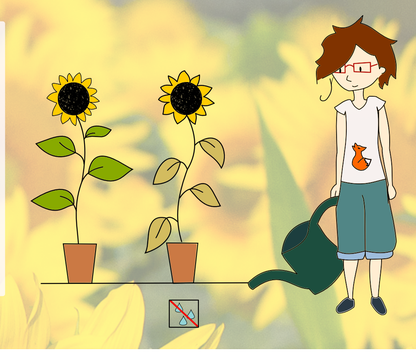
Building artificial genetical genomic datasets to optimize the choice of gene regulatory inference methods

Lise Pomiès[1], Louise Gody[2], Charlotte Penouilh-Suzette[2], Nicolas Langlade[2], Brigitte Mangin[2], Simon de Givry[1]

1 Unité de Mathématiques et Informatique Appliquées de Toulouse
MIAT INRA - UR875 - Chemin de Borde Rouge, 31320 Castanet Tolosan, France

2 Laboratoire des interactions plantes micro-organismes
LIPM INRA - UMR2594 – Chemin de Borde Rouge, 31320 Castanet Tolosan, France

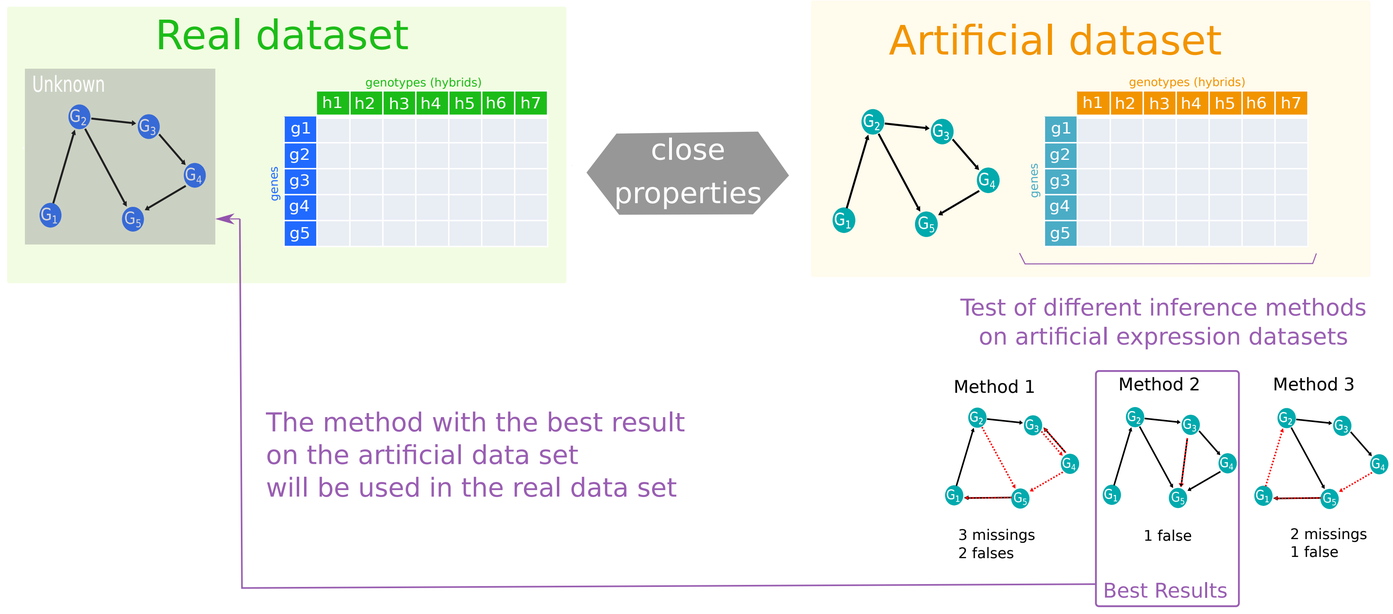


We study the **transcriptomic response of sunflower** to drought combined to the heterosis phenomenom, across **180 gene** expressions on **400 hybrids genotypes**, coming from a pool of 72 parents. SNP present on the parental genomes were measured.

Our goal is to **infer** the **gene regulatory network** among those genes. However, because of the **non-independency** of the data, accuracy of inference results is unpredictibl. Therefore, we need to test different methods, to select the best inference method for our biological question.

Select inference methods based on artificial datasets

Test different methods on an **artificial dataset** with known network, expression levels and genotypes (DNA variant). **Biological properties** of this artificial dataset must be closed to the properties of the real dataset.



How to build an artificial dataset with the same biological properties as our real one?

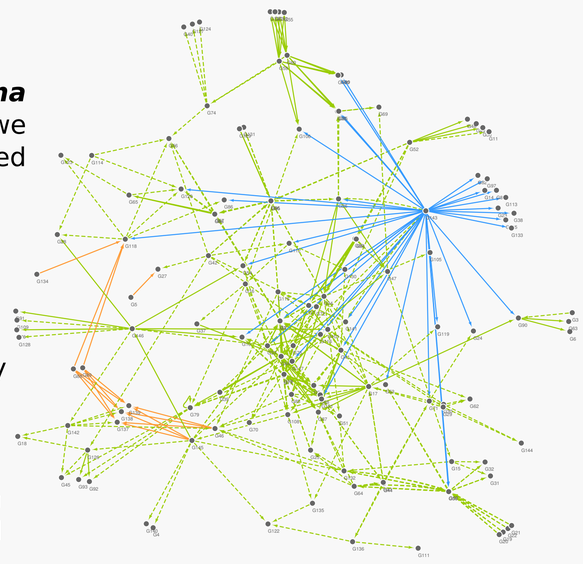
1. Build artificial network

based on real biological information available for the same biological process, on a close organism

For the **homologues** on ***A. thaliana*** of our 180 genes of interest, we selected **regulations** described between them in 3 databases.

AtRegNet [1]
AtPID [2]
PlantRegMap [3]

Regulations can be supported by experiments or predicted (----)



Our artifical network based on biological information is composed of 137 genes linked by 364 edges

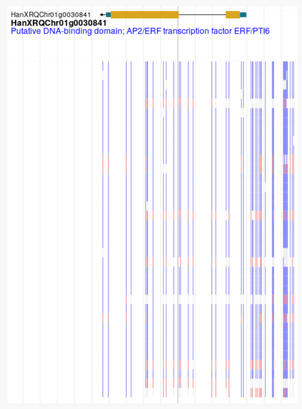
2. Create artificial hybrid genotypes

based on genomic information available for the real hybrids used in the experiment

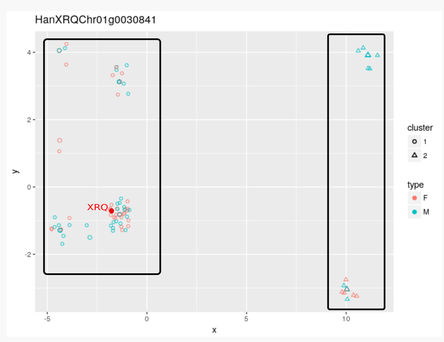
SNP for each **parental genotype** are associated to a score 0 if it is like in XRQ-line or 1 if different. The hybrids SNP are obtained by combining locus-per-locus SNP of their parents.
We created **artificial hybrid**s associated to **DNA variant** on each measured gene. We considered **one variant per gene**, those DNA variants are based on SNP of the real data.

Ex : gene *HanXRQChr001g0030841*

A. Selection of SNP on each parental genotype



B. K-medoid clustering
Manhattan distance on the SNP data



Genotypes are classified in 2 groups
Cluster with XRQ - DNA variant score = 0
Other Cluster - DNA variant score = 1

C. Artifcial hybrids
1 DNA variant per gene

DNA variant of hybrid
= mean score of it parents

3 possible value per gene

0	homozygous = XRQ variant
1	homozygous ≠ XRQ variant
0.5	heterozygous

Collection of hybrid genotypes, with known DNA variations on our genes of interest.

3. Select and adapt an existing gene expression simulator

emulating the same type of experiment that the one we performed, with steady state measurements on different genotypes

SysGenSIM simulates steady state gene expressions using ordinary differential equations. Simulation is based on a gene network topology and DNA variant for each gene.
Work only on RIL (both allele of a gene are identical) [4]

$$\frac{dG_g}{dt} = Z_g^c \cdot V_g \cdot \theta_g^{Syn} \cdot \prod_k \left(1 - A_{kg} \frac{G_k^{h_{kg}}}{G_k^{h_{kg}} + \left(K_{kg}/Z_k^t\right)^{h_{kg}}}\right) - \lambda_g \cdot \theta_g^{deg} \cdot G_g$$

basal transcription rate of gene g
SNP(c) noise
role of gene k on gene g ∈ {−1; 0; 1}
[mRNA] of gene k
degradation rate constant of gene g
degradation noise
effect of gene g SNP(c) on gene g expression
for each gene k
min [mRNA] of gene k for k to have an effect on gene g
effect of SNP(t) of k on it activity (more or less efficient regulator)
[mRNA] of gene g

We **modified the simulator** to use our **heterogenous hybrids**, and mimetized the allelic dominance caused by the **heterosis phenomenon**.

SysGenSIM
Z parameter **2** possible **values**

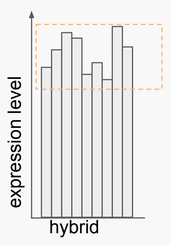
wt-wt	1
m-m	0.75

Modified SysGenSIM
Z parameter **3** possible **values**

wt - wt	1	
m - m	0.75	
wt - m	0.75 mutated dominance 0.87 additif effect 1 wt dominance	(10%) (80%) (10%)

4. Comparison of biological score

obtained on real and simulated datasets (in our case the heritability score) to adjust parameters of the simulator



heritability

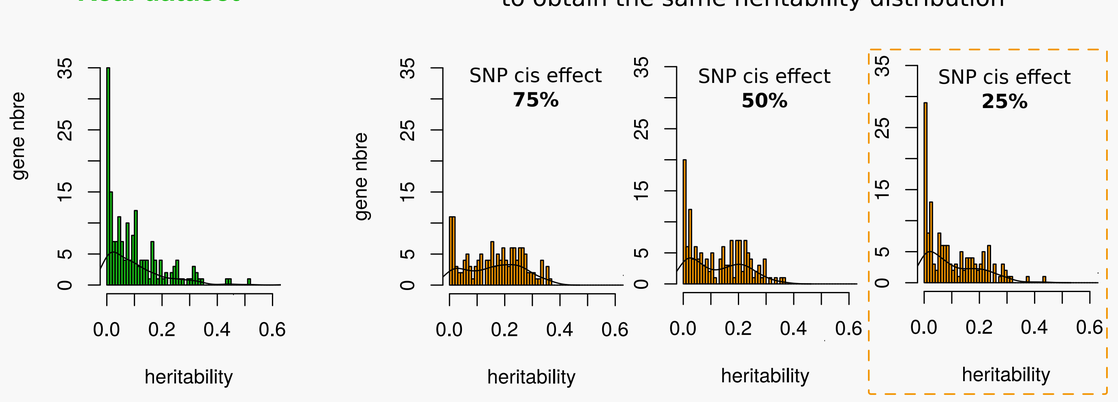
% of expression variance of a gene explained by its parent genotypes

We use a mixed model to estimate the heritability [5]

Real dataset

Artificial dataset

SysGenSIM parameters adjusted to obtain the same heritability distribution



The artificial dataset produced have the same biological properties as our real dataset. We can now test different methods of network inference and test the accuracy of these methods by comparing networks inferred by the algorithms to the artificial network. Network inference methods with the best results will be used on the experimental dataset to answer our biological question.

Funded by the SunRise project (2011-2019) : *http://www.sunrise-project.fr*

[1] Li et al. (2011). *AtPID: The overall hierarchical functional protein interaction network interface and analytic platform for arabidopsis*. Nucleic Acids Research, 39(SUPPL. 1), 1130–1133.
[2] Palaniswamy et al. (2006). *AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks*. Plant Physiology, 140(3), 818–829.
[3] Jin et al. (2017). *PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants.* Nucleic Acids Research, 45(D1), D1040–D1045.
[4] Pinna et al. (2011). *Simulating systems genetics data with SysGenSIM.* Bioinformatics, 27(17), 2459–2462.
[5] Mangin et al. (2017). *Genomic Prediction of Sunflower Hybrids Oil Content*. Frontiers in Plant Science, 8, 1–12.