SCluster represents a similarity matrix of the objects

VCluster represents the raw data, the actual multi-dimensional representation of the objects that need to be clustered.

Command line follows same format: << **vcluster [optional parameters] InputMatrixFile Nclusters** >>

For the purpose of this lab:

* **InputMatrixFile** -> avg\_E\_coli\_exp380\_clean.data.txt
* **Nclusters** = 2
* **Options**: Hierarchical clustering [ -clmethod=agglo] | Direct [ -clmethod=direct ]
* **Similarity Measurements**: - cosine angle [-sim=cos ]

correlation coefficient [-sim=corr]

Euclidean distance [ -sim=dist ]

* + Output formats: -plotmatrix | -plotcluster

Results:

I could not get the graphs to print in CLUTO using Linux. I referenced the manual, other students, and the internet but still could not produce any results. Eventually I tried it in a Windows operating system and was able to produce an output file. I tried many different commands attempting to change the -plotformat options and other suggestions from the manual, but still have not been able to view my results. I do not have the correct software to view the graphs, so I am not able to analyze the results from CLUTO.

After several days of failing with CLUTO, I tried to complete the lab purely in R studio. I spent several more hours reading the documentation and finding videos online on how to use R for cluster analysis. The files I have included in the submission file are the output files from CLUTO and from R. The graphs I was able to visualize have been included in the report below.

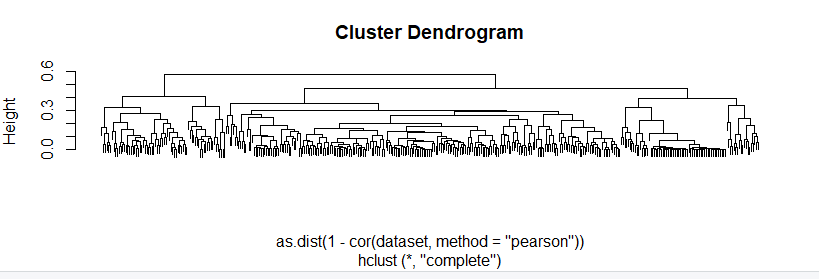


Figure : Hierarchical with euclidean distance

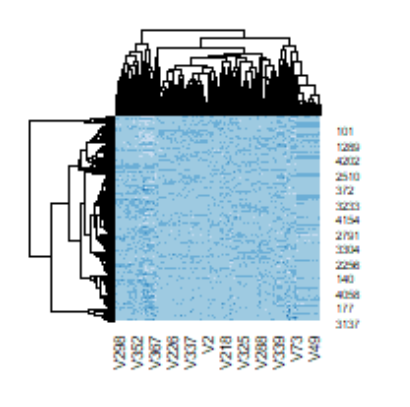


Figure 2: Hierarchical with correlation coefficient

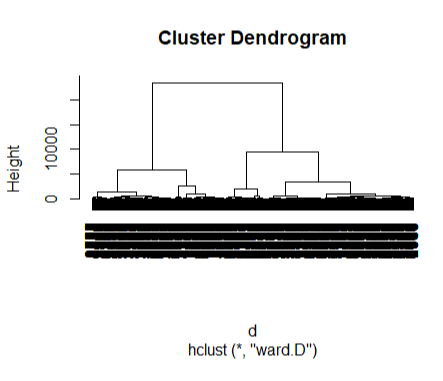


Figure 3: Hierarchical with Euclidean distance

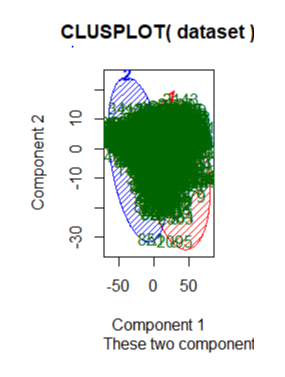


Figure 4: Direct clustering with K-means = 2

Analysis:

From figure 1, it looks like most data points can be allocated by 6 clusters. Using hierarchical clustering and Euclidean distance, this graph shows that most of the clustering nodes from the bottom are well defined as you move up the parent-child relationships to branches located at about 0.3, or 50% up the y-axis. Given the large dataset and the wide dispersion of conditions compared to genes, it is hard to recognize an obvious pattern for clustering from the other graphs. The heat map produced in figure 2 used a hierarchical approach for clustering and used correlation coefficient as the similarity measurement between clusters. We can see a small increase in relationships on the left side of the graph, but the correlation is not strong. The dispersion of gene behavior is widely expressed or unexpressed throughout the graph without much recognizable pattern of relationships. The two clusters created in figure 4 have so many data point it’s hard to distinguish where points exist in the corresponding clusters. Using 2-way clustering did not segregate the data points well enough to determine relationships. It would likely be helpful to use a higher number k to create more centroids that will cluster the data better.