Analytical Interpretation of Biological Data

BE21B037 – Assignment 2

1 Introduction

Khan et al., 2001 used cDNA microarrays to study the expression of genes in of four types of small round blue cell tumours of childhood (SRBCT). These were neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma, a subset of non-Hodgkin lymphoma (BL), and the Ewing family of tumours (EWS). Gene expression profiles from both tumour biopsy and cell line samples were obtained and are contained in this dataset.

2 Questions

Perform the below mentioned agglomerative clustering of NB, RMS, BL and EWS. Compare and contrast how the clustering changes based on distance metric and linkage metric and comment your interpretation.

2.1 Definitions

Distance Measurements:

1. **Euclidean Distance :** Represents the shortest distance between two points/vectors.
2. **Manhattan Distance :** Sum of absolute differences between points across all thedimensions (Also known as city block distance).

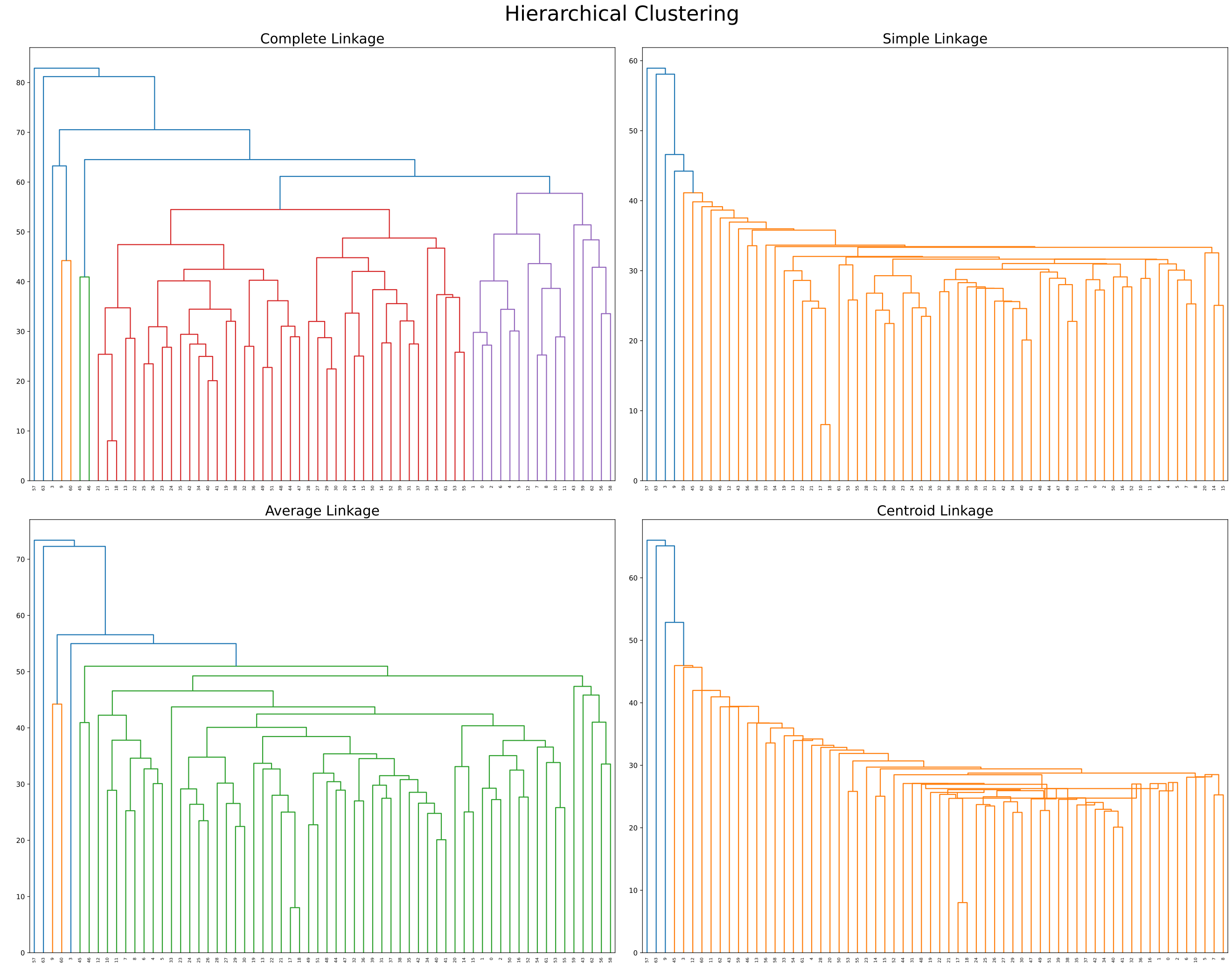
Linkage Types:

1. **Complete :** Distance between two clusters should be the maximum distance between any single data point in the first cluster and any single data point in the second cluster.
2. **Single :** Distance between two clusters is defined as the minimum distance between any single data point in the first cluster and any single data point in the second cluster.
3. **Average :** Distance between two clusters is the average distance between data points in the first cluster and data points in the second cluster

4. **Centroid :** Distance between two clusters is the distance between the two mean vectors of the clusters.

2.2 Using Euclidean Distance

We perform agglomerative clustering for the gene expression dataset using complete, single, average and centroid linkage methods. The following results are obtained when we perform clustering based on euclidean distance and simple, complete, average and centroid linkage metrics.



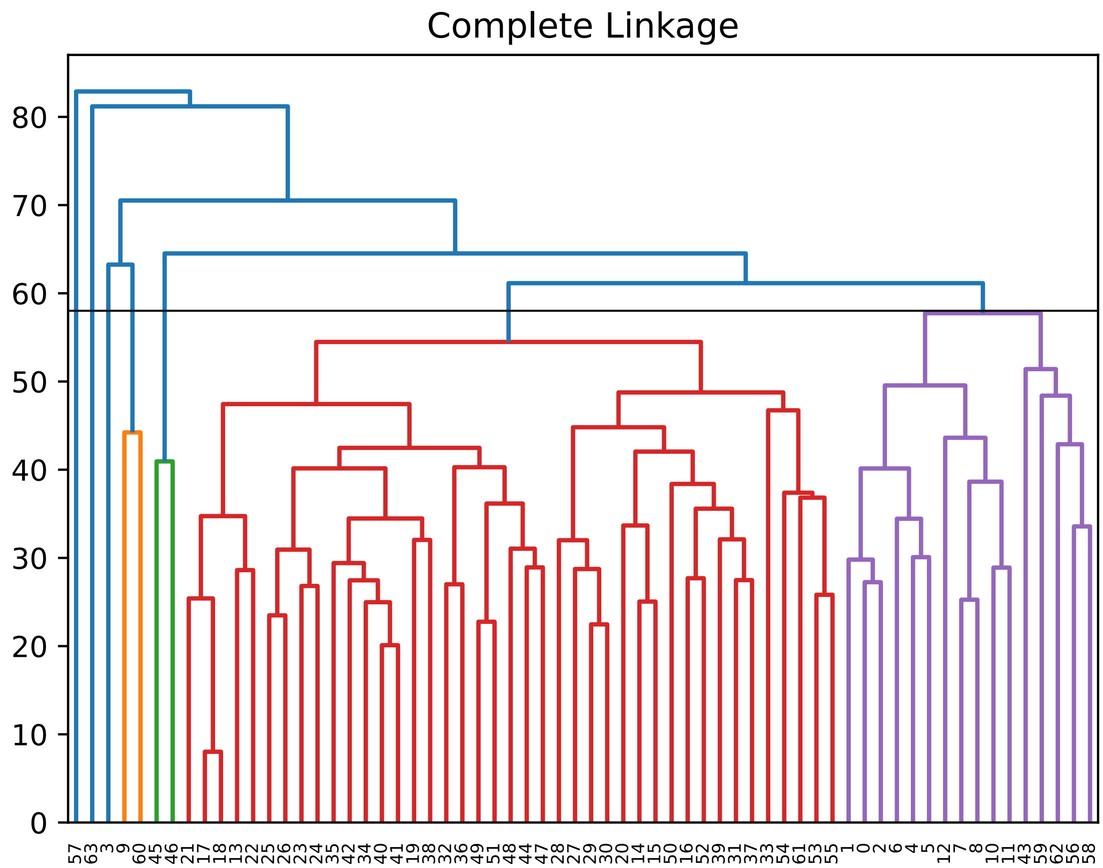
Threshold for cluster distances is set as $0.7 \times$ max(bifurcation distance) for a particular dendogram

2.2.1 Bifurcation Distances:

Different linkages have different distance values where the tree is split. The top 10 distances for each of the clustering metrics are shown below

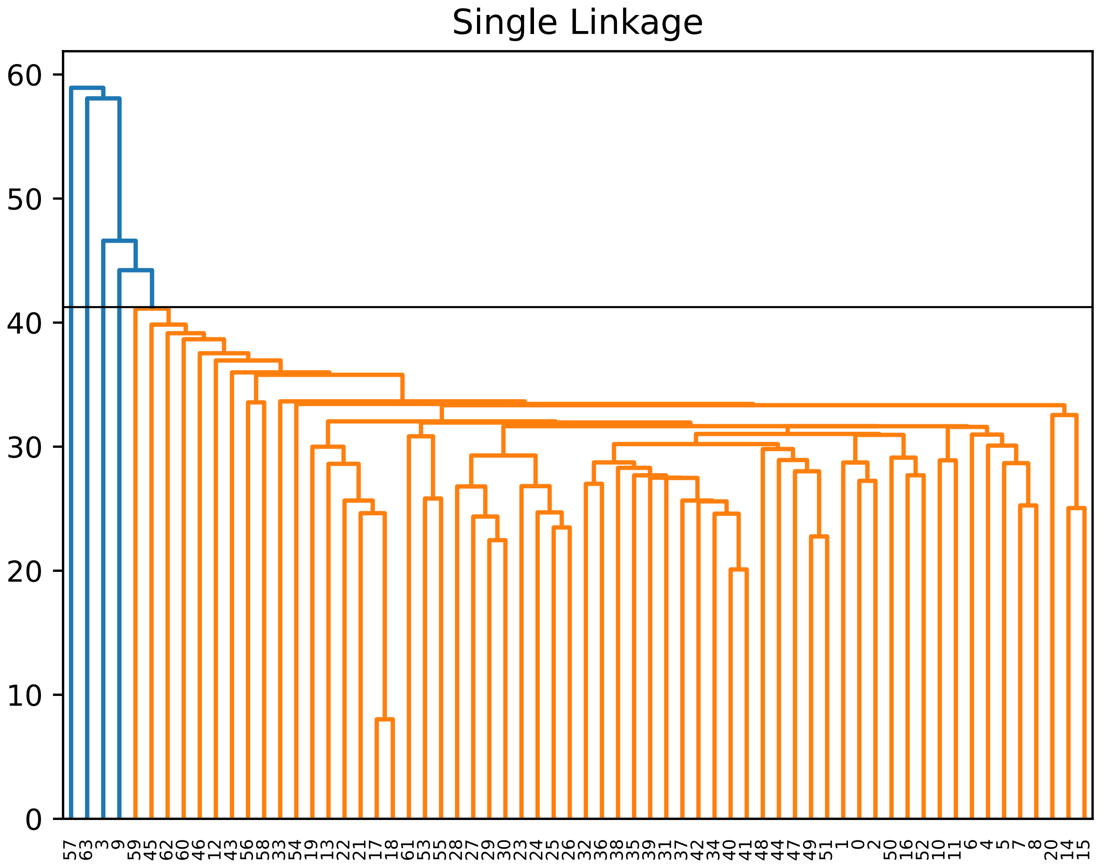
Linkage	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Threshold**
Complete	49.6	51.4	54.5	57.7	61.1	63.3	64.5	70.5	81.2	82.9	58.0
Single	36.9	37.5	38.7	39.1	39.8	41.1	44.2	46.6	58.1	58.9	41.3
Average	44.2	45.8	46.6	47.4	49.2	51.0	55.0	56.6	72.3	73.4	51.4
centroid	39.4	39.4	41.0	42.0	42.0	45.7	46.0	52.9	65.1	66.0	46.2

2.2.2 Euclidean Distance with complete linkage:



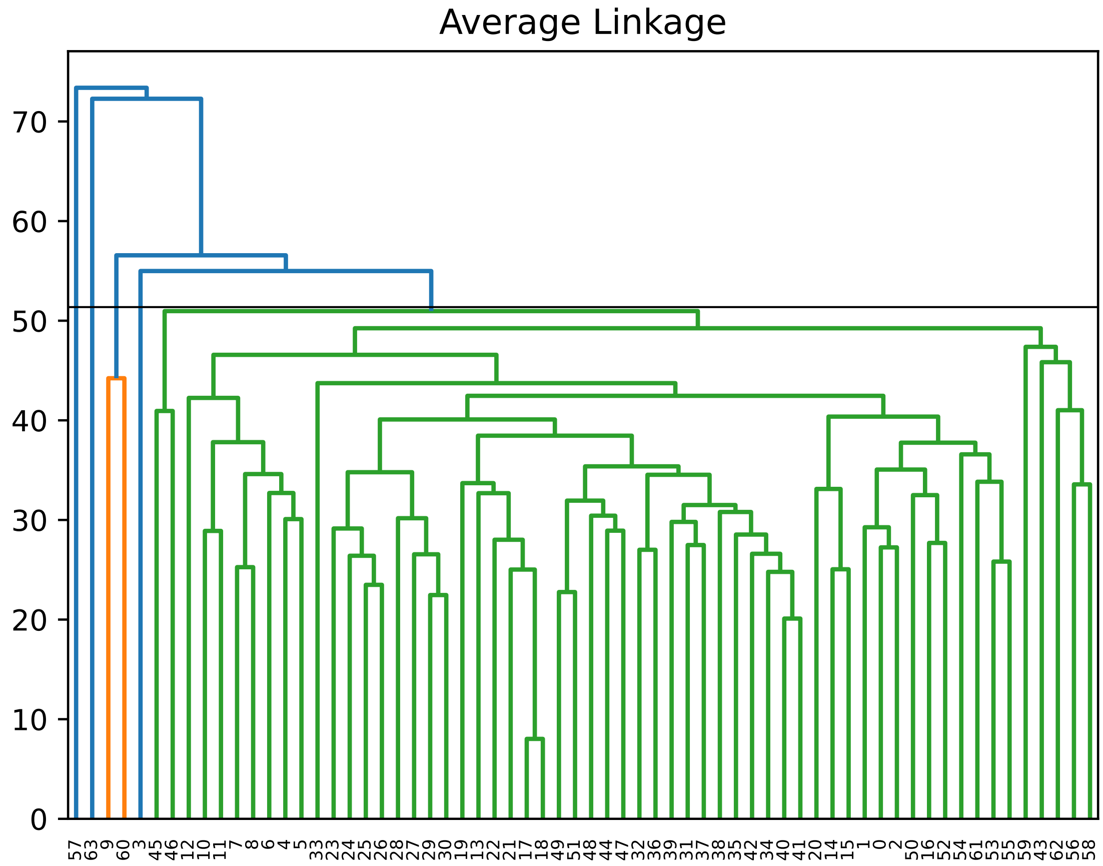
With the threshold distance set as 58, we get 4 major clusters. The data points per cluster is not evenly spread with 2 clusters containing only 3 cell types each

2.2.3 Euclidean Distance with single linkage:



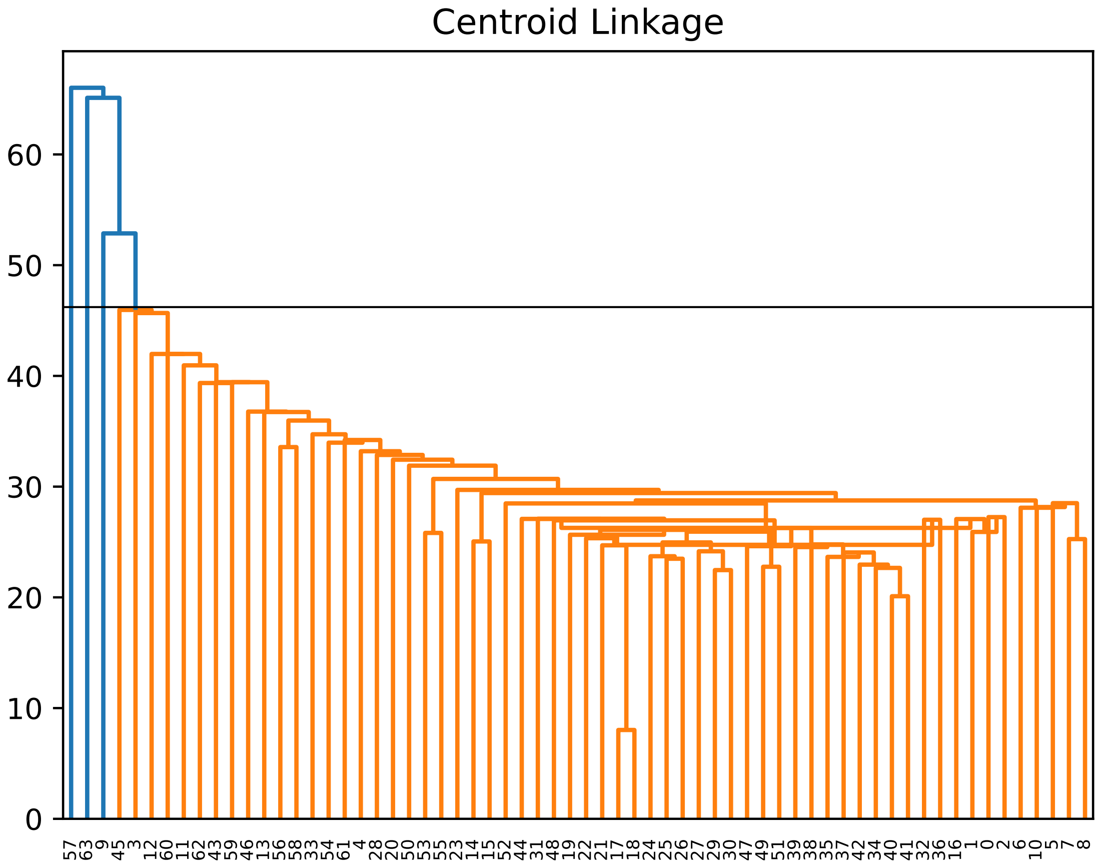
With the threshold distance set as 41.3, we just get 1 major cluster. Given that single linkage measures the minimum distance between any two clusters, we can conclude that all the data points are close together.

2.2.4 Euclidean Distance with average linkage:



With the threshold distance set as 51.4, we just get 2 major clusters.

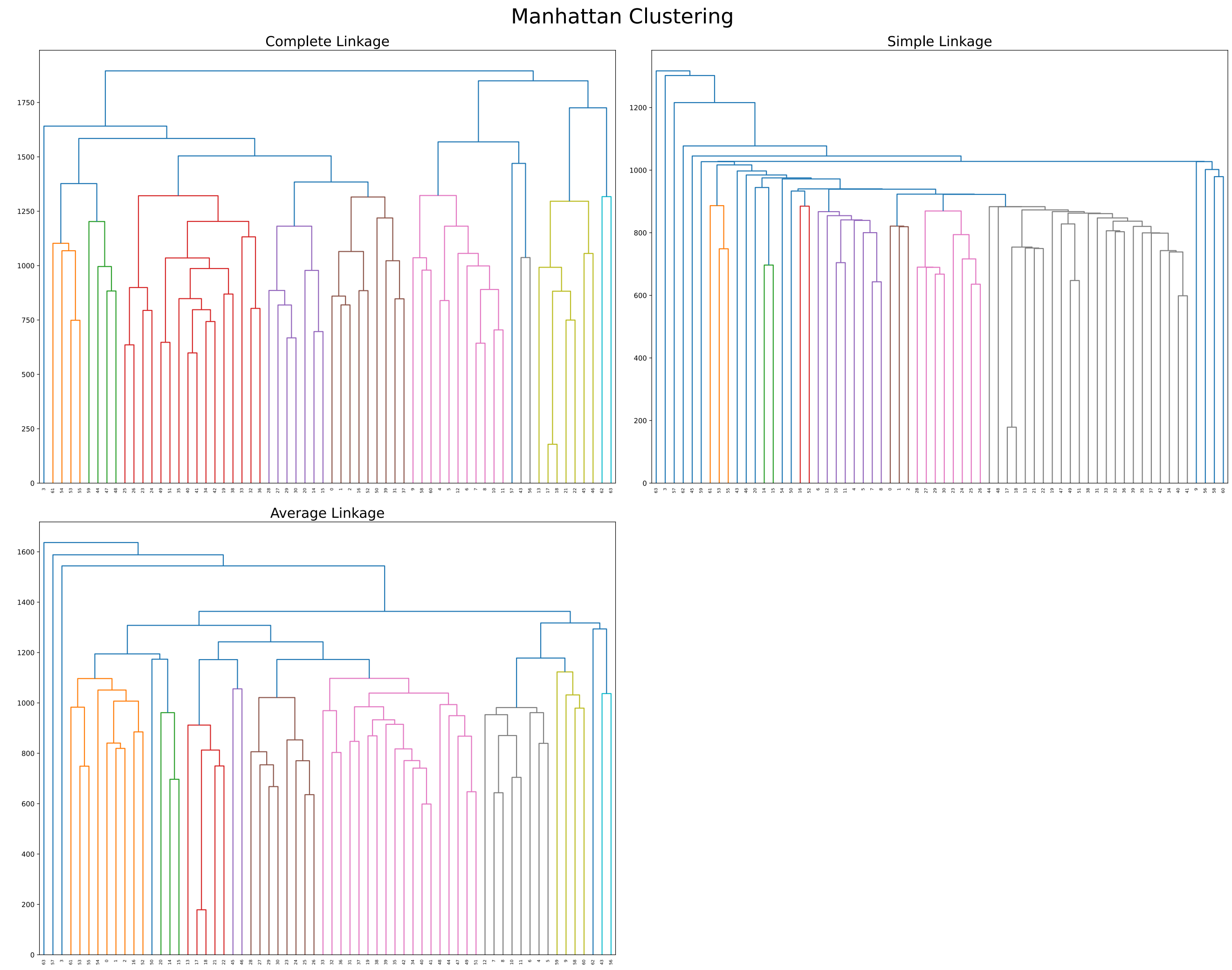
2.2.5 Euclidean Distance with centroid linkage:



With the threshold distance set as 46.2, we just get 1 major cluster.

2.3 Using Manhattan Distance

We perform agglomerative clustering for the gene expression dataset using complete, single, average and centroid linkage methods. The following results are obtained when we perform clustering based on Manhattan distance and simple, complete and average linkage metrics

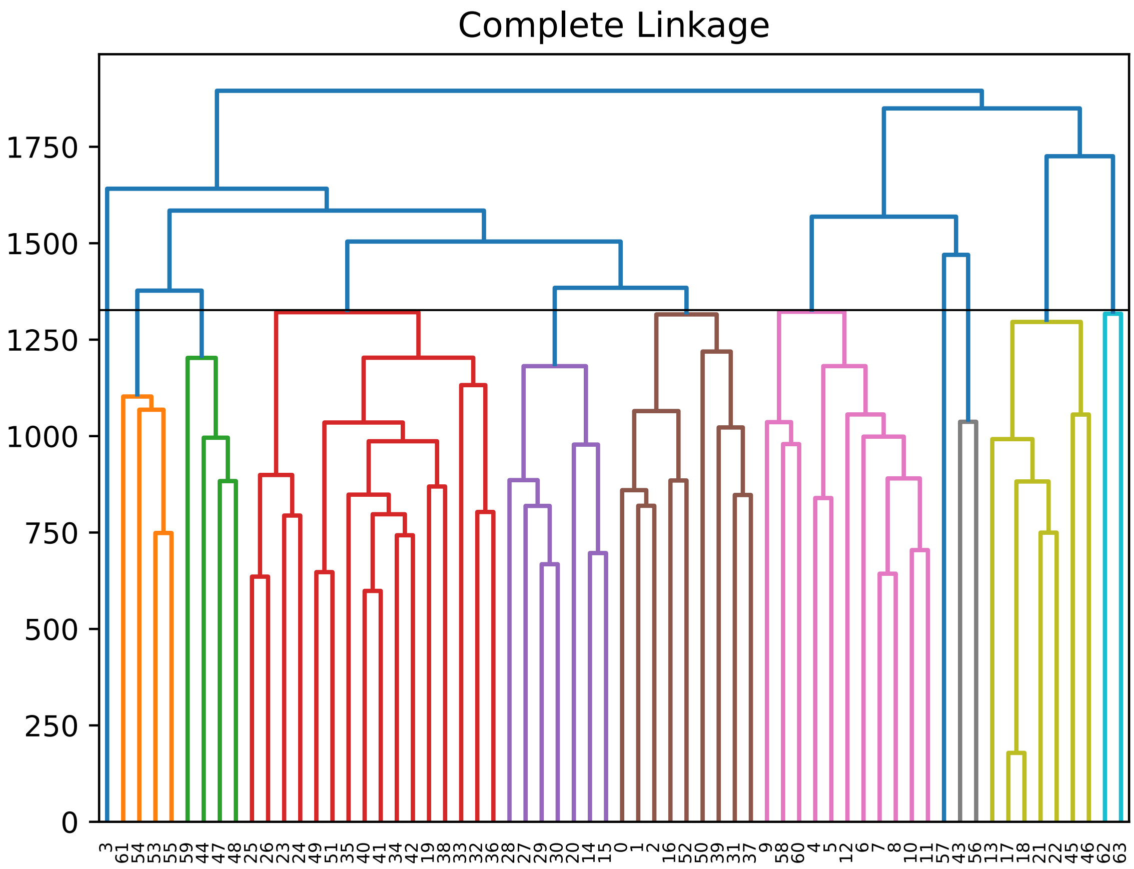


2.3.1 Bifurcation Distances:

Different linkages have different distance values where the tree is split. The top 10 distances for each of the clustering metrics are shown below

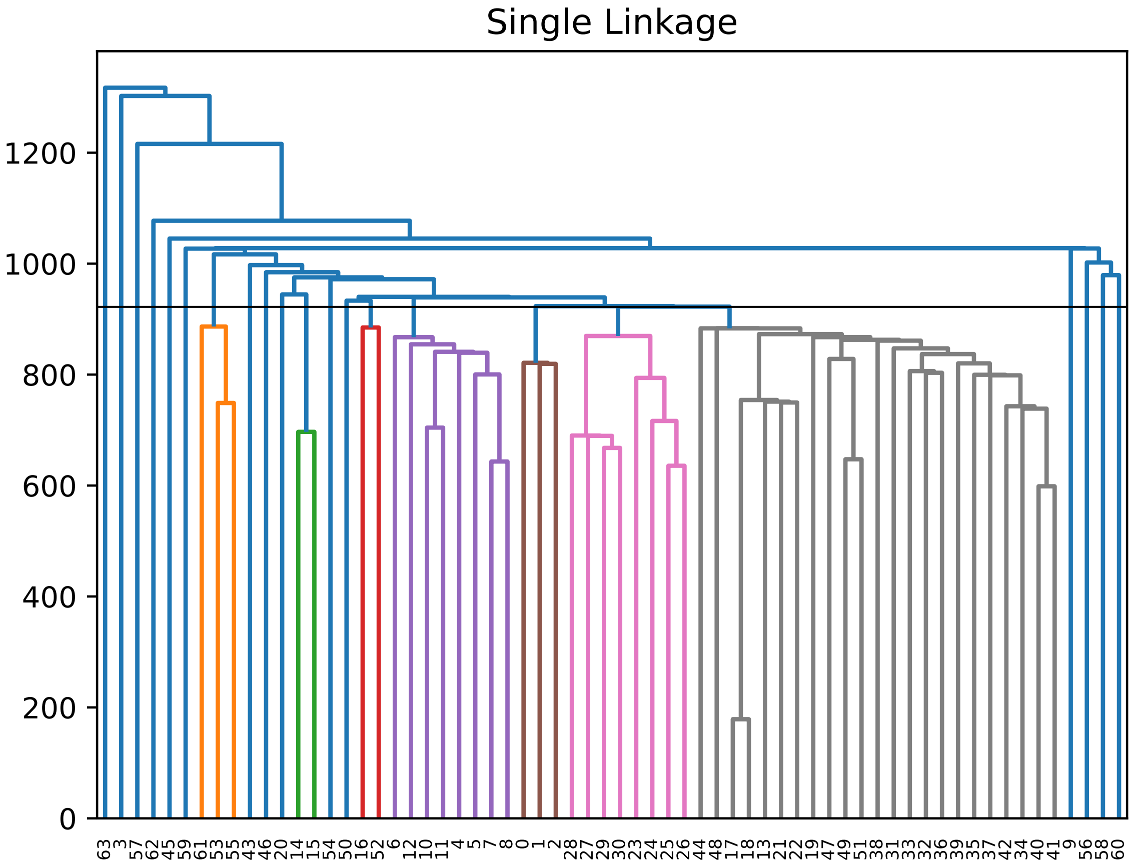
Linkage	1	2	3	4	5	6	7	8	9	10	Threshold
Complete	1377.1	1384.4	1469.9	1504.3	1568.8	1584.6	1641.3	1725.5	1849.5	1895.2	1326.7
Single	1002.1	1016.8	1027.1	1027.3	1028.0	1045.2	1077.3	1215.9	1302.3	1317.2	922.0
Average	1178.1	1194.6	1242.4	1293.7	1307.9	1317.5	1363.4	1544.0	1588.0	1636.7	1145.7

2.3.2 Manhattan Distance with complete linkage:



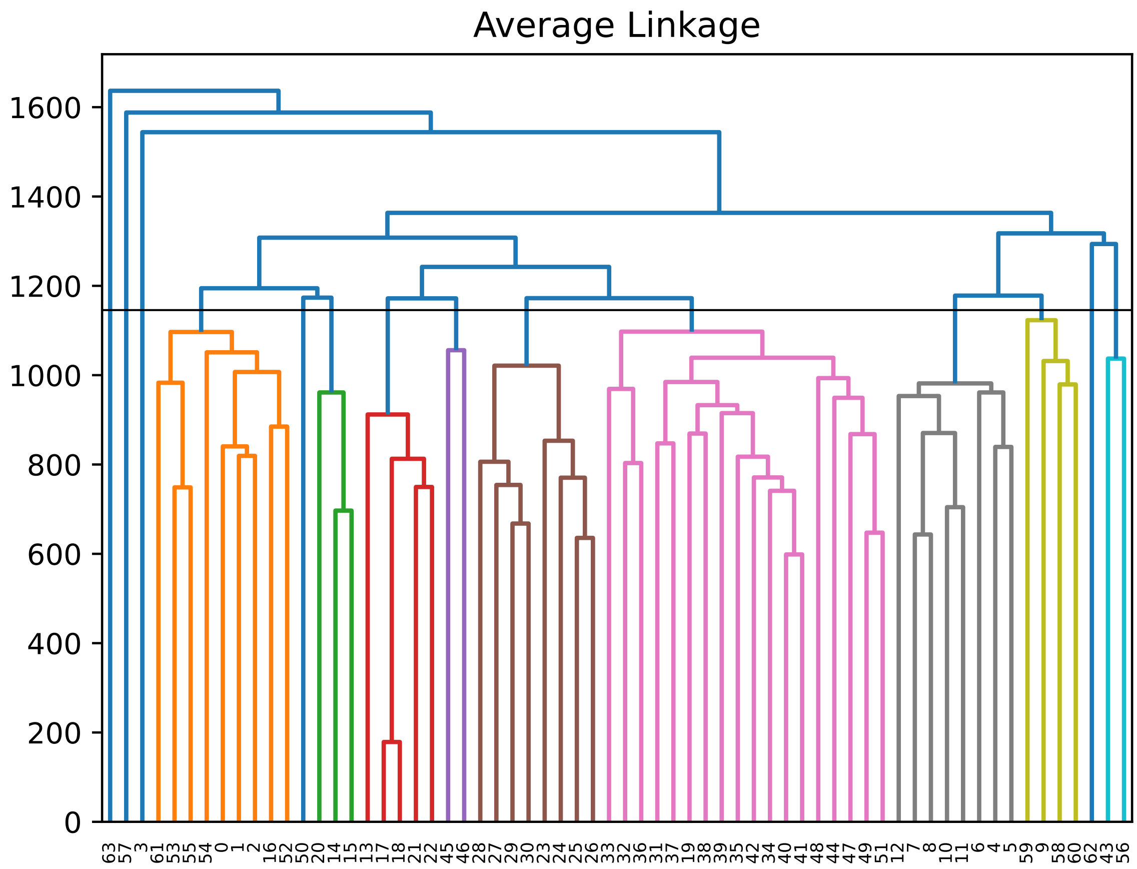
With the threshold distance set as 1326.7, we get close to 9 clusters.

2.3.3 Manhattan Distance with single linkage:



With the threshold distance set as 922.0, we just get 7 major cluster.

2.3.4 Manhattan Distance with average linkage:



With the threshold distance set as 1145.7, we just get 9 major clusters.

2.4 Interpretation

The table below summarizes the results we got from clustering with various metrics

Metrics	**Complete**		**Single**		**Average**		**Centroid**	
	Threshold	# of clusters	Threshold	# of clusters	Threshold	# of clusters	Threshold	# of clusters
Euclidean	58.0	4	41.3	1	51.4	2	46.2	1
Manhattan	1326.7	9	992.0	7	1145.7	9	–	–

Table 1: Summary

We observe that Manhattan distances worked much better than Euclidean when it came to clustering. For high dimensional vectors it is better to use Manhattan distances since it would be more sensitive to differences in individual coordinates, due to its absolute value calculations. Complete linkage, which takes into account the maximum distance between two clusters, seemed to be the more precise, in both Euclidean and Manhattan metrics. Since the gene expression values are all close in magnitude, the euclidean distance combined with single linkage or centroid linkage clustering gave the worst possible clustering.

Note: The given dataset seems to have entire columns of data points missing (Ex: BL, and NB cells have no gene expression values of Cell line sample, but only contain tumor biopsy samples).

2.5 Assumption

1. The given dataset is complete and does not contain missing gene expression values
2. The threshold for cluster distance choosing is set as $0.7 \times$ max(bifurcation distance)
3. The centroid linkage clustering cannot be used with Manhattan distances.

3 Appendix

The code to recreate all the above plots and testing can be found here.

* **