Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Machine Learning for Genomics
Fall Semester 2024

Project 2: Bulk deconvolution and single-cell clustering

Assigned on: **5:00pm on 06.11.2024** Due by: **12:00pm on 04.12.2024**

1 Exercise 1 - Theory

Please answer these questions in the final report slot on the submission website. You will not be formally graded on these questions but they should help you perform better on the practice section.

1.1 Question 1

Samples from different patients may be processed by different technicians and/or at different time points. How could this possibly affect the single cell RNA sequencing data? What type of method is supposed to correct for these potential confounding effects?

1.2 Question 2

What are the two main categories of methods for RNA deconvolution that exist? List 2 advantages and 2 disadvantages for each of these categories.

2 Exercise 2 - Practice

In this section you will have to find meaningful clusters of cells and deconvolve bulk data to obtain an estimate of cell type fractions. You will be studying immune and stromal cells coming from patients with esophageal adenocarcinoma. There are a total of 10 single-cell samples in the dataset ($n = 6$ for the train set, $n = 4$ for the test set) and 20 bulk samples.

We wish to know how the tumor microenvironment (TME) composition of esophageal adenocarcinoma patients differs between post and pre chemotherapy treated patients, as well as if we can find reliable biomarkers of the cell types across patients. To do so, we want to obtain clear clusters in our dataset that represent the underlying biology and we want to harness the bulk measurements to get estimates of TME composition across a larger amount of patients. You will perform two separate tasks. For the first task, you will have to perform clustering on your data to find biologically meaningful clusters. You are free to perform whatever transformations on your data you wish. For the second task, you will have to deconvolve the bulk data to obtain an estimate of the proportion of the nine cell types present in the data (T cells, B cells, Endothelial cells, Fibroblasts, Plasmablasts, Myofibroblasts, NK cells, Myeloid cells, and Mast cells).

You will be provided with a training set of 10 bulk measurements with associated true proportions and a single-cell train dataset (raw counts and information about cell type, patient and sample of origin, treatment condition). For the first task, you will be evaluated on a test single-cell dataset, for which you will not get the cell type annotations. For the second task, you will be evaluated on estimated proportions of 20 bulk measurements you will be given, and your estimated proportions will be compared against the true composition of the data.

You will be evaluated according to the two following tasks:

- **Clustering performance** Your clustering method must identify as best as possible the different ground-truth cell types present in the dataset. This would allow you later on to get meaningful biomarkers through differential gene expression analysis. You will thus have to pick a clustering algorithm to apply to whichever transformation of your data you choose. You will be evaluated according to 2 metrics:
  - The Adjusted Rand Index ARI
  - The V-measure score V
- **Average root mean squared error across cell types** We will be evaluating how well you estimated the proportions of different tumor microenvironment cell types in your bulk data. For this purpose we will use
  - The root mean squared error $RMSE$ between the estimated proportion of a cell type and the true proportion across patients.

    We will compute the average $RMSE$ for all cell types in the bulk dataset. Remember, for this metric the lower the better.

You will pass the project if you beat both baselines:

- $\frac{1}{2}ARI + \frac{1}{2}V$ for the clustering task.
- $RMSE$ for the deconvolution task,

A bonus of +0.25 on the semester grade will be given to the 10% students who perform the best. To choose groups getting the bonus, we will compute the Z-score associated to both subtasks (deconvolution and clustering) and sum them up (for the second task, to account for the fact lower RMSE is better, we will take the opposite of the Z-score). The 10% of students with the highest sum of Z-scores will get a bonus. As a reminder, bonuses are non cumulative for the class (e.g., if you have already received a bonus for project 1, you will only get +0.25 total even if you get a bonus for project 2).

You will have to provide the following files:

- The code you wrote for your analysis. Your code must be commented, readable and must run. You must provide a *requirements.txt* file listing the packages used in your analysis.
- For every .csv file provided, you have to ensure your rows are indexed $(0, 1, ..., n-1)$
- A file *cluster_membership.csv* containing two columns, the first the indices of the cells in the test dataset, the second the cluster membership of the cell (please ensure your cluster membership indices are 1-indexed, **not** 0-indexed). The file is expected to have a header with the column names ”index”, ”cluster”. There should be as many rows in this file as there are cells in the provided test dataset (excluding the header row).

- You will have to provide a document named *pred_props.csv* containing the estimated proportions of the nine cell types present in the data (T cells, B cells, Endothelial cells, Fibroblasts, Plasmablasts, Myofibroblasts, NK cells, Myeloid cells, and Mast cells). The file should contain meaning a matrix M where M_{ij} corresponds to the estimated proportion for cell type i for patient j. The first column of the file should be the cell types in the same order as provided here (there is a check for this in the starter code). The file is expected to have a header with the column names ”index”, and then all the samples in the bulk dataset. There should be as many rows in this file as there are cell types (excluding the header row).

You will be provided a starter code that you are free to use and that will partly check the format of the files before submission.

The files should be saved in a .zip file named *LastName_FirstName_Project2.zip*. We have provided an example of the output files you have to upload in the Project 2 archive.

You will also have to fill in on the submission website page a quick description (10000 characters max) of your work and the steps you took to perform the project. In this report, we will expect

- The answer to the two theory questions of Exercise 1,
- A quick justification of the different steps you took to perform the two tasks,
- Your specific contribution to the project.

Code and files can be shared across members of the team, however this description **must be individualized**.