Name	ID No
Avinash Rathore	H20171030069
Sourabh Sethi	H20171030072
Meghana Kumar	2014B5A70932H

DATASET USED

Human Gene DNA Sequences

The dataset constitutes of **311** DNA sequences in FASTA format

PREPROCESSING DONE ON DATA

The Proximity Matrix is preprocessed and stored in an 'out.txt' file. It is then loaded onto, when required.

FORMULAE USED

Single Linkage:
Shortest distance between two clusters is considered.
The minimum of which is taken as the metric

$$\min\{\, d(a,b) : a \in A,\, b \in B \,\}.$$

DISTANCE METRIC
Global Sequence Alignment: Using Dynamic Programming

The similarity between two DNA sequences is assessed using
1. gap
2. substitution
3. match

```
for (int i = 1; i <= sequenceA.length(); i++)
    opt[i][0] = opt[i - 1][0] + gap;
for (int j = 1; j <= sequenceB.length(); j++)
    opt[0][j] = opt[0][j - 1] + gap;

for (int i = 1; i <= sequenceA.length(); i++) {
    for (int j = 1; j <= sequenceB.length(); j++) {
        int scoreDiag = opt[i - 1][j - 1] +
                (sequenceA.charAt(i-1) == sequenceB.charAt(j-1) ?
                    match : // same symbol
                    substitution); // different symbol
        int scoreLeft = opt[i][j - 1] + gap; // insertion
        int scoreUp = opt[i - 1][j] + gap; // deletion
        // we take the minimum
        opt[i][j] = Math.min(Math.min(scoreDiag, scoreLeft), scoreUp);
```

WHERE SINGLE LINKAGE WORKS

Single linkage works with irregular shaped clusters.
But it is sensitive to outliers.

AVERAGE DISTANCE
Every pair of each cluster, is averaged out, and the distance is found.

This has been used in calculation of the Z Matrix.

Z-MATRIX

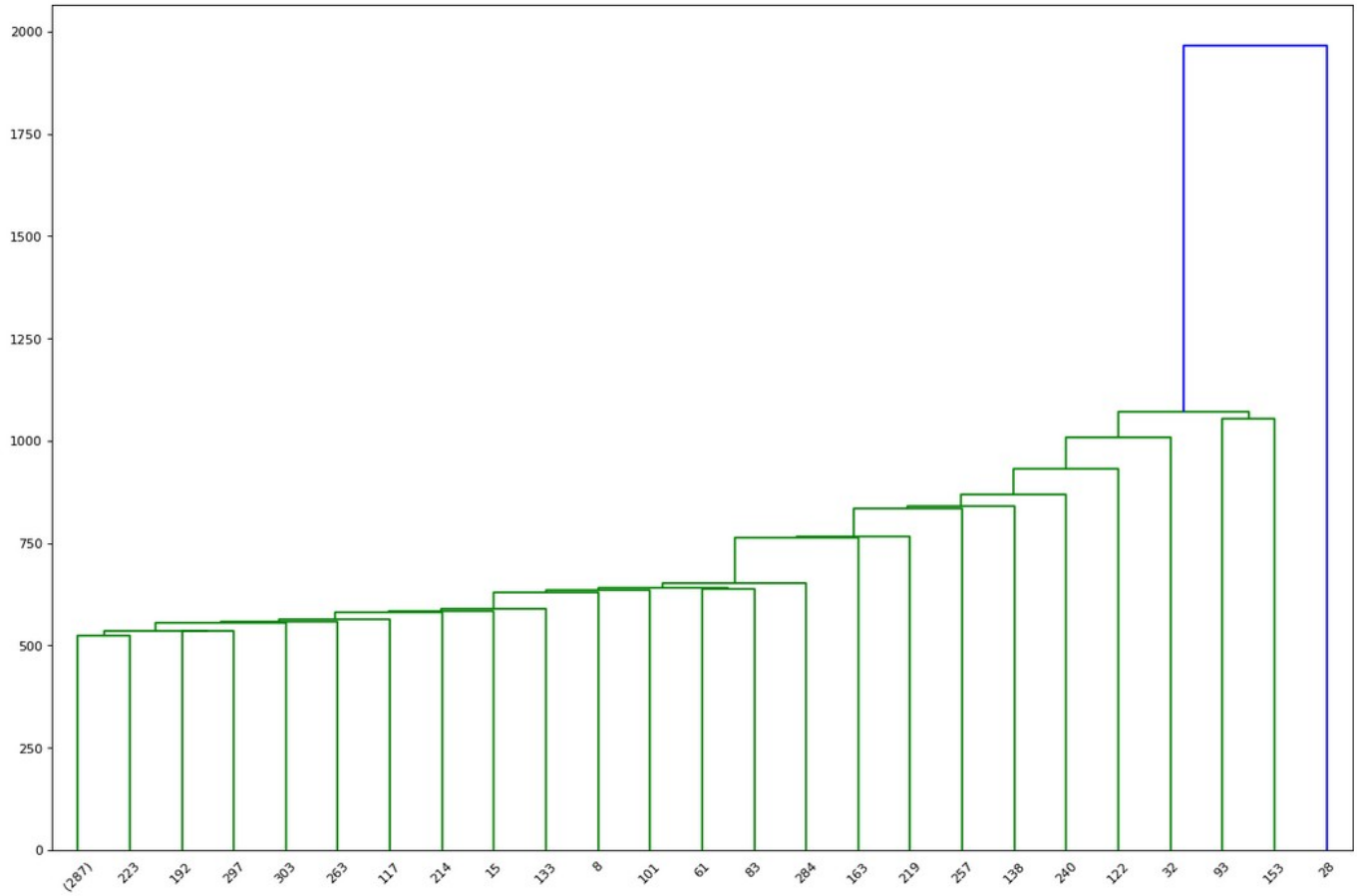
The Z-Matrix is (n-1)x4 matrix on SciPy

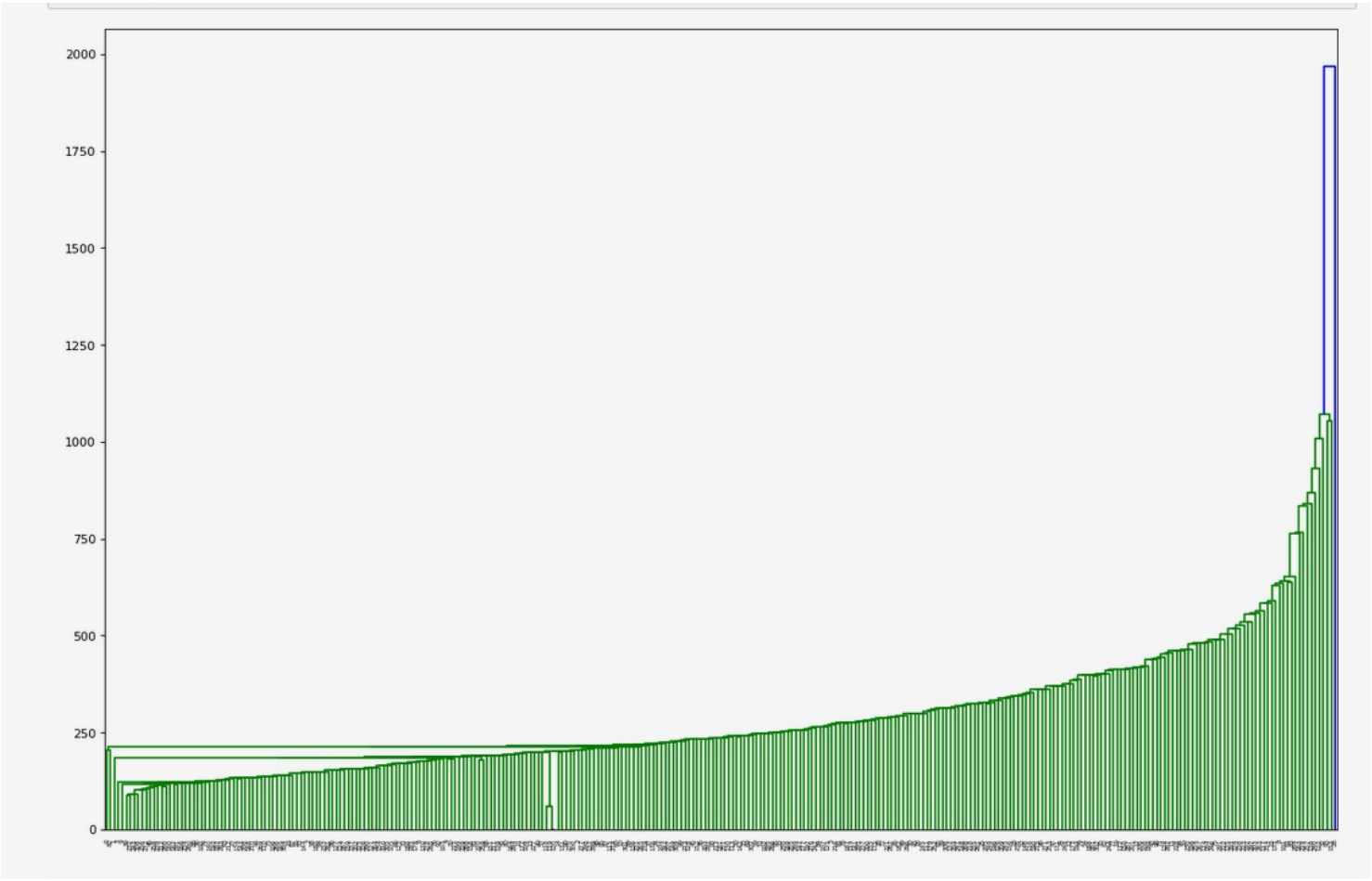
n-1 rows for every merge

Column 1: ith cluster merging
Column 2: jth cluster merging
Column 3: distance between the clusters
Column 4: Number of original points in the cluster

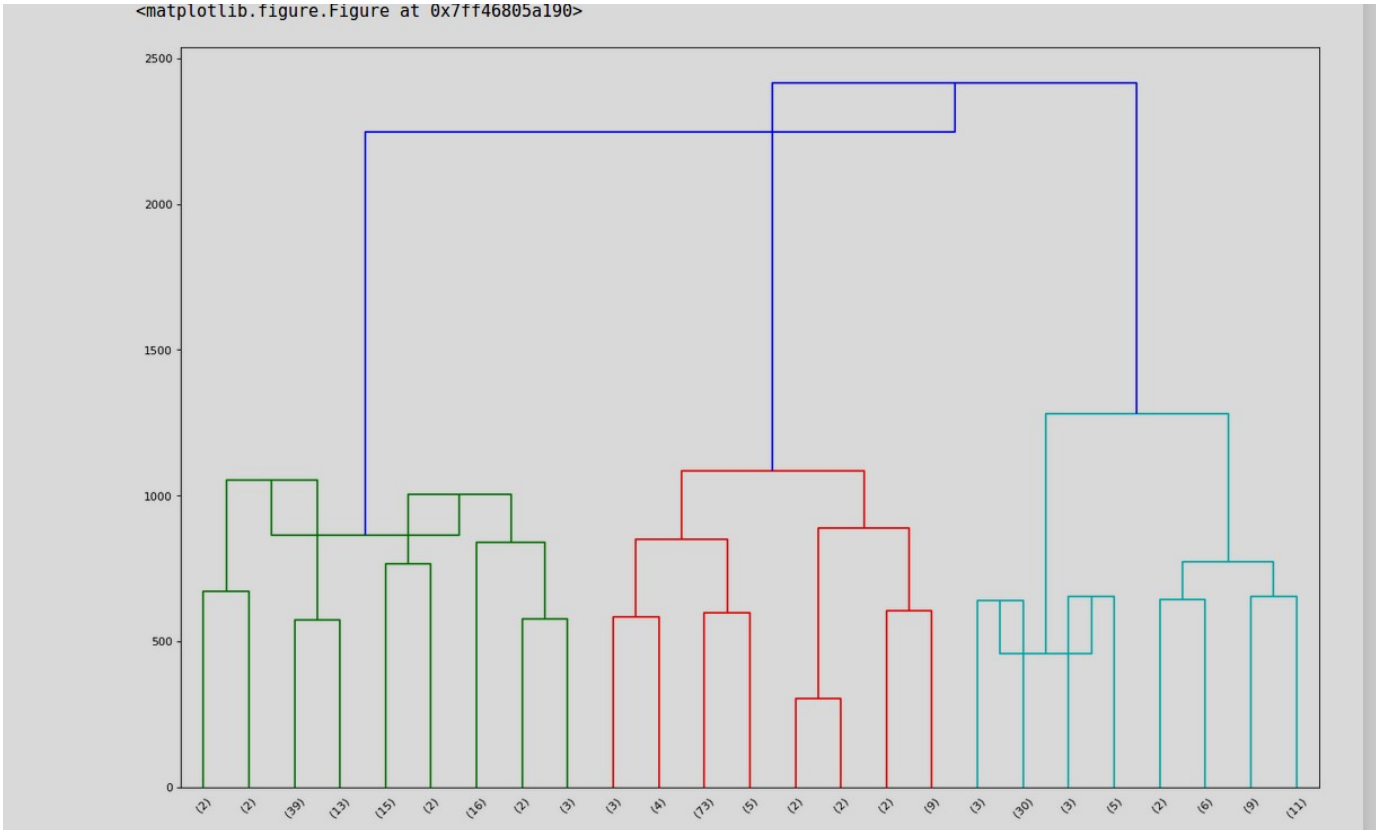
The Z matrix is an input in SciPy to plot Dendrograms.

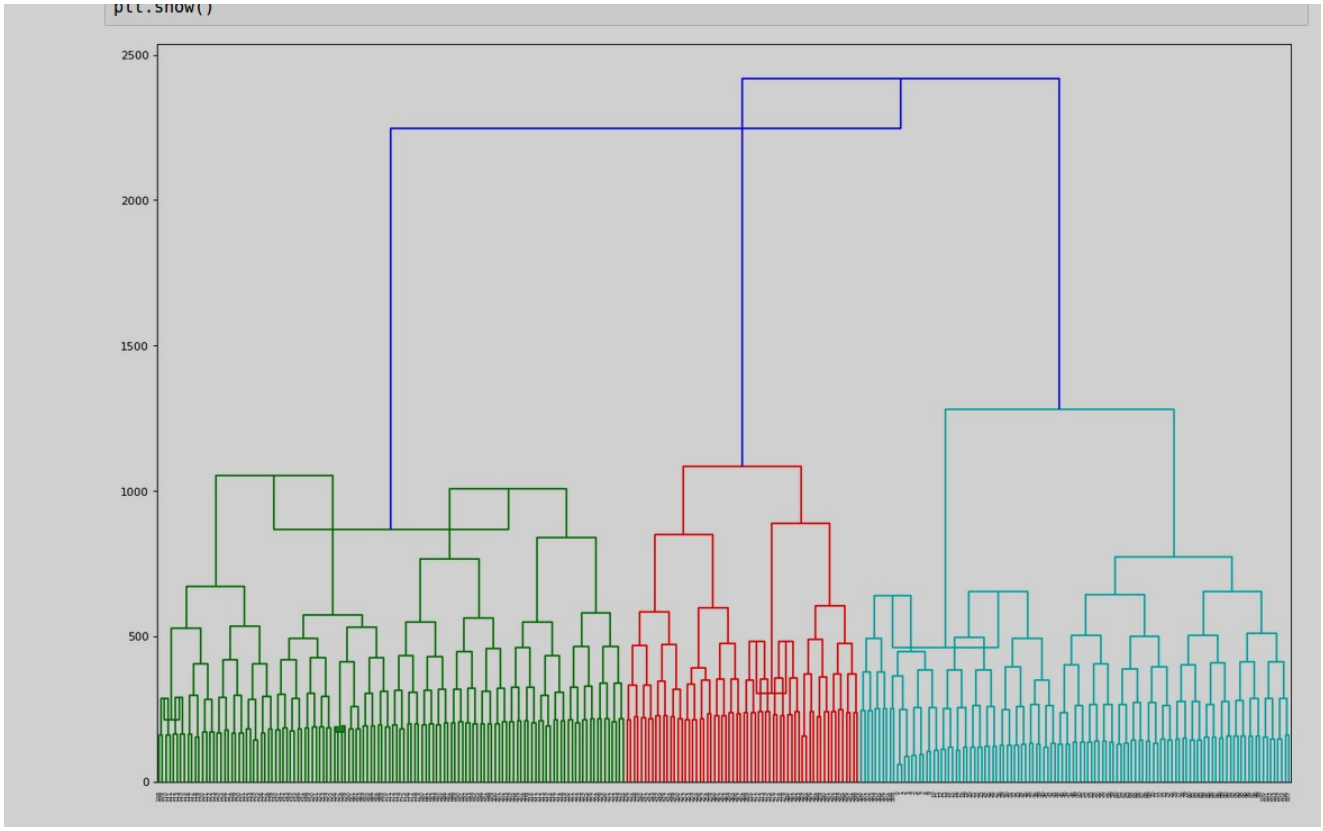
AGGLOMERATIVE CLUSTERING





DIVISIVE CLUSTERING





The Agglomerative Clustering proves to be better for larger datasets. It scales better.

In Agglomerative Clustering shows a gradual connection between consecutive clusters.

The Divisive Clustering is required for more holistic view, that converges from a single cluster.