# Molecular Classification of Cancer by Gene Expression Monitoring

Namrata Ray, Jugal Marfatia, Md Muhtasim Billah

## Abstract

Although cancer classification has improved over the last few decades, yet there are only a few approaches reported in literature for detecting new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). For this project, one of those existing studies was taken as inspiration where a generic approach to cancer classification, based on gene expression monitoring by DNA microarrays, was described and applied to human acute leukemias as a test case. In that work, a suggested class discovery technique was able to spontaneously distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. For this project, to identify sub-structures in the data and groups of genes that behave similarly, K-means clustering was performed. To determine the optimal number of clusters, the 'Elbow Method' and the 'Silhouette Method' were adopted. The derived results, in line with the literature of interest, demonstrate the feasibility of cancer classification based solely on gene expression monitoring and propose a general tactic for predicting classes for other types of cancer, irrespective of any previous information.

Keywords: gene expression, PCA, clustering, DNA microarrays, leukemia.

## Introduction

To distinguish among various pathogenetically different tumor types and targeting specific therapies to them has been one of the major challenges for cancer treatment. Improvement in cancer classification has thus been vital to achieve progresses in cancer therapeutics. Cancer classification, which was primarily based on morphological appearance of the tumor, has some limitations as tumors with similar histopathological manifestation can take considerably different clinical pathways and display dissimilar responses to therapy. Another difficulty arises for cancer classification because of its reliance on particular biological traits, instead of systematic and unbiased approaches for identifying tumor subtypes.

In the literature [1], based on which this project has been carried out, such an approach has been described depending on global gene expression analysis. They divided cancer classification into two major tasks: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes and class prediction refers to the assignment of particular tumor samples to already-defined classes. Human acute leukemias was chosen as a test case for this. Although the distinction between AML and ALL had been well established already, no single test was sufficient to establish the diagnosis till then. Rather, the clinical practice involved an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immune-phenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remained imperfect and errors did occur. But distinguishing ALL from AML is critical for successful treatment.

Though microarray studies have primarily been descriptive rather than analytical, it has been suggested [2] that such microarrays could provide a tool for cancer classification. In the literature being discussed, a systematic method was developed to tackle cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays [3-8]. They began with class prediction: How could one use an initial collection of samples belonging to known classes (such as AML and ALL) to create a "class predictor" to classify new, unknown samples? They developed an analytical method and first tested it on distinctions that are easily made at the morphological level, and then turned to the more challenging problem of distinguishing acute leukemias, whose appearance is highly similar.

For the project, the principal component analysis (PCA) and clustering has been used to classify ALL and AML cancer types using the data provided by Golub *et al* [1]. Explained variance in terms of the principal components has also been reported. The statistical analytical model was able to successfully determine the correlations of gene expressions to connect them with different cancer types which remained as the sole purpose of this project.

## Data Set

The dataset for this project comes from the proof-of-concept study published in 1999 by Golub *et al* [1]. It showed how new cases of cancer could be classified by gene expression monitoring (via DNA microarray) and thereby provided a general approach for identifying new cancer classes and assigning tumors to known classes. These data were used to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

The initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 7123 human genes (14). For each gene, a quantitative expression level was obtained. Samples were subjected to a priori quality control standards regarding the amount of labelled RNA and the quality of the scanned microarray image [9-10]. Intensity values have been rescaled such that overall intensities for each chip are equivalent. Also, there exists an independent dataset (for test, 34 samples) used in the paper.

## Method

Principal components analysis (PCA) is a common unsupervised method for the analysis of gene expression microarray data, providing information on the overall structure of the analyzed dataset. For this dataset, PCA has been implement on a set of 7123 genes of 72 individuals to reduce data dimensionality. To identify sub-structures in the data and identify groups of genes that behave similarly, K-means clustering was performed. To determine the optimal number of clusters, we relied on the ‘Elbow Method’ and the ‘Silhouette Method’.

## Results

In order to study the relation between gene expression and tumor classes, we first conduct PCA on the 7123 genes to transform them into 30 principal components. Our PCA indicates that the 30 principal components explain 70 % of the variance in the 7123 gene expression level as shown in Figure 1. Furthermore, the first two principal components explain 14.5 and 9.5 % of the variance respectively.

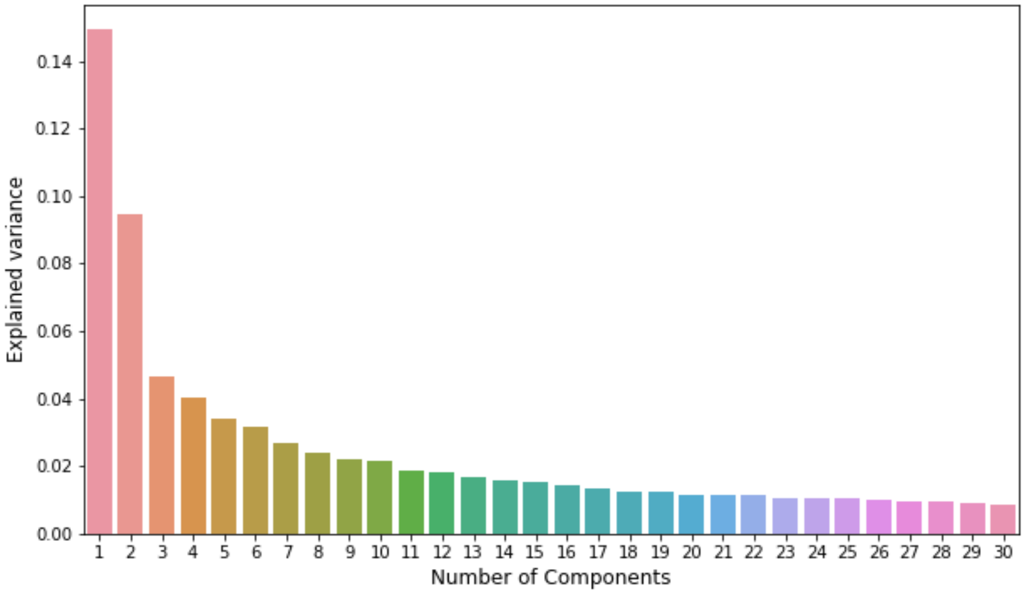

**Figure 1**: Variation explained by PC 1 through 30. PC 1 and 2 explain 14.5 and 9.5 % of the variation respectively.

Next, in order to understand if there is a pattern or a cluster in the principal components that explain the two types of tumor, we plot the principal component 1 and 2, labelled by tumor type. The simple scatter plot indicates that there is in fact some visible clustering on AML around the PC1 values 0-40 and PC2 values –40 to 0 as shown in Figure 2. This suggest that the natural step in order to understand the relation between the PC’s and tumor type is to conduct clustering analysis.

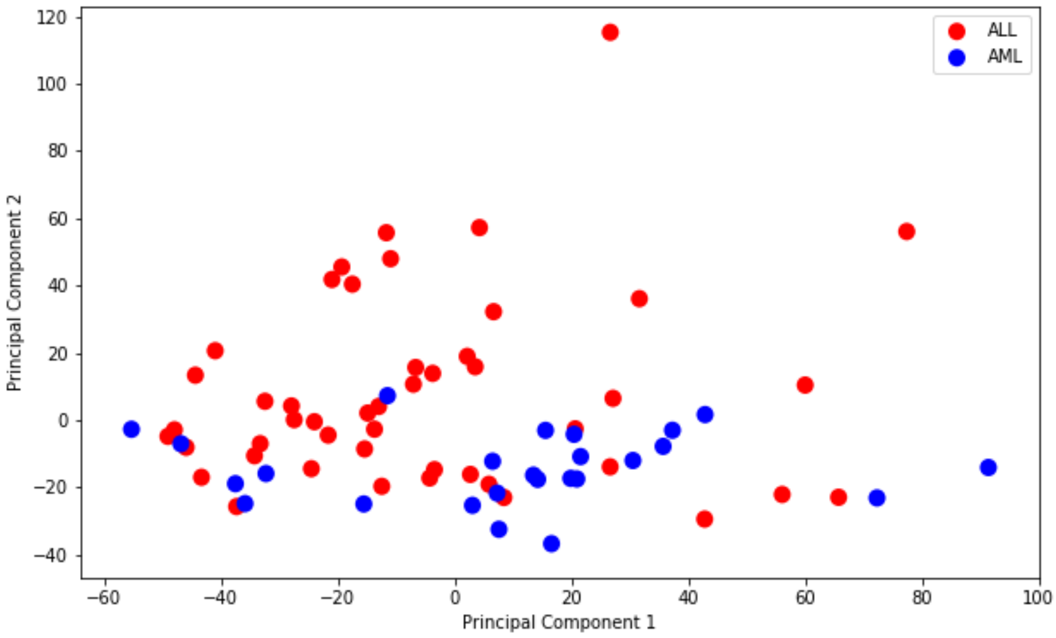

**Figure 2**: Scatter plot PC 1 vs PC 2 by cancer type. We observe some clusters for each type.

Next, the Elbow method and Silhouette method were implemented to find the optimal number of clusters to be used for analysis. Figure 3, which plots the results from the elbow method, shows no clear optimal level of cluster, but it appears that the elbow plot becomes flat after 7 cluster. With regards to the Silhouette method, there are 3 clear local optimal number of clusters at k =2, 5, and 7.

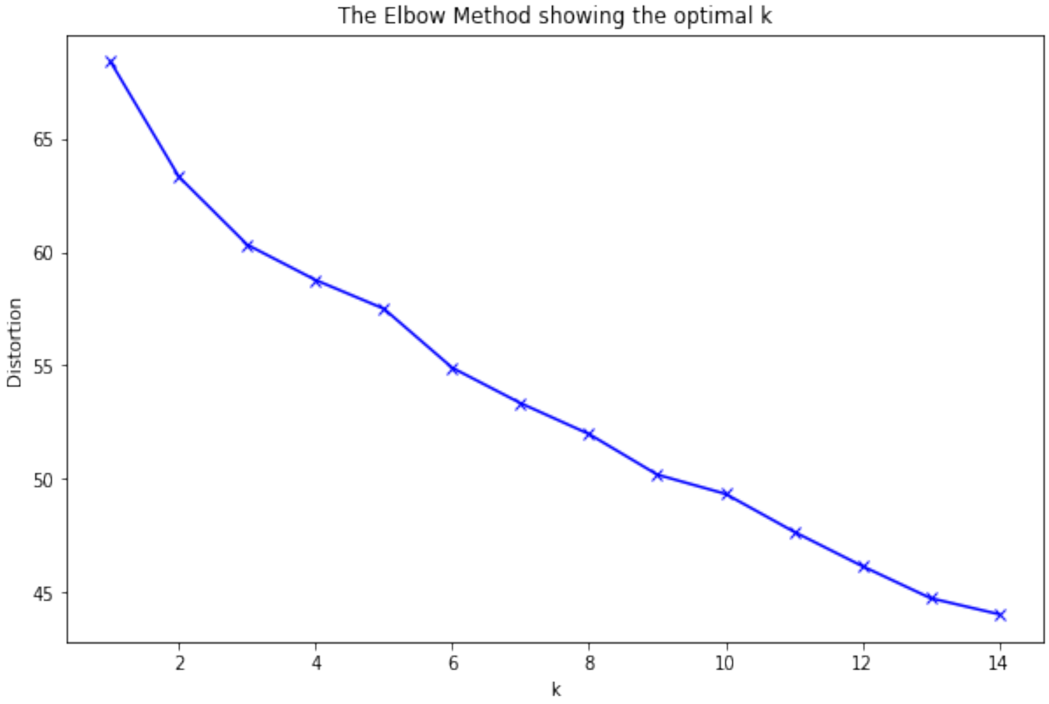

**Figure 3**: Elbow plot of k-clusters vs distortion to find optimal number of clusters.

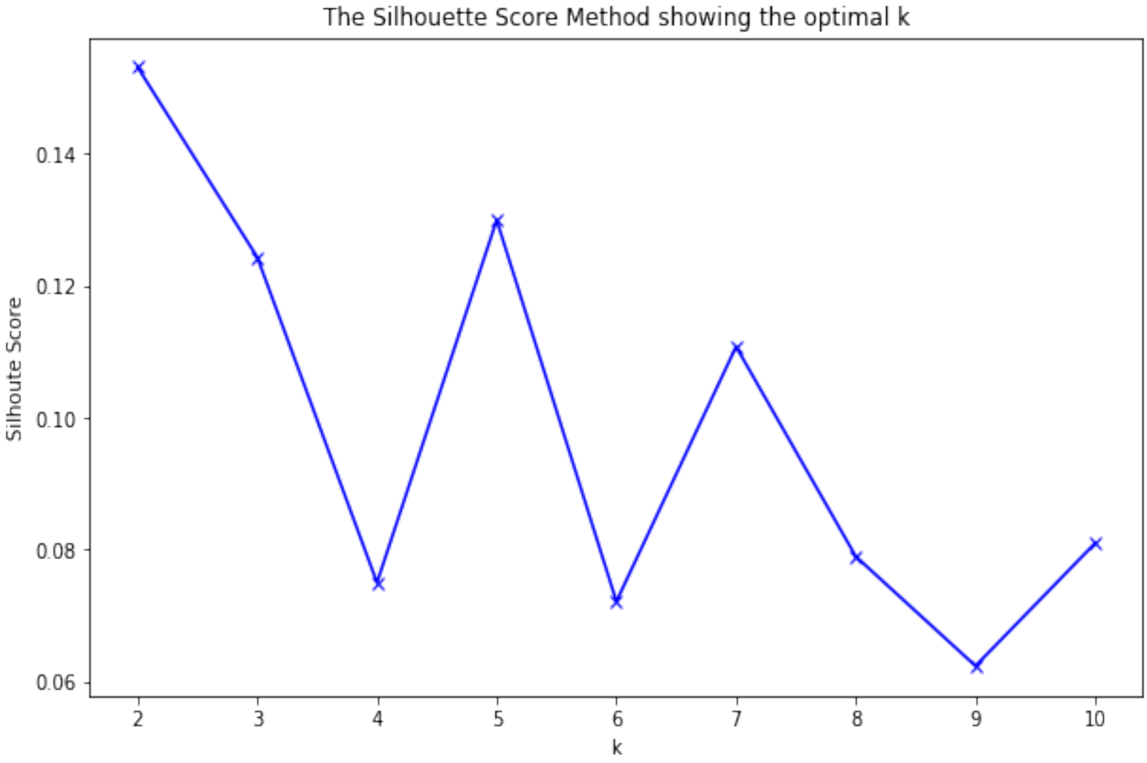

**Figure 4**: Silhouette plot of k-clusters vs Silhouette-score to find optimal number of clusters. 2, 5, and 7 clusters were found to be the local optimal.

Figure 5, 6 and 7 correspond to k-means clustering with the number of k = 2, 5, and 7 respectively. The left scatter plot is colour classified based on the clusters with PC 1 on the x-axis and PC 2 on the y-axis. The right bar plot represents the proportion of people that have AML in each cluster.

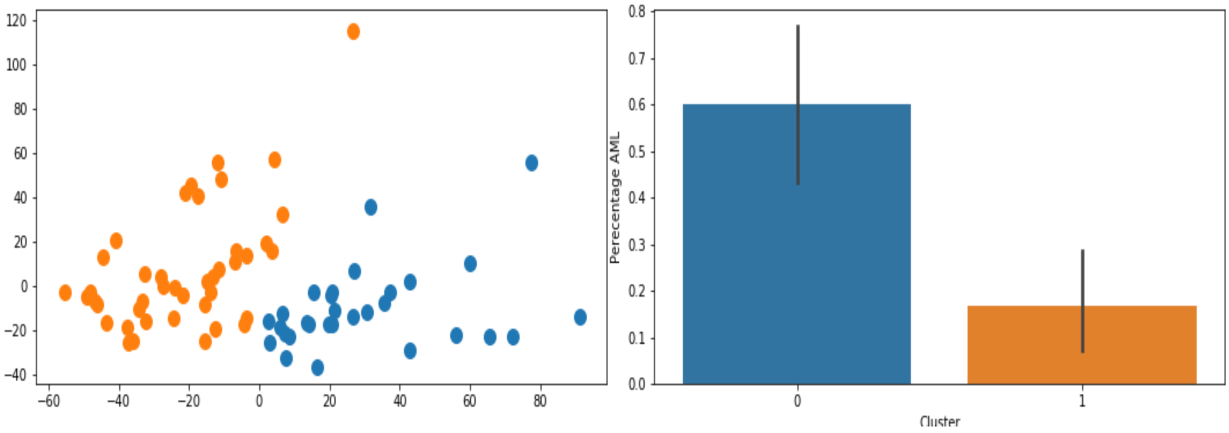

**Figure 5**: Cluster with K = 2. Left is the scatter plot of cluster. Right represents the proportion of people that have AML in each cluster.

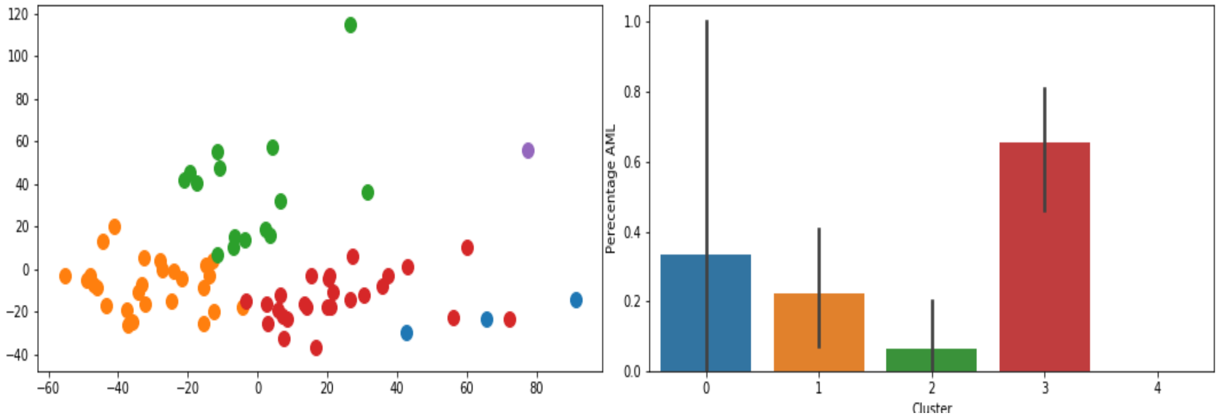

**Figure 6**: Cluster with K = 5. Left plot is the scatter plot of cluster. Right plot represents the proportion of people that have AML in each cluster.

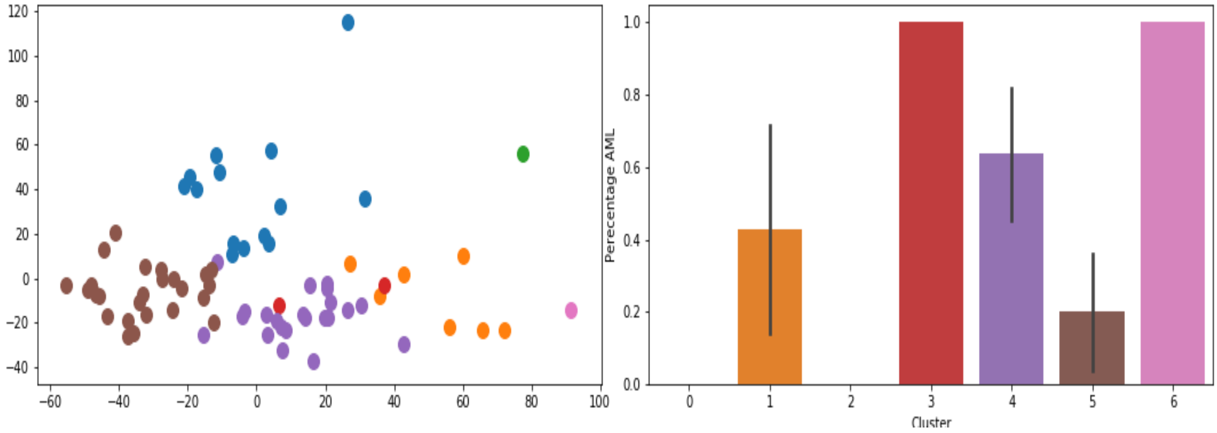

**Figure 7**: Cluster with K = 7. Left plot is the scatter plot of cluster. Right plot represents the proportion of people that have AML in each cluster.

In Figure 5, a clear difference was observed in the proportion of people that have AML in cluster 0 and the proportion of people that have AML in cluster 1. For the cluster 0 the proportion of the people that have AML is 59 % while in cluster 1 the proportion of people that have AML is 19%. It was confirmed that their difference is statistically significant by conducting a bootstrap t-test with number of sampling with replacement = 1000.

The most striking results from the clustering is presented in figure 7 where it was found that in cluster 3 and 6 the proportion of people that have AML is 100 % while in cluster 0 and 2 the proportion of people that have AML is 0 %.

## Conclusion

The sole purpose of this project was to confront the question: “is it possible to use PCA and k-means clustering to group individuals based on their genetic expression level that will provide information about which type of tumor they are likely to have". To some extent, our analyses were successful at answering that question by showing statistically significant differences in the proportion of people who have AML vs ALL in each cluster group. The results proved to be in line with the original study which proposed that there perhaps exist patterns in our genetic sequencing that make us prone to different types of tumor.

## Reference

[1]. T.R Golub *et al.*, Science 286, 531-537 (1999)
[2]. J. DeRisi *et al.*, Nature Genet. 14, 457 (1996)
[3]. D. J. Lockhart *et al.*, Nature Biotechnol. 14, 1675 (1996)
[4]. V. R. Iyer *et al.*, Science 283, 83 (1999)
[5]. L. Wodicka *et al.*, Nature Biotechnol. 15, 1359 (1997)
[6]. P. T. Spellman *et al.*, Mol. Biol. Cell 9, 3273 (1998)
[7]. M. Schena *et al.*, Proc. Natl. Acad. Sci. U.S.A. 93, 10614 (1996)
[8]. G. P. Yang *et al.*, Nucleic Acids Res. 27, 1517 (1999)
[9]. P. Tamayo *et al.*, Proc. Natl. Acad. Sci. U.S.A. 96, 2907 (1999)
[10]. L. Wodicka *et al.*, Nature Biotechnol. 15, 1359 (1997)