Classification of Cell Types through Topological Data Analysis of Gene Co-expression Distance

Cecilia Liu, Jason Chen, Zixi Yuan

06/09/2024

Abstract

The purpose of this project is to use topological data analysis (TDA) to perform hierarchical clustering and to identify relationships between cell types based on single-cell RNA sequencing (scRNA-seq) data from the Human Brain Cell Atlas. We focused on non-neural cells, employing various bioinformatics techniques to preprocess and analyze the data. Initially, we normalized and log-transformed the data, then used Pearson correlation and Euclidean distance to construct gene co-expression distance. Due to the high dimensionality of the data, we applied dimensionality reduction methods, including PCA and UMAP, to facilitate the Euclidean distance analysis. Persistent diagrams were generated from the transformed distance matrices, and the distances between cell types were calculated using Bottleneck distance and persistent images. Finally, hierarchical clustering was performed using single linkage, average linkage, and complete linkage methods to reveal the relationships between different cell types. Our findings provide a comprehensive understanding of the gene expression landscape and hierarchical structure of non-neural cells in the human brain, highlighting the intricate relationships between various cell types.

1 Introduction

The Human Brain Cell Atlas [1] provides an extensive resource for understanding the cellular composition and gene expression profiles of cells in the whole human brain. This atlas includes data from both neural and non-neural cells, offering insights into the complex cellular landscape of the brain. While much research has focused on neural cells due to their direct involvement in brain function and disease, non-neural cells also play crucial roles in maintaining brain homeostasis and supporting neural functions.

This project aims to explore the gene expression patterns and relationships between non-neural cell types using topological data analysis (TDA). TDA is a powerful tool for studying the shape and structure of data, capturing topological features that traditional methods might overlook. By leveraging TDA, we can gain a deeper understanding of the underlying biological structures and interactions within the non-neural cell population.

The primary objectives of this project are to preprocess the scRNA-seq data from the Human Brain Cell Atlas, construct gene co-expression networks, and apply TDA to identify topological features. We then aim to perform hierarchical clustering to uncover the relationships between different cell types. This approach allows us to visualize and quantify the similarities and differences between cell types, providing a comprehensive view of the non-neural cell landscape.

Specifically, we focus on:

1. Normalizing and log-transforming the data to prepare it for analysis.
2. Constructing gene co-expression distance matrix using Pearson correlation and Euclidean distance.
3. Generating persistent diagrams from the distance matrices to capture topological features.
4. Calculating distances between cell types using Bottleneck distance and persistent images.
5. Performing hierarchical clustering using single linkage, average linkage, and complete linkage methods to identify relationships between cell types.

By integrating these methods, we aim to provide a detailed analysis of non-neural cell types and their relationships, contributing valuable insights to the field of biology.

2 Methods

2.1 Data Collection

Data for this project was obtained from the Human Brain Cell Atlas [1], which provides a comprehensive dataset of brain cell types and their gene expression profiles. The initial dataset contained a mix of neural and non-neural cells. For this study, we focused on non-neural cells to reduce the sample size and focus on specific cellular populations.

2.2 Data Pre-processing

To prepare the data for analysis, we performed the following steps:

1. Select the data that excludes neural cell types, thereby retaining only non-neural cells.
2. Loading the data into ScanPy [2], which is a popular Python library for single-cell analysis.
3. Scaling the gene expression data (to counts per million, CPM) to account for differences in sequencing depth and other technical variations.
4. Log-transforming the data to stabilize variance and make it more suitable for downstream analysis.
5. Splitting the data based on cell types.
6. Selecting genes that are expressed in greater than 75% of the cells to focus on robust and consistent gene expression patterns.

2.3 Gene Co-expression Distance Construction

Using the pre-processed data, we constructed gene co-expression distance to evaluate relationships between genes. We adopt the gene co-expression network approach widely used in bioinformatics. Instead of constructing a network by setting a threshold, we use the gene co-expression distance as a point cloud data for further analysis. Two methods were employed for this construction:

1. **Pearson Correlation:** This method measures the linear correlation between the expression levels of genes across different cells. The resulting correlation matrix was transformed into a distance matrix, where each gene was treated as a vertex, and distances were calculated based on the formula $distance = \sqrt{2 \times (1 - correlation)}$.

2. **Euclidean Distance:** To address the high dimensionality of the data, we applied dimensionality reduction techniques before calculating Euclidean distances.

   - **Principal Component Analysis (PCA):** Initially, we used PCA to reduce the dimensionality, capturing 99% of the variance. However, due to the non-linear relationships between genes, PCA was found to be suboptimal.
   - **Uniform Manifold Approximation and Projection (UMAP):** We then applied UMAP, a non-linear dimensionality reduction technique, which effectively captured the complex relationships between genes. After UMAP, Euclidean distances between genes were calculated and used to create persistent diagrams. The parameters for UMAP were set to n_neighbors = 15, n_components = 30, and metric = 'Euclidean'. When choosing appropriate n_neighbors, we made sure it is smaller than the number of principal components after performing PCA.

2.4 Topological Data Analysis

Topological data analysis was applied to the distance matrices derived from both Pearson correlation and Euclidean distance methods. This involved the following steps:

1. Constructing a distance matrix where each gene was treated as a vertex.

2. Performing Vietoris-Rips filtration to build a simplicial complex from the data by connecting points (genes) within a certain distance threshold.

3. Generating persistent diagrams that plot the birth and death of topological features as a function of the distance threshold.

2.5 Comparing Cell Types via Persistent Diagrams

To compare the topological features between different cell types, we calculated distances between corresponding persistent diagrams using two methods:

1. **Bottleneck Distance:**

This is a traditional metric used to compare the distance between two persistent diagrams. It is calculated by quantifying the smallest distance to match all points in one persistence diagram with points in another diagram, including the diagonal. For each pair of persistent diagrams, we compute the bottleneck distance between 0-dimension persistent homology, 1-dimension persistent homology, and a combined distance by adding normalized 0-dim distance and normalized 1-dim distance.

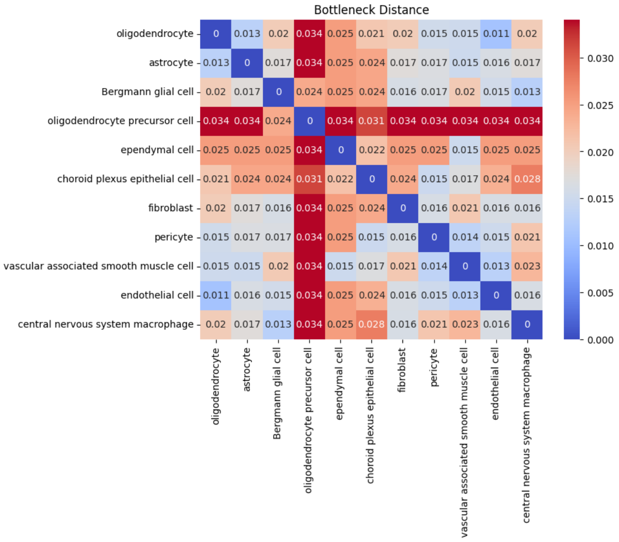


Figure 1: 1-dimensional bottleneck distance

2. **Persistent Images and p-Norms:**

   To calculate this distance,

   (a) Persistent diagrams (death v.s. birth) were rotated 45 degrees (persistence v.s. birth),

   (b) converted into persistence surface by applying a kernel (standard Gaussian kernel) to each point,

      i. adjust the weight parameters to remove points with low persistence might be crucial

   (c) and then discretized into persistent images with a fixed resolution (0.01).

   After getting persistent images for each cell type, we need to resize the images into the same shape. Two approaches to resizing the images are

   (a) zero-padding to the maximum width and maximum height

   (b) resizing to a pre-defined shape ((10, 3))

   The distances between these resized persistent images were then calculated using p-norms (2-norm).

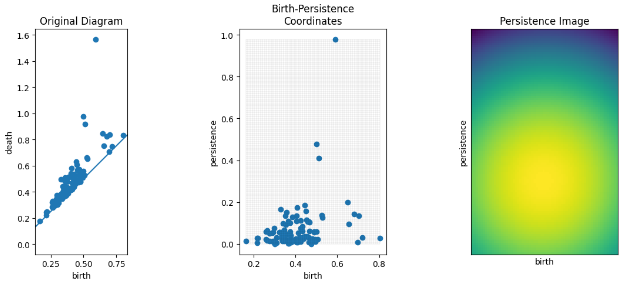


Figure 2: Persistance Image with Ependymal Cells

2.6 Hierarchical Clustering

The distances between cell types, derived from bottleneck distance and persistent images, were used to perform hierarchical clustering. Three linkage methods were employed:

1. **Single Linkage:** Clusters were merged based on the shortest distance between points in different clusters.

2. **Average Linkage:** Clusters were merged based on the average distance between all pairs of points in the different clusters.

3. **Complete Linkage:** Clusters were merged based on the maximum distance between points in different clusters.

The hierarchical clustering results were analyzed to uncover the relationships between different non-neural cell types. Dendrograms were plotted to visualize the hierarchical clustering result. The findings from the hierarchical clustering were then aligned with established biological insights for validation.

2.7 Aggregate Hierarchical Clustering

This is the final step in the attempt to acquire more stable and more accurate hierarchical clustering results. Multiple distance matrices were normalized and then summed up to get an aggregated distance matrix. Hierarchical clustering is then applied to this aggregate distance matrix. Four aggregation combinations were considered:

1. combined bottleneck distance and persistence image distance with zero padding using Pearson correlation

2. combined bottleneck distance and persistence image distance with zero padding using Euclidean distance

3. combined bottleneck distance using Pearson correlation and Euclidean distance

4. persistence image distance with zero padding using Pearson correlation and Euclidean distance

These distances were chosen because they gave more robust hierarchical clustering results.

3 Results

3.1 Hierarchical Clustering Based on Biological Knowledge

Based on previous biological knowledge, we assigned the hierarchical clustering as follows:

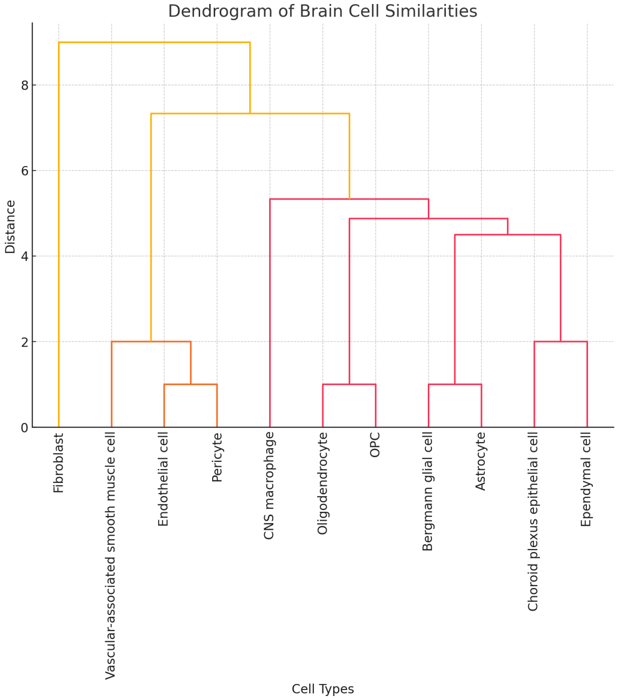


Figure 3: Dendrogram Based on Biological Knowledge

1. **Glial Cells - Bergmann Glial Cell and Astrocyte:**

   Both are astrocyte-related cells involved in supporting neuronal functions, maintaining the blood-brain barrier, and regulating synaptic activity.

2. **Oligodendrocyte and OPC:**

   OPCs are progenitor cells that differentiate into oligodendrocytes, which are responsible for myelinating axons in the CNS, facilitating rapid signal transmission.

3. **CNS Macrophage:**

   While CNS macrophages (microglia) are glial cells, their primary function is immune-related, including clearing debris and responding to injury, making them less similar to astrocytes and oligodendrocytes.

4. **Epithelial Cells - Ependymal Cell and Choroid Plexus Epithelial Cell:**

   Both are involved in the production and circulation of cerebrospinal fluid (CSF), playing essential roles in maintaining CNS homeostasis and fluid balance.

5. **Vascular Cells - Endothelial Cell, Pericyte, Vascular-associated Smooth Muscle Cell:**

   These cells are involved in the vascular system of the brain, including maintaining the blood-brain barrier, regulating blood flow, and providing structural support to blood vessels.

6. **Fibroblast:**

   Fibroblasts are primarily involved in producing the extracellular matrix and structural repair, functions distinct from both glial and vascular cells. This positions them with the highest dissimilarity to other cell types in the CNS.

3.2 Persistent Diagrams

For simplicity, only persistent diagrams of oligodendrocytes are shown.

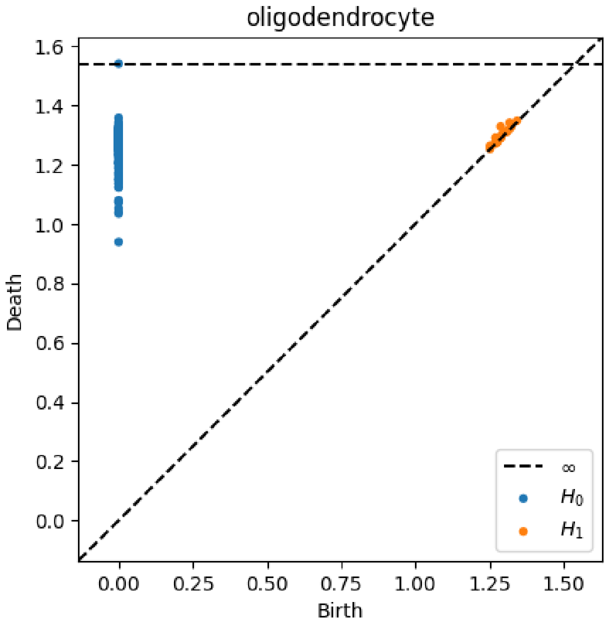


Figure 4: Persistent Diagram of Oligodendrocyte using Pearson Correlation

Significantly more simplices using Euclidean Distance than using Pearson correlation.

3.3 Hierarchical Clustering

Overall, complete linkage gives the best result. Thus, for simplicity, only dendrograms using complete linkage are shown here.

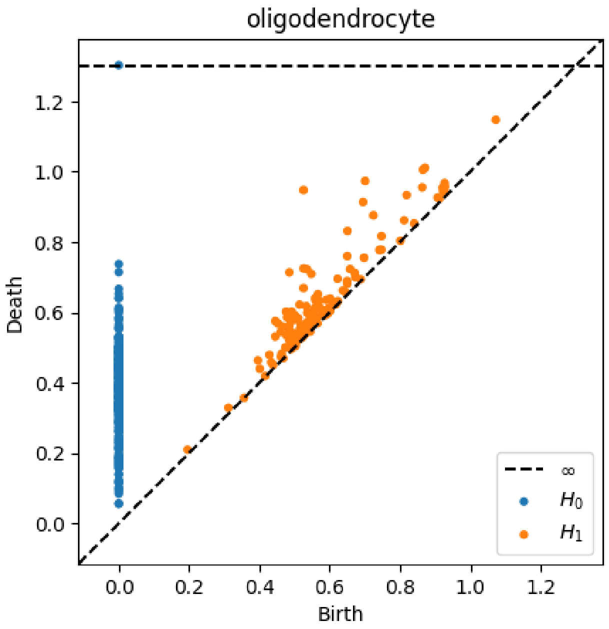
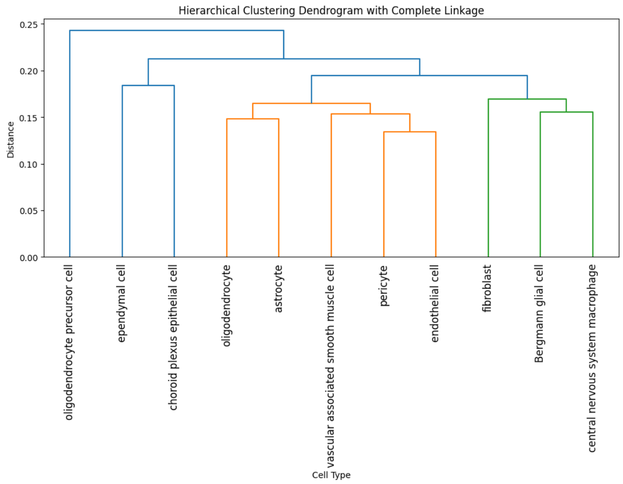


Figure 5: Persistent Diagram of Oligodendrocyte using Euclidean Distance

3.3.1 Bottleneck Distance



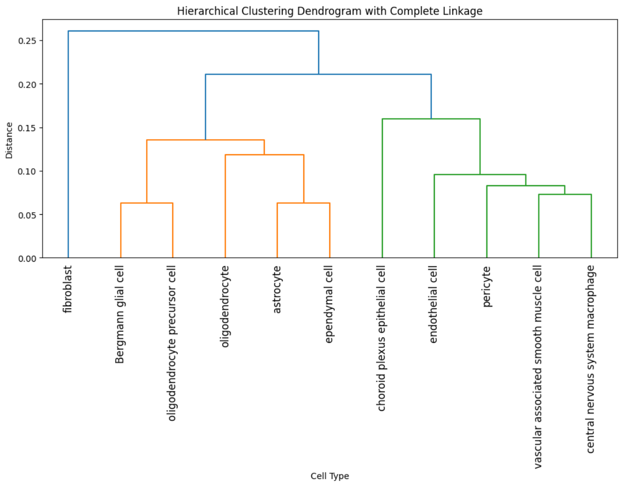


Figure 6: Hierarchical Clustering Combined Bottleneck Distance with Pearson Correlation

Figure 7: Hierarchical Clustering 1-dimensional Bottleneck Distance with Euclidean Distance

In general, the clustering result with combined distance is more reasonable compared to each dimension separately.

The main difference between the Pearson correlation bottleneck result and Euclidean bottleneck result is that they have different clustering for cells that are hard to cluster or highly specialized, for example, oligodendrocyte precursor cells, fibroblast, and central nervous system macrophage. However, they both agree on the super-clustering of vascular cells - pericyte, endothelial cell, and vascular-associated smooth muscle cell; glial cells -

astrocyte, oligodendrocyte.

An interesting finding about ependymal cells and choroid plexus epithelial cells is that their functions are very similar, but ependymal cells are closer to glial cells, while choroid plexus epithelial cells are closer to vascular cells. Pearson correlation clusters them more closely to their functions, but Euclidean distance clusters them more closely to their super-cell type.

3.3.2 Persistence Images

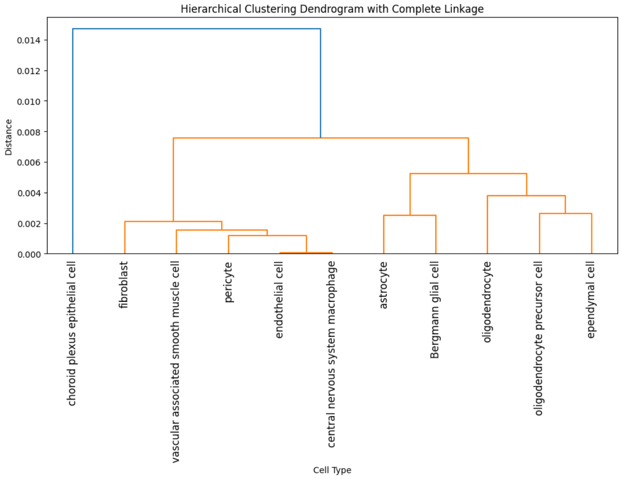


Figure 8: Hierarchical Clustering with Padded Persistence Image with Euclidean Distance

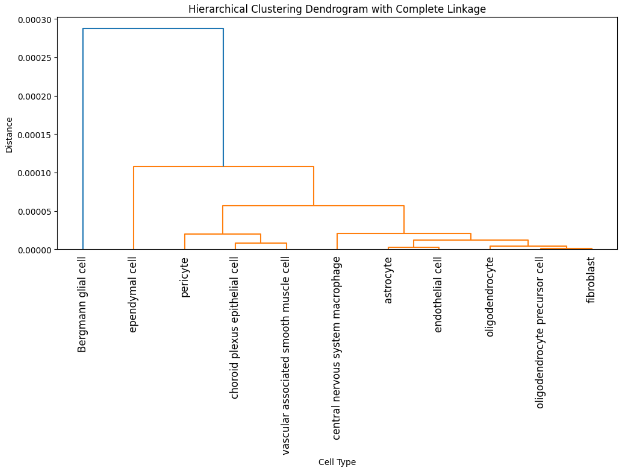


Figure 9: Hierarchical Clustering with Resized Persistence Image with Pearson Correlation

In general, bottleneck distance gives more reasonable results than persistence images. Persistence images with padding also have higher differentiation capability than directly resizing. However, this might be a result of a suboptimal choice.

3.3.3 Aggregation

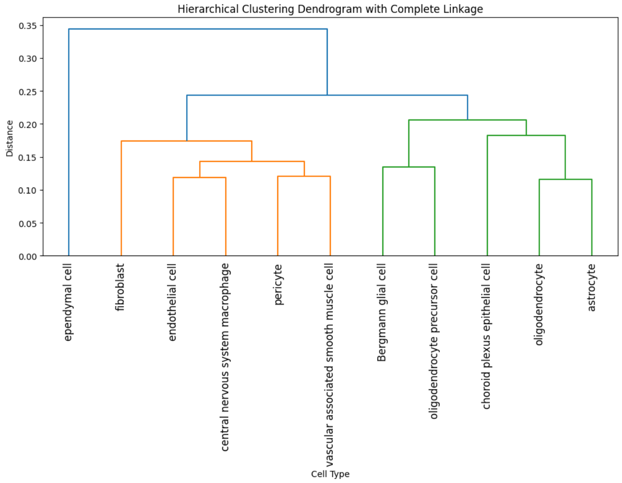


Figure 10: Aggregate Bottleneck Distance Hierarchical Distance

Aggregated distance differentiates vascular cells and glial cells more stably. Specifically, aggregated bottleneck distance gives the best aggregation result, which aligns with our previous conclusion.

4 Conclusion

To summarize, Pearson correlation and Euclidean distance give slightly different clustering results. However, non-linear dimensional reduction using UMAP before calculating Euclidean distance is very crucial, which suggests there might be non-linear features in Euclidean space.

Among all three clustering linkage methods, complete linkage gives the best results, whereas single linkage is not very useful.

When comparing the distance between persistent diagrams, bottleneck distance gives us better results than persistence images.

Our hierarchical clustering result is most stable with the clustering of the most common vascular cells and glial cells. However, it might have difficulty clustering highly specialized cells. Clustering with aggregated distance can help with this issue, e.g. clustering Bergmann glial cells to glial cells. After aggregation, it also shows that ependymal cells and choroid plexus epithelial cells are very close despite belonging to glial cells and vascular cells. This aligns with them having similar functions. But this instability can also be used to select distinctive cells like fibroblasts and central nervous system macrophages.

Our findings enhance the understanding of non-neural cell functions and contribute to the broader field of neurobiology. Future research can build on these results by further exploring the identified gene clusters and cell groups, potentially uncovering new regulatory mechanisms and functional roles of non-neural cells in the brain.

5 Discussion

There are some limitations to our work. For example, after downsizing the dataset using a 75% threshold, we have different gene counts for each cell type. This might have an impact on the comparison of persistent diagrams.

In addition, hierarchical clustering has its intrinsic limitations. For example, fibroblasts and central nervous system macrophages are very different from glial cells and vascular cells, but they are also different from each other. Hierarchical clustering is not suitable for this situation. The visualization might also have misleading results. For instance, from Fig.10 we might think the ependymal cell is a cluster by itself, but a closer look can tell us that the choroid plexus epithelial cell is the second farthest from the rest of the cells, which implies its proximity to the ependymal cells.

The future possible direction for this project is to identify what genes belong to the same homologous class, which gives more detailed information. For example, we can explore the genes that consist of the simplex with the highest persistence.

We can also adapt this project from hierarchical clustering to classification, which is useful in bioinformatics.

We can also explore Mapper to acquire a different visualization of different cell types, which can be used to compare with our current results.

References

[1] Siletti, K. et al. (2023) ‘Transcriptomic diversity of cell types across the adult human brain’, *Science*, 382(6667). doi:10.1126/science.add7046.

[2] Wolf, F.A., Angerer, P. and Theis, F.J. (2018) ‘Scanpy: Large-scale single-cell gene expression data analysis’, *Genome Biology*, 19(1). doi:10.1186/s13059-017-1382-0.

[3] Our code is available at https://github.com/cecilialmw/DSC-214-Final-Project