

PREDICTING COLON CANCER USING CLUSTERING MODELS BASED ON DNA
MICROARRAY DATA

Apoorv Saraogee
Northwestern University, Practical Machine Learning
June 10, 2022

Abstract

Colon cancer is a significant public health concern and leading cause of death in the older human population. The healthcare burden can be significantly reduced with earlier detection and preventative measures. Although genetic information has been shown to be altered in the early stage of the disease, DNA sequencing data is highly dimensional and must be analyzed using predictive algorithms. This study analyzes a DNA microarray dataset of 62 tissue samples (normal and cancerous) and 92 genes from the Princeton University Gene Expression Project. To analyze the dataset, the Rand index was compared for clustering models based on DBSCAN, kmeans and hierarchal clustering algorithms (single and ward). The kmeans model performed the best clustering of the data and had the highest Rand index of 0.796.

1. Introduction

Colon cancer is an intestinal disease that accounts for 11% of all cancer diagnoses worldwide (Bae et al. 2021). Gene expression studies provide an early detection method in patients as approximately 30% of the cases with colon cancer have a genetic predisposition (Shafi et al. 2020). However, as colon cancer is a multi-gene disease and the subsequently large number of dimensions of DNA microarray data, predictive algorithms must be used to extract meaningful information.

This study's central research topic is to analyze a subset of DNA microarray data derived from the Princeton University Gene Expression Project (Alon et al. 1999) using various clustering methods with 62 tissue samples (40 normal and 22 cancerous) across 92 genes. Analysis will first focus on finding the optimum number of clusters with three methods – silhouette/slope scoring, principal

components analysis and DBSCAN. Further research questions will use 2 clusters to compare the DBSCAN results with kmeans clustering and hierarchal clustering algorithms (single or ward linkage) using the Rand index and image representations. Other questions will also explore hierarchal clustering of tissue types instead of genes.

2. Literature Review

Many machine learning methods have been used on the full DNA microarray dataset from the Princeton Gene Expression Project including clustering, random forest, group lasso and SVM. Clustering has the highest reported accuracy of 87.1% when the full 2000 genes are used for classification (Bae et al. 2021). However, due to the high dimensionality of the data, more complex clustering algorithms are commonly used with feature selection as the first step (Bae et al. 2021; Shafi et al. 2020; Xie, Wang, and Wu 2019; Ben-Dor et al. 2004). This study also used the same subset of 92 genes used in other studies with data from the Princeton Gene Expression Project (Alon et al. 1999; Izenman 2008) and similarly analyzes it with clustering algorithms. However, instead of classification accuracy, this study uses the Rand index for comparisons.

3. Methods

This research will be conducted on a DNA microarray dataset derived from the Princeton Gene Expression Project with 92 genes or explanatory variables (Alon et al. 1999) across 62 tissue samples. Analysis is done in a Jupyter notebook using a kernel in R for visualization using imageplot and to find the number of clusters in the nClust as well as Python3 for analysis with clustering models in the sklearn package. The number of gene clusters was found using slope/silhouette scoring in R using the nClust package with 10 maximum clusters, principal

components analysis while preserving 90% of the genetic information and DBSCAN (Géron 2017) with an epsilon of 9.5 and minimum of 10 samples per cluster and silhouette coefficient scoring in Python3 using the sklearn package. Using the number of clusters=2, results from DBSCAN were compared with kmeans and hierarchal clustering methods (single or ward linkage) using the silhouette coefficients and Rand index calculated by comparing with the true tissue labels (cancerous or normal) in Python3. Additionally, tissue clustering is conducted using tissues as explanatory variables instead of genes for the hierarchal clustering models with ward and single linkage. Visualizations of the results was done with dendrograms for all hierarchal clustering models using t-SNE in the scipy package in Python3.

3. Results

The Rand index for each gene clustering model calculated using the true classification label of each tumor (cancerous or normal) were 0.779, 9,786, 0.772, 0.526 for DBSCAN, kmeans clustering, hierarchal clustering with ward linkage and hierarchal clustering with single linkage respectively. Although 18 components were required to preserve 90% of the genetic information in the PCA model, the optimal number of clusters was found to be 2 using DBSCAN with both silhouette and slope scoring in the nClust package with only 5 outlier tissue samples. The silhouette coefficient for the DBSCAN model was 0.229 with similarly calculated silhouette coefficients of 0.334, 0.332 and 0.048 for the kmeans and hierarchal clustering with ward and single linkage respectively. The silhouette coefficient for the hierarchal tissue clustering models as opposed to gene clustering was 0.393 for both ward and single linkage. The hierarchal clustering methods with ward linkage had the most separation between clusters as shown in the dendrograms in Figures 3, 4, 5 and 6. These results suggest the suitability of using kmeans clustering methods in

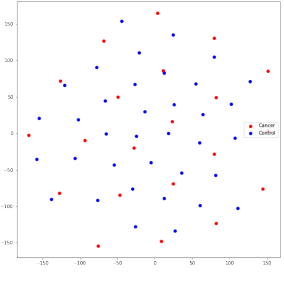
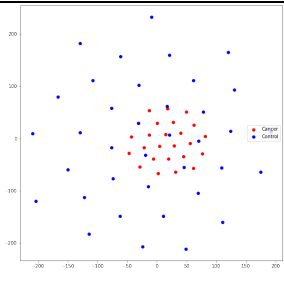
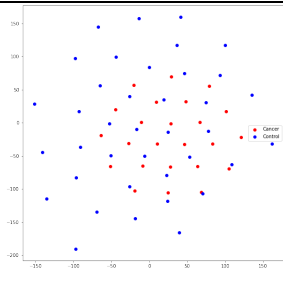
analysis of DNA microarray data. This is clearly seen in the t-SNE representation with clustering of samples seen between the cancerous and control groups as shown in Table 1. When comparing the clusters found with each method, agreement in Rand index was also seen in Table 1 with 0.895, 0.925 and 0.986 for the DBSCAN/kmeans, DBSCAN/hierarchical clustering and kmeans/hierarchical clustering with ward linkage.

4. Conclusions

Within this study, the kmeans clustering method was found to have the highest Rand index of 0.796 and corresponding highest silhouette coefficient of 0.334 for gene clustering as well as best clustering seen in the t-SNE projections as shown in Table 1. This is quite good compared to low values of 0.526 and 0.048 for the Rand index and silhouette coefficient for the hierarchical clustering method with single linkage. However, compared to two-way clustering algorithms that have up to 87% accuracy (Bae et al. 2021), this could be better. Tissue clusters are also clearly seen in the dendrograms and both have a high silhouette coefficient of 0.393. These suggest the value of two-way clustering methods for DNA microarray data. Block clustering methods could be implemented here by using the bicluster module in the sklearn package in Python3. Future studies could be with more genetic information and complete DNA sequences, because they have become increasingly common in the past decade, are capable to get better accuracy and can learn more generalizable features for other seemingly unrelated health conditions (Hurd and Nelson 2009).

5. Appendices

Table 1: Comparison of clustering algorithms DBSCAN, kmeans and hierarchal clustering with ward linkage with the Rand index and t-SNE visualizations from Python3.

	DBSCAN	Kmeans Clustering	Hierarchal clustering
DBSCAN		0.895	0.925
Kmeans clustering	0.895		0.968
Hierarchal clustering	0.925	0.968	

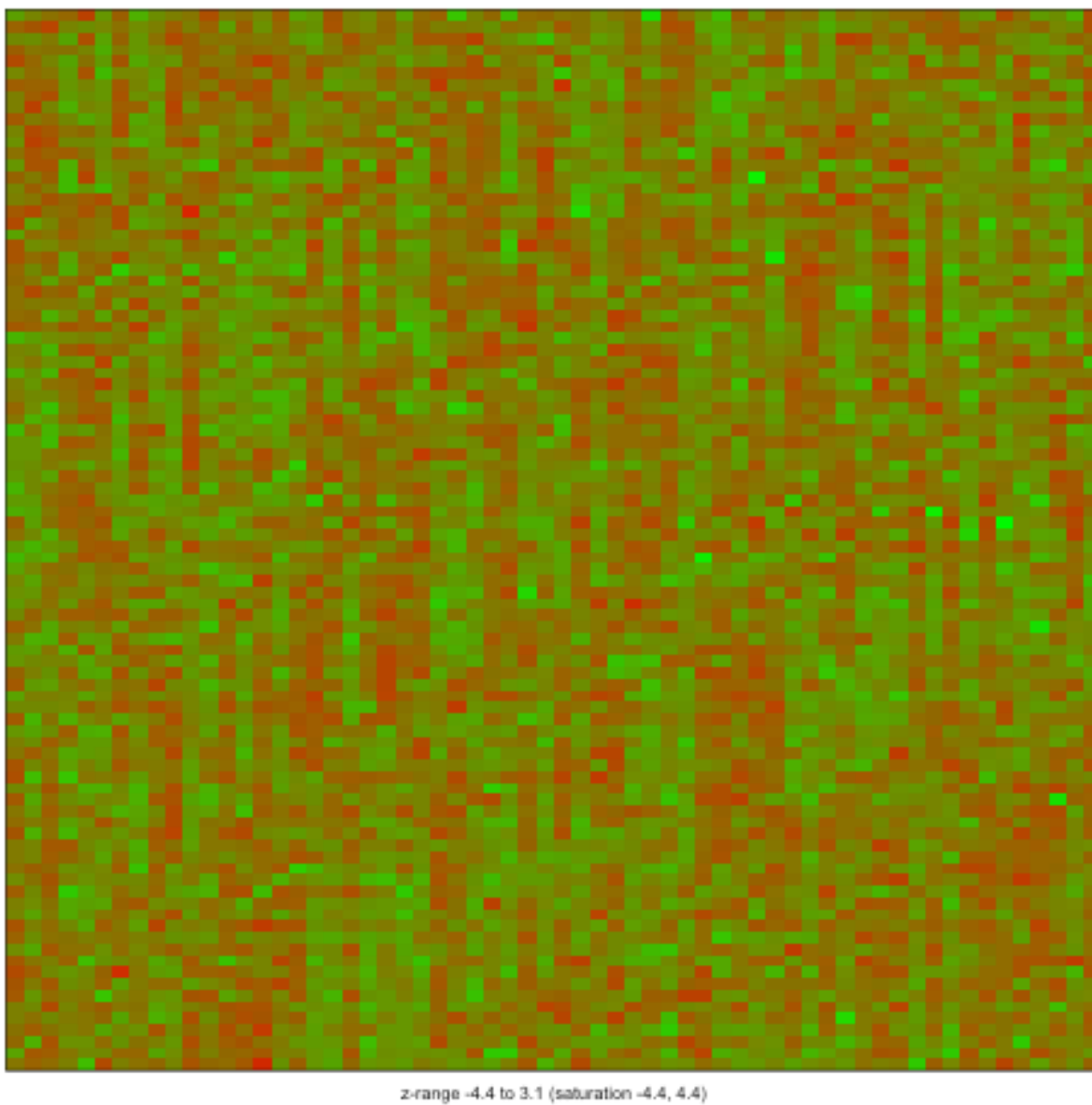


Figure 1. Visualization of the data using imageplot in R with tissues in the rows and genes in the columns.

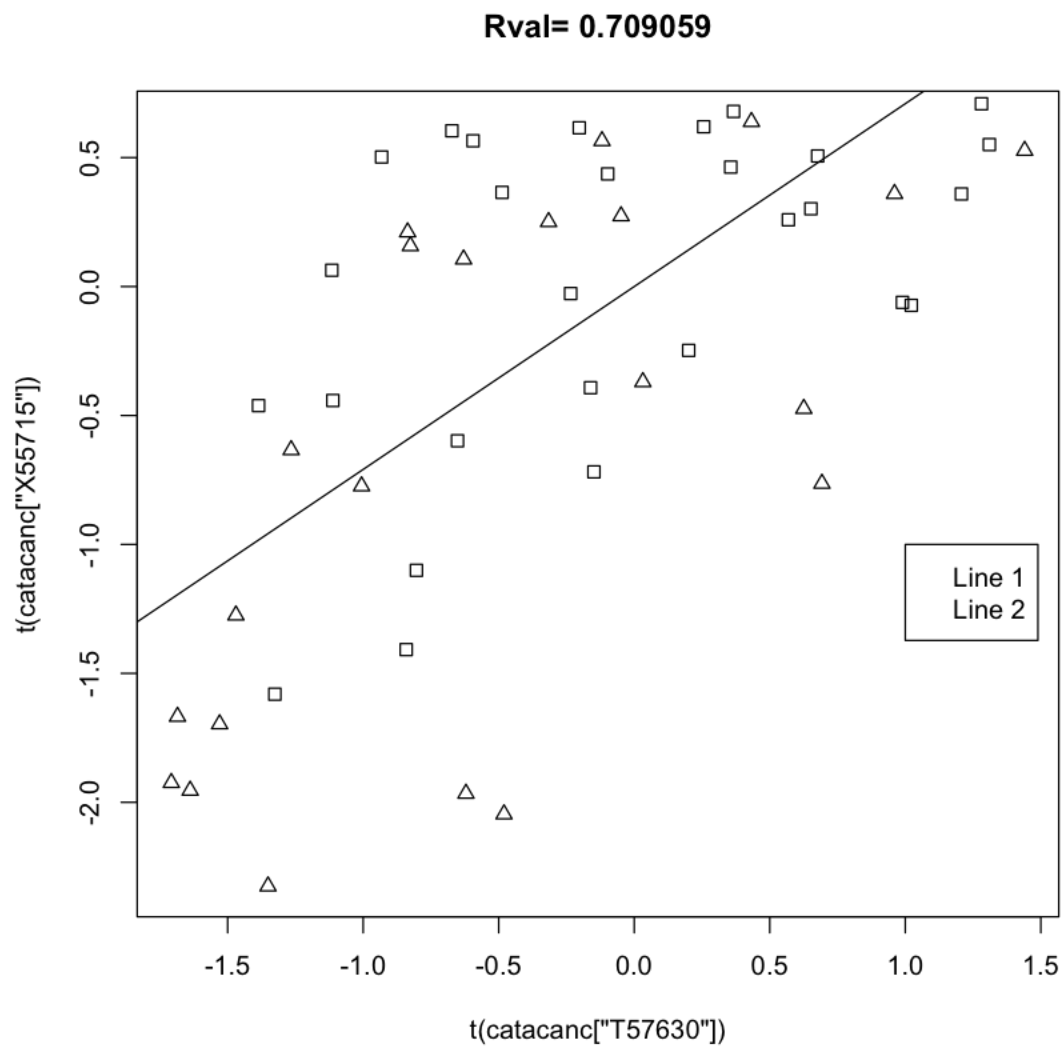


Figure 2. Reproduction of correlation R-values plot for a desired pair of genes (example EST numbers T57630 and X455716) across 62 tissue types (Alon et al. 1999).

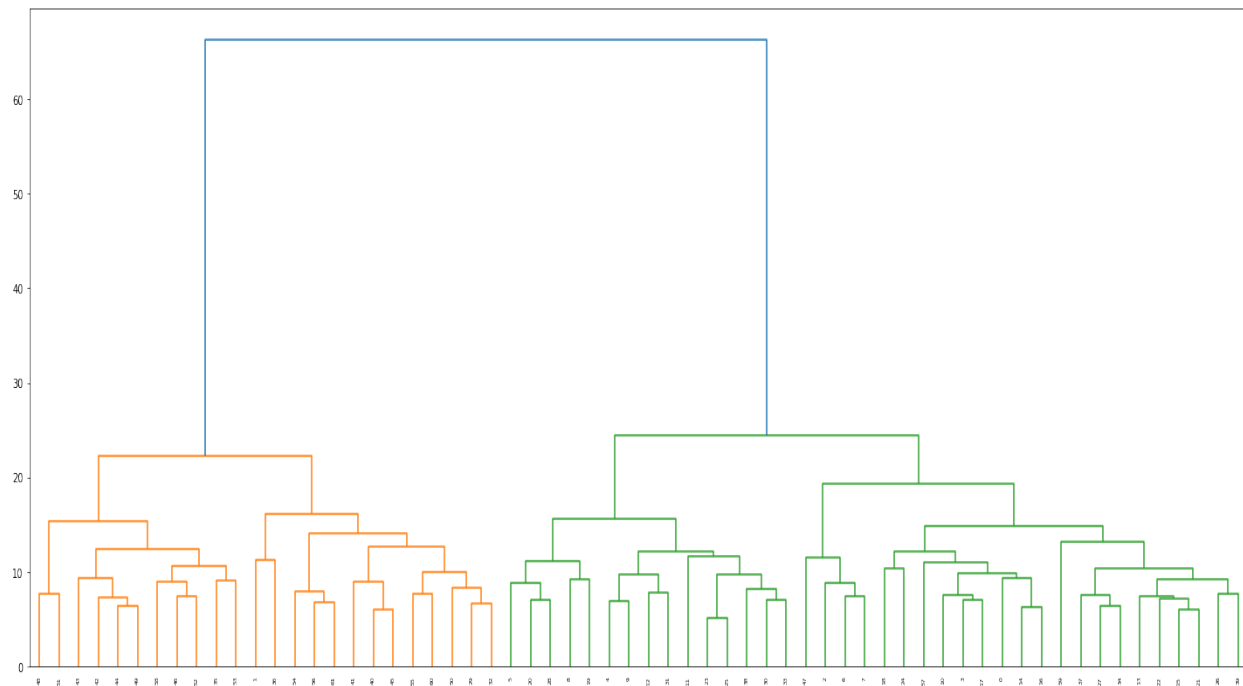


Figure 3. Dendrogram for hierarchal gene clustering method with ward linkage.

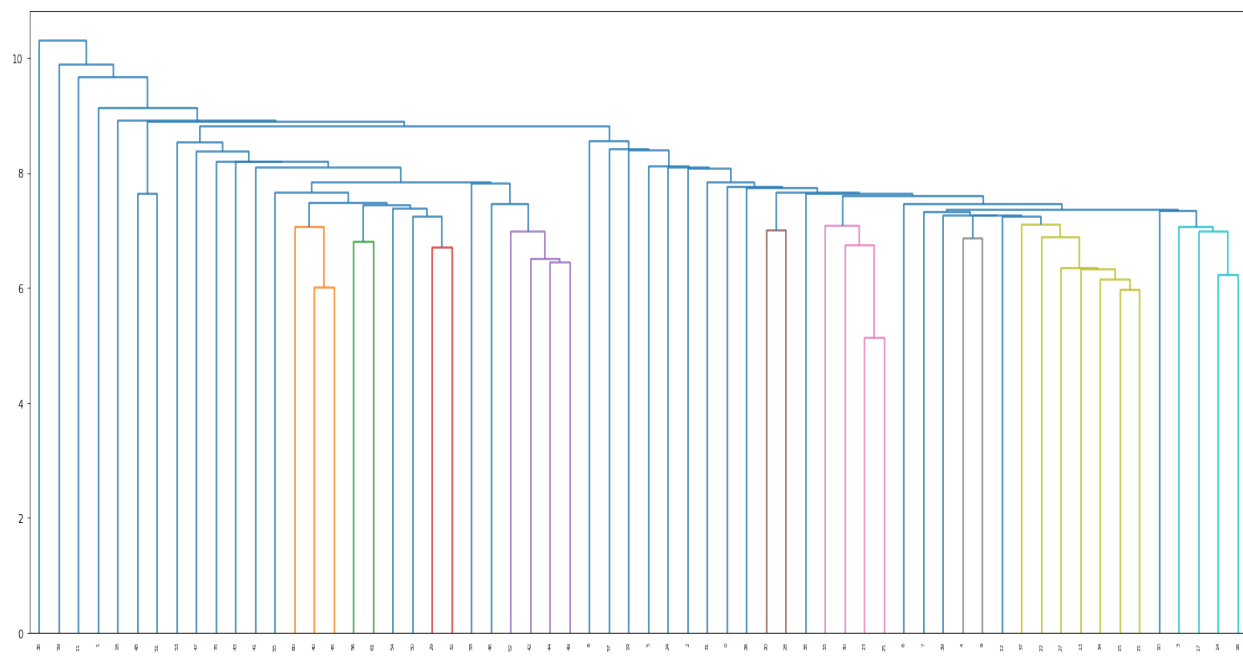


Figure 4. Dendrogram for hierarchal gene clustering method with single linkage.

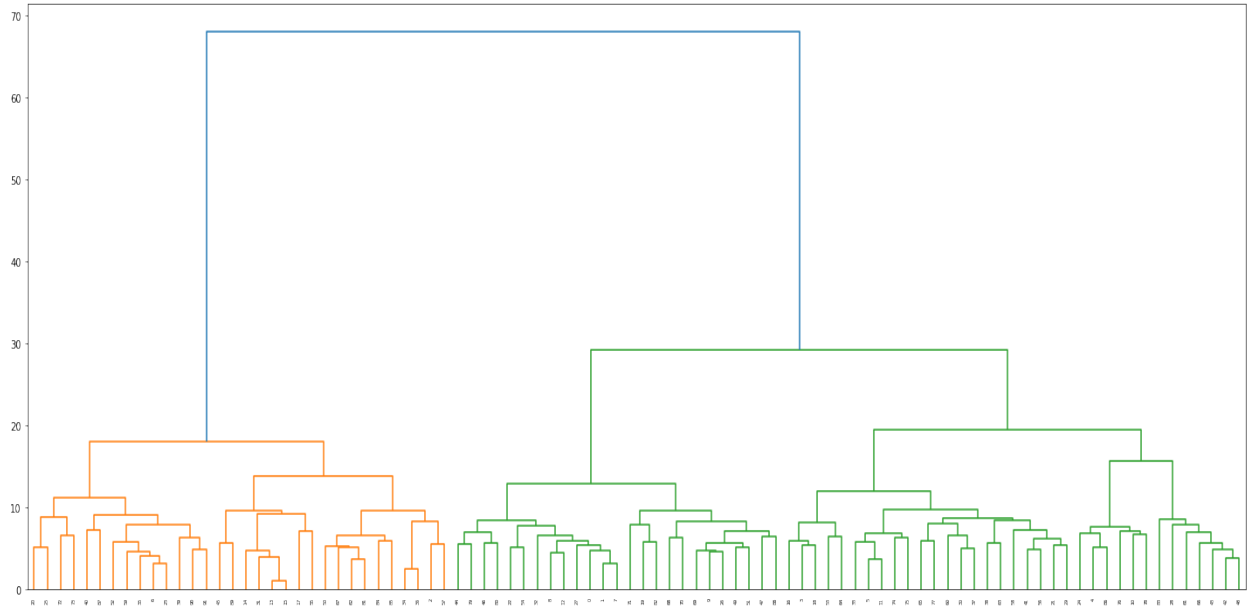


Figure 5. Dendrogram for hierarchal tissue clustering method with ward linkage.

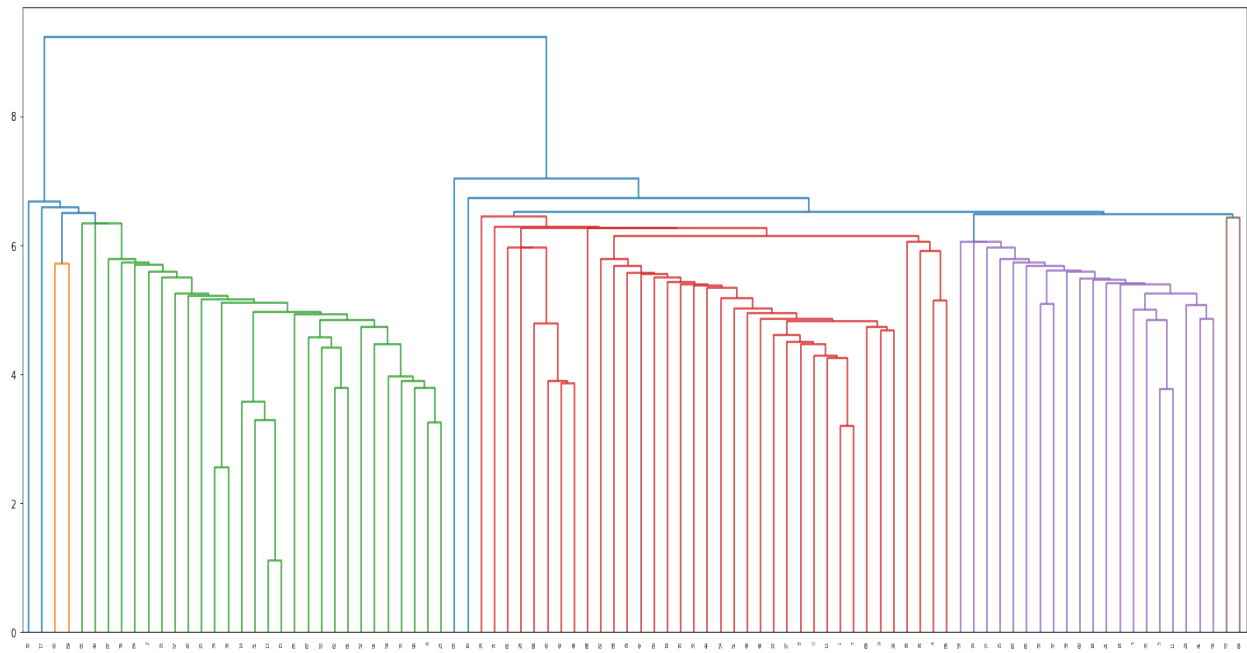


Figure 6. Dendrogram for hierarchal tissue clustering method with single linkage.

Supporting Files

- Assignment4.html
- Assignment4.py
- Assignment4.ipynb
- dna.csv

References

- Alon, U., N. Barka, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. 1999. "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays." *Proceedings of the National Academy of Sciences of the United States of America* 96 (12): 6745–50. <https://doi.org/10.1073/PNAS.96.12.6745>.
- Bae, Jin Hee, Minwoo Kim, J. S. Lim, and Zong Woo Geem. 2021. "Feature Selection for Colon Cancer Detection Using K-Means Clustering and Modified Harmony Search Algorithm." *Mathematics* 2021, Vol. 9, Page 570 9 (5): 570. <https://doi.org/10.3390/MATH9050570>.
- Ben-Dor, A., L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. 2004. "Tissue Classification with Gene Expression Profiles." *Https://Home.Liebertpub.Com/Cmb* 7 (3–4): 559–83. <https://doi.org/10.1089/106652700750050943>.
- Géron, Aurélien. 2017. "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2019, O'reilly)." *Hands-On Machine Learning with R*, 510.
- Hurd, Paul J., and Christopher J. Nelson. 2009. "Advantages of Next-Generation Sequencing versus the Microarray in Epigenetic Research." *Briefings in Functional Genomics* 8 (3): 174–83. <https://doi.org/10.1093/BFGP/ELP013>.
- Izenman, Alan J. 2008. "Modern Multivariate Statistical Techniques," Springer Texts in Statistics, . <https://doi.org/10.1007/978-0-387-78189-1>.
- Shafi, A. S.M., M. M.Imran Molla, Julakha Jahan Jui, and Mohammad Motiur Rahman. 2020. "Detection of Colon Cancer Based on Microarray Dataset Using Machine Learning as a Feature Selection and Classification Techniques." *SN Applied Sciences* 2 (7): 1–8. <https://doi.org/10.1007/S42452-020-3051-2/TABLES/9>.
- Xie, Juanying, Yuchen Wang, and Zhaozhong Wu. 2019. "Colon Cancer Data Analysis by Chameleon Algorithm." *Health Information Science and Systems* 7 (1). <https://doi.org/10.1007/S13755-019-0085-1>.