WP1: Machine learning, Visual Analytics and Network Bioscience Automation Leaders: M. Ghoniem, M. Elati

Participants: UL: W. Dhifli, J. Puig, A. Dispot, Post-doc; LIST: PhD 1; CIRAD: C. Périn, PhD 2, G. Droc

General objectives: we propose to rationalize “root regulatory network biology” through a computational cycle composed of three building “blocks”: the inference (data-learn), interrogation (visual analytics-simulation) and intervention (design-test) with regulatory networks. Network inference: to learn hypotheses (i.e., networks) from observations (i.e., data). Network interrogation: to visualize and analyze the most important patterns in terms of structure and the multi-dimensional (tabular) experimental data associated to network elements, and to make predictions about the behavior of targeted system. Network intervention: design regulatory circuits and select functional experiments.

Success indicators: Optimized GRN inference and visualization software is available.

Task 1.1 GRN inference and mapping with tissue specificity

This task is devoted to the inference and systematic analysis of the network to obtain maximal insight from the reconstructed biological networks, investigating specific types of biological issues, such as: the quantification of regulator activity and the identification of key regulators (CoRegNet (Nicolle et al. 2015)); the identification of cooperative transcription factors and the study of network cross-tissues conservation/modification. First, all public data and RNAseq for partner 1 will be integrated. We will identify rice network regions that differ between different tissues. Using an integrated regulation reference model and gene expression in different tissues or conditions, we will identify target genes with patterns of unexpected behavior on the basis of the expression of their regulators (Picchetti et al., 2016, Dhifli et al. 2019). This perturbation model will also make it possible to select key regulators by counting and ranking the minimal sets of regulators accounting for tissue specificity. This enumeration problem could be addressed using a multi-objective hypergraph-covering problem in a discrete framework as well as multi-start solutions in a linear regression approach. Efficient heuristics will be developed to address this problem. For each of the minimal sets of regulators, we will extract potential driver pathways by simultaneously maximizing ontological enrichment and subgraph connectivity based on inferred cooperative regulatory networks (Winterhalter et al., 2014, Dhifli et al., 2020).

Task 1.2 Interactive visualization of multilayer multivariate networks: HivePlot

From a network visualization perspective, many universal and biology-related tools employ node-link diagrams. In this proposal we will develop a multilayer network visualization (McGee et al., 2019). Unlike traditional node-link representations, network nodes are laid out in distinct spatial layers such that between-layer links can be distinguished from within-layer links. Layers may be defined in very creative task-specific manners. By default, layers may coincide with the omic type of the nodes, e.g. a layer for genes, another one for transcription factors, etc. To support the comparative analysis of networks corresponding to different conditions or tissues (Task 1.1), we will test the design consisting in using distinct layers for distinct conditions or tissues. Nodes in the tissue or condition layers may be further grouped based on their omic type. The semantic of links between layers depends also on the task at hand. In some cases, links have a biological meaning such as the up/down-regulation links between TFs and genes. In other cases, links have a utility function such as matching nodes occurring in several layers. This task will identify the right visual encodings to support comparative analyses of models during the model tuning phase (Task 1.1), as well as in “production mode” during the hypothesis generation/verification work as carried out by biologists. More precisely, we will reuse and extend an existing hive plot visualization developed by partner 3 in the context of the ANR/FNR

funded BLIZAAR project (2016-2019). Unlike traditional node-link diagrams, the layout stability of hive plots and the spatial separation of node types is a key quality for network comparisons. The extensions include testing the hive plots with state-of-the-art strategies to see changes in dynamic networks, such as multiple side-by-side visualizations, animation of changes and differential network views, and testing the use of chained biclustering algorithms between pairs of layers of the hive plot visualization to facilitate the identification of common and unshared features.

Task 1.3 Visualization of GRN data using the iCoVeR Tool

The iCoVer tool of partner 3 is an interactive visualization tool, originally built for the analysis of metagenomic contigs at different experimental conditions. It has since been generalized to analyze any large multivariate data set. It combines scatter plots and parallel coordinate plots, as well as data clustering and dimensionality reduction algorithms. Advanced user interactions allow the analyst to identify and select patterns of interest and subset the data accordingly. The tool was successfully used to analyze multivariate biological data, including metagenomic data (Broeksema et al., 2017), and transcriptomic data (Guerriero et al., 2017), but it does not include any network visualizations yet. In this project we will extend the tool to allow the analysis of multivariate network data. Low hanging fruits include the use of parallel coordinate plots as means to filter the network interactively by defining value ranges on node or link attributes, in order to focus on the most interesting network elements. The same interactive filtering mechanism will also be applied to higher-level constructs such as network motifs rather than simple network elements. This is very useful because all motif extraction algorithms (Task 1.1) generate many false positives, that need to be filtered out using expert knowledge.

An important contribution of the proposed research consists in the identification of an analytical workflow supporting biologists in reasoning about the simulated GRN networks, where simulation and visualization steps may alternate several times until actionable insights are found. The formalization of such a workflow will help identify gaps in the current state of affairs and help set a research agenda to close these gaps at the community level. The task will also deliver an interactive web-based data visualization software supporting such an analytical workflow. It will adopt a multiple coordinated visualizations design to support the exploratory analysis of the GRN network models (Task 1.1). Since the network models are inferred from high-dimensional expression profile data, the analyst needs to visualize both the network structure and the associated expression profile data. The combination of network visualizations and multivariate visualizations in one tool is expected to provide a powerful/expressive exploratory analysis solution.

Task 1.4 GRN intervention and selecting experiments

In this task, we aim to kick start the computational cycle by making use of knowledge generated by inference and interrogation(Task 1.1-3) to redesign of regulatory circuits and select functional experiments. This task requires to i) introduce a taxonomy of abstract features and particular standardized biological parts that preserve the relationships between features prescribed by the network, including explanation and guidelines from scientists (e.g. constraints, cost/add value of experiments). and ii) the development, adaptation, and combination of state-of-the-art active learning and experimental design algorithms (Coutant et al., 2019, Bouyioukos et al., 2016) to reason with high-level concepts, making abstraction where possible of overly specific or unobserved details (e.g. latent representations, higher-order graphs, etc.). We will extend ML tools with the capability to explain how hypotheses were produced, how experiments were selected and designed, why models they need/were revised, what hypotheses were confirmed, and which were discarded and why.

WP2: Engineering and deciphering transcriptional programs using CAS9 technologies

Leader: C. Périn (CIRAD, FR) Participants: CIRAD: Périn C., Meuiner AC. Navarro S., PhD 2, M. Bes; UL: M. Elati, A. Dispot, Post-doc; LIST: M. Ghoniem, PhD 1

General objectives: Validate and deciphering candidates TF for root cell differentiation using protoplast, CAS9 derived technology and functional analysis of the most promising candidate in planta

Success indicators: Constructs validated on protoplasts for over and under expression of TF(s); Validation of dCAS9 scaffold strategy on rice expressed genes; Validation of new TF candidates (5) for aerenchyma differentiation using dCAS9; In planta functional analysis of TF (3) involved in aerenchyma differentiation

Task 2.1 Building of a synthetic gene network control system derived from CAS9

This technology is based on the use of one or more scaffold RNAs, a dCas9 and binding-effector modules. For the duo RNA Scaffold/protein binding and effectors, several choices were possible. We have chosen to use the Scaffold MS2, PP7 and COM RNAs which are recognized by the viral proteins MCP, PCP and COM respectively (Zalatan et al., 2015). At the effector level, we have chosen to use the VP64 activator and the 3xSRDX repressor which have already been successfully tested in plants (Lowder et al., 2015). These effectors will be merged directly with dCas9. All proteins are transcriptionally fused to an NLS (Nuclear Localisation Domain). These different elements will be reconstructed in-silico and then synthesized and cloned in binary vectors. We first demonstrated functionality of the system using a fluorescent reporter (see Figure 4). All these constructions will be tested first in protoplasts with VDN6 and CAF-1 as control target genes for their ability to induce their targets into rice protoplasts by ddRTPCR (Guiderdoni et al 1992, Bes et al 2020 in press). Finally, the system's ability to induce TF effectors by combining multiple TFs will be tested by inducing simultaneous expression of these genes and using targets expression as a proxy.

Task 2.2 Assess and order candidate FTs for root aerenchyma

TF candidates from WP1 will be tested for their ability alone or in combination to induce cellular differentiation of each tissue, by analyzing by ddRTPCR the induced or suppressed primary targets. These experiments will allow us to order the TF network and these data will be added to the learning algorithms and through interactive annotation tools that have been developed (WP1). Cell autonomous and non-cell-autonomous TFs, mainly acting locally or remotely, will be predicted by comparing their influence and the presence or absence of their transcript in the tissue they are supposed to function. To sum up, a TF that has a great influence on aerenchyma differentiation but not transcribed in this tissue, is a non-cell-autonomous TF candidate, acting remotely. Validated or unvalidated data will be injected into gene network databases and gene networks will be iteratively reconstructed using machine learning algorithms (WP1, partners 2 & 3).

Task 2.3 Functional analysis of 3 TF candidates for aerenchyma differentiation in planta.

We will analyze the in planta function of 2-3 TF candidates, involved in the differentiation of the aerenchyma. Loss of function lines will be generated by CRISPR/CAS9, a well-established technology in partner 1's laboratory or loss of function lines for more than 30 genes have been obtained through this technology in the laboratory (see for instance (Nieves-Cordones et al 2017; Herbert et al Rice 2020)). Over-expressor lines and promoter lines: GFP and promoter lines: TF: GPF will also be established to confirm the expression profile and mode of action of the non-cell-autonomous and cell-autonomous TF. We have developed dynamic multiphoton imaging technology on rice root tips (Yang et al 2017 and Bureau C. et al Plant methods 2018) to follow changes at the cellular scale in mutants and in their cellular identity. Finally, the expected defects in mutants will be analysed in situ using specific dyes and a tissue clearing approach (Ursache et al 2018) which we also adapted to rice roots.

The set of tools for synthetic control of gene networks will be sufficiently generic to be used to explore other gene networks in all model and cultivated plants, for example to explore and validate new metabolic pathways. The project will bring together three laboratories with complementary expertise, including a plant molecular genetics laboratory (CIRAD), a systems and computational biology laboratory specialized in gene networks constructions (UL), and a computer laboratory expert in network visualization and exploratory data analysis (LIST). Each of these teams come with specific skills to develop comprehensive tools for the construction, operation, exploration and validation of plant gene networks. The rice root gene network for aerenchyma tissue differentiation is the perfect model to develop this strategy as it involves complex mechanisms of regulation with numerous molecular partners.

References

Bes, M., Herbert, L., Mounier, T., Meunier, AC., Durandet, F., Guiderdoni, E., Périn, C. (2020 in press in Methods in Molecular Biology) Efficient genome editing in rice protoplasts using CRISPR/CAS9 construct.

Bouyioukos, C., Bucchini, F., Elati, M., and Kepes, F. (2016). GREAT: a web portal for Genome Regulatory Architecture Tools. Nucleic Acids Res 44: W77-82.

Broeksema, B., Calusinska, M., McGee, F., Winter, K., Bongiovanni, F., Goux, X., Wilmes, P., Delfosse, P. and Ghoniem, M., 2017. ICoVeR–an interactive visualization tool for verification and refinement of metagenomic bins. BMC bioinformatics, 18(1), p.233.

Coutant A., Roper K., Trejo-Banos D., Bouthinon D., Carpenter M., Grzebyta J., Santini G., Soldano H., Elati M., Ramon J., Rouveirol C., Soldatova L., King R. D. Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. PNAS,116 (36) 18142-18147, 2019.

Dhifli, W., Puig, J., Dispot, A., Elati, M. (2019). Latent network-based representations for large-scale gene expression data analysis. BMC bioinformatics, 19(13), 466.

Dhifli, W., Karabadji, N. E. I., & Elati, M. (2020). Evolutionary mining of skyline clusters of attributed graph data. Information Sciences, 509, 501-514.

Guerriero, G., Behr, M., Legay, S., Mangeot-Peter, L., Zorzan, S., Ghoniem, M. and Hausman, J.F. (2017). Transcriptomic profiling of hemp bast fibres at different developmental stages. Scientific reports, 7(1), p.4961.

Herbert, L., Meunier, AC., Bes, M., Vernet, A., Portefaix, M., Durandet, F., Michel, R., Chaine, C., This, P. Guiderdoni, E., and Périn, C. Beyond Seek and Destroy: how to Generate Allelic Series Using Genome Editing Tools. Rice volume 13, Article number: 5 (2020)

Lowder, L.G., Zhang, D., Baltes, N.J., Paul, J.W., 3rd, Tang, X., Zheng, X., Voytas, D.F., Hsieh, T.F., Zhang, Y., and Qi, Y. (2015). A CRISPR/Cas9 Toolbox for Multiplexed Plant Genome Editing and Transcriptional Regulation. Plant Physiol 169:971-985

McGee, F., Ghoniem, M., Melançon G., Otjacques B., and Pinaud B. (2019), "The State of the Art in Multilayer Network Visualization," Computer Graphics Forum, vol. 38, no. 6, pp. 125-149.

Nicolle, R., Radvanyi, F., and Elati, M. (2015). CoRegNet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics 31:3066-3068.

Picchetti, T., Chiquet, J., Elati, M., Neuvial, P., Nicolle, R., & Birmelé, E. (2015). A model for gene deregulation detection using expression data. BMC systems biology, 9(S6), S6.

Winterhalter, C., Nicolle, R., Louis, A., To, C., Radvanyi, F., and Elati, M. (2014). Pepper: cytoscape app for protein complex expansion using protein-protein interaction networks. Bioinformatics 30:3419-3420.

Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L.A., Whitehead, E.H., La Russa, M., Tsai, J.C., Weissman, J.S., Dueber, J.E., Qi, L.S., et al. (2015). Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. Cell 160:339-350.