Project Proposal (M1)

Machine Learning Methods for scRNAseq Data

- Research Institute: Sorbonne University
- How to apply? Contact Juliana Bernardes `julibinho@gmail.com` and Nataliya Sokolovska `nataliya.sokolovska@upmc.fr`

Context The flow cytometry and single cell (sc) RNAseq techniques create cell-level data representing heterogeneous cell behavior for patients. Such data is important to understand phenotypes and find out genetic disorders. The main challenge of this project is to develop methods for extracting personalized disease signatures with implications in pathogenesis, classification, prognosis, and even treatment decisions.

Scientific Problem The focus of this project is on the scRNAseq data clustering. It is also a big data challenge since the scRNA data contain information about thousands of cells.

We will start with appropriate data integration. Patients are extremely heterogenous, and by consequence, cell-level data are also very different. Moreover, in clinical studies, a data set often includes not only the single-cell data but also data coming from different sources, such as clinical parameters, images, some omics data, etc. Such data contain precious complementary information, and proper data integration is needed.

To understand the actual context, we are planning to test the state-of-the-art multiple cells data integration methods to obtain the baseline performance and analyze their results. For example, the computer vision community has developed in past several years several practical tools allowing to reconstruct an image from its separate parts (see e.g., the Scanorama method motivated by image processing applications but developed for integration of single-cell RNA sequencing data). These batch correction approaches can be applied across patients and also across multiple datasets to reconstruct the whole picture illustrating dependencies between cells and tissues.

Next, the student(s) will study the dimensionality reduction methods for the scRNAseq data to compress the large data by keeping the most relevant information (feature selection). The main theoretical challenge is to try to compress meaningful information from cells to understand patients variability and the underlying data structure. We will also develop and implement a statistical machine learning algorithm that projects cells into a shared embedding under a constraint that similar cells (cell types) would appear in the same cluster that can also be compared across patients.

Background The student(s) will test the existing methods and implement their extensions. Knowledge of a programming language (Python/R) is required.

References

1. R. Petegrosso et al. *Machine learning and statistical methods for clustering single-cell RNA-sequencing data*, Brief Bioinform, 2020
2. A. Tran et al. *A machine learning-based clinical tool for diagnosing myopathy using multi-cohort microarray expression profiles.* Journal of Translational Medicine, 2020
3. Hie B et al. *Efficient integration of heterogeneous single-cell transcriptomes using Scanorama.* Nat. Biotechnol. 2019.
4. A tutorial on Single-cell RNA-seq clustering analysis: `https://hbctraining.github.io/scRNA-seq/lessons/07_SC_clustering_cells_SCT.html`