CS690A: Assignment 2 - Clustering of Spatial Transcriptomics Data

Hamim Zafar
hamim@iitk.ac.in

Indian Institute of Technology Kanpur — September 14, 2024

Introduction

Spatial transcriptomics techniques (SRT) spatially barcode transcriptomes with a spatial resolution larger than a single cell, ranging from 50μm to 100μm for ST to 10μm for Slide-seq. Such datasets can consist of a spot-gene matrix, where each spot corresponds to a tissue location from where the mRNA molecules have been captured. With such datasets, an important task is to **identify spatial domains** defined as regions that are spatially coherent in both gene expression and histology. Traditional clustering methods such as K-means and Louvain's method may not perform the best in such a scenario.

SRT technologies are broadly divided into 2 categories 1) Imaging-based techniques that have high spatial resolution and can capture gene expression at a cellular level but have low sensitivity in gene detection 2) Sequencing-based techniques that have low spatial resolution and can capture gene expression at spot level where a spot can have contribution from multiple cells but has high throughput mRNA capturing. The first category includes techniques such as STARmap (spatially resolved transcript amplicon readout mapping), smFISH (single-molecule fluorescent ISH), seqFISH (sequential hybridization), MERFISH(multiplexed error-robust FISH), OSMFISH (Oligonucleotide Sequential Fluorescence In Situ Hybridization). This category is based on in-situ hybridization (ISH) methods that enable the quantification of gene expression at the cellular level but are limited to hundreds of preselected genes. The second category includes techniques such as ST/Visium, Slide-seq, and High Definition Spatial Transcriptomics that spatially barcode entire transcriptomes with a spatial resolution larger than a single cell.

The goal of this assignment is to evaluate existing algorithms for clustering Imaging-based spatial transcriptomics datasets.

Info: To learn more details on the problem, read the following papers

- Hu, Jian, et al. "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network." Nature methods 18.11 (2021): 1342-1351.
- Fu, H., Xu, H., Chong, K., Li, M., Ang, K. S., Lee, H. K., ... & Chen, J. (2021). Unsupervised spatially embedded deep representation of spatial transcriptomics. Genome Medicine (2024).

1 Dataset Description

The assignment contains two datasets as described in the following

1.1 MERFISH Mouse brain preoptic hypothalamus region

The data presented here is from the mouse preoptic hypothalamus region of brain acquired with MERFISH (multiplexed error-robust fluorescence in situ hybridization) technique. For this dataset, a single h5ad file is provided which contains both the spot*gene matrix as well as the x, y coordinates for each spot. The number of spatial domains/clusters in this dataset are 8. This information might be required for running the methods. All files can be accessed from the kaggle competition page.

1.2 osmFISH Mouse brain somatosensory cortex region

This dataset represents a mouse somatosensory cortex region of brain acquired with OSM-FISH (Oligonucleotide Sequential Fluorescence In Situ Hybridization) technique. For this dataset, a single h5ad file is provided which contains both the spot*gene matrix as well as the x, y coordinates for each spot. The number of spatial domains/clusters in this dataset are 11.

2 Tasks (100 points)

The task is to run the clustering methods assigned to your team on the datasets. For each dataset, you will be able to calculate ARI as a measure of your clustering accuracy by submitting to kaggle. You will submit the solution in the format as given below

Listing 1: Format of output csv file

```
Id,Expected
AAACAAGTATCTCCCA-1,1
AAACAATCTACTAGCA-1,2
AAACACCAATAACTGC-1,3
AAACAGAGCGACTCCT-1,2
AAACAGCTTTCAGAAG-1,1
```

The solutions will be evaluated for clustering accuracy. You can number your clusters in the range $[1, K]$ where K is the number of clusters inferred by your method. You also need to generate spatial feature plots of the clusters and provide it as an output (see the deliverables section). The output file should be named as <Method>_<Dataset_i>.csv, where <Method> should be replaced by the name of the method you used for clustering and <Dataset_i> should be replaced by the name of the dataset.

Notice: In case we require a change in the format of the csv file, we will notify you. Keep an eye on the announcements.

3 Deliverables

The deliverables for the assignment are the following

1. Clustering predictions on both the datasets. These results will be evaluated on the leaderboard. You will be graded based on the successful execution and experimentation of the methods assigned to you.
2. Runnable code (in Jupyter Notebook) for the methods assigned to you. Scripts for running your code to generate the predictions on the datasets. TAs will run these scripts to reproduce the csv files you submit for the assignment.
3. Submit the environment file (.yml) for running the codes.
4. A short report describing the steps taken to solve the assignment. Describe in brief the algorithms you have used, performance of your algorithms on the datsets. The report should also contain the spatial feature plots (based on the obtained clusters) of all the datasets for both the methods assigned to you. The writeup should also contain a section describing the contribution of each member in the team. The writeup should mention the names and roll numbers of the team members.

For submission, all the deliverables should be zipped in a single file and the zip file should be named as `Team_i_CS690_ST_clustering_assignment_2.zip`, `i` should be replaced by your team number. Also, each file in the zip folder should start with the phrase `Team_i_` (`i` replaced by your team number). The file should be emailed to the instructor with the TAs copied in the email. The subject line of the email should mention the team number, [CS690A] and the phrase 'ST Clustering assignment 2'.

4 Submission Deadline

September 29^{th} 11:59 PM.

5 Kaggle Competition links

Dataset1_MERFISH - `https://www.kaggle.com/t/e657311496b14bd7898626ac45dae858`
Dataset2_osmFISH - `https://www.kaggle.com/t/16b8d025621b4e9891aa8e8d5a49fa11`