Integrating Single-Cell and Spatial Transcriptomics via Deep Learning Generative models and in a Contrastive Method

Project Management and Software Development for Medical Applications

General Info

Contact Person: Mohammad Lotfollahi

Contact Email: mohammad.lotfollahi@helmholtz-muenchen.de

Project Abstract

We aim to develop a mapping function between two RNA sequencing modalities, spatial transcriptomics and single-cell modalities using contrastive learning. Our goal is to remove the batch effect, find a mapping function and evaluate the preservation of spatial and biological information using metrics.

Background and Motivation

Single-cell analysis is a technique that involves examining the characteristics of individual cells by analyzing RNA sequencing data generated by the single-cell technology. This technology provides gene expression information for each individual cell. Variational Auto Encoders (VAEs), a type of deep learning model, have been used in single-cell analysis to map the gene expression of cells into a latent space, providing opportunities for further analysis such as identifying clusters of cells with unique biological characteristics.

Spatial transcriptomics is a cutting-edge technique that builds upon single-cell technology by adding spatial information to the RNA data of cells, incorporating the physical position of cells in the tissue as spatial coordinates. Using these coordinates, a KNN graph can be defined for the cells, assigning a set of neighbors to each cell. However, spatial transcriptomics may provide less information about gene expression than single-cell technology. While single-cell technology can extract data for up to 2000 genes in each cell's expression vector, spatial transcriptomics typically provides data on up to 1000 genes.

Consequently, computational biologists are searching for a mapping function between spatial transcriptomics and single-cell modalities. By predicting the spatial coordinates for cells sequenced by single-cell technology, exciting research opportunities can open up. This approach could enable the identification of the spatial arrangement of cells in a tissue and how cells interact with each other. Such knowledge can provide insights into the development of disease and the efficacy of therapeutic interventions. Our project is also aimed at developing a mapping function between spatial transcriptomics and single-cell modalities using contrastive learning.

Student’s Tasks Description

Our research plan includes the following steps:

1. Design two Conditional Variational Auto Encoder (CVAE) models to map single-cell and spatial transcriptomic data into the same latent spaces. We will feed a selection of mutual genes into the encoder, and concatenate the one-hotted label of their cell types to the input of the decoder.
2. Define a set of triplets (cell, positive, negative) to take advantage of a contrastive technique before training the CVAEs.
3. Recognize single-cell and spatial data as two batches of data. Merge both batches

and map them into a latent space. Each batch is represented as a group of clusters that show cell types. However, the batch effect prevents these two groups from being merged. To remove this effect, we will define a set of triplets (cell, positive, negative) to be used by a triplet loss. We will use the Mutual Nearest Neighbor (MNN) concept to define this set. We will select positives in the paired clusters of cell types in two batches and negatives randomly within the same cluster with the cell.

4. Add a triplet loss term to the total loss to remove the batch effect. We will evaluate various contrastive losses to find the best applicable loss.
5. Use metrics to evaluate how well the latent space preserves both spatial and biological information.
6. Once we achieve a well-structured latent space, we will use the KNN method to map single-cell and spatial data to each other.

Technical Prerequisites

1. Python programming language
2. Jupyter Notebook or JupyterLab for interactive data analysis and visualization
3. Pandas for data manipulation and analysis
4. Numpy for numerical computations
5. Scikit-learn for machine learning tasks
6. PyTorch for deep learning tasks
7. Scanpy or Seurat for single-cell analysis

References

1) Learning cell communication from spatial graphs of cells, David S. Fischer, Anna C. Schaar, Fabian J. Theis

2) Learning Spatially-Aware Representations of Transcriptomic Data via Transfer Learning, Minsheng Hao, Lei Wei, Xuegong Zhang

3) Integration of millions of transcriptomes using batch-aware triplet neural networks, Lukas M. Simon, Yin-Ying Wang, and Zhongming Zhao

4) Graph-based autoencoder integrates spatial transcriptomics with chromatin images and identifies joint biomarkers for Alzheimer's disease. Zhang, Xinyi, Wang, Xiao, Shivashankar, G. V. Uhler, Caroline

Please send the completed proposal to ardit.ramadani@tum.de, tianyu.song@tum.de, vanessag.duque@tum.de and shervin.dehghani@tum.de.
Please note that this proposal will be evaluated by the BMC coordinators and will be assigned to a student only in case of acceptance.