Cell Type Classification Based on Deep Learning for scRNA-seq Data

Renwei Li (rl3088) & Rong Ma (rm3707)

Background

Single-cell RNA sequencing (scRNA-seq) can detect the sequence information from individual cells with next-generation sequencing techniques[1]. Compared with the traditional bulk RNA-seq, the most prominent feature is that it can isolate every cell from samples and obtain the transcriptome information of every cell. It is widely applied in biomedicine, such as cancer study, organ development, gut microbiome analysis because researchers are able to discover heterogeneous cell composition in complex tissues and environment, and it is possible to track the dynamic process in life behavior or disease progress[2]. After data pre-processing, the downstream analysis is the key to discover the biological meaning of disease in scRNA-seq analysis. Cell type classification is a very important aspect because we need to know the characteristics of our cells in disease. For example, in the study of the tumor microenvironment, only if the cell types are correctly classified, will it be able to understand the roles of each kind of cell in tumor development in different periods. There are many methods to classify or infer cell types for scRNA-seq data, and neural networks have the ability to process complex datasets and identify hidden patterns and relationships[3]. Our project is aiming at developing a deep-learning based method for cell type classification in the scRNA-seq analysis.

Datasets

A. Reference database: 1.cellMatch database[4] (Tissue-specific cell markers reported in the literature from humans or mice); 2.Tabula Muris database[5] (scRNA-seq data from mouse); 3.Human Cell Atlas database[6] (Transcriptome data from human organs)
B. Training & Testing dataset: 1. Chen dataset, accession code GSE99701; 2. PBMC 10X Genomics; 3. Wu dataset, GSE103976; 4. Lindsey dataset, GSE102580

Methods

1. Dataset preparation: Download and organize the necessary datasets. Cell type reference will be constructed by combining the reference database for cell type inference.
2. scRNA-seq Data pre-processing: a. Quality control and normalization; b.Dimensional reduction and clustering. This step will be conducted with current *scanpy*[7] and *liger*[8] methods.
3. Training the novel model: To improve the interpretability of the results without sacrificing the accuracy of classification, we plan to introduce a novel deep learning model called CapsNet[9] which is usually applied in digital recognition and protein structure prediction. The high interpretability of the model is usually marked as the biggest advantage of it. With some modifications of the original CapsNet model, the gene expression of every single cell as the input will be processed through two parts: feature extraction and dynamic routing, while the final output will be the probability of the cell type inference for every single cell in our designed model with the combination of constructed reference. The linkage between extracted features (important gene sets) and the final output could provide a strong biological explanation of the results. Meanwhile, we will apply the KNN (K-nearest neighbors) model as the baseline of the project.
4. Performance evaluation and improvement: The accuracy of the cell type classification for our designed model will be assessed and compared with other popular methods such as CellAssign, Garnett and sc-Pred. Obviously, we have to make sure the performance of our model is outperformed than the baseline (KNN model). In the end, we will discuss the biological meaning of the results as the largest benefit of the model.

Groupwork

Data preparation; Model design and training: Rong ;
Data preprocessing, Model design and evaluation: Renwei

Reference

[1] Grün D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. Cell. 2015;163(4):799–810. doi:10.1016/j.cell.2015.10.039
[2] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175–188. doi:10.1038/nrg.2015.16
[3] Gomes T, Teichmann SA, Talavera-López C. Immunology Driven by Large-Scale Single-Cell Sequencing. Trends Immunol. 2019;40(11):1011–1021. doi:10.1016/j.it.2019.09.004
[4] Shao X, Liao J, Lu X, Xue R, Ai N, Fan X. scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience. 2020;23(3):100882. doi:10.1016/j.isci.2020.100882
[5] Tabula Muris Consortium; Overall coordination; Logistical coordination; Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature. 2018;562(7727):367–372. doi:10.1038/s41586-018-0590-4
[6] Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: from vision to reality. Nature. 2017;550(7677):451–453. doi:10.1038/550451a
[7] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. Published 2018 Feb 6. doi:10.1186/s13059-017-1382-0
[8] Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell. 2019;177(7):1873–1887.e17. doi:10.1016/j.cell.2019.05.006
[9] Sabour, S., N. Frosst, and G.E. Hinton, Dynamic Routing Between Capsules. Advances in Neural Information Processing Systems 30 (Nips 2017), 2017. 30.