Course Project Ideas for CS690A: Computational Genomics (2022-2023 Semester 1)

Project Mentors: Hamim Zafar

October 12, 2022

Introduction

This document provides some project ideas for the course project. Ideally, you will want to pick a problem in a domain of your interest, e.g., single-cell genomics, spatial transcriptomics, etc., and formulate your problem in a statistical and/or machine learning framework.

1 Generation and augmentation of single-cell RNA-seq data using denoising Diffusion-based Generative Modeling

A fundamental problem in biomedical research is the low number of observations available, mostly due to a lack of available biosamples, prohibitive costs, or ethical reasons. Augmenting few real observations with generated in silico samples could lead to more robust analysis results and a higher reproducibility rate. Previously, generative adversarial neural network models have been utilized for the realistic generation of single-cell RNA-seq data. In the vision domain, denoising diffusion models have recently emerged with remarkable results in the area of generative modeling.

The goal of this project would be to develop a diffusion-based generative model for the in-silico generation of single-cell RNA-seq data.

Papers to read

1. Marouf, Mohamed, et al. "Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks." Nature communications 11.1 (2020): 1-12.

2. Cao, Yue, Pengyi Yang, and Jean Yee Hwa Yang. "A benchmark study of simulation methods for single-cell RNA sequencing data." Nature Communications 12.1 (2021): 1-12.

3. Sun, Tianyi, et al. "scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured." Genome biology 22.1 (2021): 1-37.

4. Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in Neural Information Processing Systems 34 (2021): 8780-8794.

5. Pandey, Kushagra, et al. "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents." arXiv preprint arXiv:2201.00308 (2022).

2 Mosaic single-cell data integration using deep generative models

Data integration has emerged as a key challenge for consolidating and profiting from rich resources of single-cell datasets, with the task of integrating diverse molecular assays being known as 'mosaic data integration'. At present, many methods for mosaic data integration are typically limited to using the set of overlapping features between modalities. However, as the number and complexity of single cell datasets increase, there is a growing need to develop techniques specifically designed to perform mosaic data integration. Some existing approaches designed to tackle this problem include UINMF, which introduces a latent metagene matrix in the factorisation problem, MultiMAP, a graph-based method that assumes a uniform distribution of cells across a latent manifold structure fitted using an optimisation approach, and StabMap that projects all cells onto supervised or unsupervised reference coordinates utilising all available features regardless of overlap with other datasets, instead relying on traversal along the mosaic data topology.

The goal of this project is to develop a deep generative model for the task of mosaic data integration which can perform better than the existing methods. Methods like StabMap distinguishes the tasks of integration and batch correction. Using a deep generative model (e.g., VAE-based architecture), the two tasks can be solved using a unified approach.

Papers to read

1. Ghazanfar, Shila, Carolina Guibentif, and John C. Marioni. "StabMap: Mosaic single cell data integration using non-overlapping features." bioRxiv (2022).

2. Kriebel, April R., and Joshua D. Welch. "UINMF performs mosaic integration of single-cell multi-omic datasets using nonnegative matrix factorization." Nature communications 13.1 (2022): 1-17.

3. Jain, Mika Sarkin, et al. "MultiMAP: dimensionality reduction and integration of multimodal data." Genome biology 22.1 (2021): 1-26.

4. Lopez, Romain, et al. "Deep generative modeling for single-cell transcriptomics." Nature methods 15.12 (2018): 1053-1058.

5. Ternes, Luke, et al. "A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis." Communications biology 5.1 (2022): 1-10.

3 Integration and batch correction of single-cell data using denoising Diffusion-based Generative Modeling

Recent studies have shown that cellular features can be preserved across experimental systems from related biological contexts. The information learned from different data sources can improve the analysis and interpretation of diverse biological systems. However, the advantages of integrated data can be compromised by differences due to experimental batch, sampling (sample acquisition and handling, sample composition, reagents or media, and sampling time), or technology (sequencing depth, sequencing lanes, read length, plates or flow cells, protocol). Many methods have been established to integrate scRNA-seq studies across multiple experiments. Some methods employ supervised cross-domain transfer learning to remove domain effects with models learned from labeled datasets. Methods such as scVI and DESC that employ deep variational autoencoders for learning cellular embeddings from scRNA-seq can also integrate data from multiple batches. However, a recent benchmarking study highlighted the need for the development of new data integration methods.

The goal of this project is to develop a deep generative model that combines variational autoencoder and recently introduced denoising diffusion models for performing unsupervised as well as supervised (cell type labels are known) single-cell integration.

Papers to read

1. Xu, Chenling, et al. "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models." Molecular systems biology 17.1 (2021): e9620.

2. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.

3. Pandey, Kushagra, et al. "Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents." arXiv preprint arXiv:2201.00308 (2022).

4. Abstreiter, Korbinian, et al. "Diffusion-Based Representation Learning." arXiv preprint arXiv:2105.14257 (2021).

4 Learning latent representation of spatial transcriptomic data for denoising and better clustering

Spatially resolved transcriptomics (SRT) provide gene expression close to, or even superior to, single-cell resolution while retaining the physical locations of sequencing and often also providing matched pathology images. However, SRT expression data suffer from high noise levels, due to the shallow coverage in each sequencing unit and the extra experimental steps required to preserve the locations of sequencing. Ideally, unique molecular identifiers (UMIs) at a spot measure spot-specific expression, but this is often not the case in practice due to bleed from nearby spots, an artifact referred to as spot swapping.

The goal of this project is to develop a deep generative model for spatial transcriptomics data that will utilize the information from the physical locations of sequencing, and the tissue organization reflected in corresponding pathology images to learn a latent representation of the data which can be used for denoising the data as well as improving downstream tasks such as spot clustering, cell-cell interaction inference, etc. Graph neural network based models and deep generative models can be combined to perform such task.

Papers to read

1. Wang, Yunguan, et al. "Sprod for de-noising spatially resolved transcriptomics data based on position and image information." Nature methods 19.8 (2022): 950-958.

2. Ni, Zijian, et al. "SpotClean adjusts for spot swapping in spatial transcriptomics data." Nature Communications 13.1 (2022): 1-11.

3. Kipf, Thomas N., and Max Welling. "Variational graph auto-encoders." arXiv preprint arXiv:1611.07308 (2016).

4. Fu, Huazhu, et al. "Unsupervised spatially embedded deep representation of spatial transcriptomics." Biorxiv (2021).

5. Zong, Yongshuo, et al. "conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics." bioRxiv (2022).

5 Trajectory inference from time-series scRNA-seq data

Dynamic cellular processes such as differentiation involve cell-state transitions that are characterized by cascades of transcriptional changes. Trajectory inference (TI) methods have been developed for inferring cellular dynamics using single-cell datasets. MARGARET is one such method that uses scRNA-seq data for inferring the cell state trajectory and dynamics of cell fate plasticity. MARGARET employs an unsupervised metric learning-based approach for inferring the cell-state manifold and captures complex trajectory topologies by constructing a cluster connectivity graph. However, MARGARET does not utilize the real time point information that may be available with some datasets. In recent time, the number of time-series scRNA-seq datasets is increasing and some methods have been developed that explicitly utilizes the time point information. However, these methods have certain limitations and it is required to develop novel TI methods for time-series scRNA-seq datasets.

The goal of this project will be to develop a novel TI method that has the functionalities of MARGARET while utilizing the time point information.

Papers to read

1. Pandey, Kushagra, and Hamim Zafar. "Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET." Nucleic acids research 50.15 (2022): e86-e86.

2. Schiebinger, Geoffrey, et al. "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming." Cell 176.4 (2019): 928-943.

3. Forrow, Aden, and Geoffrey Schiebinger. "LineageOT is a unified framework for lineage tracing and trajectory inference." Nature communications 12.1 (2021): 1-10.

4. Tran, Thinh N., and Gary D. Bader. "Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data." PLoS computational biology 16.9 (2020): e1008205.

References