Cluster healthy and Diseased patients based on their gene data

Saturday, March 7, 2020 7:53 PM

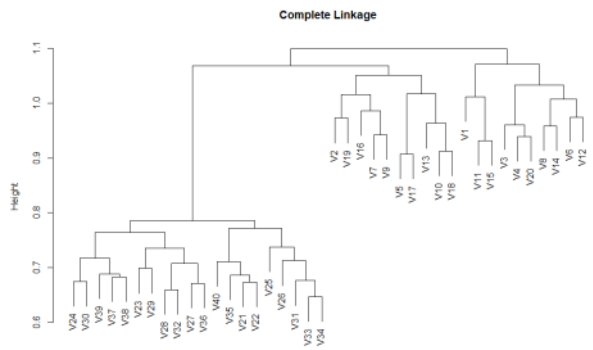
Data Source : Gene_data.csv

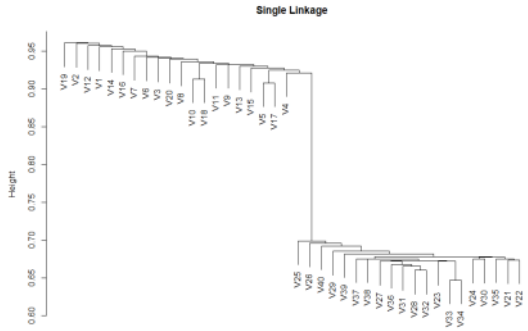
Observations :

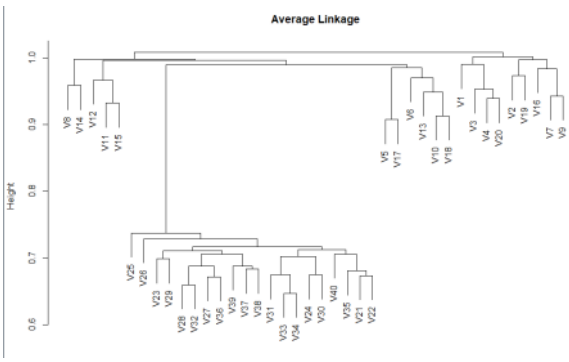
1. Data of 40 tissues for 1000 genes is provided.
2. 20 of these tissues are from healthy patients and 20 are from diseased group.

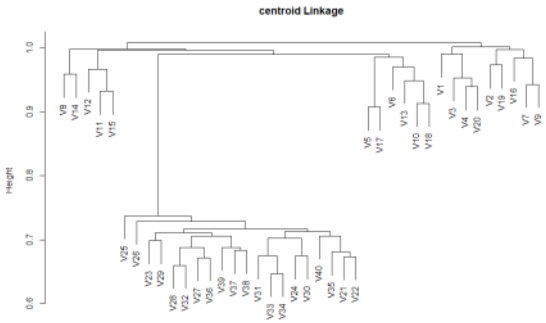
Hierarchial Clustering with different linkages :

3. Complete Linkage : Maximum intercluster dissimilarity









Inference

1. We are getting two clusters for single and complete and three clusters for average and centroid
2. Single linkage produces an unbalanced tree
3. Average and Centroid linkage dendrograms are exactly same in our case.
4. Complete, Average, Centroid dendrogram are more balanced dendrograms

Find out the gene which differs the most for healthy and diseased patients.

- In order to look for the gene which differs the most across healthy and diseased patients, we can perform Principal ComponentAnalysis and look at loading vectors to figure out which gene varies the most.
- The first principal component loading vector defines a direction in feature space along which the data varies the most. If weuse only the first principal component, we will be losing some essential information
- The loading score values for all principal component for every gene is summed up to find out the effect of each gene.
- The summed value for the 1000 gene is sorted in descending order and the top 10 genes having maximum impact on clustering arefound
- The top 10 most differing genes are **865 68 911 428 624 11 524 803 980 822**