Data Mining Project-1 (Report)

Hierarchical Clustering

Group members:

1) Nikhil Joshi : 2015A7PS0179H
2) Abhishek Savani : 2015A7PS0087H
3) Bhavesh Gawri : 2015A7PS0116H
4) Tilak Mundra : 2015A7PS0121H

Dataset Used:

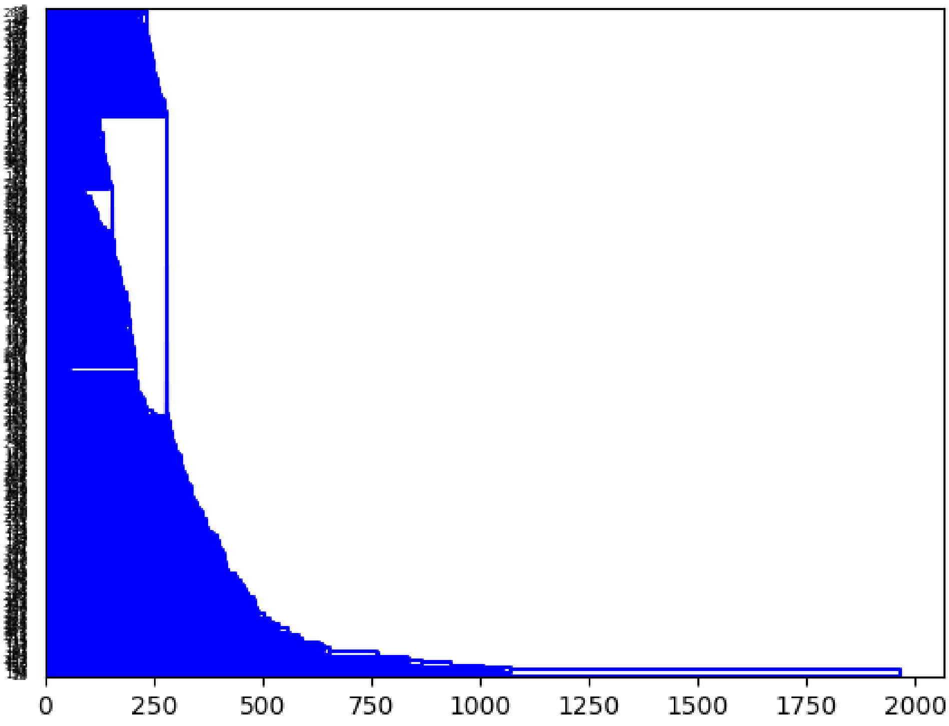
Human DNA dataset which comprises of 311 human DNA sequence which is in fasta format. In bioinformatics, **FASTA format** is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes

Data pre-processing:

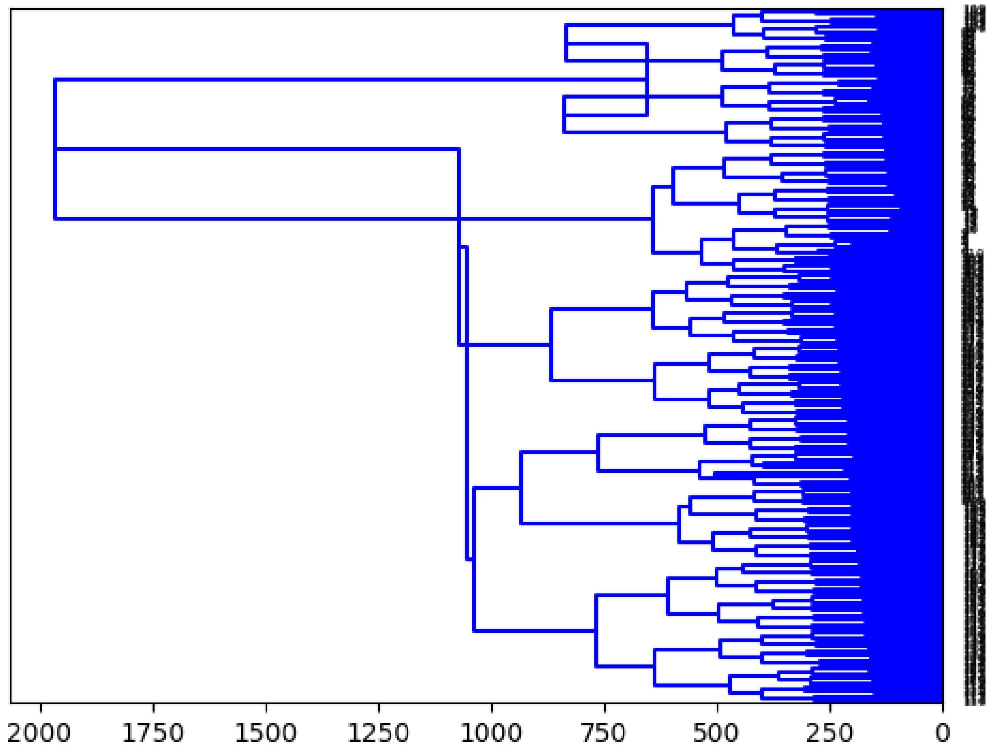
- **Parsing of dataset** : it is carried to convert dataset in 2 tuple matrix with first attribute a name and second attribute as sequence
- **Similarity Matrix Calculation:**
  Calculating weighted edit distance using dynamic programming algorithm.

Output:

Dendrogram Agglomerative:



Dendrogram divisive:



Linkage and distance metric:

- *Linkage matric consist of 4 attributes for each linkage 1st attribute is contains id which is combined with 2nd attribute.3rd attribute consist of distance between two points and 4th attribute consist number of points in new cluster formed or old cluster broken.*
- *Distance matric used is weighted edit distance which consist of penalty if not matching.*

Observations:

	Time Taken
Agglomerative	*3.73 sec*
Divisive	*5.21 sec*