Classification of Different Species Families using Clustering Algorithm

D.Chandravathi[1¥], TMN Vamsi[2†], A. M.Sowjanya[3],Allam AppaRao[4]

1. **Asst. Professor, Dept. of MCA, GVP College for Degree and PG Courses, Visakhapatnam.**
2. **Associate Professor, Dept. of MCA, GVP College for Degree and PG Courses, Visakhapatnam.**
3. **Asst.Professor, Dept of CSSE,AUCollege of Engineering, Visakhapatnam.**
4. **Vice- Chancellor, JNTUK, Kakinada.**

[¥]Correspondence author
[†]Research Scholar, JNTUH, Hyderabad.

Abstract:

The division of similar objects into groups is known as Clustering. The main objective of this implementation is classification of DNA sequences related to different species and their families using Clustering Algorithm- Leader-sub leader algorithm. Clustering is done with the help of threshold value of scoring matrix. It is another simple and efficient technique that may help to find family, superfamily and sub-family by generating sub clusters. From this analysis there may be a chance that members in sub-cluster may be affected if one of the leader clusters gets affected.

Keywords: Clustering, DNA Sequences, Sub cluster, scoring matrix, superfamily

Introduction:

Clustering is an important technique, which helps to improve classifications and also gives meaningful groupings/partitions [1]. Classification of different species can be done using Clustering techniques which helps to find family, superfamily and sub-family [2]. The Problem we have considered here is: Given a set of DNA sequences, calculate the threshold values using either of the methods for classification of DNA related sequences of different species accurately [5].

This paper is organized as follows. In section 2, two methods are used to find the threshold values and also the materials used. Section 3 contains the algorithm that is used for finding the threshold values and also choosing leaders and sub-leader for clustering different species. Section 4 contains experimental results for the proposed method. Finally, section 5 contains the conclusion and scope for further research.

Methods and Material:

In this method the input is taken in the form of DNA (nucleotide) sequences of Fasta format are downloaded from NCBI (www.ncbi.nlm.nih.gov) repository. In this method pairwise sequence alignment is used (global alignment) for finding the Scoring matrix. Global Alignment methods are based on Needleman and Wunsch algorithm [3]. Depending on the scoring matrix, threshold values are calculated [4]. We can observe that the score for the same DNA sequences (si, si) are always high and are excluded than the rest (si, sj) where i, j are number of the sequence. Further the threshold value is calculated with either of the two methods.

Method 1:

By taking the sum of least and highest value of Scoring matrix and is termed as 'T', of the two DNA sequences and excluding the similarity scores of the same sequences (si, si), we determined the threshold value termed as 't' by considering either of the following conditions:

taking the mean T
less than mean T
greater than T.

By choosing a Leader randomly among the sequences, clusters are generated by following any of the three conditions for threshold value (t) [5]. Clusters are generated by considering the values in the scoring matrix, which are greater than the threshold value (t). Similarly, Sub-clusters are also generated.

Method 2:

Threshold value is calculated by taking the average values for all the combination of (si, sj) and comparing with Leader sequence from which clusters will be generated by taking the sequences greater than threshold values (t). Similarly, Sub-clusters are also generated.

Algorithms:

Method 1 :Algorithm 1: Leader Algorithm

Step 1: Select the threshold value

a. Calculate T, T=Max+Min of the Scoring matrix (excluding (si, si))

b. Calculate the threshold value (t) of T by either taking the mean of T i.e. t=T/2 (condition 1) which is 50% of T. and normalize 't'.

or less than 't' (condition 2) i.e.,(t<50%)

or greater than 't' (condition 3) i.e.,(t>50%)

Step 2 : Choose a leader from the set of the sequences L and add it to the Leaders list.

Step 3: for all the sequences, i=2 to n
{
if (similarity Score with the nearest leader>t)
{
Assign it to the nearest leader
Mark the cluster number
Add it to the number list of the Cluster
Increment the cluster count
}
else
{
Add it to the leader list
Increment Leader count}}

Algorithm 2: Leader-Subleader Algorithm

Step 1: From the Clusters Generated choose a Leader L.
Step 2: Select 't' by considering the method 1 conditions
Step 3 :for I=1 to L
{
Choose a Sub-leader
For j=2 to members of ith cluster
{
if (similarity Score with the nearest subleader> t)
{
Assign it to the nearest leader
Mark the cluster number
Add it to the number list of the Cluster
Increment the cluster count
}else
{
Add it to the subleader list
Increment SubLeader count
}}

Method 2 : Algorithm 1: Leader Algorithm

Setp 1: Select the threshold value
a. Calculate T, T= (Max+Min) of the Scoring matrix (excluding (si, si))
b. Calculate the threshold value t= avg(T) and normalize it.
Step 2: Choose a leader from the set of the sequences L and add it to the Leaders list.
Step 3: for all the sequences, i=2 to n
{if (similarity Score with the nearest leader>t)
{Assign it to the nearest leader
Mark the cluster number
Add it to the number list of the Cluster
Increment the cluster count
}else
{Add it to the leader list

Increment Leader count
}}

Algorithm 2: Leader-Subleader Algorithm

Step 1: From the Clusters Generated choose a Leader L.
Step 2: Select 't' by considering the method 2 .
Step 3: for i=1 to L
{Choose a Sub-leader
For j=2 to members of ith cluster
{if (similarity Score with the nearest subleader> t)
{Assign it to the nearest leader
Mark the cluster number
Add it to the number list of the Cluster
Increment the cluster count
} else
{Add it to the subleader list
Increment Subleader count
}}

Results and Discussions:

The algorithms are implemented in MATLAB 7 for the species of different category and the leader , subleaders are determined .All the results are shown in table 1 and pylogenetic tree was constructed and from the tree it is found that human are nearer to mouse and rat.The constructed tree was shown in Fig 1.

Conclusions:

In this paper, by properly selecting the threshold values, Clusters and subclusters are generated which gives classification accuracy which may be used to find family, super family and subfamily relationships. We use numerical data sets, text and web document collections in the above mentioned algorithms. Further related work can be done with large data sets of DNA sequences by generating new algorithms.

References:

[1]Arun K Pujari. Data Mining Techniques, Universities Press (India) Private Limited, 2000.
[2] Peter Clote, and Rolf Backofen, Computational Molecular Biologv -An Introduction, John Wiley & Sons,Ltd., August 2000.
[3]S.B. Needleman. and C.D. Wunsch, "A genaal method applicable to the search for similarities3n the amino acid sequence of the proteins," J, of Mol. Biology. 48. pp.
[4]V. Guralnik, and G. Karypis, "'A scalable algorithm for clustering sequential data," Proc. o/ Is' IEEE confer-ence on Data Mining, 2001.
[5] An efficient incremental protein sequence clustering algorithm P. A. Vijaya, M. Narasimha Murty and D. K. Subramanian Department of Computer Science and Automation Indian lnstitute of Science, Bangalore, lndia.

Experimental Results:

(Eg. cow, carp, human, fish, frog, seal, loach, rat, mouse, chicken. Exclude (si,si) scores to find the threshold value)
The experimental result is shown in the resultant scoring Matrix (Table1) of 10 sequences:

Table:1

	Cow	Whale	Seal	Mouse	Rat	Human	Carp	Loach	Frog	chicken
Cow	16.6767	11.1300	11.5500	11.3767	11.2233	12.5600	13.0433	13.2400	13.5000	11.3533
Whale	11.1300	17.9000	11.7967	11.1333	14.1133	11.4167	11.5133	11.6267	11.5600	12.72
Seal	11.5500	11.7967	18.4000	11.8767	11.8267	11.6633	11.8200	11.6433	11.7067	12.0133
Mouse	11.3767	11.1333	11.8767	18.0567	11.2000	11.7000	11.7633	11.4433	11.7600	11.17
Rat	11.2233	14.1133	11.8267	11.2000	18.3167	11.3333	11.7867	11.6567	11.3900	12.5833
Human	12.5600	11.4167	11.6633	11.7000	11.3333	16.8533	13.8900	12.8000	12.5700	11.7567
Carp	13.0433	11.5133	11.8200	11.7633	11.7867	13.8900	16.9733	12.8567	12.9033	11.6467
Loach	13.2400	11.6267	11.6433	11.4433	11.6567	12.8000	12.8567	17.3433	13.3033	11.57
Frog	13.5000	11.5600	11.7067	11.7600	11.3900	12.5700	12.9033	13.3033	16.9367	11.7533
Chicken	11.3533	12.7200	12.0133	11.1700	12.5833	11.7567	11.6467	11.5700	11.7533	17.1233

Phylogenitic tree (Fig1) is constructed with the DNA sequences and then compared with the clusters generated with the help of threshold values. (http://www.ebi.ac.uk/Tools/es/cgi-bin/clustalw2/result.cgi)

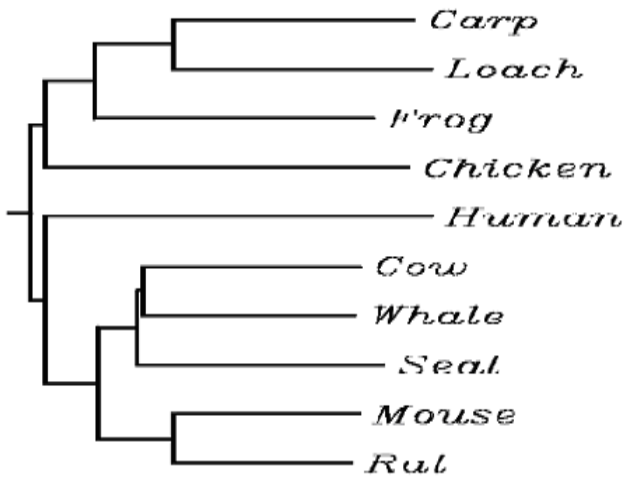

Fig:1-Phylogenetic Tree for the query sequences