Computational Biology Internship - Challenge

Internship Candidate Challenge: Building data integration pipelines

---

Introduction

Context

One of **Epigene Labs'** missions is to leverage publicly available cancer research studies to enhance the statistical power of secondary data analyses. By reanalyzing existing datasets, we aim to uncover new therapeutic opportunities in the fight against cancer. Despite the abundance of publicly available research data, much of it remains underutilized due to its complexity and heterogeneity.

The computational biology team plays a crucial role in this endeavor by developing pipelines to automatically download, pre-process, and harmonize molecular data into structured formats that enable secondary analyses.

Tasks

You can access the data located in folders on our Google Drive. There is one folder per task.

Task 1: Single-cell RNAseq data integration

Before developing scalable pipelines that can be used automatically in production, we start by an exploration phase to implement and test, on a few datasets, state-of-the-art integration methods.

Using the **literature** and available **Python open-source tools**, you are asked to implement a script to pre-process single-cell RNAseq data from the input data folder to make them comparable to the data in the output data folder. **We do not expect an exact match.**

Input Data :

- Raw gene expression data for GSE227828 along with raw clinical data

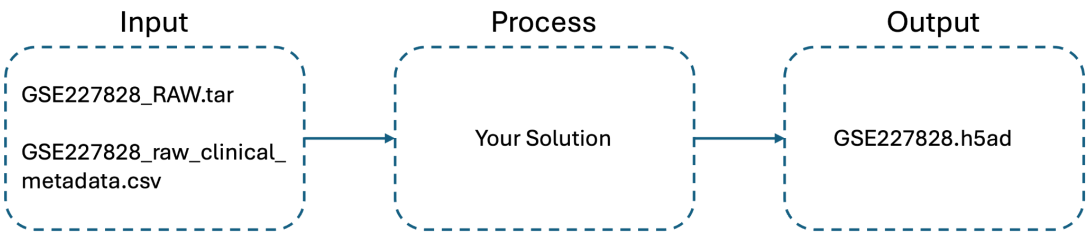
Output Data

- Pre-processed and normalized data

Deliverables

- Pre-processed dataset
- **Code**: you can create and share a Github repo or a zip file
- **Documentation:** Include documentation describing your pipeline, the tools you used, and the important steps... (report or slides). Comment on the different data formats and content.

Figure 1: Summary of the task 1



Task 2 - Improve the RNAseq data integration pipeline

Imagine it is your first day at Epigene Labs. The team presents the integration pipeline they started implementing to automatically integrate bulk RNAseq data.

Comment on the pipeline

- Update the workflow to give more details on the processing you think is followed
- How could we enhance the pipeline to ensure the consistency of the data?
- What challenges do you foresee?

Comment on the different output files

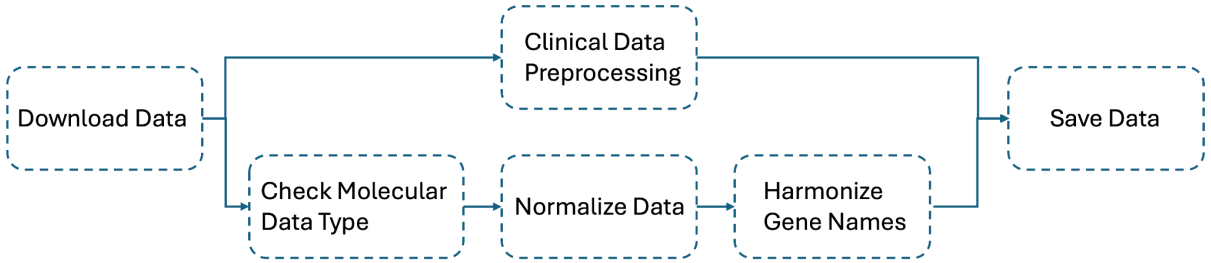
- Why do we have different files for one dataset?
- How would you use these different files for downstream analysis?

Tips

To help you answer these questions, we provide you with two folders of data - one corresponding to the data downloaded from the public database and the other one to the data once pre-processed

- Notice that the column names and values vary from one dataset to another both for genes and samples.

Figure 2: RNAseq data integration pipeline



NB: In the figure below, the clinical data integration was simplified on purpose to one step. As it is not the core focus of the challenge, we do not provide you with the clinical data. We represented it to highlight the possible interaction between clinical and molecular data.

Deliverable

- A **report or slides** explaining your approach. This task is focused on design and is more open-ended than the first one. We do not expect any implementation.

Instructions and Logistics

Communication

- **Questions**: Feel free to reach out with any questions or to discuss your preliminary thoughts.
- **Contact**: You can email us through Welcome to the jungle platform

Submission Details

- **How to Submit**: Email your completed materials through Welcome to the jungle platform
- **What to Include**: Your code, documentation, and any supporting materials.
- **Formats Accepted**: Jupyter Notebooks, GitHub repositories, PDF documents, slide presentations, etc.

Timeline

- **Duration**: No deadline but keep in mind that people in average send their challenge after **one week** from the date they receive it.

Expectations

- **Understanding**: Demonstrate a clear grasp of the context, problems, and potential tools.
- **Problem-Solving**: Show how you break down complex problems into manageable sub-problems.
- **Approach**: Choose a few promising solutions and explain why you selected them.
- **Benchmarking**: Where appropriate, compare different methods to highlight their strengths and weaknesses.

Final Note

There are no right or wrong answers. We are most interested in your analytical approach, problem-solving skills, and how you communicate your ideas.

Good luck, and we look forward to seeing your work!
