Jayam Sutariya

Dr. Nguyen

CS 661

9th December 2022

Project 2 Report

The purpose of the project is to compare the t-score values for the GSE19804 dataset using two different methods. The project is essentially separated into eight parts. The first part is simply to gather the dataset and create a data matrix and the corresponding patient data information. The second part is to perform a t-test to compare the control and condition groups that were gathered in the last step. Further calculations are performed on these vectors and ultimately a data frame is constructed with valuable information of the results. The third step is to extract the differentially expressed genes from the data frame and displaying the results through a volcano plot. The fourth step is to simply calculate the t-score between the two groups.

The next two steps are where the two methods in question are experimented with. The fifth step is the first method where empirical p-values for the t-scores are calculated using a specific permutation analysis. The same process is repeated for the sixth step using the second method where this time the Euclidean distance is used instead using a similar permutation analysis. The seventh step is simply to visualize the results of the two methods by plotting the values from the methods. From the hist-pT and hist-pE (attached) histograms, the two graphs are very similar. However, further statistical analysis must be done to ensure that hypotheses. In step eight, the correlation between pT and pE vectors is calculated. The resulting correlation value is 0.9900595. The correlation between the empirical p-values is very high, meaning that they very correlated even though different methods were used to calculate the p-values.