I. **Titre du Projet en anglais** : Statistical inference of tomato gene regulatory networks under climate change

II. **Description du projet en anglais** (1 page maximum, Arial 10)

Elevation of atmospheric CO2 concentration is a process that will continue for a large period of the century, leading to predictable CO2 levels between 750 to 1000 ppm around 2100. Independently of consequences on climate change, elevated atmospheric CO2 level (eCO2) will greatly impact plant growth, development and physiology. Indeed, eCO2 leads to a degradation of plant nutritional quality, with most plants growing under eCO2 showing a decline in their nutritional content, especially nitrogen (N) and iron (Fe). As plants are a major source of N and Fe in human food, the drastic decrease of their content under **eCO2 represents a major threat for human health in the coming decades**.

We recently generated several transcriptomic datasets from plants in the context of eCO2, in order to understand the effect of eCO2 on plant physiology, and to identify candidate genes or pathways able to improve plant nutritional content under eCO2. Using a Machine-Learning approach, we reconstructed gene regulatory networks of Arabidopsis under eCO2, which allowed us to identify and experimentally validate in this model plant several genes involved in biomass and nutritional content under eCO2. In order to explore these responses in a major crop, we have now generated a transcriptomic dataset under eCO2 in tomato. This combinatorial dataset consists of contrasted growth conditions in CO2, iron and nitrate levels, for a total of 24 RNA-seq transcriptomes.

The main objective of this internship is to analyze the transcriptomic dataset underlying the tomato response to eCO2, in order to identify in this species genes or signaling pathways involved in the interaction between eCO2 and nutrition. More specifically, the M2 student will have to :

-**process the 24 RNAseq samples** using standard mapping, normalization and counting tools to the tomato reference genome, and identify differentially expressed genes (DEGs) in response to eCO2.

-**perform a clustering analysis** on DEGs to identify the most relevant patterns inside the transcriptomes.

-**reconstruct, using selected Machine-Learning network inference approaches, the regulatory network associated with the response of eCO2** under nutrient starvation,

-evaluate the robustness of the inferred network(s) and possibly **consider the combination of several inference approaches through a methodological development**

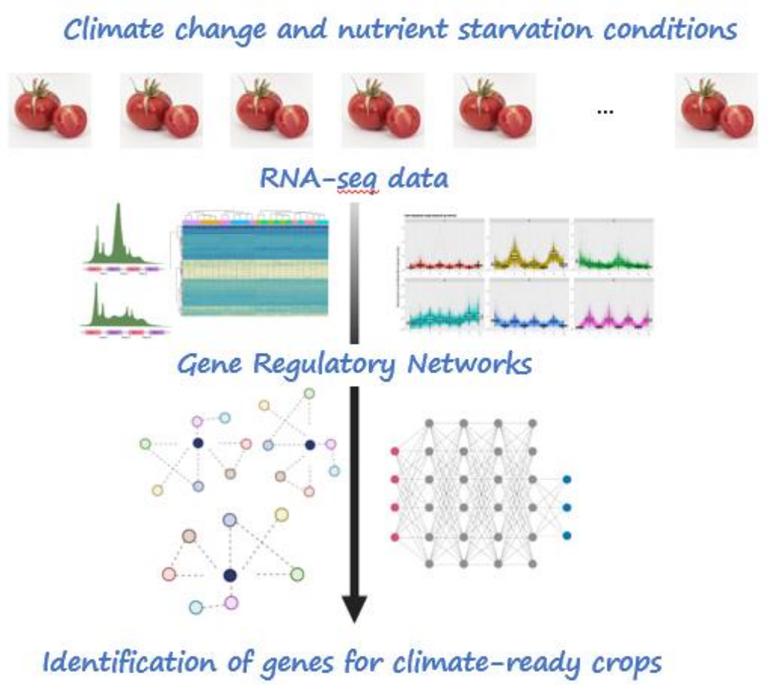
-**identify candidate genes from the inferred network(s)** (topological clustering, connectivity, analysis of associated functions, etc.).

- perform a **comparative analysis of regulatory networks** between Arabidopsis and tomato.

This work will be performed in the frame of a collaborative project between biologists and mathematicians that started 2 years ago, involving 2 researchers and 1 third-year PhD student. This successful collaboration recently led to the publication of a research paper introducing a complete and interactive suite for the inference and analysis of gene regulatory networks, and is currently the object of the writing of a second paper around plant's response to eCO2.

This work will be based on manipulation of bioinformatics tools, statistical analysis of genomic data, machine-learning for network inference, visualization of large datasets, and will possibly involve experimental work with biological samples. This internship is addressed to students with a strong motivation to work in an interdisciplinary environment, with a formation in bioinformatics, statistics and/or biology.

III. GRAPHICAL ABSTRACT EN ANGLAIS (facultatif) :


Climate change and nutrient starvation conditions

RNA-seq data

Gene Regulatory Networks

Identification of genes for climate-ready crops

IV. Description **précise,** en anglais, du programme de travail de l'étudiant

(taches, techniques utilisées) (1/2 page maximum, Arial 10)

The student will be first in charge of implementing a pipeline to analyze RNA-Seq data, from the raw fastq files to gene expression levels, and from gene expression levels to biological hypotheses supported by statistical analyses. To that end, the bioinformatic workflow consisting in quality control, reads mapping to the reference genome, and genes quantification will be performed using mainly command line software such as fastQC/fastp, STAR/bowtie, htseq-count. Following the steps proposed in the DIANE package, the student will leverage state of the art statistical procedures in R (and the RStudio IDE) to normalize the expression data, and establish the global behavior of gene expression changes (visualisation, dimensionality reduction). Differential expression analyses and the exploration of their outputs will be performed. (TCC, EdgeR, DESeq2, ggplot2, plotly, clusterProfiler packages).

The clustering of genes will be carried-out and tuned using the Coseq R package, relying on Gaussian/Poisson mixture models. Then, **a great interest will be taken in regulatory pathways reconstruction from expression values**. To do so, techniques based on regressions between regulators and target genes will be considered, such as GENIE3 or TIGRESS, employing respectively **Random Forests or regularized linear regressions** to extract the strongest regulatory interactions between genes. As pointed-out at the occasion of the DREAM5 challenge, combining predictions from several inference strategies often leads to more robust and accurate results. As a consequence, the student will be encouraged to **test, benchmark and even combine different gene regulatory network methods**. The inferred networks will then be the object of a topological study to identify candidate master regulators that could be later experimentally tested. Finally, to compare the results from this study on tomato to previous findings on Arabidopsis, comparative genomics tools will be employed.