Master Internship offer
Statistical methods for Single Cell Genomics

Keywords: single cell analysis, high dimensional statistics, clustering, variable selection, optimal transport.

Single Cell Analysis

Until recently almost all of our current knowledge regarding genomes and their regulations was based on studies carried out at the population level, typically thousands to millions of cells averaged out in bulk experiments to reach sufficient molecular material. This averaging process, thought necessary, masked our view of genomic diversity and resulted in misrepresenting signals of molecular variations that vary between individual cells. Fortunately, recent technological advances in massively parallel sequencing and high-throughput cell biology technologies have paved the way for a better investigation of the suspected but inaccessible cell-to-cell variability of molecular profiles, based on DNA, RNA, chromatin states and conformation, or proteins. The use of these techniques, which we collectively refer to as single-cell genomics, allows the study of cell-to-cell variability within a biological sample and investigate new questions that were out of reach for classical bulk genomics. With the development of single-cell technologies, a modern cellular taxonomy is on its way, with the ambition to provide a comprehensive catalog of all types of cells within an organism, and to elucidate the molecular mechanisms underlying this diversity.

The methodological challenge is huge to analyze single cell data, since the technology now provides the distribution of gene expression among a population of cells. Consequently, our project is to propose distribution-based methods to catch, represent, and analyze the complex variability that structure population of cells. Linear methods such as PCA or k-means methods, are not necessarily the most efficient to analyze such data. Indeed, thanks to the development of high dimensional statistical learning, we known that the Euclidean distance is not suitable in the high dimensional setting, as large datasets possess an inner geometry that is far from the Euclidean one. Non-linear methods have emerged as a powerful alternative to standard PCA.

Project description

This internship will be dedicated to the investigation of non-linear dimension reduction methods, with a particular emphasis on Wasserstein-based methods, since the nice properties of Wasserstein distances are suitable to catch complex features in high dimensional spaces. The internship will also be focused on the development of methods to interpret the link between gene expression and the low-dimensional non linear representation of the data based on re-sampling strategies. These methods will be applied to better characterize population of lymphocytes that differentiate following a yellow fever vaccine shot (J. Mold, Karolinska Institutet Sweden). Applications will also concern the study of tumor clones of Multiple myeloma (CRCINA/CHU Nantes).

This internship is for master students in statistics/computational biology interested in methods developments and analysis of single cell datasets.

Details

Supervisors : B. Michel (ECN-LMJL, Nantes) and F. Picard (LBBE, Lyon)

Duration: 4 to 6 months (start: March-April 2019)

Location: Nantes, Laboratoire de Mathémtiques Jean Leray.

Skills: Master students in statistics or computational biology, experience in R is mandatory.

Contact: Bertrand MICHEL, Email: bertrand.michel@ec-nantes.fr