Deep learning foundation models applied on bulk and single-cell sequencing data for cellular state classification

Environment: QARMA Team at Laboratoire d'Informatique et Systèmes (LIS)
Location: Ecole Centrale Méditerranée (ECM), Technopôle de Château-Gombert, 13013 Marseille
Supervisors: R. Sicre (LIS – ECM), B. Habermann (IBDM – AMU), M. Haugland (LIS – IBDM)
Salary: legal minimum
Keywords: RNA-seq, deep learning, interpretability
Contact: ronan.sicre@lis-lab.fr

Large Language Models (LLM) and more generally foundation models have a very large impact in the deep learning research community. These models are now applied on more diverse fields and particularly several transformer-based models, inspired from LLMs, are applied to sequencing data [1]. Such models can already differentiate cells from different cancer types, for instance [2].

In this internship, we want to explore how such models can classify cellular states based on gene expression data. All cells contain a specific pattern of expressed genes and this pattern determines how the cell will behave: a nerve cell will have a different gene expression pattern than a heart cell, a cancer cell will show a different set of expressed genes from a healthy cell [4]. Deep learning techniques inspired from LLMs have recently emerged to classify cells based on gene expression data [1,2,3]. However, their lack of interpretability renders them difficult to use in this context and unable to identify genes and pathways that determine cell fate.

We want to take advantage of the large amount of gene expression data available from bulk- and single-cell sequencing to train and evaluate deep learning models for cellular state classification. We also seek to interpret these models' decisions in order to identify important genes regarding various categories. In a second part of the project, we want to adapt these models to a smaller gene set: the 1200 genes that are important for mitochondrial function. This will help us identify mitochondrial states in cells and therefore, its metabolic profile.

[1] Transformers in single-cell omics: a review and new perspectives, Szalata et al. (2024)
[2] scGPT: Towards building a foundation model for single-cell multi-omics using generative AI (Cui et al. 2024)
[3] Large-scale foundation model on single-cell transcriptomics (Hao et al. 2024)
[4] On knowing a gene: A distributional hypothesis of gene function (Kwon et al. 2024)