INTERNATIONAL SCHOOL, 30 SEPT – 03 OCT 2024, MONOPOLI (BA), IT
AI-PHY

Contribution ID: **33**

Type: **Hackathon project proposal**

Integrating Spliced and Unspliced Gene Expression Data for Improved Cell Type Annotation in Single-Cell RNA Sequencing

Friday, 4 October 2024 10:40 (20 minutes)

Cell type annotation is one of the primary tasks in single-cell RNA sequencing analysis and it is of a significance importance and difficulty in computational biology. This difficulty arises from two primary factors. First, gene expression levels typically exist on a continuum rather than being discrete, making it harder to draw clear boundaries between different cell types. Second, variations in gene expression don't always correlate with functional differences at the cellular level, adding complexity to the classification process.
Given these challenges, a more precise and reliable method for cell type annotation is always welcomed. We propose an approach that integrates additional layers of biological data, mainly spliced and unspliced expression, to look at gene expression estimates from a different perspective. By combining these pseudo-multimodal layers of information, we might achieve more accurate cell type classification and better capture the functional nuances of different cell populations.

Project proposal: general context

This project proposal focuses on computational biology, particularly the primary analysis of single-cell RNA sequencing (scRNA-seq) data and cell type annotation. scRNA-seq is essential for exploring gene expression at the single-cell level, enabling the identification of distinct cellular populations and their functional roles.

Project proposal: description of the problem

The primary challenges in cell type annotation using single-cell RNA sequencing (scRNA-seq) arise from two main factors. First, gene expression levels exist on a continuum rather than discrete categories, making it difficult to delineate clear boundaries between different cell types. Second, variations in gene expression do not always correlate with functional differences at the cellular level, adding complexity to the classification process.

These challenges lead to potential inaccuracies in identifying cell types, which can hinder our understanding of cellular functions and the underlying biological processes. Developing a more reliable method that integrates diverse data types, such as spliced and unspliced gene expression, is essential to address these issues effectively.

Machine learning methods

We will use community detection algorithm such as Louvain or Leiden algorithm to do unsupervised clustering for cells. Then we run statistical tests to retrieve genes associated to each cluster and annotate the cell accordingly.

Input dataset

We will provide dataset for the hackathon. The data type is gene expression matrices and metadata, all stored and encoded in a well optimized data structure.

Goal and FOM

We will use an already annotated dataset and compare the accuracy of annotation to that

Primary authors: SANGUINETTI, Guido (SISSA); KOUADRI BOUDJELTHIA, Idris

Session Classification: Hackathon