A Comparative Analysis of Clustering Gene Expression Data for Cancer Subtypes

Leila Michal (lmichal)

Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University

Abstract

- **Aim:** Evaluate clustering methods for identifying cancer subtypes from gene expression profiles using dimensionality reduction.
- **Methods:** Applied PCA, t-SNE, and UMAP followed by K-means, Hierarchical, and Leiden clustering on high-dimensional RNA-Seq data.
- **Results:** K-means and Hierarchical on t-SNE and UMAP achieved top ARI, Jaccard, and Silhouette scores, outperforming Leiden. PCA with Leiden gave the best performance among linear methods.
- **Further Directions:** Expand to multi-omics analysis and assess clustering reproducibility and clinical relevance.

Introduction

Focus: identifying biologically meaningful cancer subtypes by clustering gene expression profiles across multiple tumor types.

- Gene expression profiling enables unsupervised discovery of subtypes with potential clinical relevance [1].
- Clustering helps reveal hidden patterns not apparent through traditional classification methods.

Goal: evaluate the effectiveness of unsupervised clustering methods following dimensionality reduction.

- Compare K-means [2], hierarchical clustering [1], and Leiden algorithms on PCA, t-SNE, and UMAP embeddings [3].
- Assess cluster quality using silhouette score, Adjusted Rand Index (ARI), and Jaccard Index.

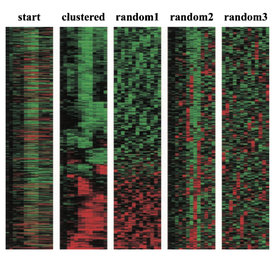

Figure 1:Figure 3 from Eisen et al. (1998) [1] showing the power of clustering. The "clustered" heatmap arranges genes with similar expression patterns together. Randomly ordered matrices (random1–3) destroy this structure, highlighting that patterns seen in gene expression data reflect genuine biological processes rather than noise.

Data Description

Dataset: obtained RNA-Seq (HiSeq) PANCAN dataset via UC Irvine Machine Learning Repository.

- **Samples: 801** patients diagnosed with one of five cancer types: **BRCA, KIRC, COAD, LUAD, and PRAD**
- **Features/Predictor Variables: 20,531** RNA-Seq gene expression levels per sample
  - Numerical features measured using the Illumina HiSeq platform
  - Each sample is stored as a row with gene expression as columns
- **Preprocessing:** Filtered out low-variance genes, retaining **20,264** features for analysis
- **Task Type:** Suitable for unsupervised clustering and classification

Dimensionality Reduction

- **PCA** retained 50 components; first 10 accounted for **45.9%** cumulative variance.
- **t-SNE** and **UMAP** embedded **801** patients into 2D while preserving neighborhood structure.
- Visualizations suggest **separable clusters** but depend on method and **perplexity/resolution**.

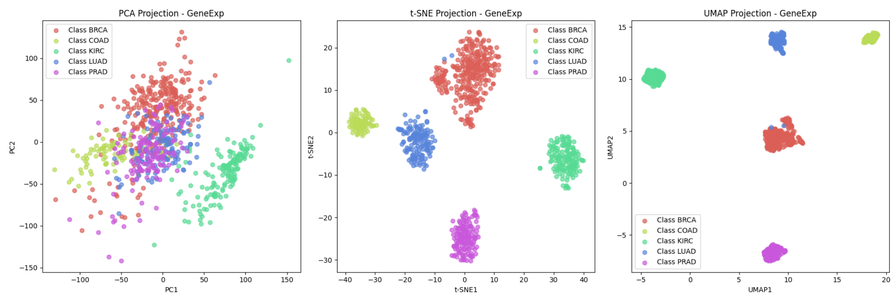

Figure 2:Comparison of PCA (left), t-SNE (middle), and UMAP (right) projections of the gene expression dataset. UMAP and t-SNE revealed more visually distinct clusters compared to PCA.

Clustering Methods

- **K-means:** Centroid-based algorithm that partitions data by minimizing intra-cluster variance.
- **Hierarchical:** Recursively merges data points based on distance; visualized via dendrogram.
- **Leiden:** Graph-partitioning method that detects communities by optimizing modularity over a nearest-neighbor graph.

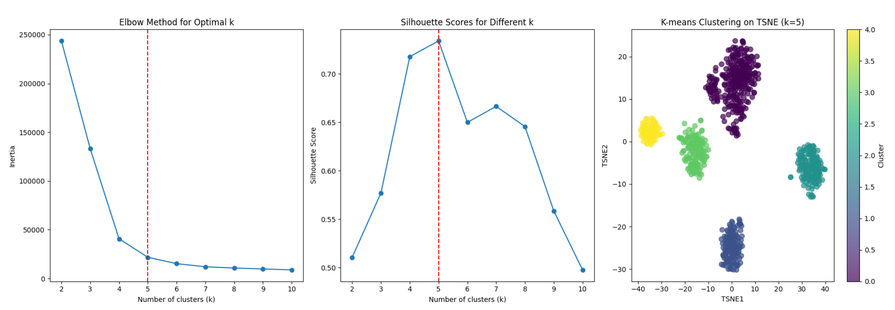

Figure 3:K-means clustering on t-SNE-reduced gene expression data. **Left:** Elbow method shows diminishing returns in inertia after **k=5**. **Middle:** Silhouette analysis confirms **k=5** as the optimal choice with highest average cohesion and separation. **Right:** Final clustering on 2D t-SNE space reveals five compact, well-separated clusters.

Results

Evaluation Metrics: The clustering performance was evaluated across all combinations of dimensionality reduction and clustering methods. Each method was assessed using internal (Silhouette) and external (Adjusted Rand Index and Jaccard Index) metrics. The table below summarizes the performance for each configuration.

Dim Reduction	Method	Clusters	ARI	Jaccard	Silhouette
PCA	K-means	7	0.7979	0.7241	0.2462
PCA	Hierarchical	7	0.8258	0.7587	0.2457
PCA	Leiden	7	0.8366	0.7720	0.2432
t-SNE	**K-means**	**5**	**0.9925**	**0.9888**	0.7341
t-SNE	**Hierarchical**	**5**	**0.9925**	**0.9888**	0.7341
t-SNE	Leiden	7	0.7635	0.6807	0.6598
UMAP	**K-means**	**5**	**0.9925**	**0.9888**	**0.8891**
UMAP	**Hierarchical**	**5**	**0.9925**	**0.9888**	**0.8891**
UMAP	Leiden	7	0.7635	0.6807	0.7667

Table 1:Clustering performance metrics.

UMAP combined with K-means or Hierarchical clustering yielded the best silhouette scores (0.8891), indicating strong intra-cluster similarity and clear separation. These combinations also achieved the highest Adjusted Rand Index (0.9925) and Jaccard scores (0.9888), outperforming Leiden clustering across both nonlinear and linear projections. Although PCA with Leiden produced the highest ARI among PCA-based methods (0.8366), it was outperformed by both K-means and Hierarchical on t-SNE and UMAP.

Implementation Details

- Gene expression data was log-transformed and standardized. Low-variance genes **(variance < 0.1)** were removed.
- **PCA (up to 50 components)**, **t-SNE (perplexity=30)**, and **UMAP (n_neighbors=15, min_dist=0.1)** were applied using `scikit-learn` and `umap-learn`.
- K-means and agglomerative clustering used Euclidean distances (`sklearn.cluster`); Leiden clustering used `leidenalg` + `igraph` on UMAP-based graphs.
- Optimal clusters were selected via elbow and silhouette analysis. Results evaluated using `silhouette_score`, `adjusted_rand_score`, and `jaccard_score`.

Discussion

Goal Evaluation:

- Dimensionality reduction enabled meaningful separation of cancer subtypes, supporting its integration in unsupervised workflows.
- Nonlinear methods (t-SNE, UMAP) enhanced cluster visibility and interpretability over PCA.

Model Metrics:

- K-means and Hierarchical clustering on t-SNE and UMAP outperformed Leiden in ARI, Jaccard, and Silhouette scores.
- **Leiden** on **PCA** achieved moderate **ARI (0.8366)** but was surpassed by nonlinear projections paired with distance-based methods.

Limitations:

- Leiden's performance varied by embedding, reflecting sensitivity to resolution and graph structure.
- External metrics relied on known subtype labels; methods were not evaluated for discovery of novel subtypes.

Future Direction:

- Evaluate clustering reproducibility and parameter sensitivity across random seeds and bootstrapped datasets.
- Expand framework to include integrative clustering with other clinical outcomes.

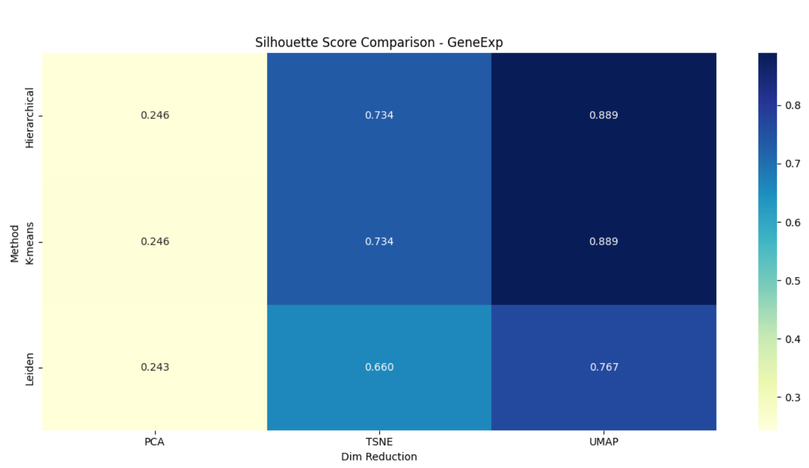

Figure 4:Silhouette heatmap showing cluster compactness across method and dimensionality reduction pairings. UMAP-based approaches exhibit the strongest cohesion, particularly with K-means and Hierarchical.

References

[1] M. B. Eisen et al.
Cluster analysis and display of genome-wide expression patterns.
Proceedings of the National Academy of Sciences, 95(25):14863–14868, 1998.

[2] J. MacQueen.
Some methods for classification and analysis of multivariate observations.
In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[3] L. van der Maaten and G. Hinton.
Visualizing data using t-sne.
Journal of Machine Learning Research, 9:2579–2605, 2008.