Global Journal of Computational Intelligence Research.
ISSN 2249-0000 Volume 2, Number 1 (2012), pp. 21-29

http://www.ripublication.com/gjcir.htm

Clustering of Different Species based on mtDNA Sequences by Frequent Codons

Dr. Appavoo Kumaravel

Professor and Dean, School Computing Sciences,
Bharath University. Chennai-600073, India
E-mail: drkumaravel@gmail.com

Abstract

DNA is considered as the basis set to scan the whole storage repository for the information for any cell to function and acts as the building block of life. It is a molecule made from sugar, phosphate and bases called adenine (A), guanine (G), cytosine (C) and thymine (T). The various combinations of these four bases make up the DNA in plants, animals, bacteria, yeast, fungi etc. The frequency of occurrence of the codons in the DNA sequence can be used as a basis for determining how closely related the different species are. Clustering is a process of grouping similar objects so that similarity among the objects within the same cluster is high whereas similarity among objects in different clusters is low. Therefore, clustering can be used to classify DNA sequences into meaningful groups. Classifying Mitochondria DNA (mtDNA) sequences in this way has a wide variety of applications including disease diagnosis, DNA forensics, identifying paternity and to determine how closely two different species are related on an evolutionary scale. In this paper, we are going to extract the pattern features based on the frequency of occurrence of codons from different mtDNA sequences and then measure the dissimilarity among those sequences using the extracted pattern features. Then the UPGMA algorithm is applied to the adjacency matrix representing the dissimilarity among species to cluster the given DNA sequences and the result is represented using a phenogram. The main objective of this paper is to identify the predominant features of a species instead of considering all the 64 features for this purpose. The search methods proposed here are compared for identifying maximum reduction of features.

Keywords: Mitochondria DNA sequences, Hierarchical clustering algorithms, Search Algorithms, Key Feature Selection, species classification, Dendrogram, Data mining, Bioinformatics, Euclidean distance.

Introduction

Contributions from machine learning and soft computing to Bio-informatics are evolving day by day serving the society by giving important clues for identification of contents of genome sequences. Mitochondria DNA (mtDNA) have been essential for the evolution of animals. It is generally believed that the energy-converting organelles of eucaryotes evolved from procaryotes that were engulfed by primitive eucaryotic cells. Mitochondrial DNA can be regarded as the smallest chromosome, and was the first significant part of the human genome to be sequenced. The DNA sequence of mtDNA has been determined from a large number of organisms and individuals (including some organisms that are extinct), and the comparison of those DNA sequences represents a mainstay of phylogenetics, in that it allows biologists to elucidate the evolutionary relationships among species. It also permits an examination of the relatedness of populations, and so has become important in the domain of anthropological studies.

Illustrating our paper we need to understand the basics such as nDNA, mtDNA, translation, genetic code and codon. Compared with Traditional nuclear nDNA analysis, Mitochondrial mtDNA offers three primary benefits to the scientists. (i) Its structure and location in the cell make mtDNA more stable,(ii) mtDNA is available in larger quantities per cell and (iii) mtDNA can be extracted from samples in which nDNA cannot, especially hair shafts and bone fragments.

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person’s body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences.

A gene is a sequence of DNA that contains genetic information and can influence the phenotype of an organism. Within a gene, the sequence of bases along a DNA strand defines a messenger RNA sequence, which then defines one or more protein sequences. The relationship between the nucleotide sequences of genes and the amino-acid sequences of proteins is determined by the rules of translation, known collectively as the genetic code. The genetic code consists of three-letter 'words' called codons formed from a sequence of three nucleotides (e.g. ACT, CAG, TTT). These codons represent the structural signature of an organism. Therefore, features extracted from the DNA sequences based on the frequency of occurrence of codons can be used to measure similarity among the species. Mining the genomic database is a complex task since it contains large number of sequences of varying length corresponding to several species.

Clustering is a data mining technique which is used to determine the similarity among the data on predefined attributes. The most similar data are grouped as clusters. Hierarchical clustering algorithms are especially used in the generation of phylogenetic trees because they will be producing sets of clusters as output. Initially each and every species will be present in its own cluster. It iteratively merges clusters till all the species are merged to form a single cluster. The output will be a phenogram

representing the relationship among the species. This is nothing but the needed evolutionary tree.

UPGMA and NJ are the two widely used algorithms in bioinformatics especially used in the construction of phylogenetic trees given the distance matrix representing the dissimilarity between each and every pair of species considered. UPGMA or Unweighted pair group method assumes equal rate of mutation across all branches i.e all the leaf nodes are at equidistant from the root. The NJ or Neighbour-Joining algorithm allows for unequal rate of evolution, so that the branch lengths are proportional to amount of change. If the rates on different branches are not markedly unequal, the branching orders produced by the two methods will not differ.

The dataset for studying the efficiency of our approach is collected from NCBI[8], a database containing the complete genome sequence of thousands of species. We are going to collect the mitochondria genome sequences for 11 species to test the proposed approach. The mitochondria genome sequences of different species have length around 16,000 base pairs.

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 UPGMA algorithm and NJ algorithm used for clustering the sequences based on the dissimilarity measure obtained by extracting the key features, Section 5 represents the architecture of the Proposed system, Section 6 describes the input to the system and the experimental results obtained by implementing the system in Java, Section 7 details about the conclusion and future work and Section 8 lists the references.

Background

In bioinformatics, CLUSTALW and Muscle are widely used multiple sequence alignment programs that are used to generate distance matrix given the DNA sequences of the species to be clustered. The limitation of these programs is that they are time consuming because they are based on the Smith-Waterman algorithm which performs a pair-wise alignment between any two given DNA sequences. It compares segments of all possible lengths and optimizes the similarity measure. It uses dynamic programming and it generates the optimal alignment score. It may not work for larger and more complex sequences. These drawbacks lead to the need for finding some approximate procedures for finding distance matrix representing dissimilarity among the sequences that produces quick results. One such approximation that is found in the existing work [1] 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 the proposed work, instead of considering the position, the pattern of codon usage based on the number of occurrences of each codon (AAA, AAC, AAG,... TTT) is extracted from N random samples of a species and various search methods available in Machine Learning or Artificial Intelligence used to project the key features of the species. Attribute Selection or Key Feature Extraction is a statistical technique that has found application in fields such as face recognition and image compression and is a common

technique for finding patterns in data of high dimension. So, key feature vector of dimension (1X64) is obtained for each species representing the overall feature possessed by the species. Once the key feature matrix is obtained, the distance matrix representing the dissimilarity among the species is obtained by using Euclidean distance formula in multi-dimensional space. Then, clustering is performed using hierarchical clustering algorithms such as UPGMA. Comparison can be made with the existing system like CLUSTAL and Muscle in terms of running time and clustering accuracy.

Proposed Algorithm

Input: The mitochondria genome sequences corresponding to different species that are to be clustered in FASTA format[7] stored in a flat file.
For example,
>human
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATT
TGGTATTTTCGTCTGGGGGGTGTGCACG.....
>mouse
GTTAATGTAGCTTAATAACAAAGCAAAGCACTGAAAATGCTTAGATGGAT
AATTGTATCCCATAAACACAAAGGTTTGG...
>…..
Output: An NXN adjacency matrix representing dissimilarity among the N species.(N=11)

Main Steps:

1. Extract 16 samples from the mitochondria sequence of the species
2. Calculate the number of occurrence of each codon. There are 64 codons. So, we get a 16X64 matrix for a species by AFOC (Refer Fig 1)
3. Choose any search algorithm and apply to get 16XC matrix where C is the reduced number of codons using Weka tool. (Refer table 1)
4. Repeat steps 1 to 3 for all species.
5. Measure the Euclidean distance between every species with every other species.
6. Give the distance matrix as input to DendroUPGMA program to get the Phenogram.

```
Algorithm for frequency of occurrence of codons: [AFOC]

Read the input line by line.
If the line starts with '>' then the string following it represents the name of the species
Otherwise it represents the mitochondria sequence of the species.
for i = 1 to n do // n is the number of species
  for j = 1 to 64 do
       freq[i][j] = getcount(i,codon[j]) // returns the number on occurrence of jth codon in ith sequence

function getcount(sequence,searchfor)
begin
  count = 0;
  start = sequence.indexof(searchfor)
// returns the first position of occurrence of the substring searchfor in the string sequence
  repeat
            count++; //count indicates the number of occurrence of the substring 'searchfor' in the string 'sequence'
            start = base.indexOf(searchFor, start+1);
           // from the current position search for the next occurrence of the substring and return the position if found
  until start=-1 // loop is terminated no more occurrence of the substring can be found
  return count
end
```

Figure 1: AFOC Pseudo Algorithm

Table 1: Elimination of Codons by various search methods made in the attribute space.

S.No	Search Method	Cut-Off usage percentage for selection	Selected Attributes for removal	Total Number (cumulative) of Attributes to be removed with respect to cut-off
1	Best First Search	0%	CTC	1
		10%	GAA,GAT,TTA	4
		20%	GCG,CAG	6
		30%	CGC,CCC,TCA,TCC,TTG	11
2	Genetic Search	0%	-	0
		10%	GCG	1
		20%	ACC	2
		30%	AGC,ATA,GAA,GAC,CAG,CCA,TAT,TCC,TTG	11
3	Greedy Stepwise Search	0%	CTC	1
		10%	GAA,GAT,TTA	4
		20%	GCG,CAG	6
		30%	CGC,CCC,TAG,TCA,TCC,TTG	12
4	Linear Forward Search	0%	GAC, GCT, GTA, CAA, CAC	5
		10%	AGA,ACA,GAA,GAT,CAG,CCA	11
		20%	AAA,GCG,GCC,GTT,CGC	16
		30%	ACC,CAT,CCC,CTC,TAG	21

Description of Methods

Our main objective is to search for the subsets of C = {c_1.,c_2,…c_{64}} where each c_i is a codon such that every subset maintains the key relationship while clustering as the parent set C.

Extraction of frequencies of codons of a given mtDNA sequence can be made easier in java implementation as reflected in the following pseudo code. The input mitochondria DNA sequences corresponding to different species that are to be clustered in FASTA format[7] stored in a flat file. The output is an NX64 matrix representing the number of occurrence of the codons for each species. The following search methods are applied to identify the codons as in Table 1. These methods use a search algorithm to search through the space of possible codons and evaluate each subset by running a model on the subset.

Best First Search

This method[11] searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions (by considering all possible single attribute additions and deletions at a given point).

Genetic Search

This method [12] performs a search using the simple genetic algorithm described as encoding of the problem in a binary string. random generating of a population representing a group of possible solutions. reckoning of a fitness value for each subject, selecting of the subjects that will mate according to their share in the population global fitness. making crossover and mutations and iterate these for expected convergence

Greedy Stepwise Algorithm

This method [14] performs a greedy forward or backward search through the space of attribute subsets. It may start with no/all attributes or from an arbitrary point in the space. Stops when the addition/deletion of any remaining attributes results in a decrease in evaluation. It can also produce a ranked list of attributes by traversing the space from one side to the other and recording the order that attributes are selected.

Linear Forward Search

This method [13] is the extension of Best First Search. Takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected. The search uses either the initial ordering to select the top k attributes, or performs a ranking (with the same evalutator the search uses later on). The search direction can be forward, or floating forward selection (with optional backward search steps).

Experiment setup and Evaluation of Search Methods

We used java implementation for Codon extractions from the given original sequence and enumerated all the frequencies stored as attribute relation flat file format. These files are used to generate the list of Condons based on the usage frequencies for classification into the species type. The input sequence which is a mitochondrial genome sequence pertaining to 11 different species has length around 16K. After pre-processing the input data for attribute selection in Weka tool which is again implemented in Java [15], the results are tabulated as shown in the Table-1.

The remaining key features (codons), after eliminating the features as prescribed in the table, are used to draw the final Phenogram and it remains invariant for all the subsets as in Fig 2. From the results we find linear Forward Search finds smallest subset by eliminating most number of codons from the original set C.

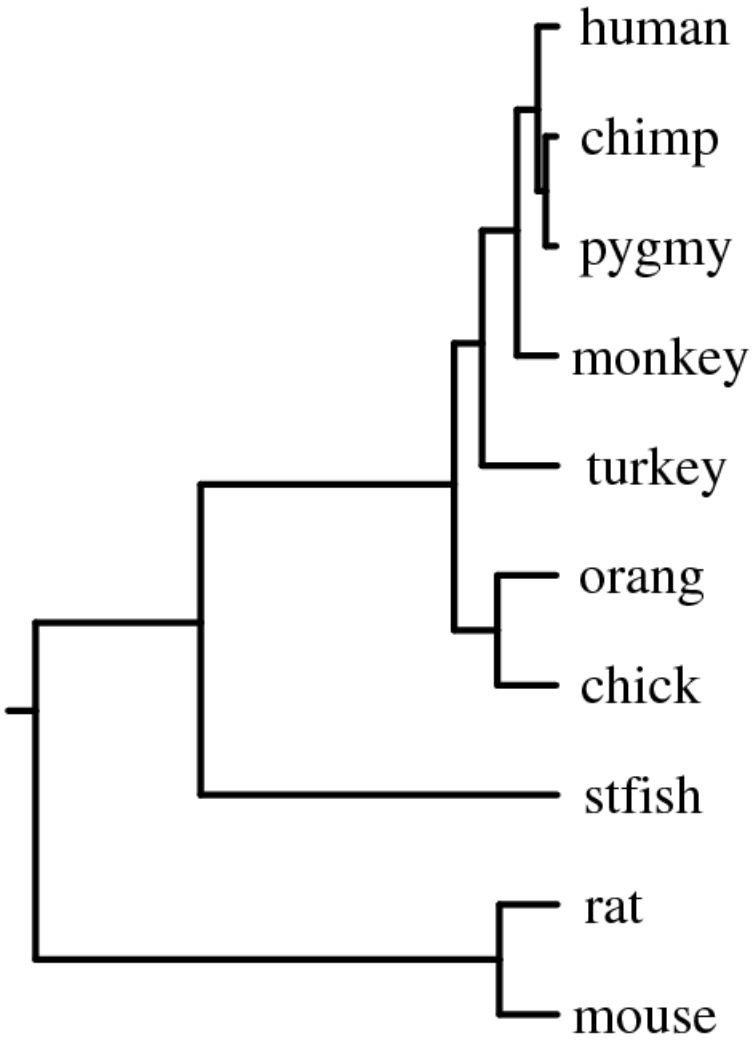

Figure 2: Phenogram representing the resulting clusters

Conclusions & Future work

In this paper a new approach based on the frequencies of codons in the mitochondria sequence and the selecting the key codons (predominant features) is introduced for clustering. Experiments are reported 16 samples of sequence for 11 different species. The clustering quality from this approach happens to be comparable with the standard methods.

The search methods proposed here further tuned with additional parameters of the base algorithms for better performance in the way of minimizing errors or improving the accuracies. Here, mitochondrial genome sequences pertaining to different species are used for building the classifier. As a future work the same approach may be applied by considering the nuclear DNA sequences of more number of different species.

Acknowledgements

The author would like to convey profound thanks to the management of Bharath University for their support and encouragement for this research work.

References

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

[2] UPGMA[online],Available: http://www.cs.cornell.edu/courses/cs426/2003fa/week10%20phylogenetic%20trees.pdf

[3] The**Needleman-Wunsch-algorithm** for sequence alignment [online] Available: *home.olemiss.edu/~pasaxena/Needleman-Wunsch.ppt*

[4] Arthur M. Lesk , Introduction to Bioinformatics , ISBN (Pbk)0 19 925196 7,United States by Oxford University Press Inc, 2002.

[5] Margaret H.Dunham (2006): Data Mining-Introductory and Advanced Concepts, Pearson Education.

[6] http://blast.ncbi.nlm.nih.gov/Blast.cgi

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

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

[9] "Special Issue on Bioinformatics, Part I:Advances and Challenges," Proceedings of the IEEE, vol90,November2002.

[10] Tajunisha and Saravanan, "An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-Means", International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011.

[11] Russell S., Norvig P(2003): Artificial Intelligence – a modern approach Pearson Education .
[12] David E. Goldberg (1989).:Genetic algorithms in search, optimization and machine learning.. Addison-Wesley.
[13] Martin Guetlein, Eibe Frank, Mark Hall, Andreas Karwath(2009): Large Scale Attribute Selection Using Wrappers. In: Proc IEEE Symposium on Computational Intelligence and Data Mining, 332-339.
[14] Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford (2009). Introduction to Algorithms (3rd ed.). MIT Press
[15] Source for Weka Available www.cs.waikato.ac.nz/ml/weka/index_documentation.html