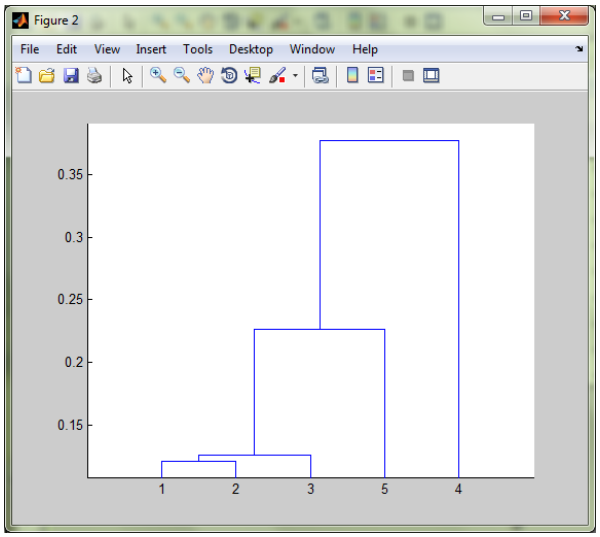

Figure 10. Creating similarity matrix and drawing dendrogram (p-distance).

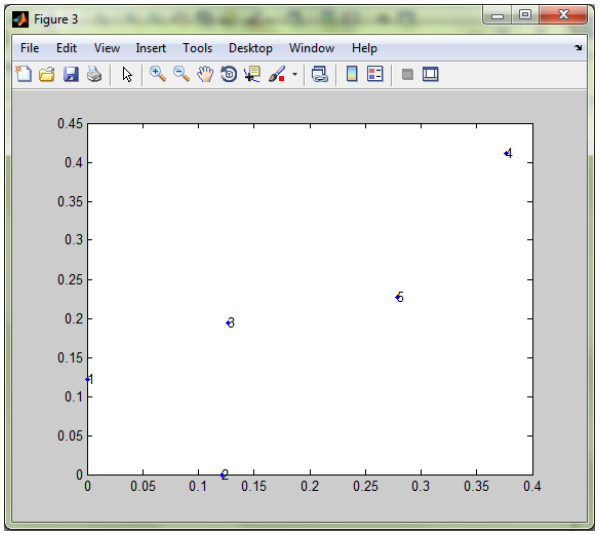

Figure 11. Converting pairwise distances to square form (p-distance).

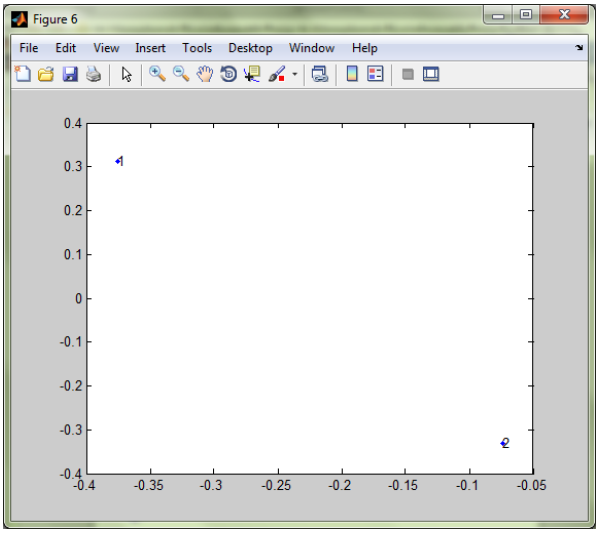

Figure 12. Clustering using K-means algorithm (p-distance).

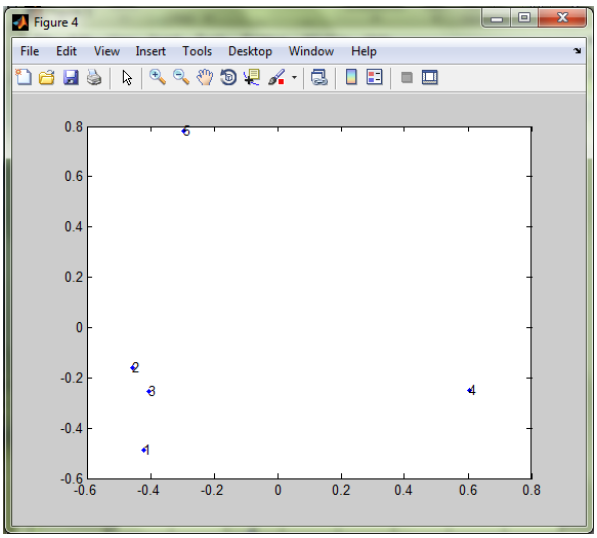

Figure 13. Performing PCA on square form (p-distance).

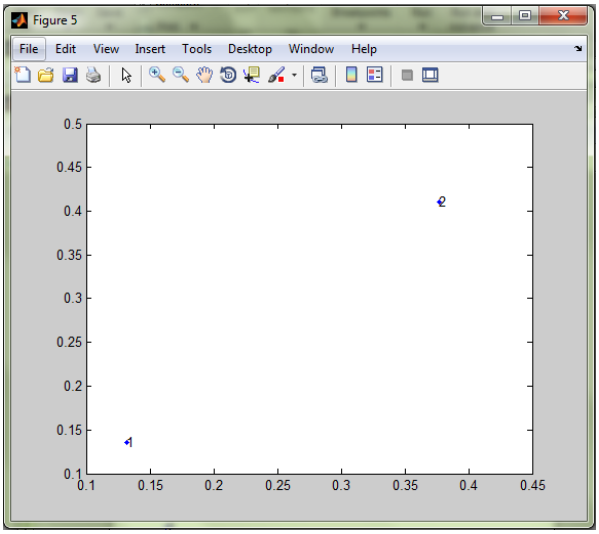

Figure 14. Clustering using K-means algorithm after applying PCA (p-distance).

As seen in Figure 3–Figure 14, clustering of mitochondrial D-loop sequences isolated from different hominid species is successfully performed using similarity matrix, PCA and K-means algorithm. It is observed that both Jukes-Cantor and p-distance methods are practical for computing pairwise distances.

4. Conclusion

In recent years, clustering has become a significant research topic in the area of machine learning. In this paper, when clustering mitochondrial D-loop sequences isolated from different hominid species, similarity matrix, PCA and K-means algorithm are used. First of all, K-means algorithm is used alone and then it is utilized with PCA in order for extracting features of the pairwise distances located in the similarity matrix. Besides, pairwise distances are calculated using Jukes-Cantor and p-distance methods. According to the study results, it is seen that the mitochondrial D-loop sequences are successfully clustered using similarity matrix, PCA and K-means algorithm.

References

[1] H. Zischler, H. Geisert, A. Von Haeseler, and S. Pääbo, "A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans," Nature, vol. 378, no. 6556, pp. 489–492, November 1995.

[2] W. M. Brown, E. M. Prager, A. Wang, and A. C. Wilson, "Mitochondrial DNA sequences of primates: tempo and mode of evolution," Journal of Molecular Evolution, vol. 18, no. 4, pp. 225–239, July 1982.

[3] D. R. Maddison, M. Ruvolo, and D. L. Swofford, "Geographic origins of human mitochondrial DNA phylogenetic inference from control region sequences," Systematic Biology, vol. 41, no. 1, pp. 111−124, 1992.

[4] A. R. Hoelzel, J. M. Hancock, and G. A. Dover, "Evolution of the Cetacean Mitochondrial D-Loop Region," Molecular Biology and Evolution, vol. 8, no. 3, pp. 475−493, 1991.

[5] W. M. Brown, "The mitochondrial genome of animals," MacIntyre RJ (ed) Molecular Evolutionary Genetics, Plenum Press, New York, pp. 95−130, 1985.

[6] A. C. Wilson, R. L. Cann, S. M. Carr, M. George, U. B. Gyllensten, K. M. Helm-Bychowski, R. G. Higuchi, S. R. Palumbi, E. M. Prager, R. D. Sage, and M. Stoneking, "Mitochondrial DNA and two perspectives on evolutionary genetics," Biological Journal of the Linnean Society, vol. 26, no. 4, pp. 375−400, December 1985.

[7] W. B. Upholt and I. B. Dawid, "Mapping of mitochondrial DNA of individual sheep and goats: rapid evolution in the D loop region," Cell, vol. 11, no. 3, pp. 571−583, July 1977.

[8] M. W. Walberg and D. A. Clayton, "Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA," Nucleic Acids Research, vol. 9, no. 20, pp. 5411−5421, October 1981.

[9] D. Chang and D. A. Clayton, "Priming of human mitochondrial DNA replication occurs at the light-strand promoter," Proceedings of the National Academy of Sciences of the United States of America, vol. 82, no. 2, pp. 351−355, January 1985.

[10] C. Eyupoglu, "Implementation of Color Face Recognition Using PCA and k-NN Classifier," 2016 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference (ElConRusNW), pp. 199−202, St. Petersburg, Russia, 2–3 February 2016.

[11] X. Xiang, J. Yang, and Q. Chen, "Color face recognition by PCA-like approach," Neurocomputing, vol. 152, pp. 231−235, March 2015.

[12] D. Wei and Q. Jiang, "A DNA Sequence Distance Measure Approach for Phylogenetic Tree Construction," 2010 IEEE Fifth International Conference Bio-Inspired Computing: Theories and Applications (BIC-TA), pp. 204−212, Changsha, 23–26 September 2010.

[13] P. Bhambri and O. P. Gupta, "Development of Phylogenetic Tree Based on Kimura's Method," 2012 2nd IEEE International Conference on Parallel Distributed and Grid Computing (PDGC), pp. 721−723, Solan, 6–8 December 2012.

[14] S. S. Patil, V. Kumar, V. R. Pai, and A. K. Patil, "Constructing phylogenetic tree and analysis using information retrieval approach for MYB tfr's of rice genome," 2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 523–529, Dhaka, 19–20 December 2015.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, Berkeley, University of California Press, 1967.

[16] W. K. Daniel Pun and A. B. M. Shawkat Ali, "Unique Distance Measure Approach for K-means (UDMA-Km) Clustering Algorithm," 2007 IEEE Region 10 Conference (TENCON), pp. 1–4, Taipei, 30 October–2 November 2007.

