COMPUTATIONAL INFERENCE OF CELL DIFFERENTIATION DYNAMICS FROM HIGH-DIMENSIONAL BIG DATA

Supervisor: **Olivier Martin** **Co-supervisor: Thomas Blein**
mail: olivier.c.martin@inrae.fr mail: thomas.blein@cnrs.fr
Institute of Plant Sciences – Paris-Saclay, Bâtiment 630, rue de Noetzlin, 91405 – Orsay, France
Web: Team REGARN: Regulatory non-coding RNAs in root plasticity

INTERNSHIP DESCRIPTION

Single-cell transcriptomics [1], i.e., the quantitative measurement of abundances of RNAs simultaneously in thousands of individual cells, have generated a revolution in our knowledge of cell types and of their associated developmental trajectories that originate in stem cells and then branch out in a cascade of differentiations until terminally differentiated states are reached. Teams working in single-cell RNA-seq (scRNA-seq) have now developed intuitive computational tools [2,3] thanks to which they have constructed expression atlases in a number of different species [4,5]. Developmental trajectories for gene expression patterns can now be produced for the tens of thousands of genes in the genome, in both natural conditions and when these dynamical systems are perturbed using mutations or drugs. Although this descriptive approach is informative, it remains unfinished for two reasons. First, the algorithmic approaches need to be improved, for instance by relying on more modern inference methods including machine learning tools since these have not been explored much in the present context. Second, current work involves only transcriptomic quantification, so one does not know whether a gene is inactive because of an absence of activators or because the gene has been rendered "heterochromatic" (inaccessible to its activators because the DNA is too condensed). Our goal in this M2 internship is to address those two challenges, improving the computational methods and performing integration of data on DNA accessibility produced thanks to single cell ATAC-seq. The result will be a robust inference of developmental trajectories that will realize Waddington's "epigenetic landscapes" [6] with associated branching processes.

To mine scRNA-seq data, three main computational tools have been brought in from other fields [2,3]. The first is dimensional reduction, necessary because of the very high dimension (from 20,000 to 40,000, corresponding to the number of genes) needed to represent the transcriptomic profile of each cell. The dimensional reduction methods currently used [2,3] are PCA, t-SNE and UMAP. Using better methods, such as Minimum Distortion Embedding, would enhance the algorithmic performance. The second is unsupervised clustering to group cells of comparable expression profiles, leading one to identify different cell types. The algorithms most used for this are referred to as Louvain [7] and Leiden [8]; they come from statistical physics and are based on community detection and have been extensively used in research communities working on Complex Networks. The great *flexibility* of these algorithms has never been exploited in the scRNA-seq field, leaving significant opportunities for quantitative-minded researchers. The third is the construction of a developmental trajectory within a cluster. It turns out that the field is not very mature regarding this particular challenge. For instance, a major defect of the scRNA-seq analyses to date is that they perform trajectory inference after the clustering rather than simultaneously. Ideally the synergistic identification of clusters and trajectories should allow one to generate reliable developmental trajectories with branching in the form of a tree, going from pluripotent stem cells (the tree's root) all the way to terminally differentiated cells (the tree's leaves) with stochastic transitions between different intermediate cell types.

In addition to developing better inference approaches, it is necessary to integrate DNA condensation data into these analyses because this mechanism of turning off of genes is omnipresent in cellular differentiation. To do so, we will rely on single-cell ATAC-seq data sets that are now becoming widely available across different organisms. The integration will require crossing the two different types of profiles (transcriptomic and "epigenetic") and inferring the association between DNA accessibility and expression of the downstream gene(s), leading to a complete description of the branching differentiation cascade. With this work, we will provide the conceptual and computational tools to generate state of the art "atlases" of cellular differentiation. Biologists will be able to use these to unravel the genetic and molecular processes driving these complex dynamical systems while modelers will study them to reveal general principles of gene regulation [9].

TECHNIQUES USED DURING THE INTERNSHIP

The computational work will be done at least partly in the language R to be able to exploit the Seurat [2] and Monocle [3] packages that are easy to use and allow the identification of cell types and pseudo-time trajectories. The work will require working with high dimensional data and exploiting algorithms for dimensional reduction, community detection, and trajectory inference. This will involve introducing novel inference methods, testing algorithmic choices, quantifying inference reliability and interacting with biologists for confirming the inferred complex cellular dynamics on publicly available datasets.

This M2 project forms a stepping stone for integrating more heterogeneous datasets, a topic that will be considered in the follow-up doctoral work (expected funding from the LabEx SPS). The goal there will be to integrate single-cell/nucleus and bulk measurements and to compare the topologies of the regulatory networks across homologous organs and across species, of importance from an evolutionary perspective and for real world applications such as improving plant architecture.

REFERENCES

[1] Chen et al. (2019), Single-Cell RNA-Seq Technologies and Related Computational Data Analysis: https://doi.org/10.3389/fgene.2019.00317

[2] Seurat computational tools: https://satijalab.org/seurat/index.html

[3] Monocle computational tools: http://cole-trapnell-lab.github.io/monocle-release/

[4] Mouse Organogenesis Cell Atlas. https://oncoscape.v3.sttrcancer.org/atlas.gs.washington.edu.mouse.rna/landing

[5] Plant sc-Atlas, https://bioit3.irc.ugent.be/plant-sc-atlas/root

[6] Waddington, C.H. *The Strategy of the Genes*. London: Geo Allen & Unwin, 1957

[7] Blondel et al. *Fast unfolding of communities in large networks*. *J. Stat. Mech. Theory Exp.* **10008**, 6 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008

[8] Traag et al. *From Louvain to Leiden: guaranteeing well-connected communities*. *Sci Rep* **9**, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

[9] Subbaroyan et al. *Minimum complexity drives regulatory logic in Boolean models of living systems*. *PNAS Nexus 1 (1), pgac017 (2022). https://doi.org/10.1093/pnasnexus/pgac017*