A Comparative Analysis of Feature Selection Methods for Clustering DNA Sequences

B.Umamageswari *umamage@gmail.com*
Department of Computer Science & Engg
Bharath University,
Chennai – 600 073, India.

B.Karthikeyan *karthikeyan512@gmail.com*
School of Computing,
National University of Singapore,
Singapore.

T.Nalini *nalinicha2002@yahoo.co.in*
Prof, Department of Computer Science & Engg
Bharath University,
Chennai – 600 073, India.

Abstract

Large-scale analysis of genome sequences is in progress around the world, the major application of which is to establish the evolutionary relationship among the species using phylogenetic trees. Hierarchical agglomerative algorithms can be used to generate such phylogenetic trees given the distance matrix representing the dissimilarity among the species. ClustalW and Muscle are two general purpose programs that generates distance matrix from the input DNA or protein sequences. The limitation of these programs is that they are based on Smith-Waterman algorithm which uses dynamic programming for doing the pair-wise alignment. This is an extremely time consuming process and the existing systems may even fail to work for larger input data set. To overcome this limitation, we have used the frequency of codons usage as an approximation to find dissimilarity among species. The proposed technique further reduces the complexity by extracting only the significant features of the species from the mtDNA sequences using the techniques like frequent codons, codons with maximum range value or PCA technique. We have observed that the proposed system produces nearly accurate results in a significantly reduced running time.

Keywords: Evolutionary Tree, Data Mining, Bioinformatics, Euclidean Distance, PCA, Hierarchical Clustering Algorithms, MtDNA, Codons.

1. INTRODUCTION

DNA sequences are made of the four bases ‘A’,’G’,’C’ and ‘T’. Any three bases out of four constitute a codon. Altogether there are 64 codons. Not all codons are significant. Therefore, we make use of any of the three methods namely frequent codons (refer Section 3), codons with maximum range values (refer Section 3) and PCA (refer Section 3) to determine the key codons. The dissimilarity among the species is calculated using the Euclidean distance formula based on the extracted ‘C’ key features. Therefore, once the dissimilarity among the species are extracted based on the frequency of occurrence of codons, clusters can be formed by applying any Hierarchical clustering algorithms. UPGMA is one of the widely used algorithms for clustering DNA sequences based on the dissimilarity or distance measure. The resulting clusters are then represented using a Phenogram (term used in bioinformatics for Dendrogram). The dataset for studying the efficiency of our approach is collected from NCBI, a database containing the complete genome sequence of thousands of species. We have collected mitochondrial DNA sequences for 45 species to test the proposed approach.

The paper is organised as follows: Section 2 details the background, Section 3 describes the algorithm for extracting the key features from the input mtDNA sequences, Section 4 describes the input to the system and the experimental results obtained by simulating the system in Matlab, Section 5 details about the conclusion and future work and Section 6 lists the references.

2. BACKGROUND

In Bioinformatics, CLUSTALW is the program, which is used to find regions of similarity between biological sequences. It uses a heuristic approach that approximates the Smith-Waterman algorithm[1] to achieve faster alignment at lower accuracy. The well-known techniques for sequence alignment include Smith-Waterman algorithm[1] for local alignment, Needleman-Wunsch algorithm[2] for global alignment and evolutionary algorithm for multiple sequence alignment[3] presented by Chellapilla et al. etc. The drawback of all these algorithms are, when they are applied for comparing larger sequences with sizes varying from hundred to thousand pairs its highly time consuming. These results in the need for finding procedures that may be somewhat approximate in nature and produce quick results. One such approximation that is found in the existing work[4] is to make use of position of occurrence of the A, G, C and T for plotting the sequences in 2-Dimensional space and to measure the similarity among the sequences using the Euclidean distance measure. In our previous work[5], instead of considering the position, the key features based on the number of occurrences of each codon(AAA,AAC,AAG,... TTT) is used to find similarity among the Hemoglobin Beta sequences extracted from different species and clustering is done using UPGMA algorithm[6]. Here, we are going to consider only key codons (i.e only those codons that highly distinguishes one species from the other) to find dissimilarity among the mtDNA sequences extracted from different species and clustering is done using UPGMA algorithm[6].

UPGMA (Unweighted Pair Group Method with Arithmetic Mean)[6] is a simple agglomerative/hierarchical clustering method widely used in bioinformatics for the creation of phylogenetic trees (phenograms).

3. ALGORITHM FOR KEY FEATURES EXTRACTION

Input to this algorithm is the flat file containing the mtDNA sequences whose length around 16k of the species that are to be clustered in FASTA format[7]. This algorithm extracts the frequency of occurrence of the codons (AAA, AAC...TTT) from each sequence. Suppose if there are N sequences, we get an NX64 matrix. It then finds the discriminant codons (NXC) matrix using any one of the methods such as frequent codons, codons with maximum range values and PCA[8] where C is the number of discriminant or key codons and measures dissimilarity among the species only based on those key features using the Euclidean distance formula for multiple dimensions and the result is stored in the form of an adjacency matrix.

Input: The input mtDNA sequences corresponding to different species that are to be clustered in FASTA format[8] stored in a flat file.
For ex,
>human
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCT
GGGG.....
>mouse
GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGATAATTGTATCCCATAAA
CACAAAG...
>…..

Output: An NXN adjacency matrix representing dissimilarity among the N species.

Algorithm for Key Features Extraction:

1. Read the input line by line.
2. Let C be the number of codons to be selected.
3. If the line starts with '>' then the string following it represents the name of the species
4. Otherwise it represents the mtDNA sequence of the species.
5. For each species extract the number of occurrence of each codon (AAA, AAC, AAG, AAT ....TTT). Altogether, there are 64 features. So, we will be getting an NX64 matrix where N is the number of species.
6. Normalize the feature matrix: Let the i^{th} input vector be $F_{i=}$ [F_{i1} F_{i2}…..F_{i64}].To normalize the vector the following transformation is used.

   k^{th} entry value of the i^{th} sample is given by $$\mathbf{F_{ik}} = \frac{\mathbf{F_{ik}}}{\sum_{j=1}^{64} \mathbf{F_{ij}}} \quad (1)$$
7. Key codons can be determined using any one of the techniques given below:

7a. Steps in Determining Frequent Codons:

1. Find the sum of frequency for each codon by considering its value for all species.
2. Sort the codons in descending order based on the sum of frequency computed in step(1). Only the initial 'C' codons are considered further.
3. The columns in the normalized feature matrix corresponding to the 'C' codons are considered further to form the NXC matrix.

7b. Steps in Determining Codons With Maximum Range Values:

1. Find the range of values for each codon. i.e difference between maximum and minimum value it can take for all the species.
2. Sort the codons in descending order based on the range value computed in step(1). Only the initial 'C' codons are considered further.
3. The columns in the normalized feature matrix corresponding to the 'C' codons are considered further to form the NXC matrix.

7c. Steps in PCA:

1. Calculate the Mean adjusted data: This step is used get the data adjusted around zero mean.

   k^{th} entry value of the i^{th} sample is given by $\mathbf{M_{ik}} = \mathbf{M_{ik}} - \overline{\mathbf{M_i}}, \forall \mathbf{i,k\ where} \overline{\mathbf{M_i}} = \frac{\sum_{j=1}^{64} M_{ij}}{64}$ (2)

2. Calculate the covariance matrix: CM (64X64)

   Covariance between any two codons a_i, a_j is given by $\mathbf{CM_{ij}} = \frac{\sum_{k=1}^{N}(a_{ik}-\overline{a_i})(a_{jk}-\overline{a_j})}{(M-1)}$ (3)

3. Calculate the Eigen value and the Eigen vector: Computing the roots of the equation $|CM - \lambda I| = 0$, the Eigen values of the Covariance matrix C are then obtained for every species.
   a. A 64X64 covariance matrix produces 64 Eigen Values.
   b. Corresponding to each Eigen Value there is an Eigen vector of dimension 64X1.
   c. Find the C largest Eigen Values and find the corresponding Eigen Vectors. This corresponds to the Principal Components.
   d. Key Features matrix = $(\text{mean adjusted data})_{NX64}$ X $(\text{principal components})_{64XC}$. This results in an (NXC) matrix representing the key feature of the species.

8. Calculate the Euclidean Distance using the formula:

$$\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\ \ldots+\ (x_{iC}-x_{jC})^2} \quad (4)$$

This result is an NXN adjacency matrix representing the dissimilarity among the species.

4. EXPERIMENTS AND RESULTS

Simulation of the proposed system is done in Matlab. The mtDNA sequence corresponding to different species are obtained from NCBI[9]. The mtDNA sequences[12] corresponding to 45 species are clustered by determining the key codons using any of the selected three methods namely frequent codons, codons with maximum range values, PCA[8] and then clustering is done by applying the UPGMA algorithm[6]. The phenogram generated by the proposed key feature selection method and the existing multiple sequence alignment method can be seen in Figure(1) & Figure (2).

By looking at these figures, we see that the phenogram obtained using proposed method is consistent and close to the phenogram generated using existing method based on multiple sequence alignment however there are still some differences. Analysing real phylogenetic trees are more complex. Understanding such trees requires visual inspection, structural comparison, and interactive manipulation and exploration, and thus present a number of visualisation challenges. Therefore we have used COMPONENT application[10] to aid our comparison process.
We have used the following metrics obtained from COMPONENT application[10] to compare the phenograms generated using existing and proposed method.

(i) Greatest Agreement Subtree (GAS): An agreement subtree is a subtree that is common to two or more trees. A **Greatest Agreement Subtree** (GAS) is an agreement subtree with the greatest number of leaf nodes. Agreement subtrees are useful for identifying trees that differ in the placement

of one or more taxa but are otherwise very similar. The distance dGAS (T1, T2) as the number of leaves removed to obtain a greatest agreement subtree.

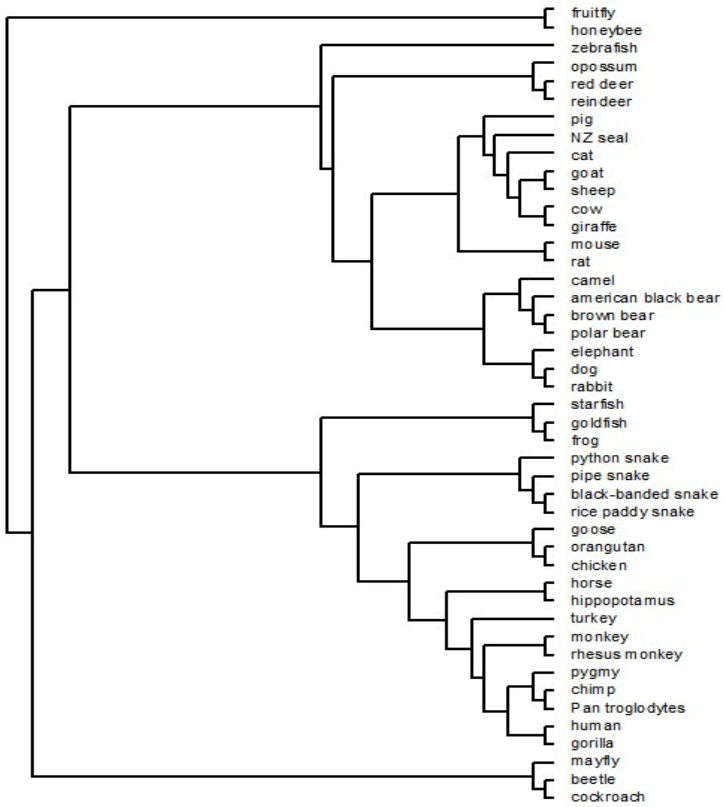

FIGURE 1: Phenogram representing the resulting clusters using KFEA using PCA and considering 80% codons + UPGMA

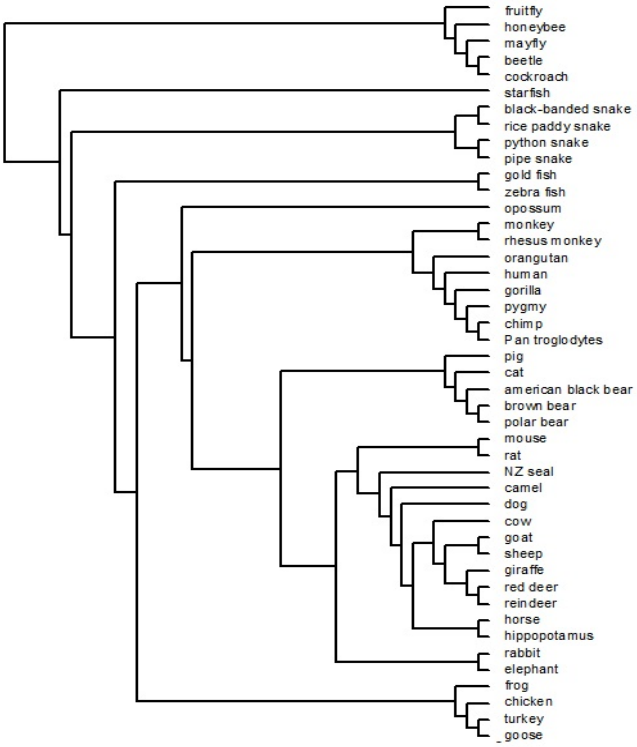

FIGURE 2: Phenogram representing the resulting clusters using multiple sequence alignment and UPGMA

(ii) Nearest Neighbour Interchange (NNI): The **Nearest Neighbour Interchange** (NNI) technique is also used to quantify the difference between trees. A nearest neighbour interchange is an interchange of two "nearest neighbour" branches. The NNI difference between two trees is the minimum number of such interchanges needed to convert one tree into the other.

(iii) Quartet: A quartet is the smallest possible informative subtree of an unrooted tree, and contains just four leaves. To measure the similarity of two unrooted trees, T1 and T2, is to compare their quartets. Each pair of quartets from two trees belongs to one of five classes:
s - resolved and identical; d - resolved and different; r1- resolved in T1 but not in T2;
r2 - resolved in T2 but not T1and u - unresolved in both T1 and T2
Note that Q = s + d + r1 + r2 + u, and that for two binary trees, r1 = r2 = u = 0.
From these classes of quartets the following dissimilarity values can be derived:
Do not **C**onflict (DC) = d; **E**xplicitly **A**gree (EA) = d + r1 + r2 + u;
Strict **J**oint **A**ssertions (SJA) = d / (d + s); **S**ymmetric **D**ifference (SD) = (2d + r1 + r2)/(2d + 2s + r1 + r2). Graph shown in Figure 5 is plotted based on Symmetric Difference measure.

The above 3 comparison metrics are used to compare the phylogenies generated by the proposed feature selection methods with the reference (sequence alignment + UPGMA) phylogeny. Different input samples containing 25, 45 and 68 species are considered for the experiment. The comparison is done by initially selecting all the 64 features (100%) and then gradually reducing the number of selected features by 10%. The results are plotted as graph as shown in Figure 3, Figure 4 and Figure5. From the graphs, we infer that the accuracy of results obtained does not vary much even when the percentage of selected features is reduced considerably. Thus, all 64 features are not needed to perform clustering. PCA produces consistent results irrespective of number of codons selected. The other two techniques namely frequent codons and codons with maximum range values produce poor results in the presence of insignificant features (i.e on selecting 80% - 100% codons) and also in the absence of significant features(i.e on selecting less than 50% of codons). Thus, the other two techniques perform well only in the presence of relevant features i.e presence of irrelevant feature affects clustering accuracy.

The first metric namely GAS is used to measure the structural similarity between the two trees. The second metric namely NNI is based on minimum number operations required to transform one tree to

another. The third metric quartet represents the two trees as sets of simpler structures (such as clusters or quartets) and then uses various measures of similarity between sets. From the figures (3)(4) and (5), we note that the behaviour of the feature selection techniques namely frequent codons, codons with maximum range values and PCA remain the same irrespective of the metric used to compare the trees.

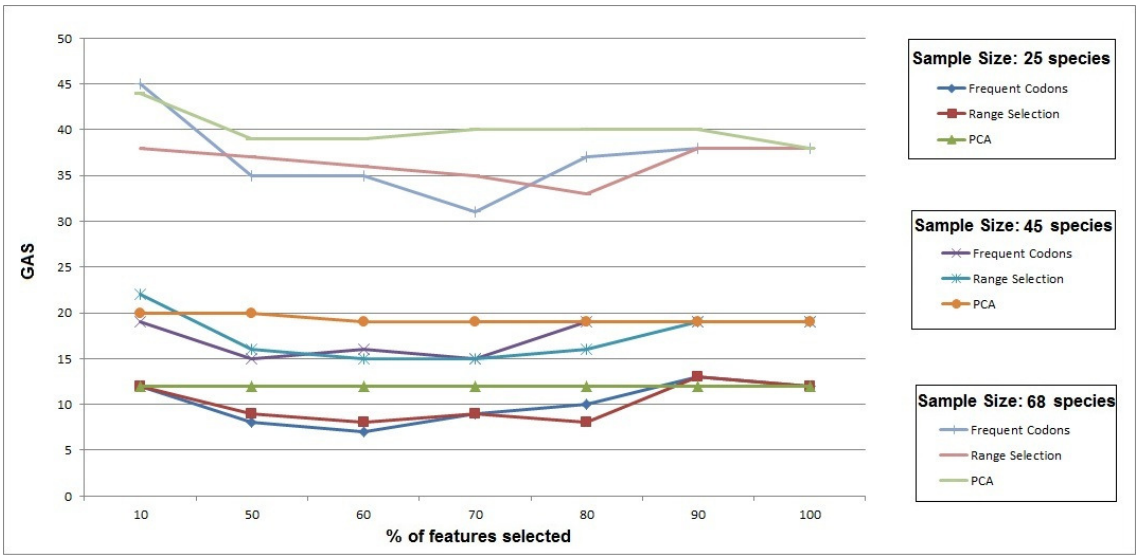

FIGURE 3: % of features (codons frequency) selected vs. GAS

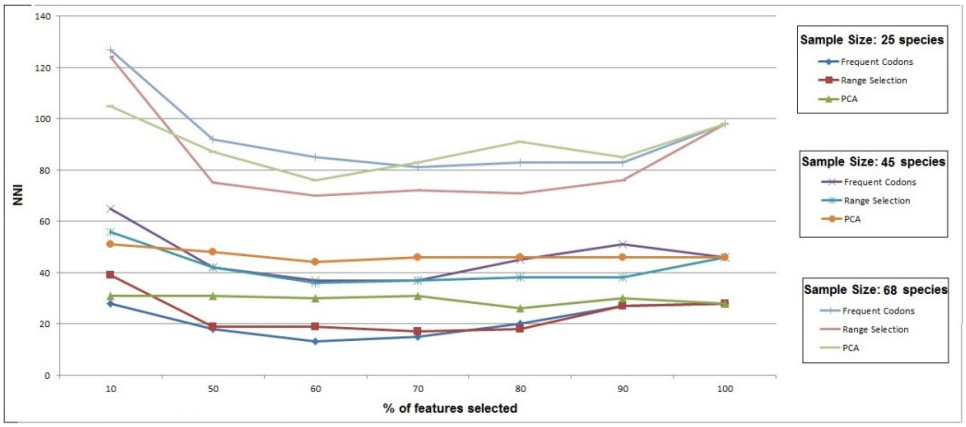

FIGURE 4: % of features (codons frequency) selected vs. NNI

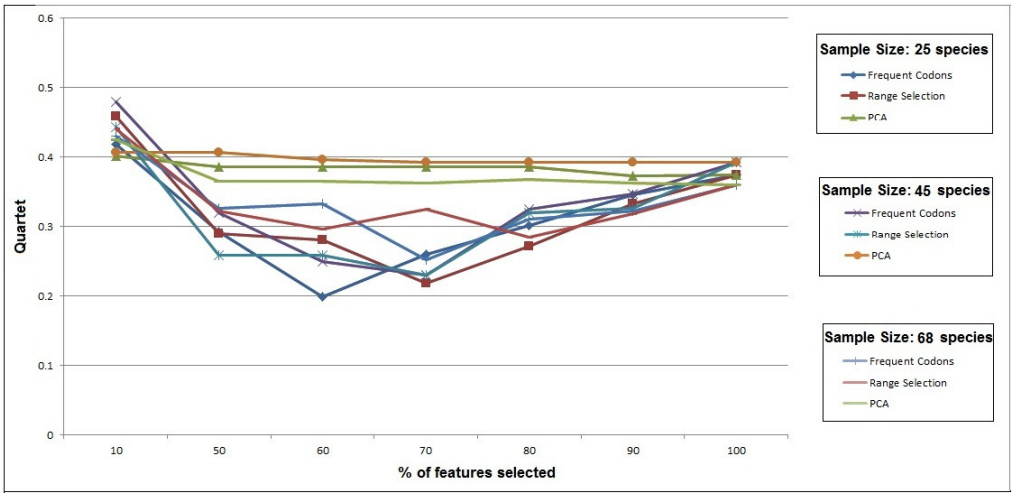

FIGURE 5: % of features (codons frequency) selected vs. Quartet

The main advantage of the proposed techniques is that the running time is reduced significantly compared to Sequence Alignment + UPGMA. A comparative study is done between the running times of Sequence Alignment + UPGMA and KFEA + UPGMA by varying the number of input sequences and the results are tabulated (Table1). It can be seen that the running time for Sequence Alignment + UPGMA varies from few minutes to several hours based on the input data set. It takes around 5 minutes to cluster 10 species while taking almost 4 hours to cluster 68 species. But the running time for KFEA + UPGMA increases only by few seconds even if the input data set is increased considerably. From the below table, it can be seen that the running time increases only by 8.5 sec when the number of species to be clustered is increased from 10 to 68.Thus, the proposed technique would be handy in situations where it is necessary to get instant results for large input data sets since the existing system will take several hours to compute results for such data.

No. of Species clustered	Sequence Alignment + UPGMA (seconds)	KFEA+UPGMA (seconds)	Ratio
10	299.80	2.6795	112
25	1922.36	5.7652	333.7
45	5378.48	9.5651	562.3
68	15481.36	11.0944	1394.7

TABLE 1: Comparison of phylogenies obtained using sequence alignment and KFEA based on running time

5. CONCLUSION AND FUTURE WORK

In Bioinformatics, identification of species given the DNA sequence is a challenging task. The existing system like CLUSTALW[11] deals with aligning the sequences using well-known heuristic approach like Smith-Waterman algorithm[1] and Needleman-Wunsch algorithm[2]. Though these systems produce accurate results, the process is extremely time consuming even for smaller sequences and may not work with larger and more complex sequences.

DNA sequences consist of three-unit patterns (ATG, AGC, etc.) called codons which are translated to a specific amino acid. The frequency of codon usage is generally similar among closely related species. Therefore, we are extracting the frequency of occurrence of codons as an approximation for identifying similarity among the species. Altogether there are 64 codons based on which the dissimilarity among the species can be determined. Instead of considering all the 64 features, in this paper technique such as PCA, frequent codons and codons with maximum range values are employed to reduce the number of features so that only the significant features are used to measure the dissimilarity among the species. We found that PCA works better than the other two techniques because the accuracy remains consistent irrespective of the percentage of selected

features. For the other two methods, presence of irrelevant features (considering 80%-100% codons) and the absence of relevant features (considering < 50% codons) affect the accuracy of clustering. Also, this method produces nearly accurate results in significantly reduced time compared to the existing systems based on sequence alignment algorithms. Also, the proposed system is guaranteed to produce output irrespective of length and number of sequences, since the comparison is made only based on the extracted key features. In the future work, we will try to examine feature selection techniques such as Best First Search, Greedy Stepwise Search, Genetic Search and Linear Forward Search to obtain the significant features and comparison of the techniques can be made to identify the best that can be used to select the key features based on which the clustering can be done.

6. REFERENCES

[1] Smith, Temple F. and Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences".Journal of Molecular Biology147: 195–197.

[2] Needleman, S. B. and Wunsch, C. D. 1970 " A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology, 48: 443-453.

[3] Chellapilla, K. and Fogel, G. B. 1999. "Multiple sequence alignment using Evolutionary Programming", Proceedings of the 1999 Congress on Evolutionary Computation, Washington D. C.:445-452.

[4] Elhadi, G.F., Abbas, M.A.,"Clustering DNA sequences by selforganizing map and similarity functions", In proceedings of the 7th International Conference on Informatics and Systems (INFOS)", Publication Year: May 2010.

[5] B.Umamageswari,T.Nalini,A.R.Arunachalam, "Clustering DNA sequences by Extracting Pattern Features and using Hierarchical Clustering Algorithm", presented at the National Conference on Recent Trends in Data Mining and Distributed Systems(NCTD2S), September 2011.

[6] Sneath & Sokal (1973). "UPGMA (Unweighted Pair Group Method with Arithmetic Mean) Numerical Taxonomy". W.H. Freeman and Company, San Francisco, pp 230-234

[7] Lindsay I Smith.(2002, February 26). "A tutorial on Principal Component Analysis" [online] Available: www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[8] FASTA Format Description [online], NGFN-BLAST by NationaleGenomforschungsnetz. [online] Available: http://ngfnblast.gbf.de/docs/fasta.html

[9] Source of DNA Sequences [online], National Center for Biotechnology Information. Available: http://www.ncbi.nlm.nih.gov/mapview

[10] Roderic, D. M. (1993): Component 2.0 – User Guide, [online] Available: http://taxonomy.zoology.gla.ac.uk/rod/cplite/Manual.html

[11] CLUSTALW[online], Available: http://toolkit.tuebingen.mpg.de/.

[12] About mtDNA [online] Available: http://en.wikipedia.org/wiki/Mitochondrial_DNA