Data Mining

Assignment 2 Report

Team members:

C Shri Akhil	2015A3PS0314H
Aditya Desai	2015A3PS0211H
Amey Athale	2015A7PS0077H
Ahraz Rizvi	2015A7PS0012H

The goal of this assignment is to construct the phylogenetic tree based on Amino Acid sequences using both Agglomerative (bottom-up) and Divisive (top-down) Clustering methods. We then contrast between the results of the two.

Dataset Used: Amino Acid Sequence of Human Gene

Preprocessing Done: *pointsDistanceMatrix.csv* file was generated using *Assignment2.cpp* since it was a severe bottleneck to compute in Python. It takes around 3 minutes to get generated.

Linkage and distance metric used: Average Linkage, Global Sequence Alignment

Type of data it can cluster properly: Any kind of data that has a measure of similarity between any two clusters. Measure of similarity here means a quantifiable value which can be used for comparison.

Comparison of dendrogram plot of top-down and bottom-up clustering: The dendrograms are quite different since bottom-up clustering is based on nearest neighbours (local minima) while top-down is based on farthest points (global maxima). Naturally, the top-down division relatively more uniformly split.