Leukemia Class Classification by Gene Expression Monitoring

Bioinformatics | Assignment 1 | Group 6
Tania González-Robles, Sion Hau, Nicholas Stitt, Garrett Yoon

A gene expression dataset for 38 randomly selected samples of bone marrow aspirates from human acute leukemia patients was used to demonstrate the viability of automatic class discovery and prediction using only gene expression monitoring. Of the 38 mononuclear cell-derived RNA samples, 27 were from childhood Acute Lymphoblastic Leukemia (ALL) patients and 11 were from adult Acute Myeloid Leukemia (AML) patients.

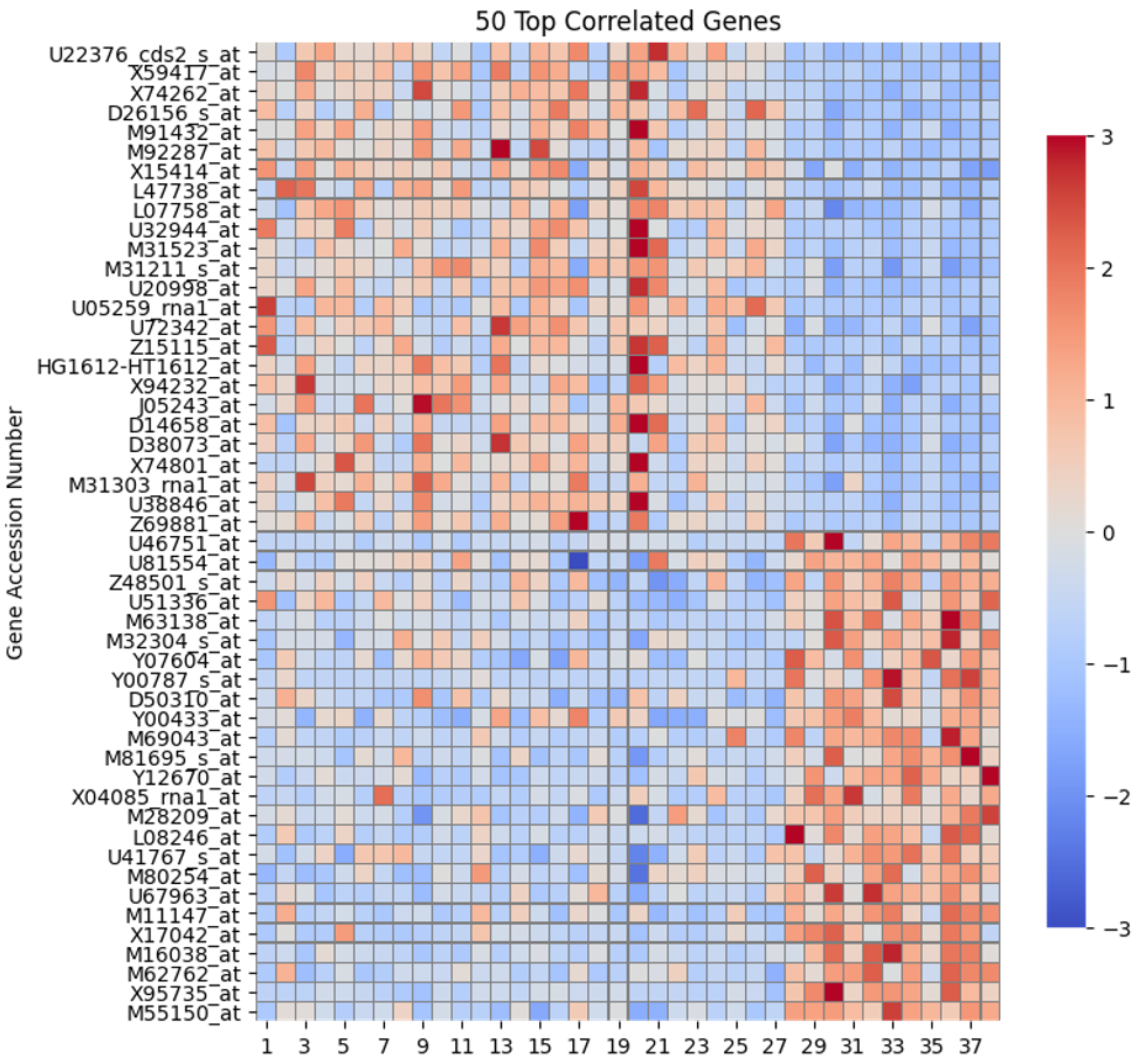
1. Recreate Figure 3b - See how many overlaps you have with their top correlated genes

The top 50 genes most correlated with the ALL-AML class distinction can be represented on a heatmap following data realignment, filtering, transformation, validation, ranking and selection of the initial 7129 genes.

In order to include only the genes most relevant to the class distinction, probes that returned an 'absent' (A) call for more than 77% of the samples were excluded from the dataframe. Endogenous control probes, vital for experimental validation but not relevant to the leukemia class separation, were also excluded by targeting gene accession numbers containing "AFFX". The two filtering processes described left behind 3045 genes for analysis.

By applying the z-score formula to the gene expression values and thereby standardizing the datasat, the features of the dataset are represented more equally and any issues with outlier data-points are avoided, these properties of z-score normalization mean that the data is more easily represented and visualized on a heatmap. Following normalization, the dataframe was checked for missing values before the mean and standard deviation of each of the remaining gene's expression values for the 38 samples was computed and added to the dataset.

To allow separation of data and classification of genes, a Welch t-test is applied to the dataframe using the newly created mean and standard deviation values to compare the means between samples to assess the significance of their difference. The data was then resorted based on the returned t-value which gives an indication for how different the gene expression values for that gene is compared to other genes. By selecting the first 25 and last 25 genes from the sorted list, essentially the top 50 most differentially expressed genes and therefore the genes most likely to contribute to the difference in the leukemia samples are chosen. What would be expected is a separation of the data into two distinct classes, ALL and AML, half of the genes should show high expression (red) in approximately 70% of the samples and significantly lower expression (blue) in the other 30% (ALL) whilst the other half of the genes would show the opposite (AML). This distinction in classes is expected due to the split in leukemia samples described in the beginning.



When compared to the Golub et al. paper, 29 out of the 50 genes exactly matched the gene accession numbers that were identified using the aforementioned processes. Multiple factors could have contributed to the discrepancy between the two heatmaps. To begin with, the number of genes that were included in the analysis could have been different given that the paper started with only 6817 probes, whereas the filtering done to produce the heatmap above began with the complete set of 7130 probes. Although there were steps taken here to filter out absent and control probes, the authors of the paper could have excluded probes based on other criteria too. Other differences in data normalization processes could have significantly contributed to the different results, where Golub et al. used a measure of correlation P(g,c), where $P(g,c) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g)+\sigma_2(g)]$ which emphasized the signal-to-noise ratio when g represents the log of each expression level, the processes described here do not normalize the data through a log transformation but rather standardization. Standardization returning z-values in this instance was preferred over log transformation due to the overwhelming presence of negative values, however using standardization does not decrease any existing skew in the data probably explaining the difference in heatmaps.

Some of the genes identified in processes described here, although not present in the top 50 genes selected by Golub et al., have been emphasized in other studies of class distinction between AML and ALL. For example, the ALDR1 Aldehyde reductase 1 gene (X15414_at) and PRG1 Proteoglycan 1 gene (X17042_at), among others, are known to be affected in AML patients specifically. Genes such as MPL gene (thrombopoietin receptor) (U68162_cds1_s_at) and RAS-RELATED PROTEIN RAB-1A (M28209_at) have been known to be differentially expressed in other types of leukemias. This result aligns with the suggestion in the Golub et al. paper that the genes are not only markers of hematopoietic lineage but also related to cancer pathogenesis.

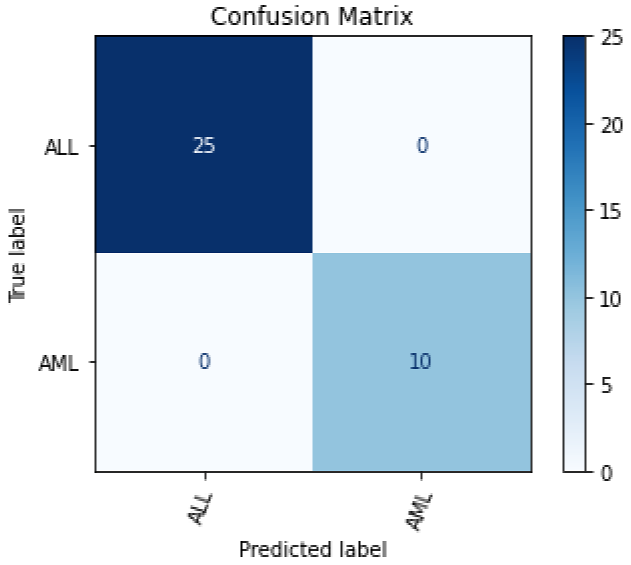
On the other hand, there were other genes selected through our analysis that seem to be (thus far) unrelated to leukemia classification, hematopoietic lineage or cancer pathogenesis. This feature along with the lack of uniform expression across classes, again, highlights the value of multigene prediction methods for class distinction.

2. **Train a model using the 50 most highly correlated genes (25 AML and 25 ALL) to predict the class of each sample using cross-validation.**

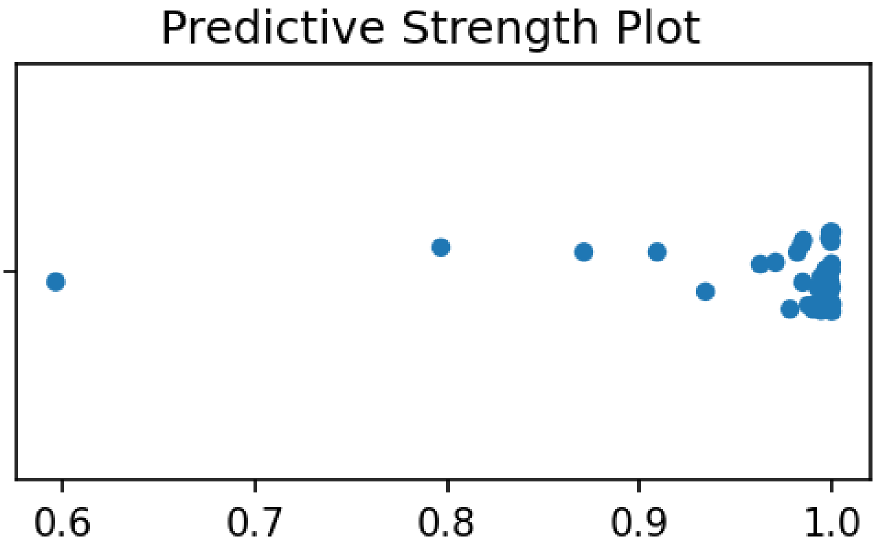
Once we found the top 50 most correlated genes characteristic of the two leukemia subtypes, we used this information to train a predictive model leveraging logistic regression. The logistic regression was used to estimate the performance of the algorithm when predicting if a sample is AML or ALL. The Train Test Split function was used to mediate the splitting of the data into training and test sets for the logistic regression. The logistic regression algorithm then bins the test sets and returns the results of the algorithm in the form of two arrays: the predicted class and the actual class. In our results the predicted class matched the actual class in all but one case.

We next needed to test the predictive strength of our classification. We employed the 'Leave One Out' cross-validation technique to give an insight on how well the model will generalize to an independent data set. This technique moves through the data set, pulling one sample out at a time to test and using the remaining samples to train the model. This process is then repeated for every sample.

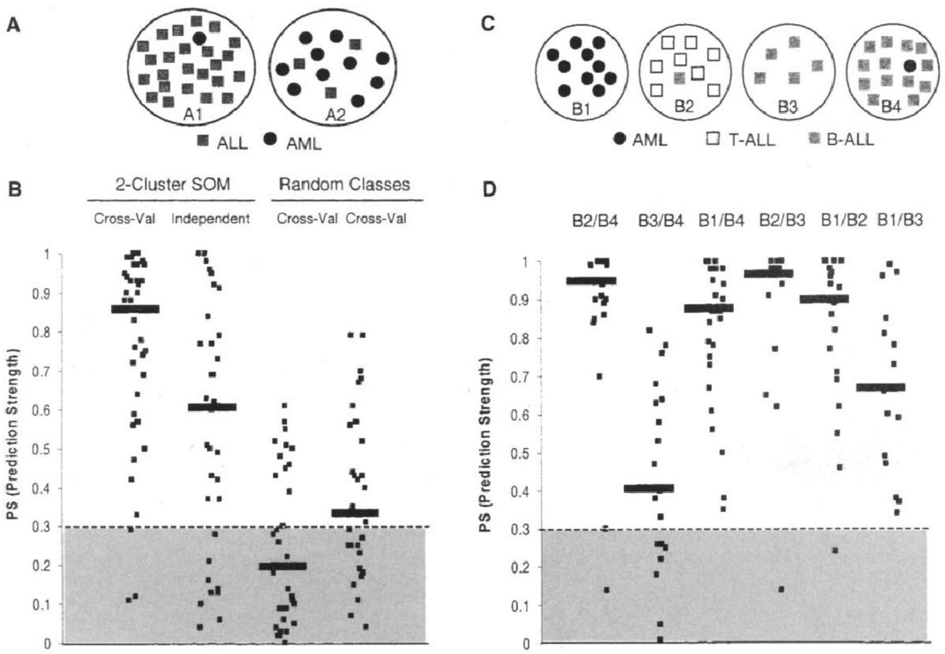
A confusion matrix was created to visualize the results of the classification. This confusion matrix shows the true labels on the y-axis and the labels predicted by the linear regression on the x-axis. We can see all of the true AML and ALL samples were correctly predicted by the model.



The predictive strength of this data set is visualized by sample in the plot below. Each point on the plot represents the predictive strength of a class predictor that was generated by the leave one out technique.



We can compare the predictive strength of our technique to that of Golub et. al (1999) by looking at the "Cross-Val" column in the 2-Cluster SOM of figure 4B. They show that their model produces a median predictive strength of 0.61.


A
A1
A2
ALL
AML
C
B1
B2
B3
B4
AML
T-ALL
B-ALL
B
2-Cluster SOM
Random Classes
Cross-Val
Independent
Cross-Val
Cross-Val
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
D
B2/B4
B3/B4
B1/B4
B2/B3
B1/B2
B1/B3
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

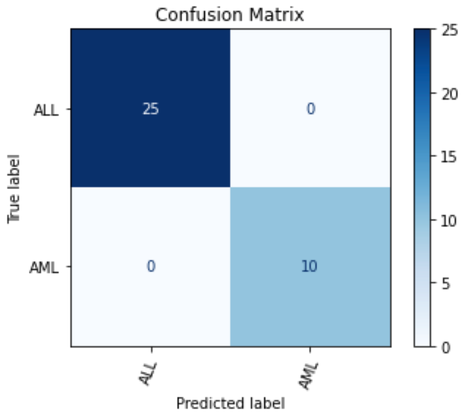
3. **Try differing amounts of genes and compare the results eg:(top 2, 10, 100, 1000)**

The prediction model proved to be robust using the top 50 most differentially expressed, so we next investigated creating prediction models using different numbers of genes. We started by expanding the number of genes used to train the model to 100.

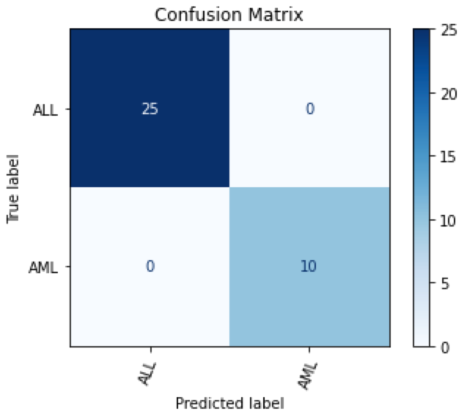
After running the same training and validation operations, we found a decrease in the predictive strength of the model. One sample was incorrectly labeled by the logistic regression instead of all samples being correctly labeled. This variation may be explained by the inclusion of more genes that were less differentially expressed than the original top 50. The increased overlap of signal for the added 50 genes between ALL and AML samples may lead to a less significant difference for the algorithm to use for classifying samples.

We then further increased the set of genes to 1000. In this case, the logistic regression misidentified two of the samples, one for each class. The confusion matrices for these runs is shown below.

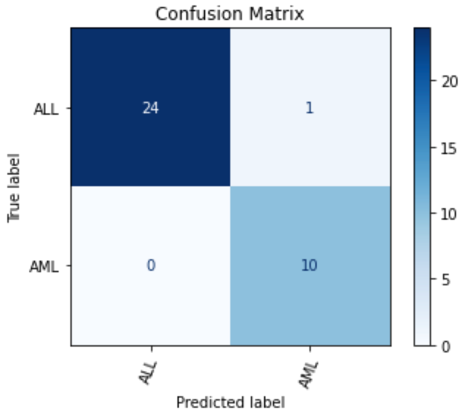
2 probes



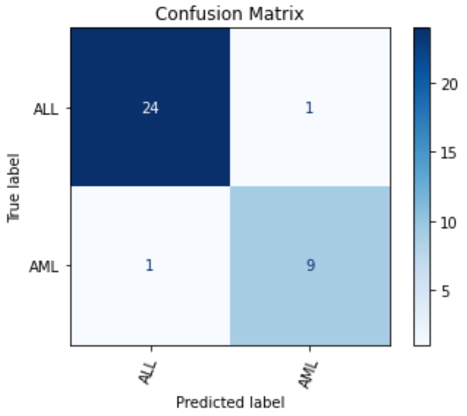
10 probes



100 probes

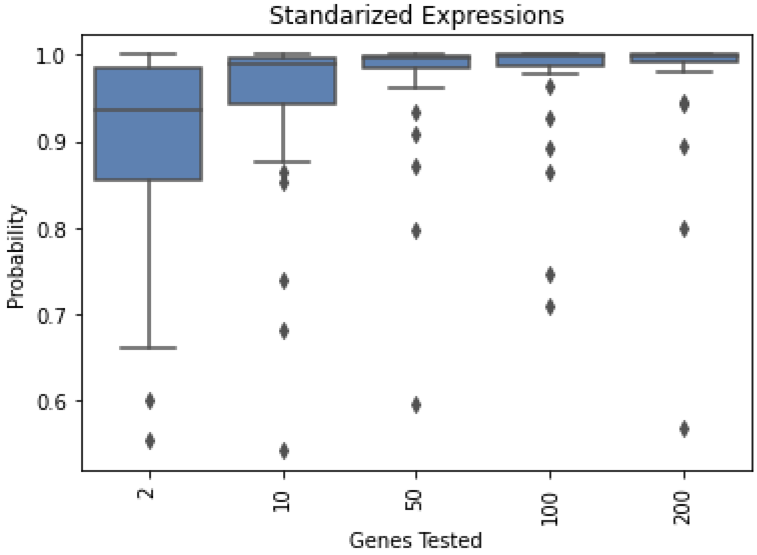


1000 probes



Prediction strength was then investigated using less than 50 genes for the analysis. We started by testing the predictive strength of the model using just two genes. In this scenario, the algorithm correctly classified all of the genes. The parameters were then changed again to use ten genes. The model correctly classified the samples when using the 10 most differentially expressed genes.

The predictive strength of the different sets is visualized side by side below. Using just two genes has the highest spread in the deviation of probability. The deviation decreases as the number of genes used is increased, suggesting that the probability of accurately assigning the samples to the correct groups might increase with more genes per sample. However, when we increase the number of genes to 100 and 1000, we observe an increase in false-positive (type-I errors) and false-negative (type-II errors) classification.

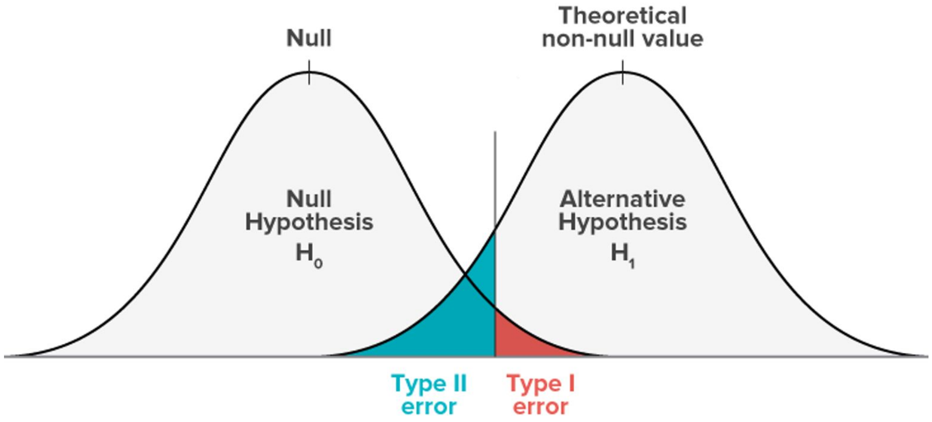


The results, as seen in the confusion matrices, show the opposite of this. The classifications using two, ten, and 50 genes correctly assigned all of the subjects, while the classification with 100 genes had one misassignment and the classification with 1000 genes had two misassignments.

This can be explained through the algorithm used to determine predictive strength. Even though using two genes lead to an accurate assignment for our data set, it does not have the statistical power to be used for being used to predict other data sets. The power increases the more genes that are added as seen in the predictive strength comparisons.

Increasing the number of genes leads to a greater possibility of misassignment. The accuracy of the assignment becomes increasingly diluted as more genes are added to the data set, with more samples sharing similar gene expression profiles. This concept is shown in the example below.

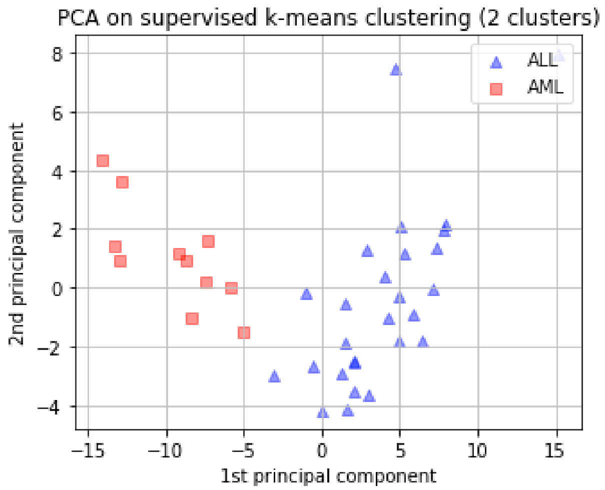
Errors

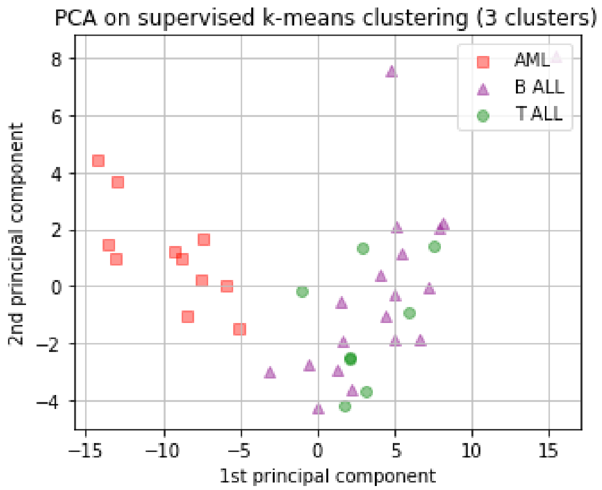


A middle ground must be chosen between the two phenomena which is why we, and possibly Golub et al., chose 50 genes to train the predictive models.

4. **Cluster the data using any method you like (K-Means, Hierarchical). What are the defining features of the clusters?**

Using the K-means clustering method with PCA visualization, first the number of clusters for K-means was set to 2 clusters (left) to determine if it is able to distinguish between the gene expression analysis profiles of the 2 classes of acute leukemia (ALL/AML). As can be seen, the algorithm was able to successfully cluster the ALL samples away from the AML samples, this is represented by the Euclidean distance between the two types of data-points (red/blue). The AML samples (represented by the red squares) defined by their gene expression profiles, are most separated from the ALL samples (represented by the blue triangles) based on the 1st principal component, as expected because this component accounts for the highest variance in the dataset.





However, when given the challenge of grouping into three distinct clusters (right), this method is unable to distance the samples for the ALL subtypes (B-ALL/T-ALL). The representative data-points (purple triangle/green circle) are mixed within the same cluster and virtually remain unchanged from the previous plot. There is not enough distance created between the samples to form a third distinct cluster based on their gene expression profiles. Multiple reasons could account for the failed clustering of ALL subtypes including and not limited to, sample size being too small, lack of variance in gene expression profiles between B-ALL and A-LL, underlying assumptions of K-Means not aligning with the dataset etc. It can also be noted that there was an outlier in both plots separated from its cluster largely based on the 2nd principal component.

GitHub Link to code:

https://github.com/TaniaJes/bioinformatics/blob/master/module-2/notebooks/reports/assig1-group6.ipynb

Group breakdown:

Garrett: Coding 5/5

Tania: Coding 5/5

Nick: Report Writing 5/5

Sion: Report Writing 5/5