Anirudh Rao
BE21B004

BT3041 – Analysis and Interpretation of Biological Data

Assignment 2

Report

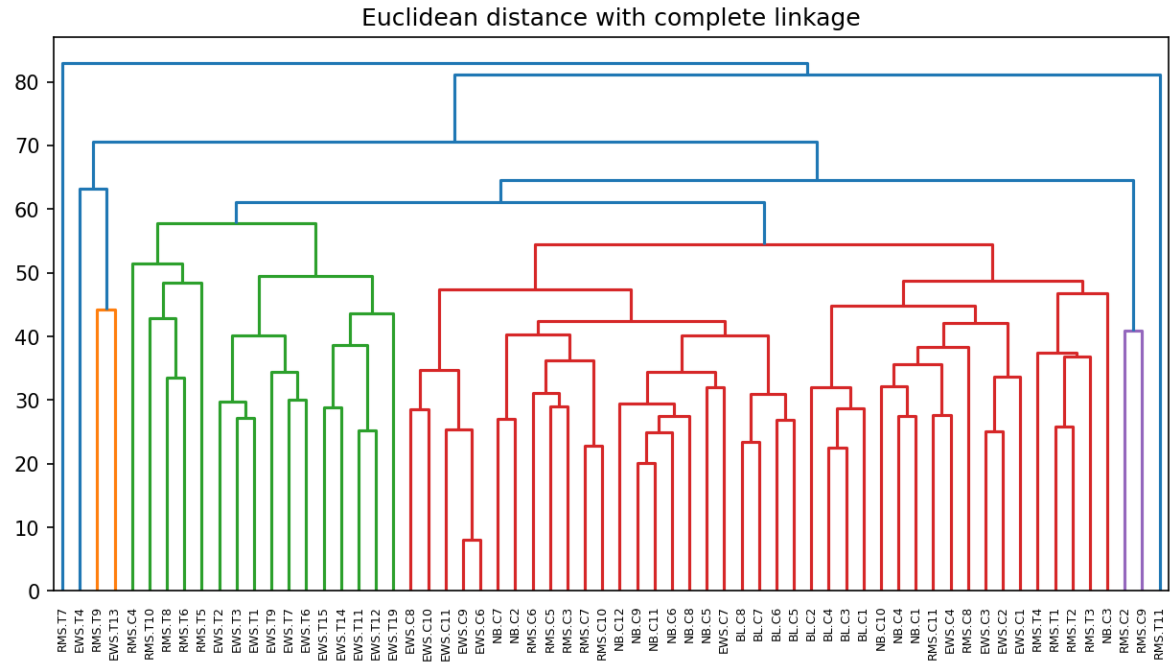
The link to the Python notebook containing the code to generate the dendrograms in this report can be found here –

https://drive.google.com/file/d/1Vroh_pUv_ftylSOv-IzBkGVTXKUoyJUr/view?usp=sharing.

- The goal of this analysis is to perform agglomerative clustering with different distance metrics and linkage methods on the given dataset.
- We hope to uncover the different cancer types by clustering them based on their gene expression patterns.
- As the expression levels have the same order of magnitude across genes, we do NOT scale the given data.

1

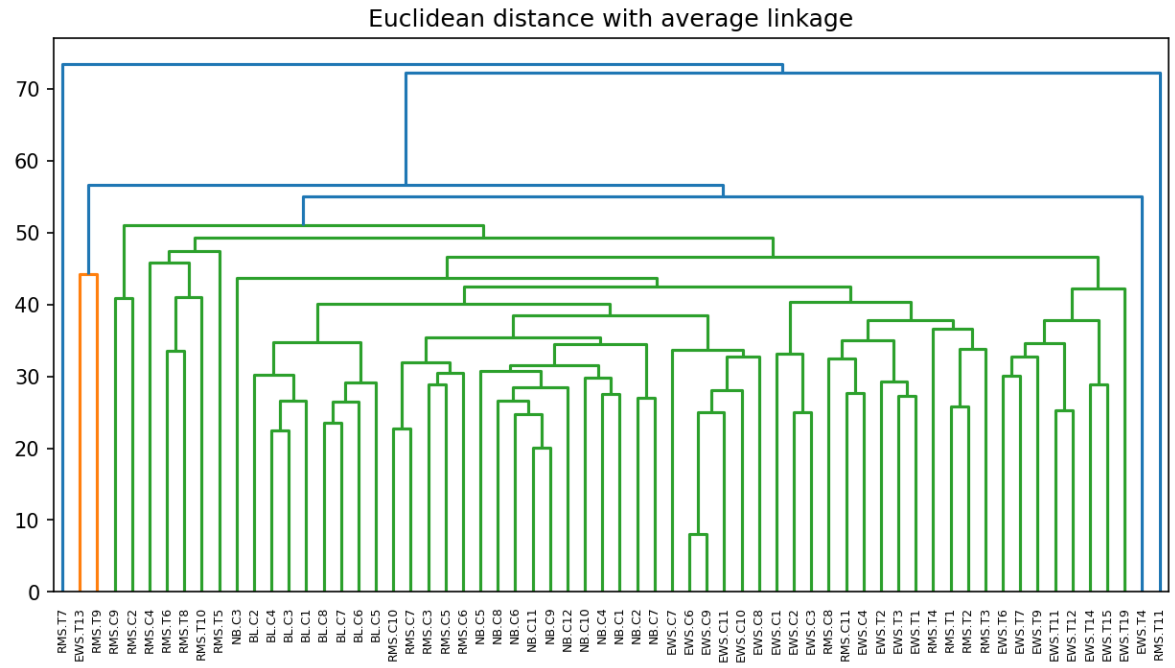
When Euclidean distance and complete linkage are used, the dendrogram looks like:



The clustering algorithm identifies 5 clusters. The **green** cluster contains only rhabdomyosarcomas (RMS) and Ewing tumours (EWS). However, the **red** cluster contains all cancer types. This is a decent clustering.

2

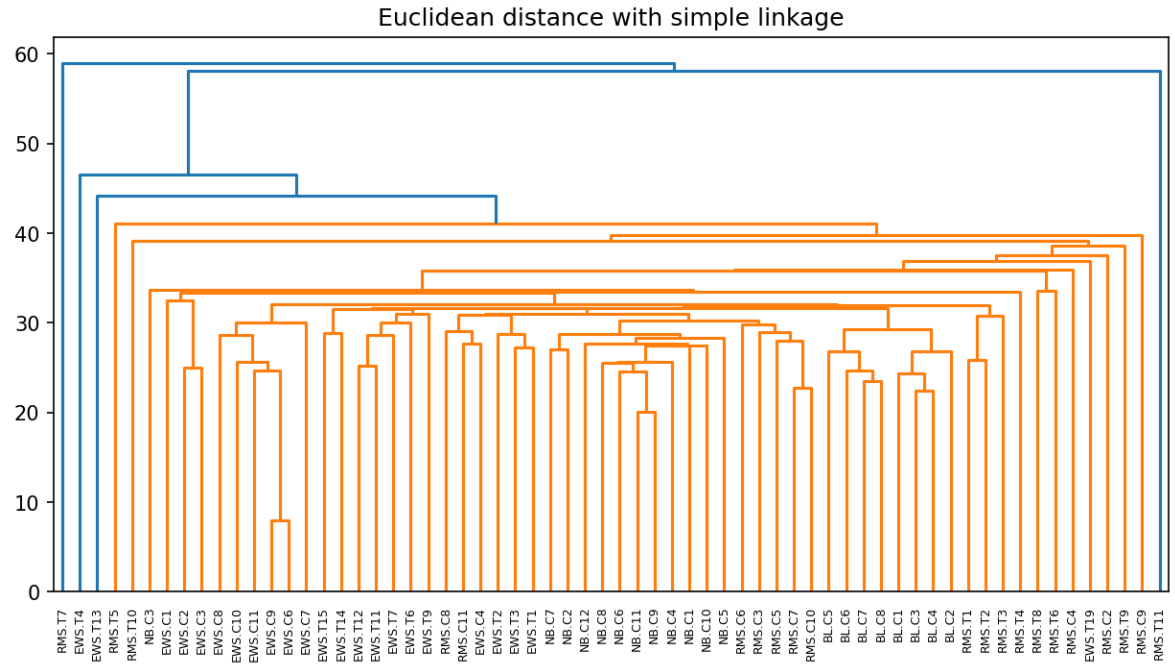
When Euclidean distance and average linkage are used, the dendrogram looks like:



The clustering algorithm identifies 3 clusters. The green cluster contains almost all the data points. This is not a good clustering.

3

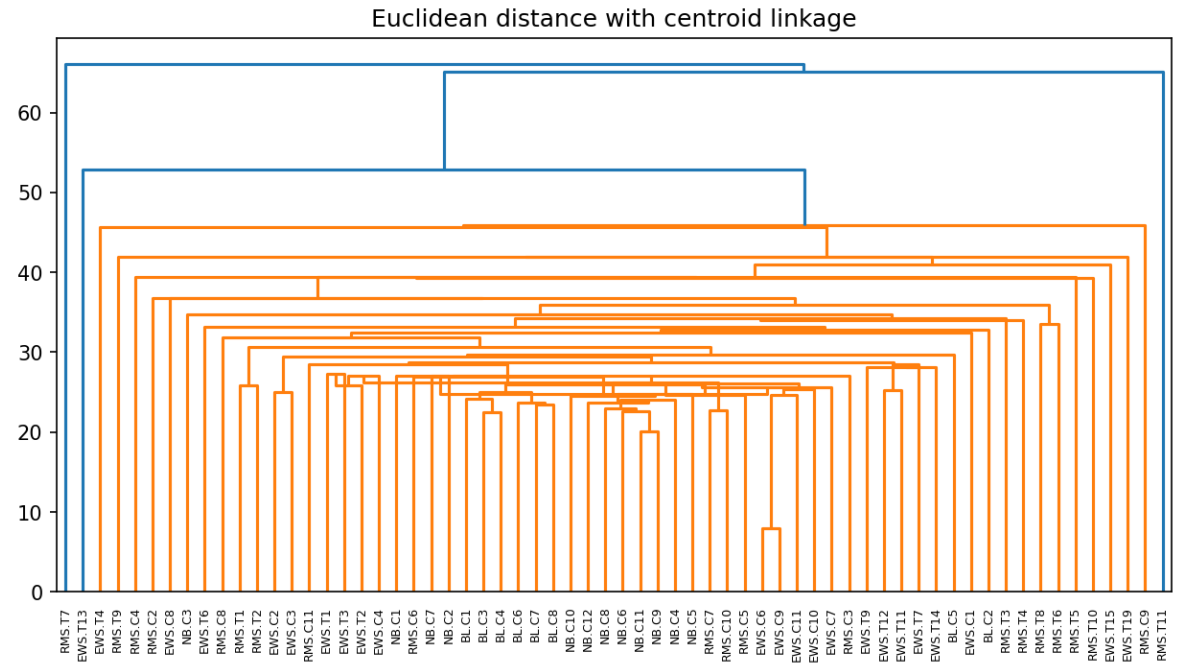
When Euclidean distance and simple linkage are used, the dendrogram looks like:



The clustering algorithm identifies 2 clusters. The orange cluster contains almost all the data points. This is not a good clustering.

4

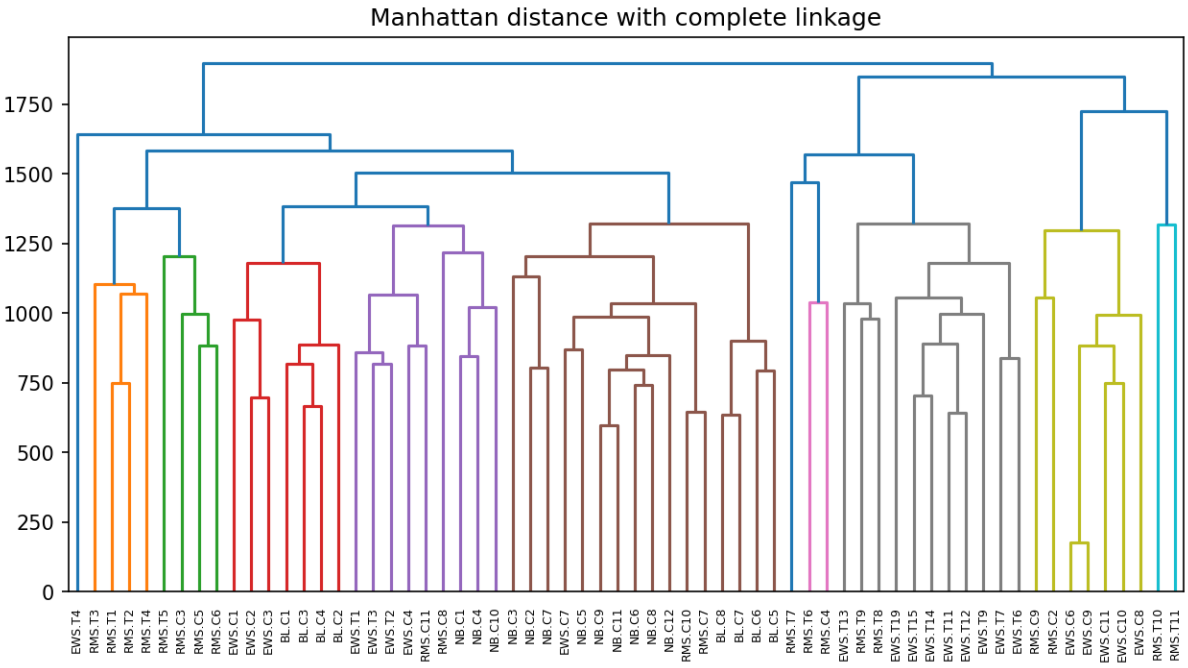
When Euclidean distance and centroid linkage are used, the dendrogram looks like:



The clustering algorithm identifies 2 clusters. The orange cluster contains almost all the data points. There are also a lot of overlaps in the dendrogram. This is not a good clustering.

5

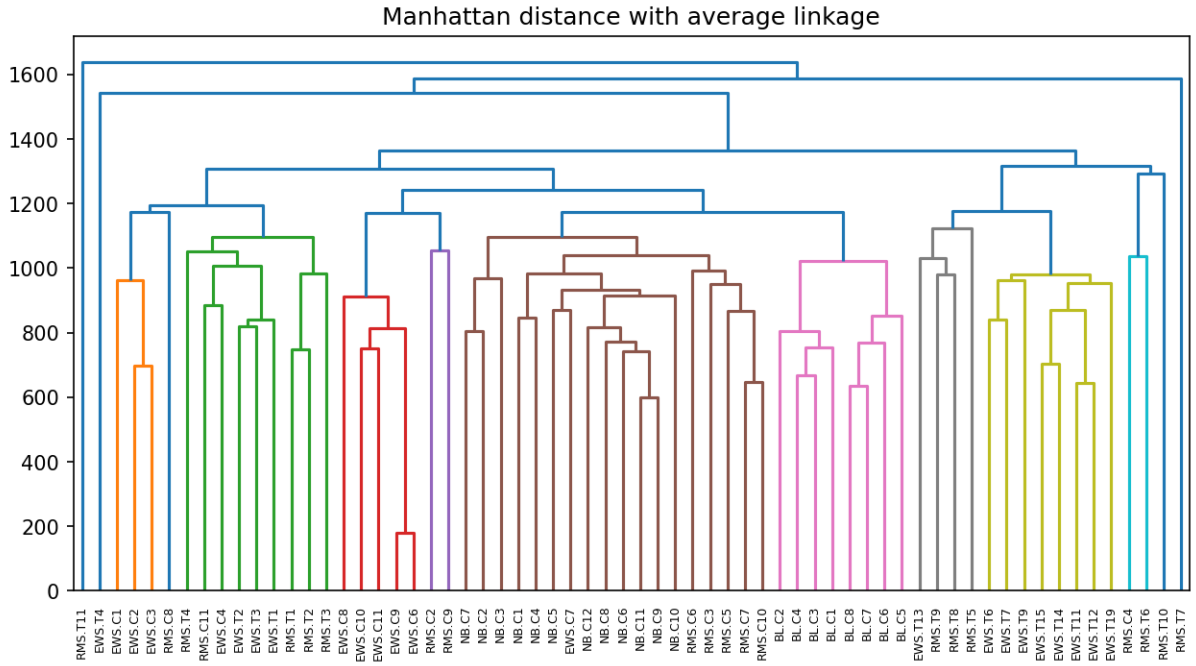
When Manhattan distance and complete linkage are used, the dendrogram looks like:



The algorithm identifies 10 clusters. However, most of these are composed of multiple cancer types. RMS and EWS are usually clustered together. This is a decent clustering.

6

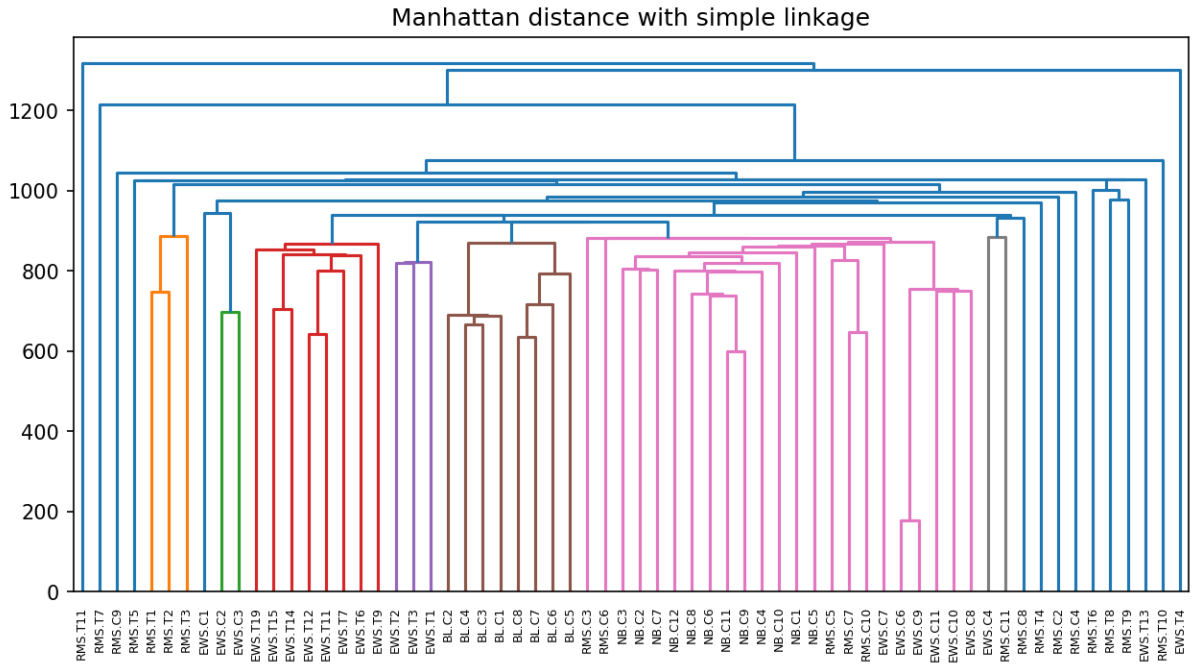
When Manhattan distance and average linkage are used, the dendrogram looks like:



The algorithm identifies 10 clusters. However, most of these are composed of multiple cancer types. The pink cluster contains Burkitt lymphomas (BL) exclusively and the lime cluster contains EWS exclusively. All the neuroblastomas (NB) are concentrated in the brown cluster. This is better than the previous clustering with Manhattan distance and complete linkage.

7

When Manhattan distance and simple linkage are used, the dendrogram looks like:



The algorithm identifies 8 clusters. However, most of these are composed of multiple cancer types, with RMS and EWS clustered together frequently. This is a slightly messy clustering.

When Manhattan distance and centroid linkage are used, a dendrogram cannot be produced. Centroid linkage requires that Euclidean distance be used as the metric.

```
ValueError                                Traceback (most recent call last)
Cell In[17], line 1
----> 1 Z = linkage(
      2     y = df,
      3     method='centroid',
      4     metric='cityblock',
      5     optimal_ordering = True
      6 )
      8 plt.figure(figsize=(10, 5),dpi=100)
      9 dendrogram(Z,labels=df.index)

File ~\anaconda3\Lib\site-packages\scipy\cluster\hierarchy.py:1052, in linkage(y, method, metric, optimal_ordering)
   1050 elif y.ndim == 2:
   1051     if method in _EUCLIDEAN_METHODS and metric != 'euclidean':
-> 1052         raise ValueError("Method '{}' requires the distance metric "
   1053                          "to be Euclidean".format(method))
   1054     if y.shape[0] == y.shape[1] and np.allclose(np.diag(y), 0):
   1055         if np.all(y >= 0) and np.allclose(y, y.T):

ValueError: Method 'centroid' requires the distance metric to be Euclidean
```

Conclusions

- Manhattan distance appears to be a better metric than Euclidean distance for this particular dataset.
- Manhattan distance with average linkage is able to cluster the cancer types better than the other combinations.
- Clustering with simple linkage or centroid linkage is quite messy with this dataset.
- Euclidean distance works well when complete linkage is used.