BMI203 – Winter 2020
Homework 2
Erin Gilbertson

Assignment Questions

1. Explain your similarity metric, and why it makes sense biologically.
    a. I chose to use Euclidean distance between a one-dimensional vector representation of active sites. The 1-D representation is essentially just counts of amino acids, which made Euclidean distance a straightforward way to calculate the difference between two active sites. I think that this makes sense biologically because after looking at the active sites more closely I realized that these aren't actually in sequence order so it seemed more relevant to look at how many of each amino acid are present at each active site. While it would probably be interesting to look at 3-D structure of the active sites, I have no expertise in this area and decided that my time was better spent understanding the algorithms.
2. Explain your choice of partitioning algorithm.
    a. I chose to do a K Means algorithm because it is a fast, and straightforward partitioning algorithm. The basic process of K-Means is shown below. I started with this outline and started building up my functions to complete the tasks. More specific documentation on these is in my code.
        i. Randomly initialize k cluster centroids
        ii. Assign data points to nearest cluster
        iii. Recalculate cluster centroids
        iv. Assign data points to nearest cluster
        v. Repeat steps 3-4 until centroids stop moving significantly
3. Explain your choice of hierarchical algorithm.
    a. I used a basic frame work of an agglomerative clustering algorithm as outlined below.
        i. Create an all sites by all sites similarity matrix
        ii. Find the two most similar clusters
        iii. Combine the two most similar clusters
            1. Recalculate cluster average for use in average linkage comparisons
        iv. Repeat steps 2-3 until everything is in the same cluster (or you reach the desired number of clusters.)
4. Explain your choice of quality metric. How did your clusterings measure up?
    a. I used silhouette score for my quality metric, it works by XXXXXX…..
    b. My K – Means clustering resulted in a maximum silhouette score of 0.79 at 2 clusters, which is reasonably high so I feel pretty good about my K-Means clustering given the simple distance metric that I

used. This is shown below in figure 1. Increasing the number of clusters consistently decreases the silhouette score.

c. My agglomerative clustering resulted in a maximum silhouette score of 0.72 at 2 clusters, which is reasonably high so I feel pretty good about my agglomerative clustering given the simple distance metric that I used. I did not optimize agglomerative clustering to choose the number of clusters with the highest silhouette score due to time constraints, but chose the number of clusters that performed the best with K means.

5. Explain your function to compare clusterings. How similar were your two clusterings using this function?
    a. I used Jacaard Index to compare my two clusterings, Jacaard works by asking whether two given points are contained in the same cluster as each other between two different clustering method results. This serves to compare how similar the clusters between methods. Jacaard index ranges from 0-1.
    b. The Jacaard similarity score between my two cluster methods was 0.97 which is pretty high. I don't really think this tells me all that much since the vast majority of the points in both sets were in one large cluster. I think this is probably due to my vector representation and similarity metric not actually picking up any signals that exist in the data. Given more time I would have used a more detailed representation of the data to hopefully get more meaningful results.
6. Did your clusterings have any biological meaning?
    a. Realistically, the only biological meaning that can be derived from my clusters is that active sites that cluster more closely together have similar amino acid compositions. If I were to be using this in research and wanting to learn more details about the biology, I would want to incorporate something about the 3D structure of the active sites.

Figure 1:

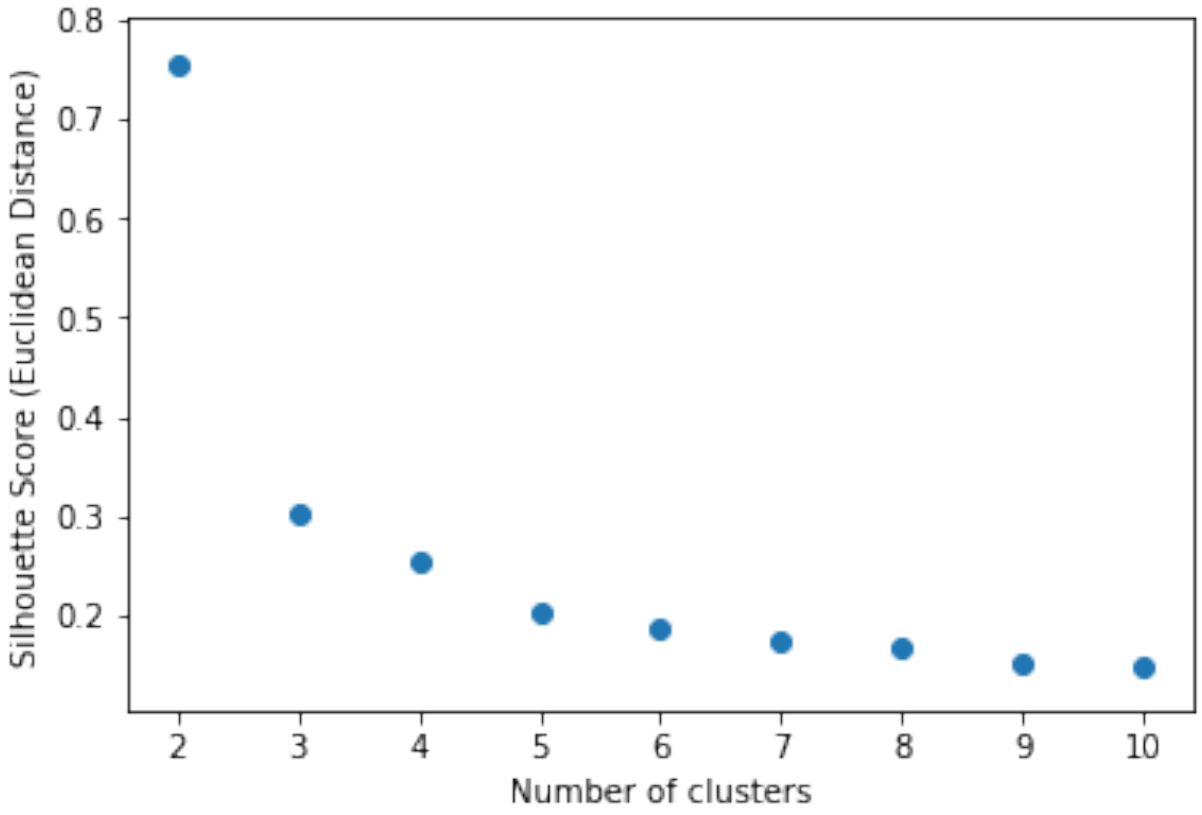
Silhouette Score (Euclidean Distance)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2
3
4
5
6
7
8
9
10
Number of clusters

Figure 2: K-Means Cluster Colored PCA of active sites; based on vector representation of each active site

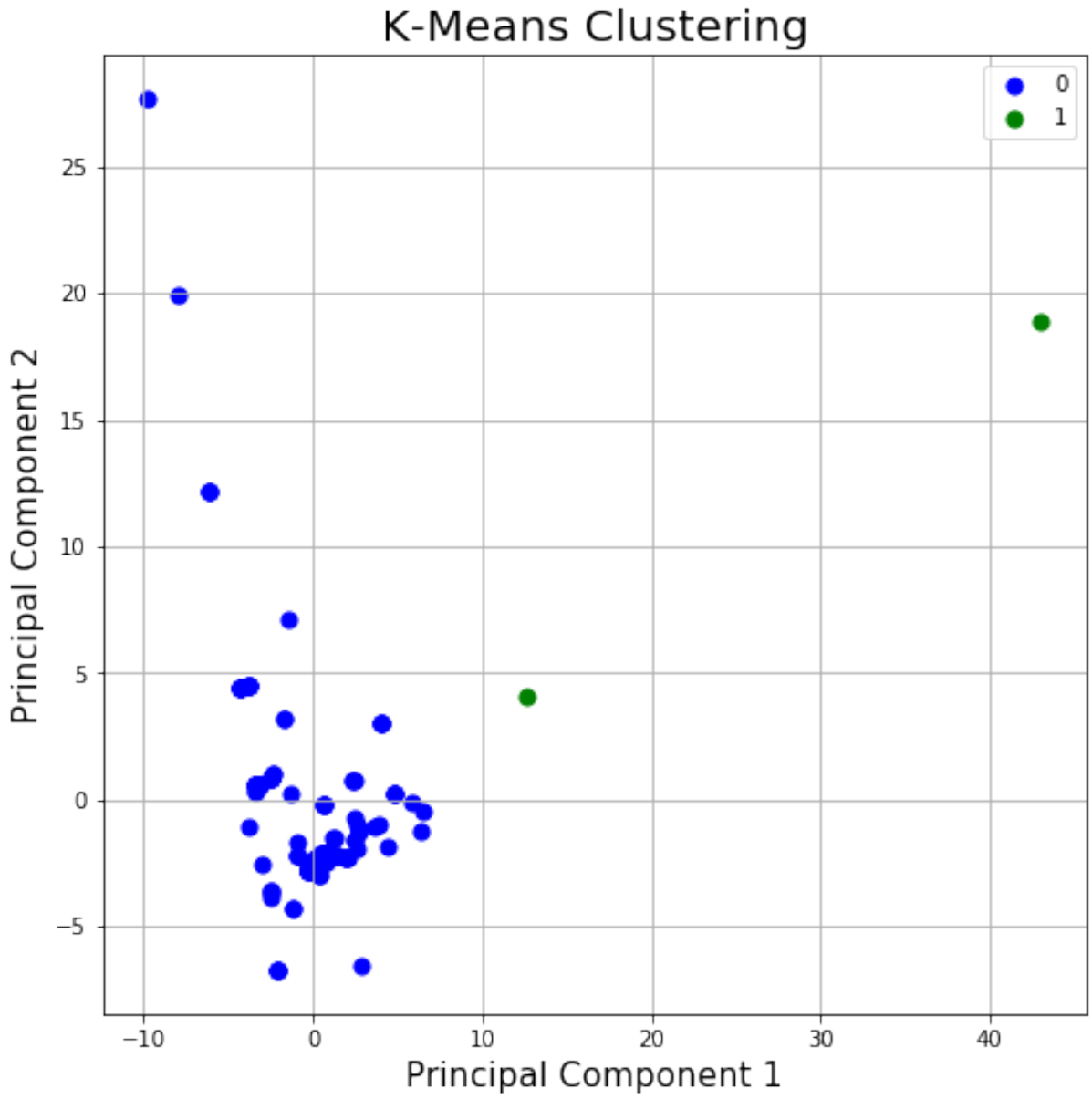

Figure 2: Agglomerative Cluster Colored PCA of active sites; based on vector representation of each active site