Finding the ROC Cell in Xenopus Laevis Tadpoles

Abstract:

The purpose of this report is to investigate tail regeneration in xenopus laevis tadpoles and to identify the regenerative organizing cell (ROC) that facilitates this process. Using single-cell RNA sequencing (scRNA-seq) data, a variety of unsupervised machine learning techniques were used to visualize the data and locate the ROC, including dimensionality reduction via UMAP and t-SNE and clustering via K-Means and PCA+Leiden. The clustering algorithms were evaluated against a variety of metrics, including rand-index, mutual information, and the silhouette score, with the ultimate goal being to identify the genes that were differentially expressed during regeneration. Of the evaluated clustering algorithms, PCA + Leiden achieved the holistic best performance as evaluated by the aforementioned metrics and a qualitative examination of performance based on UMAP and t-SNE plots. PCA + Leiden achieved a rand-index of 0 9182 and an adjusted rand-index of 0.5395 when evaluated against the ROCs identified in Aztekin et al.'s 2019 paper.

Introduction:

Tissue regeneration is an intricate biological process in which various cell types coordinate their efforts to replace lost or damaged structures. Xenopus laevis has proven to be an invaluable model organism for studying regeneration, particularly because of its ability to regenerate tissues like the spinal cord, tail, and limbs [Borodinsky, 2017]. Understanding the cellular and molecular underpinnings of regeneration is vital for advancing regenerative medicine and tissue engineering.

ScRNA-seq data was used to investigate the cellular composition of regenerating Xenopus laevis tails at different stages post-amputation. The aim of the analysis was to distinguish cell types involved in regeneration and explore the expression of genes specifically associated with the regenerative process.

Methods:

Analysis was based on a publicly available scRNA-seq dataset of Xenopus laevis regenerating tail tissue. To accurately identify cell populations and study their transcriptional profiles, the following preprocessing and analytical steps were taken:

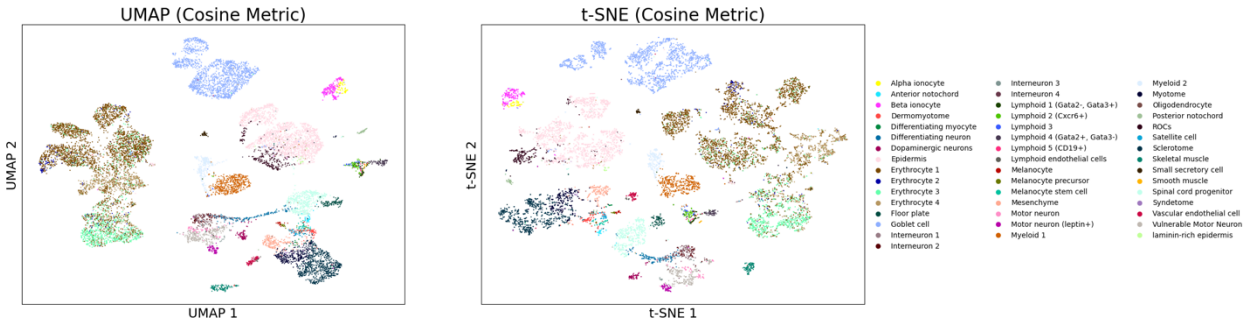
Preprocessing:

- Cleaned scRNA-seq data stored was loaded and log-normalized with scanpy.
- Inspired by Aztekin et al.'s paper, highly variable genes were selected based on a dispersion threshold of 0.65 and a mean expression quantile between 0.05 and 0.8.

Dimensionality Reduction:

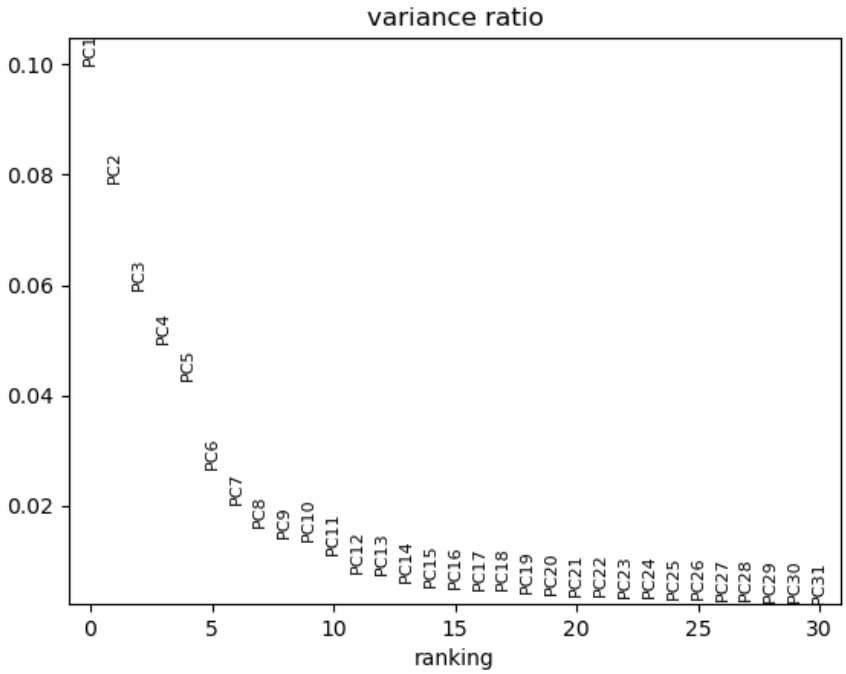
- UMAP and t-SNE methods were used to visualize the data in two dimensions. Several distance metrics — including Euclidean, Manhattan, and cosine distances — were employed to gain more understanding of the data's structure. The cosine metric was preferred for later analysis for its performance in separating clusters in the exploratory UMAP plots.

Figure 1: UMAP and t-SNE for Cosine Metric



- PCA was also performed on the highly variable genes, which showed solid explanatory power from the first 11 principal components. This PCA data was used explicitly by Leiden and Louvain clustering methods.

Figure 2: Explained variance ratio for PCA



Clustering:

- Several clustering algorithms were applied: K-Means, Leiden, Louvain, and topological clustering using a random walk algorithm.
- The topological clustering algorithm uses the high-dimensional UMAP graph as a basis for similarity. It computes clusters from this graph by initiating a random walk along the graph. Points that are visited frequently in the same path are presumed to be from the same cluster.

Clustering Performance Metrics:

- Silhouette score, adjusted rand index, rand index, mutual information, and adjusted mutual information were computed for the various clustering methods.

Gene Expression Analysis:

- To identify marker genes for each cluster, differential gene expression analysis was conducted using both t-test and logistic regression methods. The t-test results were adjusted for multiple testing using Bonferroni correction.
- The identified marker genes were cross-referenced with the cluster identities outlined in Aztekin et al.'s paper.

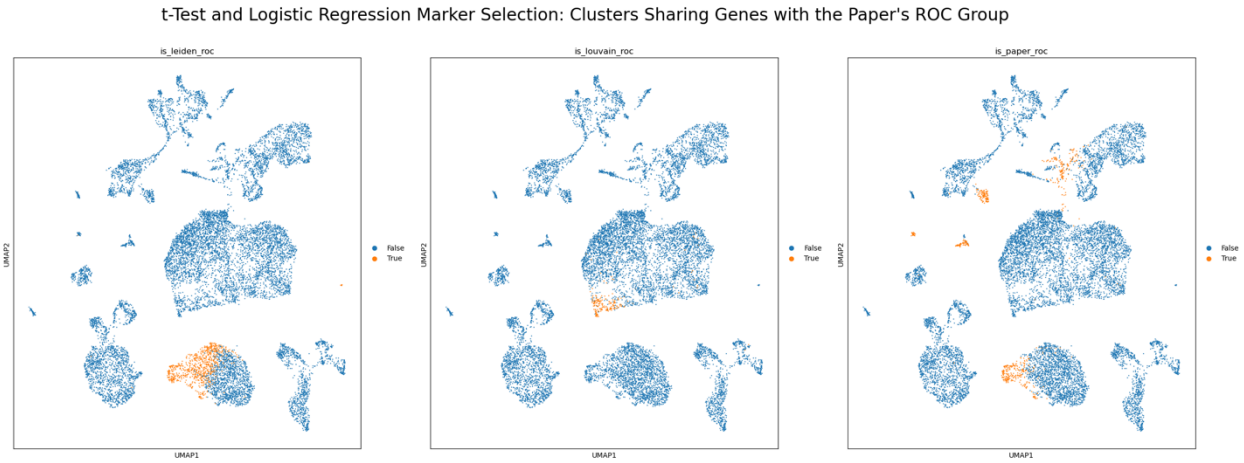
Code Availability:

The complete pipeline for data preprocessing, clustering, and gene expression analysis is available at https://github.com/wvietor/stat4243_project1.

Results:

Metric	K-Means	Topological	Leiden	Louvain
Adjusted Rand Index	0.4952	0.3625	0.5395	0.5422
Rand Index	0.8725	0.7707	0.9182	0.923
Silhouette Score	0.0998	0.1364	0.1284	0.1156
Mutual Information	1.9148	1.4099	1.9875	2.0127
Adjusted Mutual Info	1.9148	1.4099	1.9875	2.0127

From the above analysis and visual inspection of the plots available in the Jupyter notebook above, Leiden clustering was selected as the overall optimal clustering metric. Leiden cluster 12 (shown below via UMAP) showed an identity matching the ROCs discussed in Aztekin et al.'s paper.



Conclusion:

Thus, the analysis techniques described above were able to generally reproduce the results described in Aztekin et al.'s paper. Across several clustering metrics and upon visual

inspection, leiden clustering most closely recreated the clustering which revealed the ROC cells. In addition to using popular clustering algorithms like K-Means, less familiar ideas (namely the random walk clustering on the UMAP graph) were also explored and generated solid results. One way to improve the analysis would be to more intelligently choose the values of the various parameters. Due to compute limitations and the variety of metrics considered, it was difficult to arrive at globally optimal values, though attempts were made to choose intelligent values. Lastly, the analysis largely hinges on the results of Aztekin et al.'s paper. The results are still fascinating, but removing this explicit dependency would allow for more convincing "first principles" results instead of mere confirmation.

Works Cited

Borodinsky L. N. (2017). Xenopus laevis as a Model Organism for the Study of Spinal Cord Formation, Development, Function and Regeneration. *Frontiers in neural circuits,* 11, 90. https://doi.org/10.3389/fncir.2017.00090.

C. Aztekin et al. (2019). Identification of a regeneration-organizing cell in the Xenopus tail. *Science*, 364, 653-658. https://doi.org/10.1126/science.aav9996.