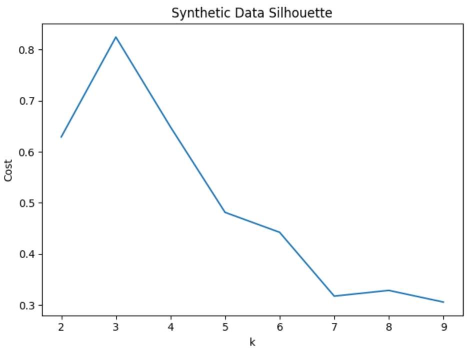
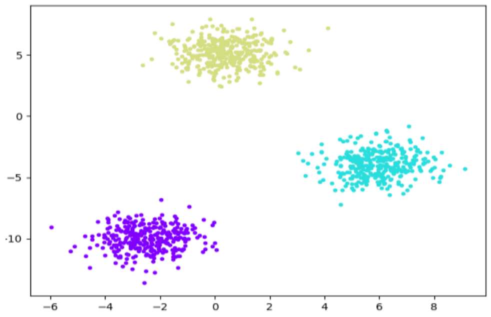
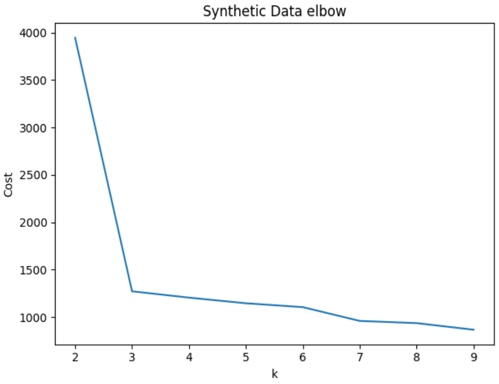
Ex4 APML
Saji.Tibi
207905407

1- In order to get correct cluster number for synthetic data we ran both elbow and Silhouette methods, I used center = 3 (make_blob) , so we expect to see k =3.
Note that in silhouette we can see the maxima in k=3, and in elbow the big drop in k=3 as it should be.

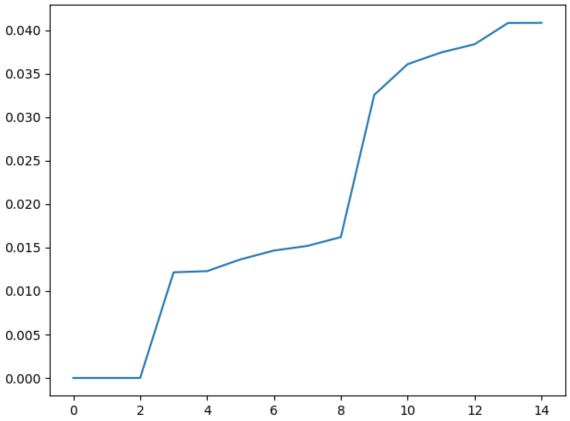




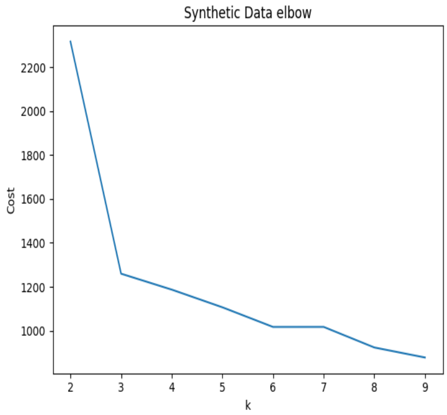


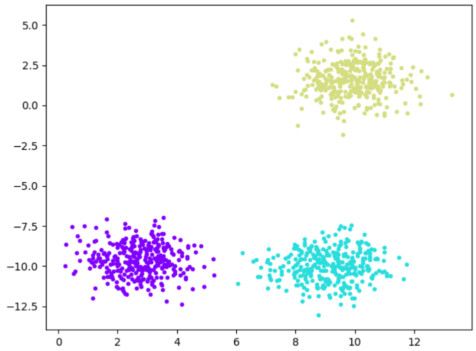
Now we move to spectral clustering on same data, in order to obtain correct sigma and m i’ve ran eigengap with mnn, and gaussian kernel for different values, since we know that k=3, i’ve looked for sigma and m that points to k=3, those m and sigma were my best match.

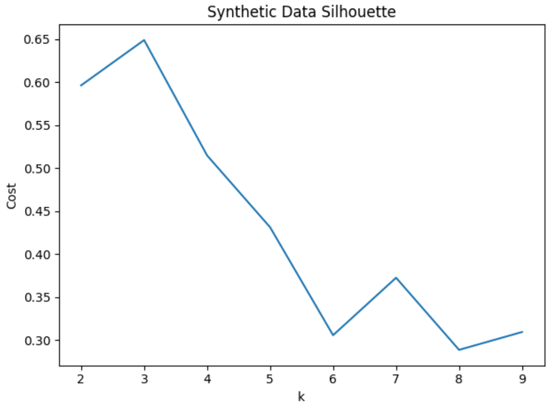
With m = 10 we get this eigengap graph:



to verify we use both of k selection methods and see if we got k=3 as match, and we got :



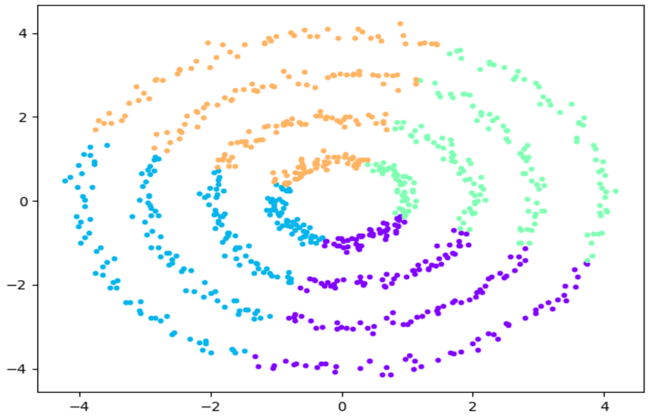




same with sigma but i’ve omitted since this data isn’t too hard to cluster, as we can see above both kmean++ and spectral clustering performed well.

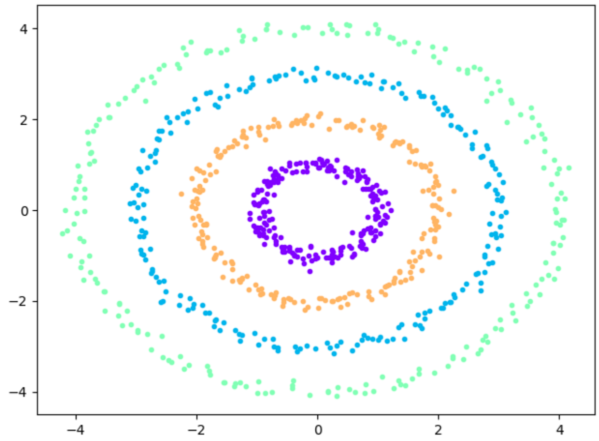
Now we will look in more interesting data:

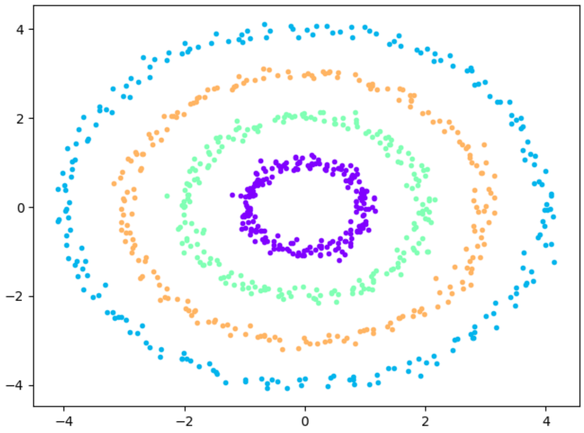
Circles: we will see how kmean++ performs compared to spectral clustering in circles.
Running kmeans++ with k=4 (as its obvious) gives:



Now we will try spectral clustering: (k= 4)
using m =6

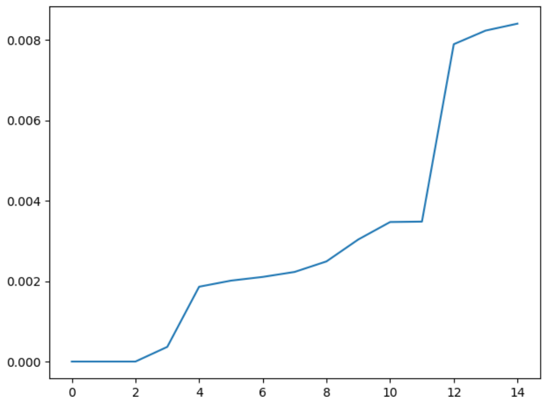
and with sigma = 0.1



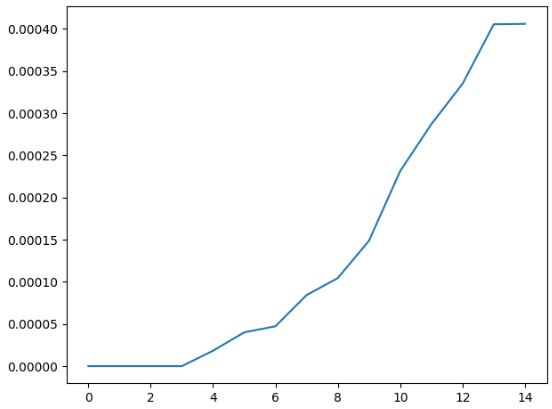


Note the huge difference between spectral clustering and kmean++, kmean++ performs terribly as explained in class with rings, in the other hand spectral clustering with its algorithm of punishing points that close to other clusters and rewarding points with minimal distance to their clusters gave better results here.
Choosing this values comes from same explanations above.

Eigengap for mnn =6



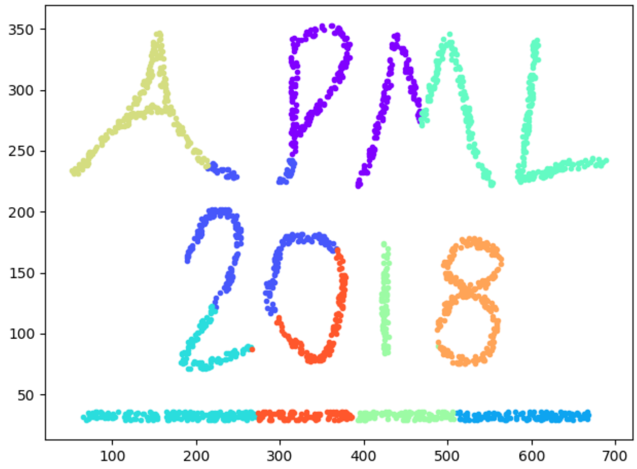
and for gaussian_param = 0.1



As we can see above that both values also pointed towards k =4.

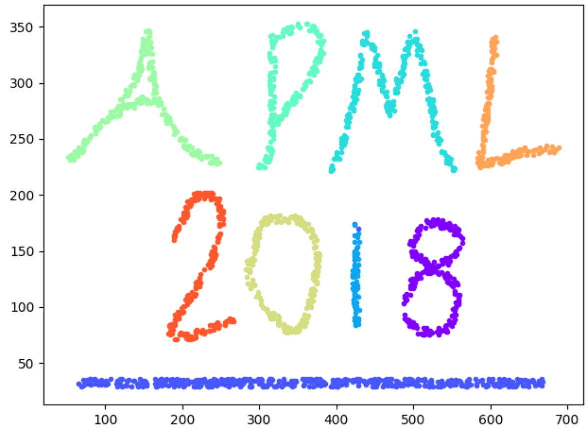
Now will try APML picture:

With kmean++ we get: (k=9, counting letters and underline)

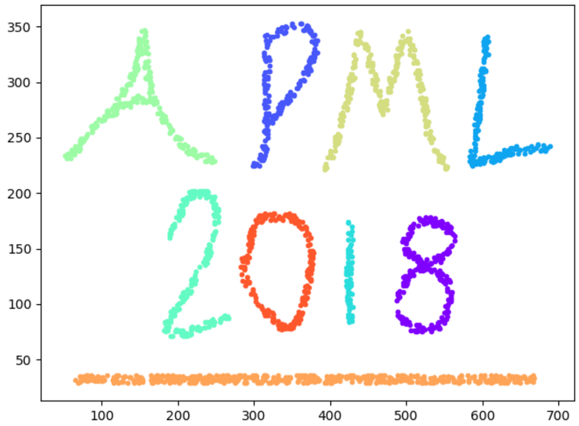


Now with clustering we get:

With sigma = 3.4



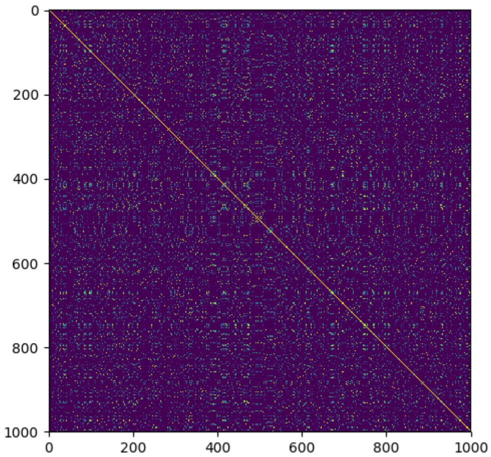
and with m = 16



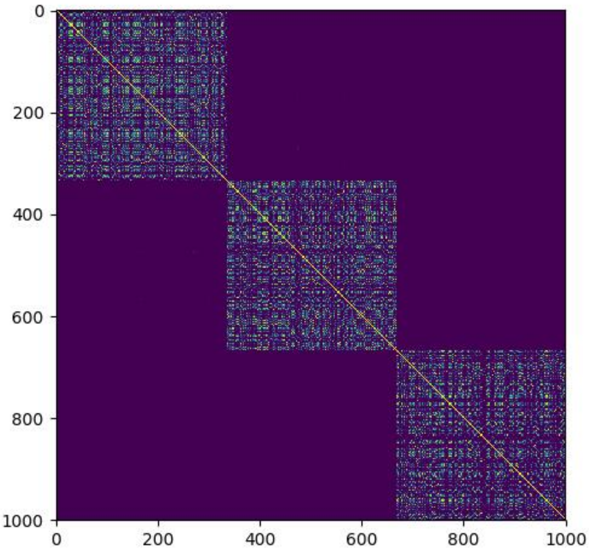
<u>Similarity :</u>

We draw 1000 sample with 3 centers of synthetic data, then we compare similarity matrix of shuffled data and sorted data according to spectral clustering and compare the results:

similarity matrix of shuffled data, using Gaussian kernel with parameter 0.6 and get:

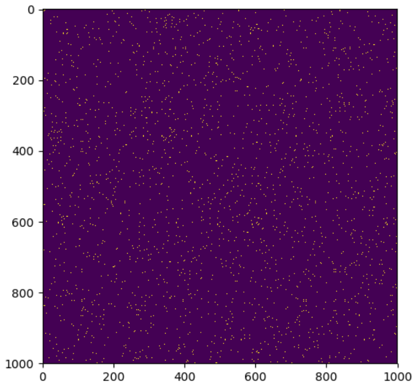


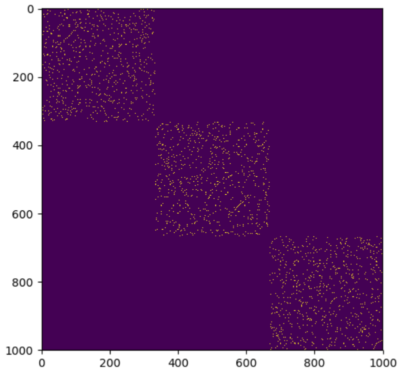
Now after sorting the data we get:



Note that how before sorting the matrix showed almost uniform distribution and after we can clearly see the thee clusters and the rest of matrix is clear, so it clustered the data correctly.

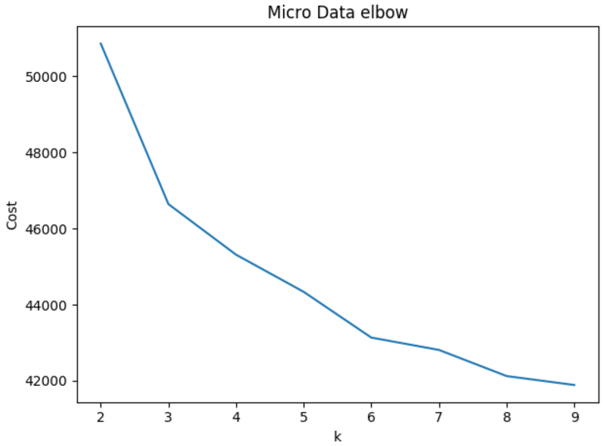
We get almost the same result in mnn method as seen below: (m = 10, as in first section)

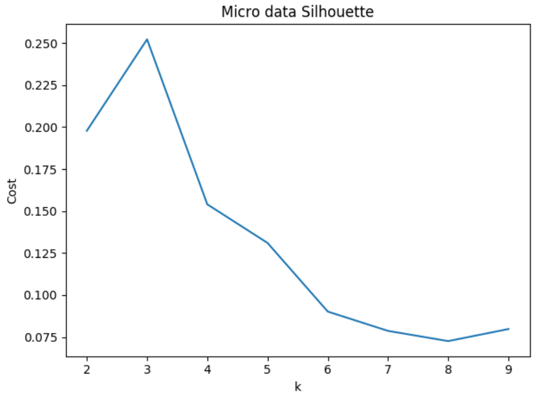


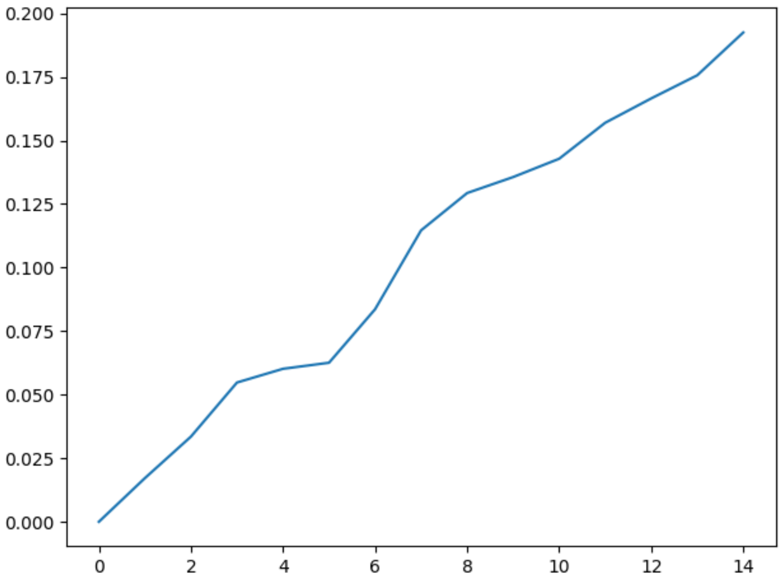


Microdata Data Set:

In order to select K we apply elbow, Silhouette and eigengap methods



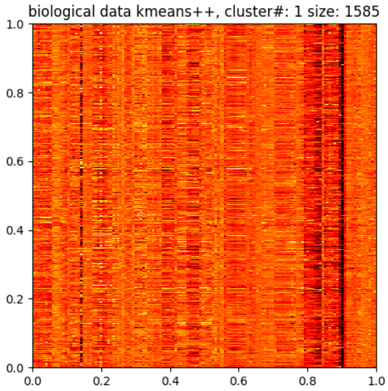


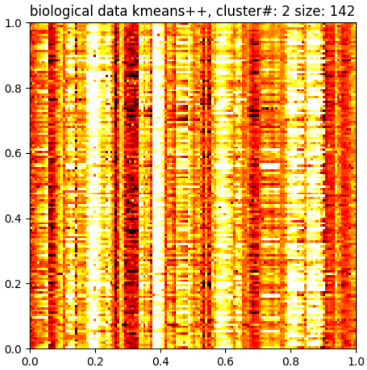


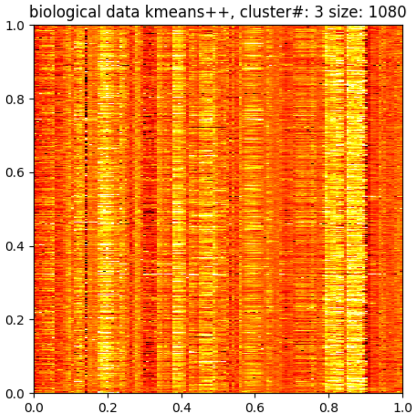
Eigengap for mnn with m = 14, microarray.

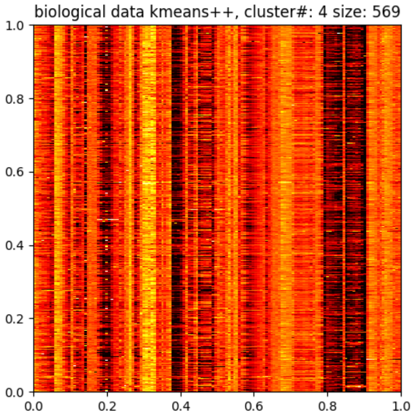
Selecting K is more complicated than the former cases, we can see inelbow method it points towards k= 6, in sihouette k =3, eigengap with mnn and m=14 gave k =6 or other values such as 9 and 3 (as elbow) so in order to decide which k to choose I ran clustering with multiple k’s and choose to select k = 6 (as in elbow and eigengap)

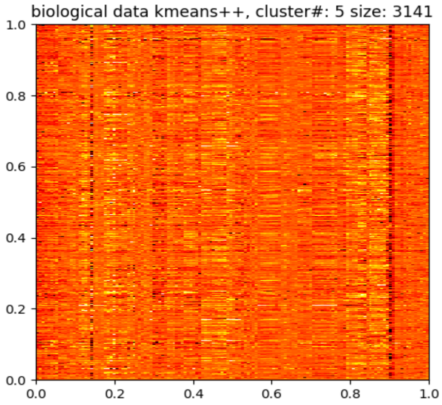
Kmean++ :

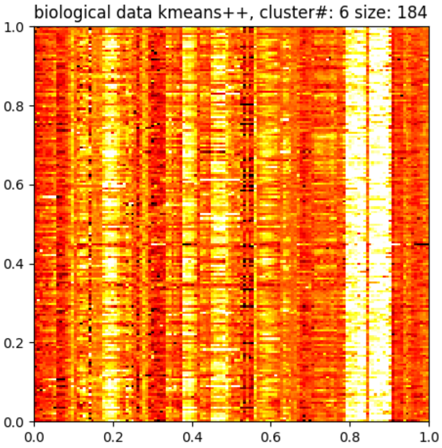

biological data kmeans++, cluster#: 1 size: 1585


biological data kmeans++, cluster#: 2 size: 142

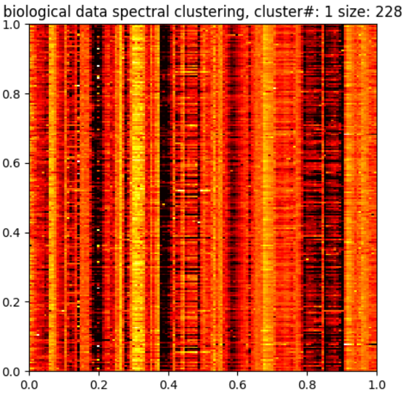

biological data kmeans++, cluster#: 3 size: 1080

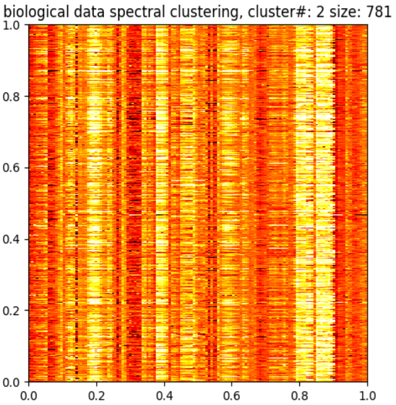

biological data kmeans++, cluster#: 4 size: 569

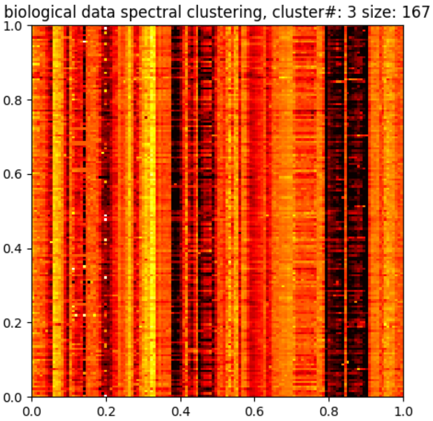

biological data kmeans++, cluster#: 5 size: 3141

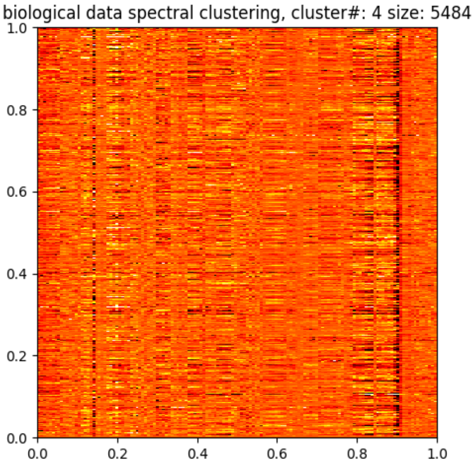

biological data kmeans++, cluster#: 6 size: 184

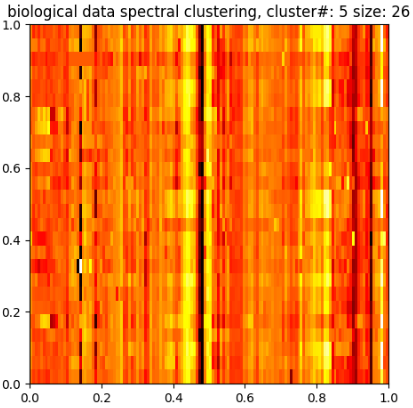
Spectral Clustering, using mnn with m= 14, k =6:

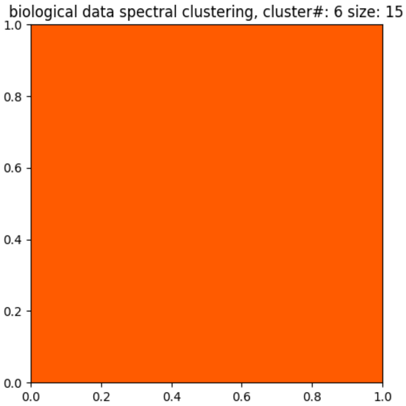








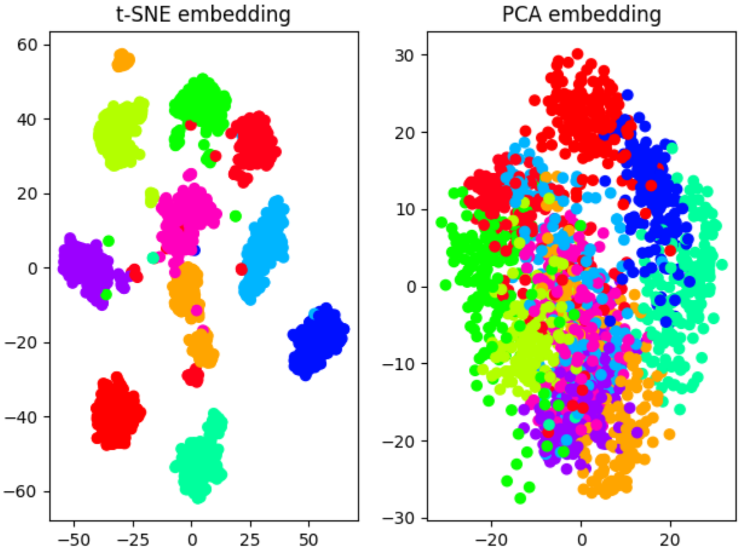




As we can see the clustering was more successful, note the lines and distribution of among the clusters specially #6 , #2, #3

t-SNE :

We compared in ex3 the embedding of our network to PCA, here we compare tSNE to PCA and we get:



From the comparison we can tell that t-SNE gives better results, the clusters are more far away and the colors(clusters) doesn't interfere with each other.