LEARNING HIDDEN DYNAMICAL RULES: INTEGRATIVE CAUSAL INFERENCE IN GENE REGULATORY NETWORKS

Supervisor: **Olivier Martin**
mail: olivier.c.martin@inrae.fr

Co-supervisor: Thomas Blein
mail: thomas.blein@cnrs.fr

Institute of Plant Sciences – Paris-Saclay, 91405 – Orsay, France

Web: Team REGARN: Regulatory non-coding RNAs in root plasticity

INTERNSHIP DESCRIPTION

Regulatory processes are ubiquitous in living systems, ensuring homeostasis, adaptive responses and proper development. Their operation involves the coordinated action of multiple molecular components that are often assembled into "Gene Regulatory Networks" (GRN) [1]. The framework to describe such networks is sometimes reduced to its simplest form, for instance via a list of pairwise interactions as provided in transcriptional co-expression networks. More sophisticated frameworks seek to infer the "who controls who" and also the associated molecular mechanisms, thereby providing biological insights within causal modeling. The most direct way to address this causality issue is to examine the system's evolution in time, either following a perturbation or as part of a spontaneous developmental process. For instance, upon induction or inhibition of a gene or some cellular machinery process, it is possible to generate a time series of the transcriptional response to such a perturbation and thereby follow the cascade of changes. The order of these transcriptional changes informs us on the possibilities for "who controls who" from which one can infer candidate GRNs. In the past few years single-cell RNA sequencing [2] has seen huge advances due to improved technologies but also analysis methods, some coming from physics [3]. This approach thus provides high throughput transcriptomic cascades in particular in the context of developmental programs of tissues, organs and even whole organisms.

A number of computational approaches have been published for exploiting these transcriptomic time series and machine learning tools just beginning to be used [4]. The first challenge of this internship will be to integrate the results coming from different algorithms to enhance predictive power. The second challenge will consist in integrating multiple transcriptomic datasets, again to improve the reliability of the inference and to shine light on the molecular mechanisms.

The *tasks* specific to the M2 internship can be broken down as follows:

(1) Building a simulator of GRNs from which *in silico* time series will be produced. (Real biological networks remain quite uncertain and so cannot be used to validate the inference algorithms.)

(2) Quantify the accuracy of inference of 3 published algorithms (SCODE, GENIST, TDCor) on simulated data. (These algorithms are among the most used ones for biological studies.)

(3) Design, test and benchmark strategies to integrate GRN inference using these 3 algorithms. (In particular we want to quantify the gains achieved by algorithmic integration and provide insights into the strenghts and weaknesses of each algorithm.)

(4) After optimizing this *algorithmic integration* methodology, apply it to experimentally measured time series, in particular ones from our team that follow the time dynamics of root formation in *Arabidopsis*.

(5) Develop *data integration inference* methodology whereby several transcriptomic datasets are to be exploited. Specifically, we have time series for root development in Wild Type and in several mutants. The goal is to provide a methodology to infer GRNs across these different genotypes given that they share almost all the same components of an underlying network.

We plan to extend this project and continue the work via a Ph.D. It will consist in building on these first developments to tackle *multimodal* dataset integration. Indeed, transcriptomics provide just one window into a system's response; other modalities such as ChIP-Seq and ATAC-Seq are generally highly complementary. Although the very different nature of these datasets remains a major challenge today for their integration, succeeding in doing so holds the promise of being a game changer in GRN inference.

TECHNIQUES USED DURING THE INTERNSHIP

The internship period will lead to a mastering of different methods from statistical inference and machine learning. The production and management of computational runs will be organized within the Linux OS.

REFERENCES

[1] de Luis Balaguer et al. (2017), Predicting gene regulatory networks by combining spatial and temporal gene expression data in *Arabidopsis* root stem cells. PNAS. https://doi.org/10.1073/pnas.1707566114

[2] Wendrich et al. (2020), Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions. Science. https://doi.org/10.1126/science.aay4970

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

[4] Ding et al. (2020), Analysis of time-series regulatory networks. Curr. Opin. Syst. Biol. https://doi.org/10.1016/j.coisb.2020.07.005