Prof. Dr. Manfred Claassen
Matthias Bruhns & Jan Schleicher
Machine Learning for
Single Cell Biology 2024

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Assignment 2

Task 1 – Non-linear dimensionality reduction (70 points)

Single-cell RNA sequencing data has high dimensionality due to the number of genes expressed in a typical cell. Non-linear embedding is often used to visualize the high-dimensional data in 2D. The various algorithms used for this task have different strategies to represent the relevant properties of the high dimensional data in 2D.

A. Locally linear embedding (LLE) was an early technique that represents each data point as a linear combination of its neighbors. Explain the algorithm step by step and also implement standard LLE as described in the reference below. Use your LLE implementation to generate a 2D embedding of the data: *expression_data_1.txt*

Roweis, Sam T., and Lawrence K. Saul. "Nonlinear Dimensionality Reduction by Locally Linear Embedding." Science, vol. 290, no. 5500, American Association for the Advancement of Science, 2000, pp. 2323–26, http://www.jstor.org/stable/3081722.

Cell type identity for *expression_data_1.txt* have been provided in *metadata_1.txt*. Create a plot of the 2D embedding estimated above and annotate it with the cell type identity. (40 points)

B. Next use the umap-learn package to generate a 2D UMAP embedding of the same data: *expression_data_1.txt*. Create a plot of the UMAP embedding annotated with the cell type identity. (5 points)

C. We will now investigate the role of the neighborhood parameter on the embeddings. For both LLE (use the sklearn implementation of LLE for this subtask) and UMAP generate embeddings for a range of neighborhood sizes from 5 to 100 in intervals of 5. Using each embedding and the provided cell type labels compute the Davies-Bouldin Index. Plot the dependence of the index on neighborhood size for both methods in a single plot. (10 points)

D. Answer the following: (15 points)

- What are the differences between the two embeddings?
- Which embedding is better suited to visualize cell populations and why?
- Based on your plots, what do you think are the conceptual differences between the two embedding techniques?

Task 2 – Batch effects in scRNA seq data (30 points)

Differences between single-cell RNA sequencing datasets generated independently can be ascribed to biological and non-biological sources of variation. Typically, the non-biological sources of variation can lead to incorrect conclusions, and therefore batch effect correction is an important processing step in the analysis of multiple single-cell RNA sequencing datasets.

A. Explain the algorithm step by step and implement quantile normalization as a batch correction strategy for single-cell RNA sequencing data. Use your implementation to correct batch effects present in the dataset: *expression_data_2.txt*. (20 points)

B. Batch and celltype labels for each cell are provided in *metadata_2.txt*. Create two plots visualizing pre-corrected and batch-corrected data using UMAP embedding. Annotate with the provided batch labels. Is the batch correct satisfactory? Explain your assessment for either outcome. (10 points)