Jessica Miron

BME 598: Applied Programming

Homework 11

Executive Summary

The given dataset is highly temporally resolute data on the activation of regulatory T cells (Treg) and effector T cells (Teff). Analysis is being done to determine the number of clusters needed to show a pattern of differing expression after activation. With the correct number of clusters, there should be a pattern of increasing expression over time meaning that the genes were specifically responsive to the activation. The analysis steps taken can be seen below:

1. Load the Python packages
2. Using GEOParse, load GSE11292 into a variable called 'gse' and load expression values into a variable called 'expr'
3. Remove samples that are measured on ThGFP and ThGARP
4. Make a boxplot of the log2 transformed signal versus gene
5. Make a boxplot of the normalized log2 transformed signal versus gene
6. Select the top 3000 most variable features and scale them using the StandardScaler
7. Compute silhouette scores and plot them for k-means clustering between 2 and 20 clusters. Also plot the average silhouette scores for each number of clusters
8. Choose the best number of clusters based on average silhouette scores and which clusters show increasing expression over time
9. Make a clustermap of the eigengenes for the chosen number of clusters

Biological Interpretation

After analyzing the data, it is found that six clusters should be discovered in the data. Looking at the average silhouette score plot in Figure 1, there is a peak at six clusters. This peak signifies that the silhouette scored jumped up at six clusters, meaning the clusters fit well at six clusters. Five and seven cluster clustermaps were also made to ensure that six clusters was the correct number.

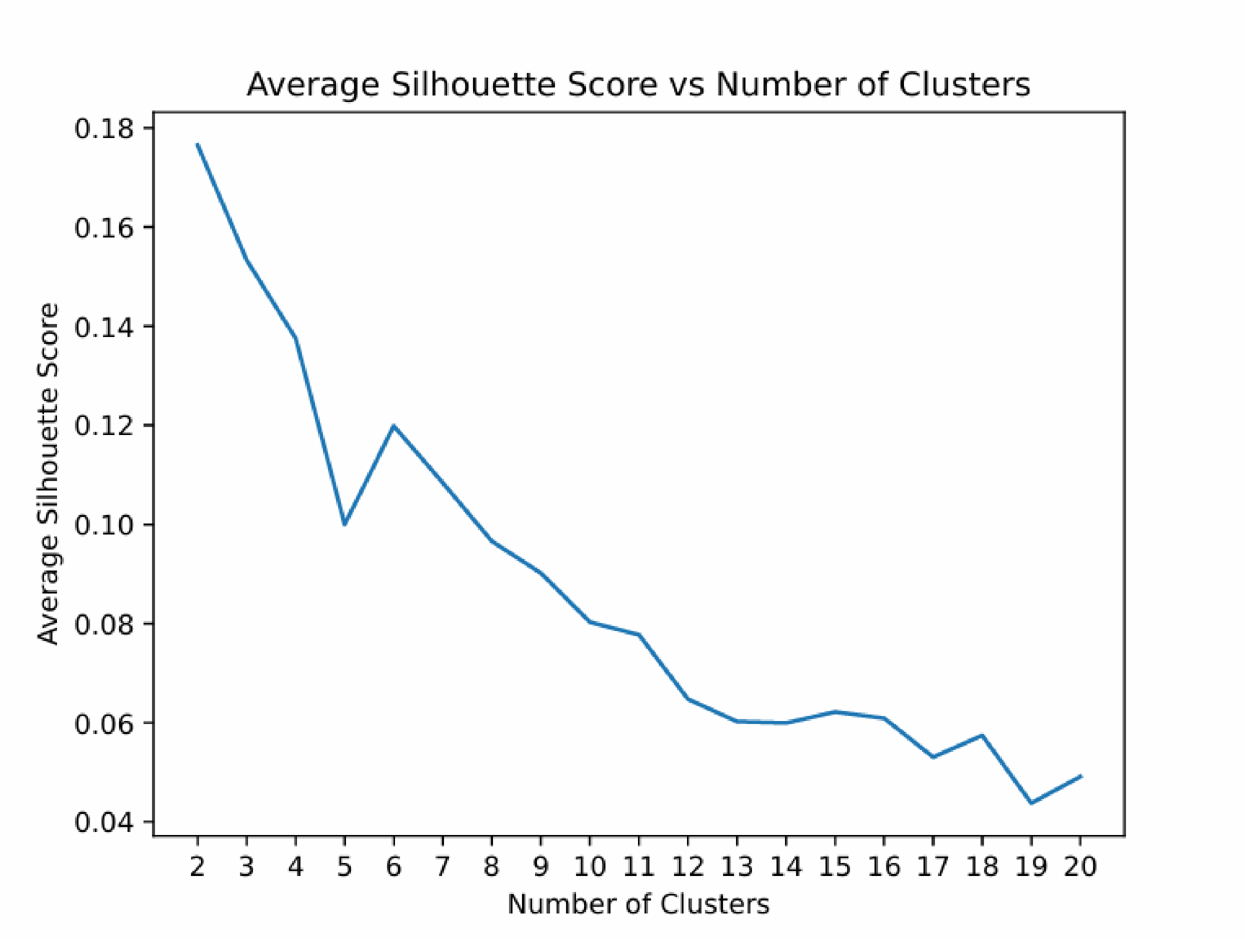

Figure 1: Average silhouette score versus the number of clusters. Using k-means clustering with two to twenty clusters the average silhouette score was computed and plotted. A peak is seen at six clusters signifying that six clusters fit the data well.

To determine the correct number of clusters from the clustermaps, clusters were assumed to be correct if a pattern was seen for all Treg and Teff timepoints across the cluster. For example, if all Treg samples are expressed at all time points then Teff should either be all expressed or all not expressed. The cluster that is most useful would be the cluster that has genes that become expressed over time in both Treg and Teff. Starting with the five cluster clustermap in Figure 2, it can be seen that some clusters fit well while others don't. The left two columns are Treg samples and the right two columns are Teff columns with time going from left to right. Cluster 0 has genes that are always expressed in Treg cells and always not expressed in Teff cells. Cluster 1 has genes with the opposite expression. Cluster 2 shows genes that are changing from not expressed to expressed over time in Treg but then are expressed in one set of Teff and not expressed in the other set of Teff. Cluster 3 has Treg that is close to no change in expression while Teff has one set that is not expressed and the other is. Finally, cluster 4 has no obvious pattern off expression within groups or across the cluster. Due to all the clusters not following the same pattern within the cluster, five clusters is not the correct number of clusters.

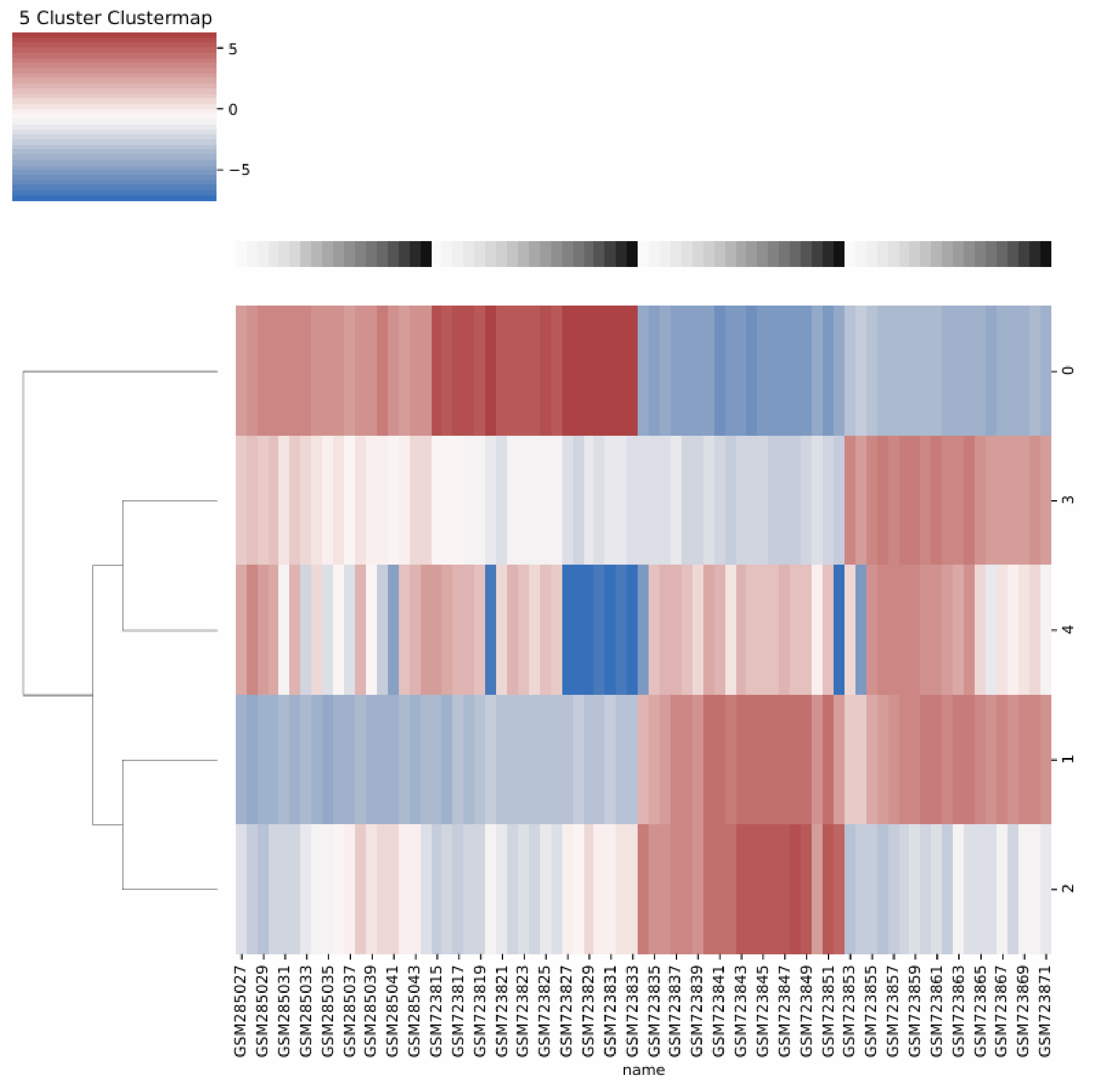

Figure 2: Five cluster clustermap. Treg samples are the two columns on the left and Teff samples are the two columns on right. Time moves from left to right in each column. Red shows genes that are expressed more after activation and blue shows genes that are expressed less after activation.

Moving to six clusters in Figure 3, there is a greater pattern within clusters. Cluster 0 has little pattern within the cluster. Cluster 1 has one Treg and one Teff set that are expressed and one not expressed each. Cluster 2 has genes that are under expressed in Treg and over expressed in Teff. Cluster 3 is the opposite with genes that are over expressed in Treg and under expressed in Teff. Cluster 4 has genes that are under expressed except in one Teff set where they are over expressed. Cluster 5 is the cluster of interest. Across all Treg and Teff, the genes start under expressed and become more expressed as time goes on. These are genes that are affected by the activation and become more expressed after activation. This cluster of genes that are equally affected in both Treg and Teff means the six clusters are most likely the correct number of clusters. To verify, seven clusters were also run to ensure that six clusters are the correct number of clusters.

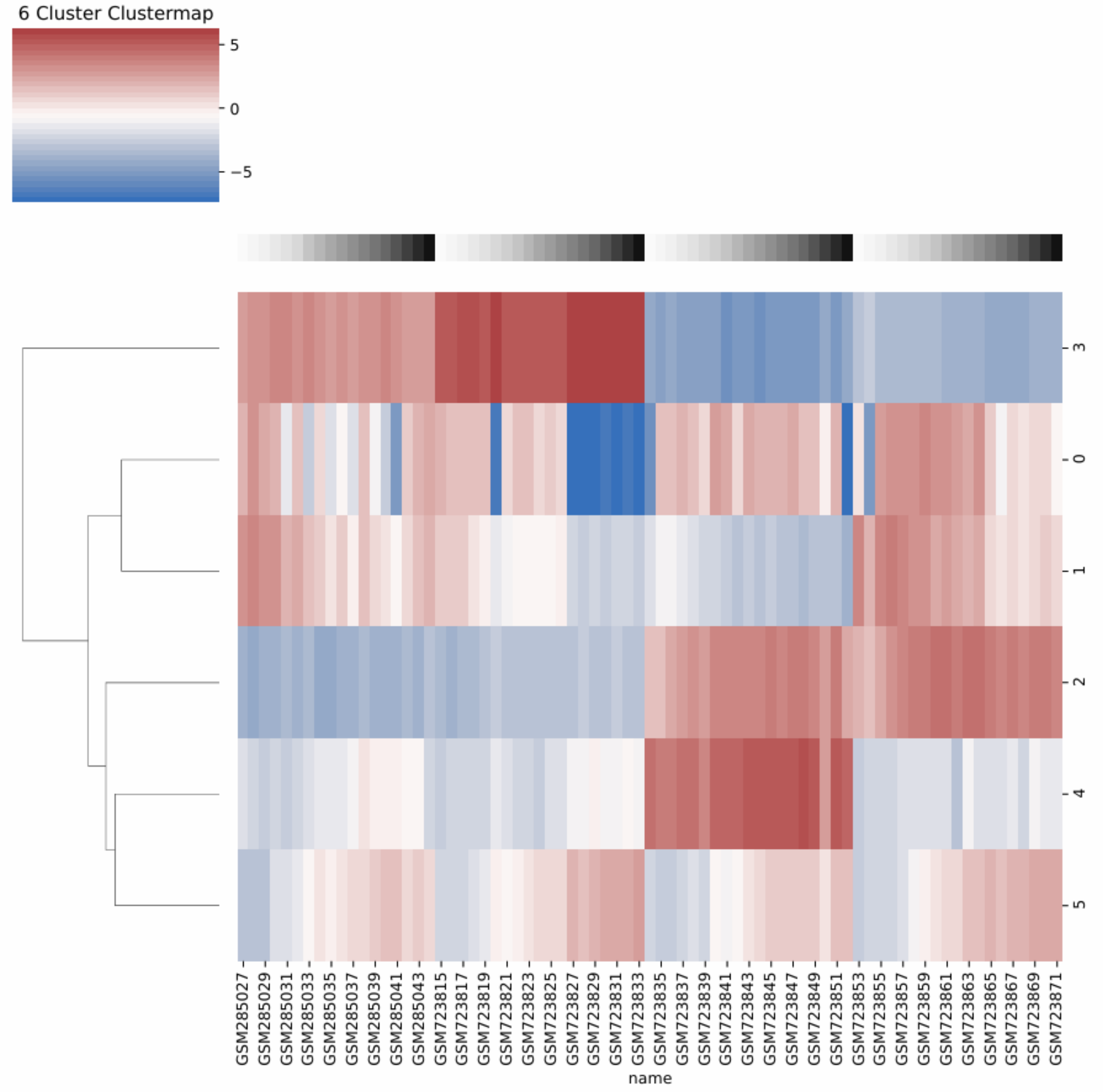

Figure 3: Six cluster clustermap. Treg samples are the two columns on the left and Teff samples are the two columns on right. Time moves from left to right in each column. Red shows genes that are expressed more after activation and blue shows genes that are expressed less after activation.

Seven clusters is seen in Figure 4. Since the cluster of interest has genes that become more expressed after activation, focus is placed on groups that show this pattern. Cluster 1 has the obvious pattern of increased expression over time. However, cluster 5's Treg samples also increase expression over time while the Teff are all over expressed in one set and all under expressed in the other. This mismatch means that there are too many clusters and genes that should be together have been wrongly separated across multiple clusters. This shows that six clusters is the correct number to identify genes that have increasing expression over time after activation. Knowing that six clusters is the correct number, the genes in cluster 5 of the six clusters can be identified to know which genes are being affected by the activation agent.

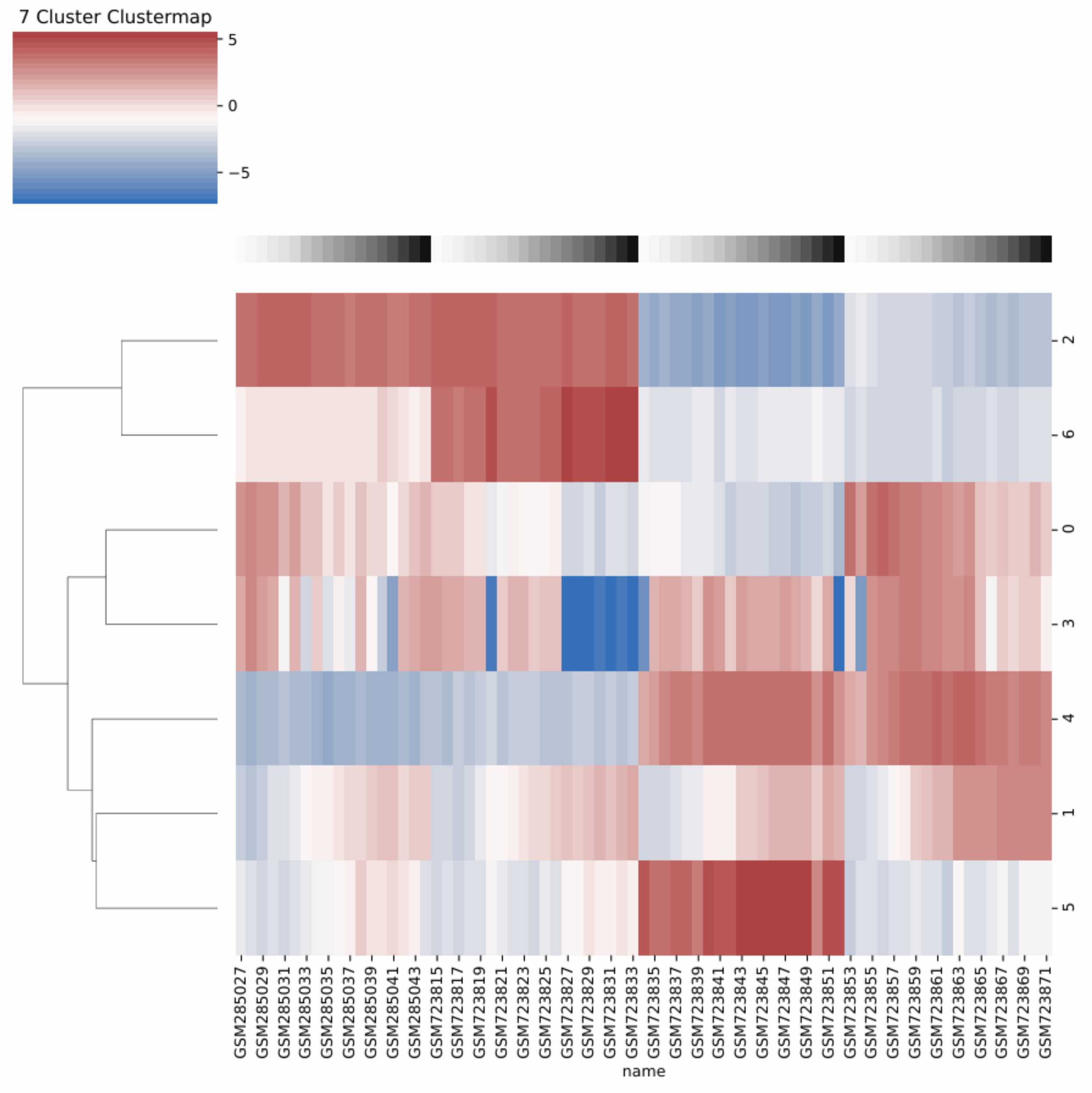

Figure 4: Seven cluster clustermap. Treg samples are the two columns on the left and Teff samples are the two columns on right. Time moves from left to right in each column. Red shows genes that are expressed more after activation and blue shows genes that are expressed less after activation.