Microarray Based Tumor Classification- Programmer Role

Introduction

This task involves filtering the normalized microarray gene expression data. The output of this is a reduced set of genes that can then be used for hierarchical clustering. In this normalized expression data, there were 135 samples as columns and 54,675 genes as rows. When there are that many more genes than samples, calculating statistics over the entire dataset is not very informative, and therefore requires cutoff thresholds for filtering out genes that do not meet those thresholds. Careful selection of these methods and cutoff values are required for proper noise filtering. After noise filtering, clustering is used as an unsupervised method for creating groups of similar "objects", or in this case, distinguishing expression data by disease status of the colon cancer is able to visualized.

Methods

The first step of filtering required keeping genes that contain at least 20% of its expression values being above log2(15). For the second step of filtering, the variance of expression values for each gene along with the median value of all these variances were calculated to be used for a chi-squared test. A threshold of $p<0.01$ was also required for this test. $T=(N-1)(s/\sigma_0)^2$ was the formula for the chi-squared test, where N is the number of samples, s is the sample standard deviation, and σ_0 is the median standard deviation. If the variance to median variance ratio for each gene was at or below the chi-squared test result, those were the next genes to be filtered out. Lastly, the coefficient of variation for each gene was calculated, and genes having a value less than 0.186 were filtered out as the last noise filtering step.

The clustering method used on the filtered data samples was hierarchical clustering, which was the most feasible computational method appropriate for this data. A dendrogram was the output of the clustering, which was then cut into 2 clusters.

Using the expression data and the cluster memberships, a Welch t-test was performed to identify genes that were differentially expressed. A data frame was produced from this result, with the calculations of t-statistic, p-value, and adjusted p-value for each probeset gene ID. The genes significantly expressed had a p-value less than 0.05.

Results

The 54675 genes in the data fell to 39661 after the first expression value filtering threshold, then fell to 24,622 genes after the chi square test, and lastly fell to 1558 genes after the coefficient of variation threshold. After clustering, one cluster had 58 samples, and the other had 76. Figure 1 below shows the heatmap of the gene expression across all samples, with the C3 subtype falling under the red cluster and the C4 subtype falling under the blue cluster. What can also be seen in the heatmap are the 2 blue stripes within the red cluster that shows that those samples may have been misidentified. There were 1261 significantly expressed genes that fell under an adjusted p-value threshold of less than 0.05.

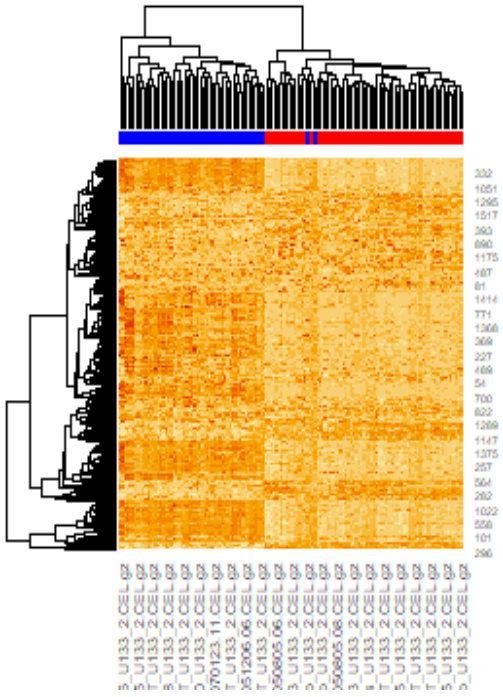

Figure 1. Heatmap of gene expression across all samples.

Discussion

The goal of this analysis was to filter a very noisy microarray expression data set, to cluster the samples and look at expression levels, and identify genes that are significantly expressed at an adjusted p-value threshold. There were 1558 genes left after noise filtering, and 2 clusters to group by with 58 samples in one and 76 samples in the other.

Results from this analysis had some difference with the associated paper's results. The Chi-squared test done could have been different since it wasn't quite clear in the paper is the researchers had performed a lower, upper, or 2-tailed test. Consensus clustering was the method

used by the researchers, and the clustering done here was hierarchical clustering, which could have also caused slightly different results, as well as the heatmap looking slightly different.