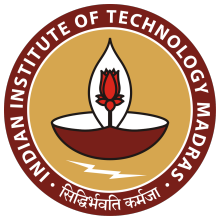
BT3041: Analysis and Interpretation of Biological Data

Assignment 2

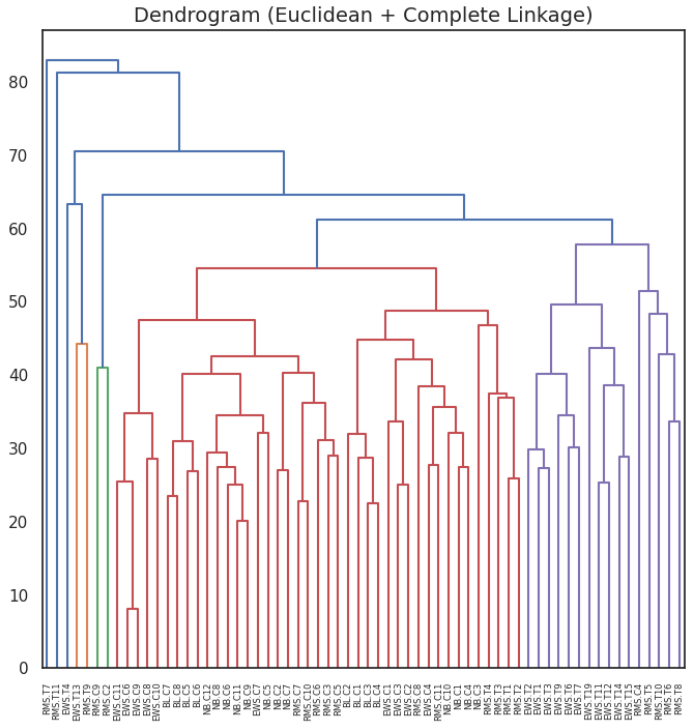
Sumedh Sanjay Kangne (BE21B040)
Department of Biotechnology
Indian Institute of Technology, Madras

A BT3041: Analysis and Interpretation of Biological Data Assignment



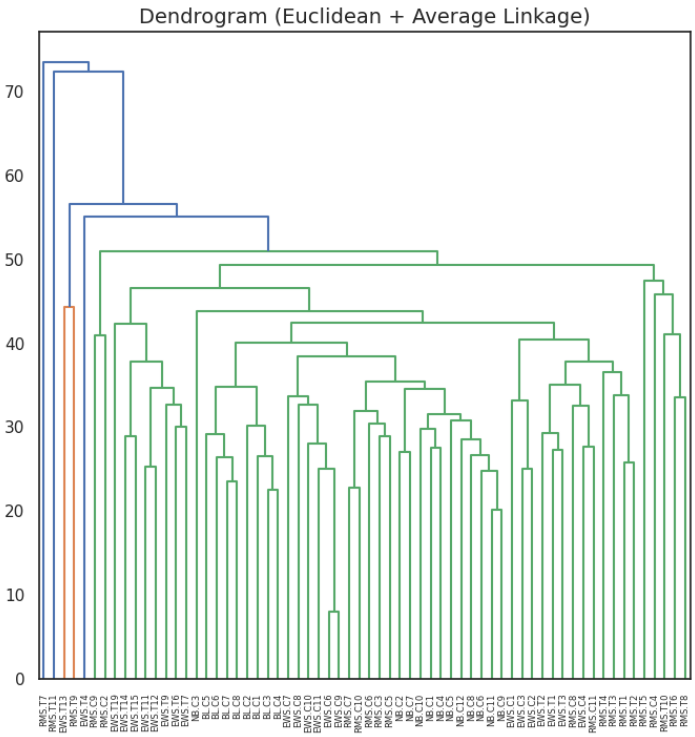
April 27, 2024

1. Using Euclidean distance and complete linkage



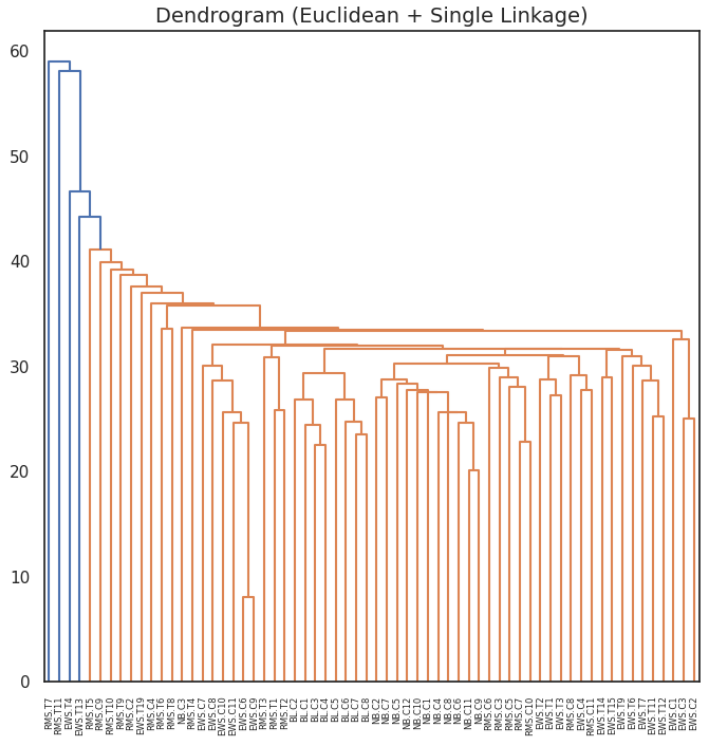
We can see 2 prominent clusters, the red one and the purple one. The red one seems to have clustered together most of the 'cell line samples' whereas the purple one has clustered most of the 'tumour biopsy' samples. Some smaller clusters with arbitrary values can also be seen, although their result seems quite insignificant.

2. Using Euclidean distance and average linkage



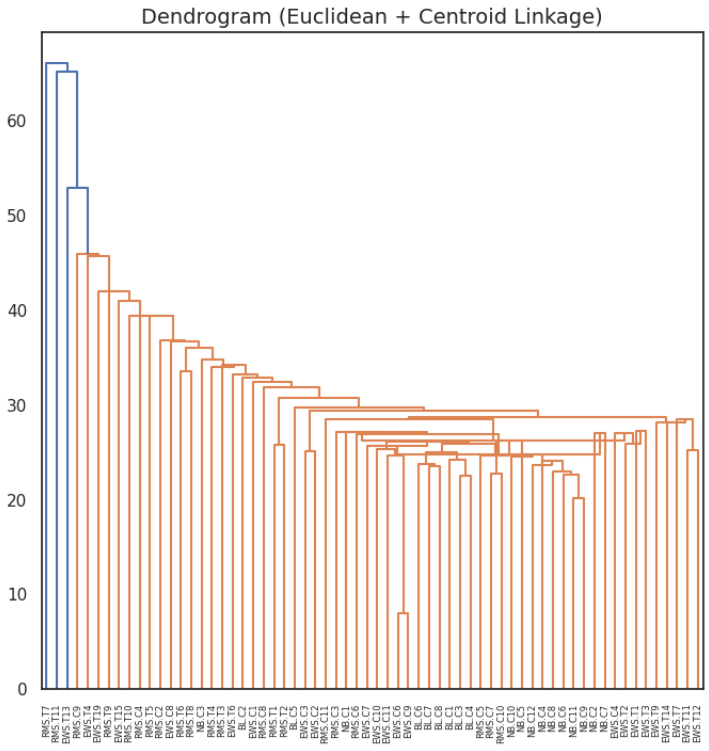
One huge cluster (green) can be seen. It seems to have clustered together both the cell line samples as well as the tumour biopsy samples for all the 4 different kinds of tumours. Thus, this clustering can be disregarded.

3. Using Euclidean distance and simple linkage.



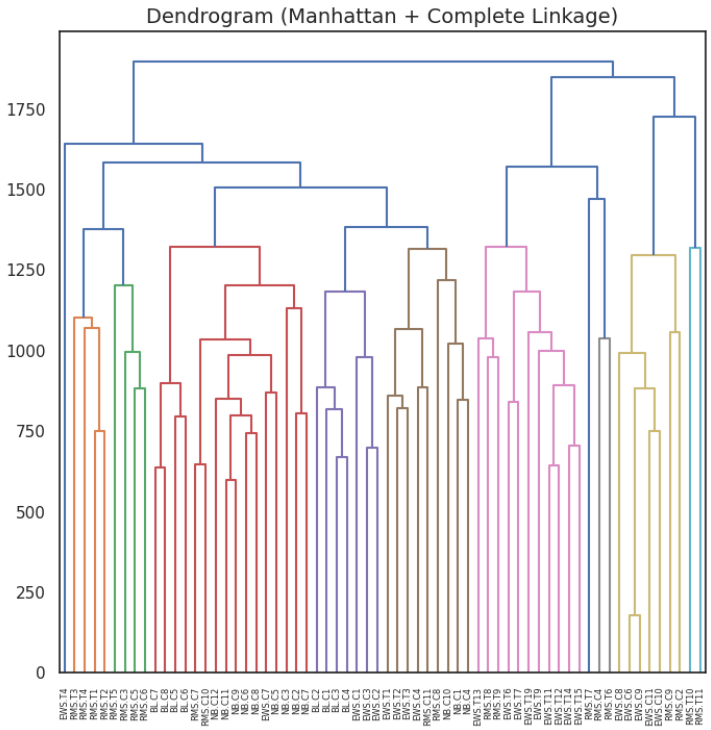
The result is quite similar to the above-mentioned case. No proper clustering is observed and thus, this clustering can also be disregarded.

4. Using Euclidean distance and centroid linkage



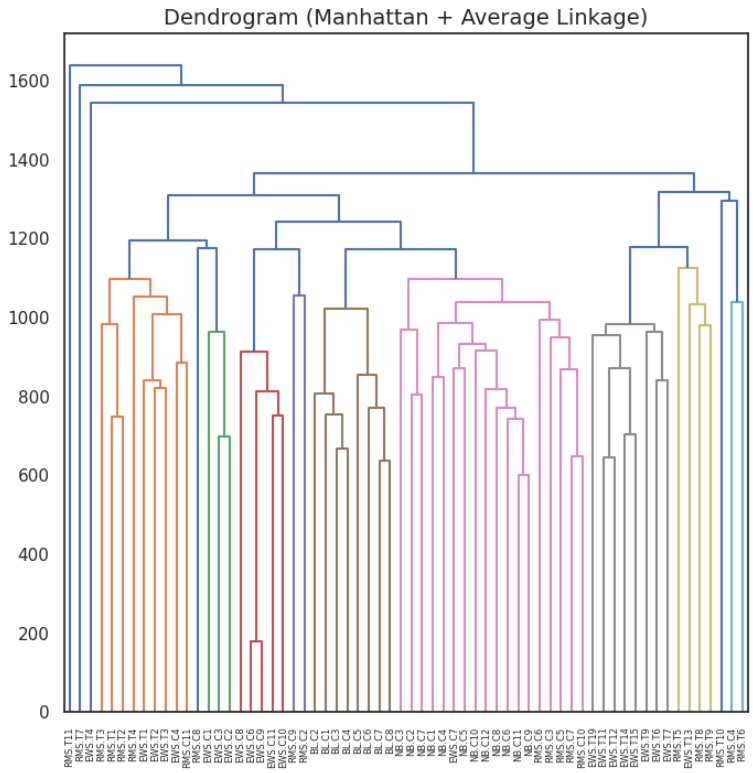
The result is quite similar to the above-mentioned case. No proper clustering is observed and thus, this clustering can also be disregarded.

5. Using Manhattan distance and complete linkage



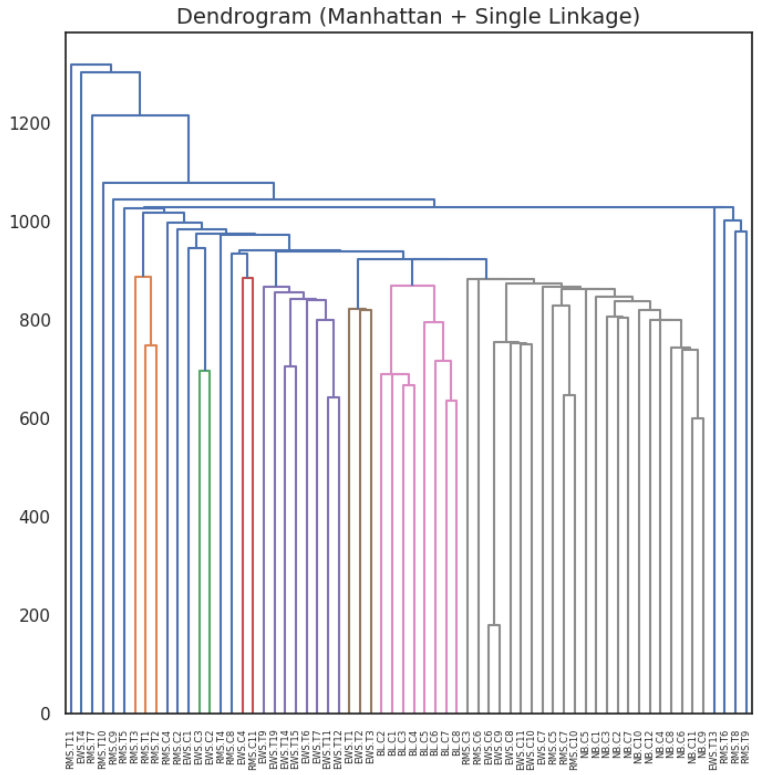
This method seems to have worked the best with reasonably good clustering results. The pink cluster seems to have properly clustered together most of the tumour biopsy results from the Erwing family of tumours (EWS). The orange and green clusters seem to have clustered together a few cases of rhabdomyosarcoma (RMS). The red, purple and brown clusters seem to have clustered together most of the cell line sample results. Thus, in general, we can see a difference between the cell line samples and the tumour biopsy samples with distinct overlapping between the different types of tumours.

6. Using Manhattan distance and average linkage



This method seems to have worked the best based on the differentiation of cell line samples from tumour samples. No overlap can be seen between the two. Although, amongst the four types of tumour samples we can still see quite a bit of overlap.

7. Using Manhattan distance and simple linkage



The clustering in this one is quite contrary to the previous ones. There is a much lower overlap between the 4 types of tumour cells. Moreover, it has been able to properly cluster the Burkitt Lymphoma (BL) and Neuroblastoma (NB) cell line samples.

Although, there is quite a bit of error in terms of clustering the RMS and EWS samples.

Final Analysis

After comparing the overall results, we can see that clustering with the 'Manhattan' distance metric has given much better results compared to the 'Euclidean' distance metric. We can also say that the expression values for the genes with RMS and EWS tumour cell types are quite similar and thus end up in the same cluster. Hence, other types of clustering methods like DBSCAN should be used which excel in grouping together points that are packed closely together (points with many nearby neighbours), and marking points that lie alone in low-density regions as outliers.

For the noticeable difference in the cell line sample expression values and tumour samples' expression values, we can suggest something like k-means clustering as it is a much simpler algorithm which excels at binary/categorical classifications and also takes the number of clusters as an input.

Code

The link for the code and all the images used can be found here:

https://drive.google.com/drive/folders/1ySFPD5L_9d6sR1o7ktTNgQTFef90KIZJ?usp=sharing

The End