Unsupervised Analysis of High-Dimensional Gene Expression Data for the Classification and Visualization of Acute Leukemia Subtypes

Abstract

The advent of high-throughput sequencing technologies has led to gene expression datasets containing thousands of features, presenting a significant challenge for traditional analysis methods. The primary objective of this study was to explore the intrinsic structure and natural groupings within a high-dimensional acute leukemia gene expression dataset (ALL and AML subtypes) using unsupervised machine learning techniques. We applied two primary clustering methods, K-means and Hierarchical Clustering, to identify natural groupings in the data. Principal Component Analysis (PCA) was utilized for dimensionality reduction and data visualization. Both K-means (validated by the Elbow Method) and Hierarchical Clustering consistently identified two robust and distinct clusters. Furthermore, PCA demonstrated that the first principal component (PC1) alone captured over 90% of the total variance, successfully separating the identified clusters along a single dimension. These findings confirm that the underlying molecular profiles of the patient samples possess a strong, well-defined two-group structure, which is highly reflective of the known ALL and AML biological subtypes. This project validates the effectiveness of combining unsupervised learning and dimensionality reduction for uncovering clinically relevant patterns in complex genomic data.

1. Introduction

Acute Leukemia is a group of aggressive hematological malignancies categorized primarily into Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). Accurate and timely classification of these subtypes is critical, as treatment protocols and prognoses differ significantly. Genomic technologies generate massive gene expression profiles, providing a molecular fingerprint for each patient. However, the resulting datasets are characterized by high-dimensionality (thousands of genes) and significant noise. This study addresses the challenge of identifying the core biological structure in this high-dimensional space without prior knowledge of the true labels, by employing unsupervised machine learning methods.

Unsupervised Learning is an essential tool in genomics, as it allows researchers to uncover hidden patterns and intrinsic groupings in data. The core assumption is that the gene expression differences driving the ALL and AML classifications are substantial enough to form two distinct, non-overlapping clusters discoverable by these methods.

- **K-means Clustering:** A partition-based algorithm that aims to partition observations into clusters by minimizing the distance to the cluster centroid.
- **Hierarchical Clustering:** A connectivity-based method that builds a hierarchy of clusters, revealing nested structures without requiring a predetermined target variable.
- **Principal Component Analysis (PCA):** A linear transformation technique that identifies the directions (principal components) that maximize the variance in the data, thereby reducing dimensionality.

2. Methodology

2.1 Data Preprocessing

The analysis utilized the gene expression data for ALL and AML patients (data_set_ALL_AML_train.csv). The data was processed to ensure suitability for distance-based algorithms:

1. **Data Cleaning:** Rows with missing values were removed, and non-numeric columns were excluded.
2. **Scaling:** All gene expression values were standardized (Z-scored) using the scale() function. This step is critical for clustering and PCA, ensuring highly expressed genes do not artificially dominate the distance calculations.

2.2 Clustering Analysis

- **K-means Clustering:** The Euclidean distance metric was used. The optimal number of clusters () was determined via the Elbow Method (WCSS minimization).
- **Hierarchical Clustering:** Performed using the Ward's method (minimizing within-cluster variance) on a sampled subset of the data. The output was visualized using a Dendrogram.

2.3 Dimensionality Reduction

Principal Component Analysis (PCA) was performed on the standardized data. The explained variance was assessed using a Scree Plot to identify the most significant principal components.

3. Results and Discussion

3.1 Clustering Consistency

Both clustering methodologies provided strong evidence for a two-cluster structure in the data. The Elbow plot (Figure 1) clearly suggested two optimal clusters, where the rate of decrease in WCSS (the total within-cluster variance) slowed down significantly. The k-means clustering plot (Figure 2) showed a distinct grouping of data points into two subgroups.

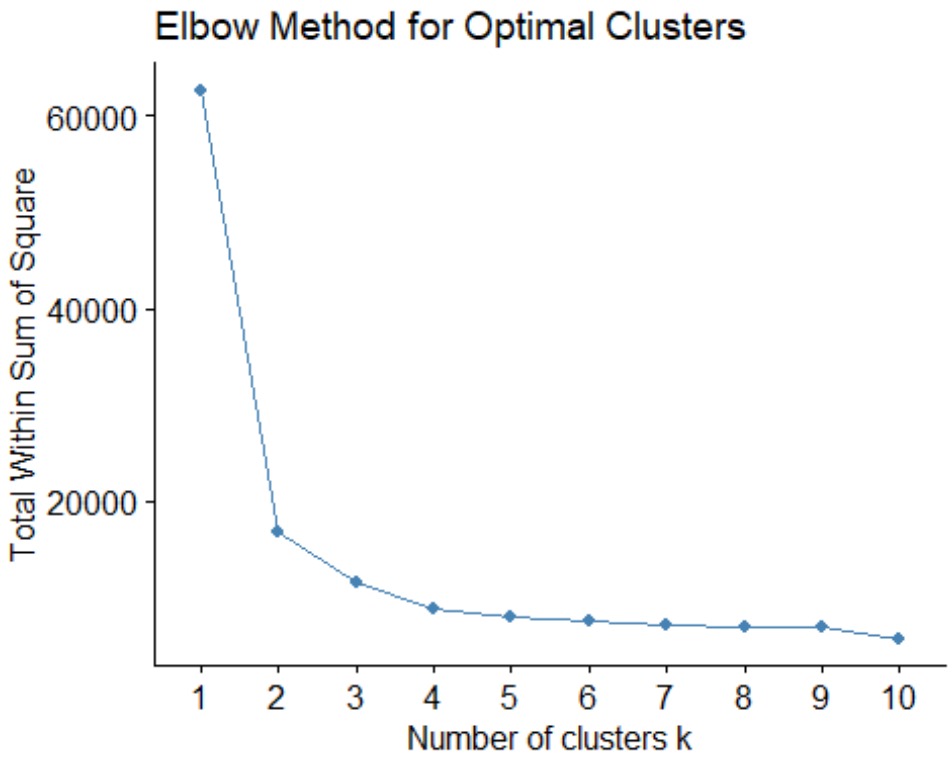

Figure 1: Elbow Method for Optimal Clusters. Shows the distinct bend at k=2, indicating the optimal number of clusters.

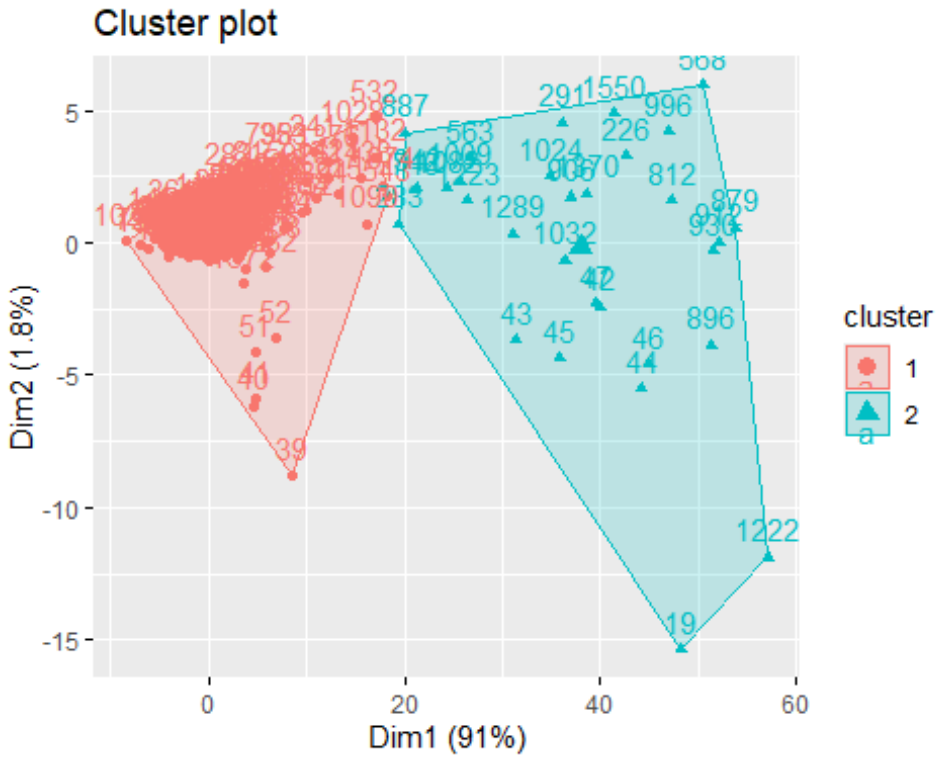

Figure 2: K-means Cluster Visualization. A scatter plot showing the data points partitioned into two groups.

The complete linkage dendrogram (Figure 3) exhibited a clear, large vertical jump in linkage distance (height) when moving from two clusters to one cluster, strongly supporting the natural separation into two primary groups.

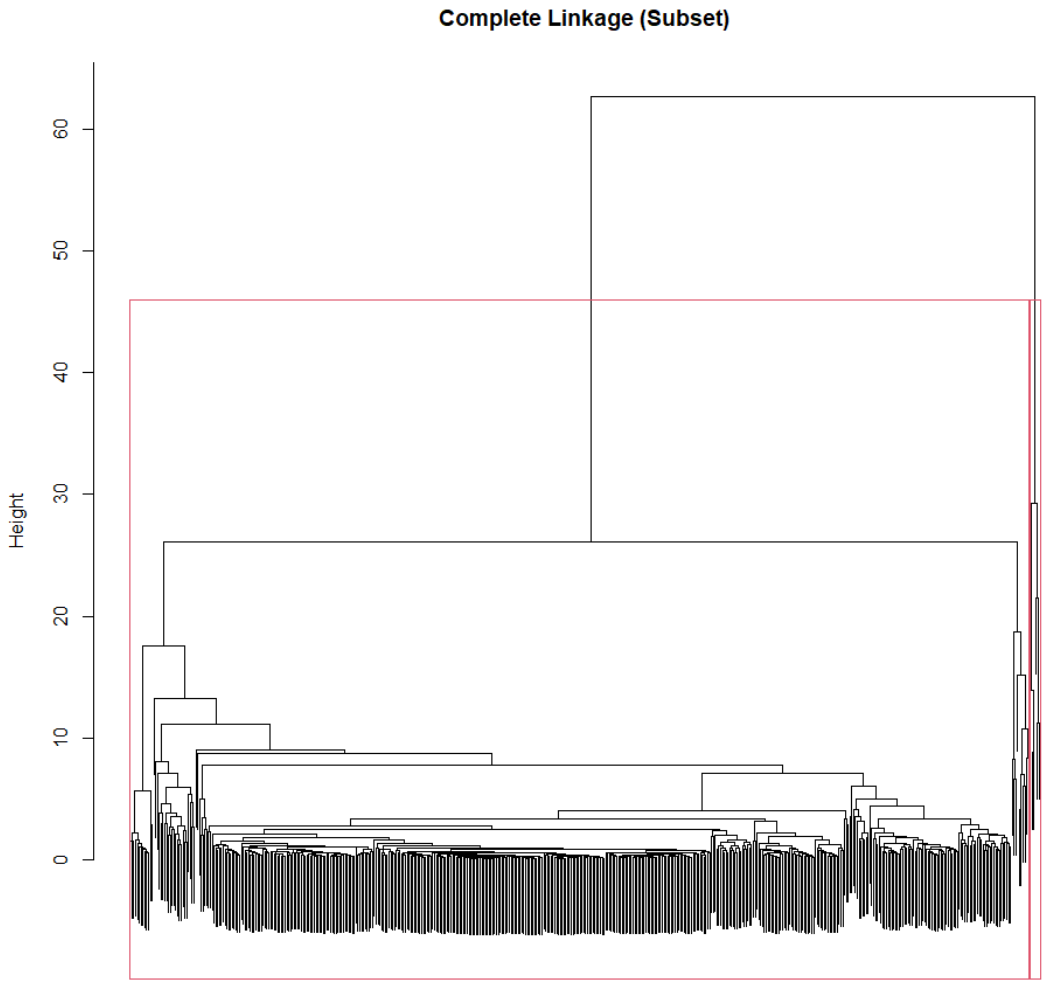

Figure 3: Complete Linkage Dendrogram. Illustrates the hierarchical structure with a clear division into two main clusters.

3.2 Dimensionality Reduction by PCA

PCA revealed that a single underlying factor largely dictates the dataset's high-dimensional structure. The first principal component (PC1) explained more than 90% of the total variance (Figure 4 and 5).

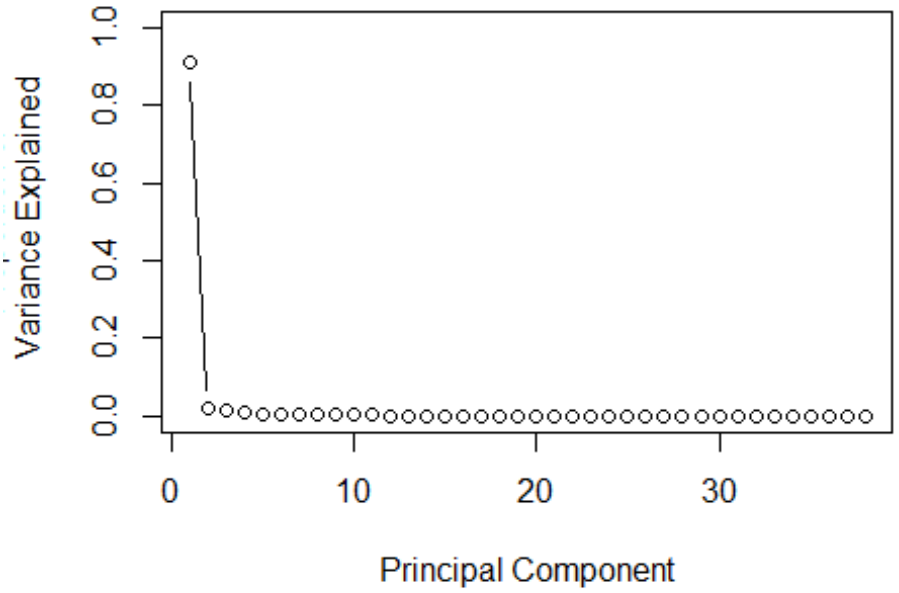

Figure 4: Proportion of Variance Explained by Principal Components. Highlights the significant variance captured by PC1.

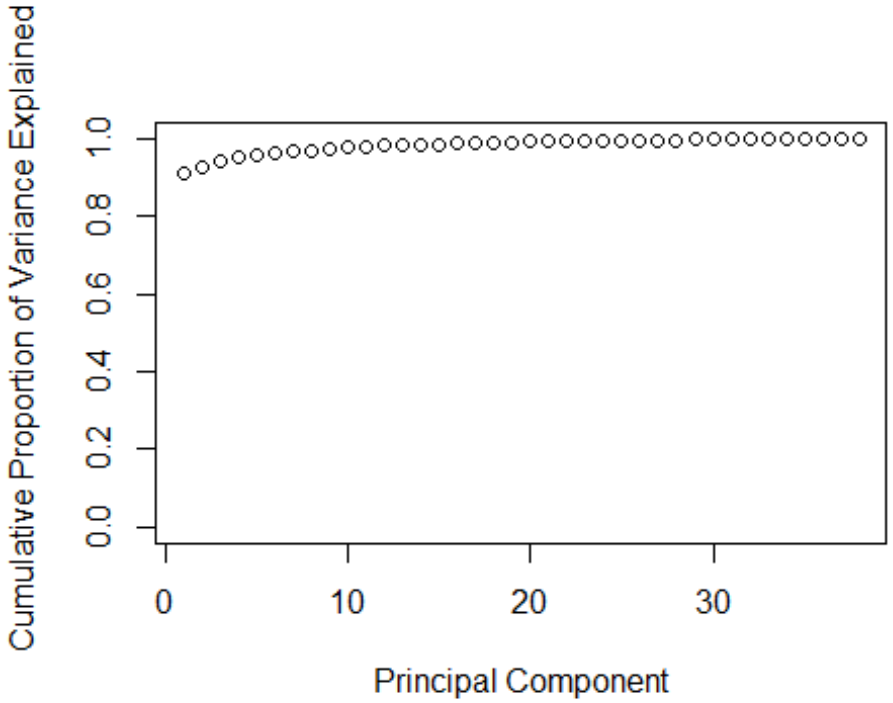

Figure 5: Cumulative Proportion of Variance Explained. ***Confirms that cumulative variance quickly plateaus after PC1.***

3.3 Visualization of Clustering on PCA

The PCA scatter plot demonstrates that the two K-means clusters are perfectly separated along the PC1 axis.

Figure 6: PCA Scatter Plot Colored by K-Means Clusters. Visually confirms the separation of the two patient groups along the primary variance dimension (PC1).

Appendices

```
### Preprocessing
# Load libraries (omitted for brevity)
# Load data
train_data_raw <- read.csv("data_set_ALL_AML_train.csv")
# Handle missing values and remove non-numeric columns
train_data_clean <- train_data_raw %>% drop_na()
train_data_numeric <- train_data_clean %>% select(where(is.numeric))
# Scale the data
scaled_data <- scale(train_data_numeric)

### Clustering
# a. K-means Clustering + Elbow Method
# CODE FOR FIGURE 1: Elbow Method
fviz_nbclust(scaled_data, kmeans, method = "wss") +
  labs(title = "Elbow Method for Optimal Clusters")
# Apply k-means
set.seed(123)
kmeans_result <- kmeans(scaled_data, centers = 2, nstart = 25)

# CODE FOR FIGURE 2: K-means Cluster Visualization
fviz_cluster(kmeans_result, data=scaled_data)
# b. Hierarchical Clustering
set.seed(123)
# Sample data for visualization (due to dataset size)
sample_indices <- sample(1:nrow(scaled_data), 500)
subset_data <- scaled_data[sample_indices, ]
# Compute distance matrix and perform clustering (Complete Linkage)
dist_subset <- dist(subset_data, method = "euclidean")
hc.complete.sub <- hclust(dist_subset, method = "complete")
```

```
# CODE FOR FIGURE 3: Dendrogram Plot
par(mfrow = c(1, 1))
plot(hc.complete.sub, main = "Complete Linkage (Subset)", xlab = "", sub = "", labels = FALSE, cex = 0.6)
rect.hclust(hc.complete.sub, k = 2)

### Dimensionality Reduction (PCA)
# Perform PCA
pr.out=prcomp(scaled_data, scale=TRUE)
pr.var=pr.out$sdev^2
pve=pr.var/sum(pr.var)

# CODE FOR FIGURE 4: Proportion of Variance Explained Plot
plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,1),type='b')

# CODE FOR FIGURE 5: Cumulative Proportion of Variance Explained Plot
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", ylim=c(0,1), type='b')

# CODE FOR FIGURE 6: PCA Scatter Plot with K-Means Clusters
# Get first 2 principal components
pc_data <- data.frame(pr.out$x[, 1:2])

# Perform K-means clustering with 2 clusters on PCs
km.out <- kmeans(pc_data, centers = 2)
pc_data$Cluster <- as.factor(km.out$cluster)

# Plot the PCA scatter plot with clusters
plot(pc_data$PC1, pc_data$PC2,
     col = pc_data$Cluster,
```

```
pch = 20, cex = 2,
xlab = "PC1", ylab = "PC2",
main = "PCA Scatter Plot Colored by K-Means Clusters")
```