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

Lise Pomiès*[1], Louise Gody[2], Charlottte Penouilh-Suzette[2], Nicolas Langlade[2], Brigitte Mangin[2], and Simon De Givry†[1]

[1]Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT INRA) – Institut National de la Recherche Agronomique : UR875 – Chemin de Borde Rouge, 31320 Castanet Tolosan, France
[2]Laboratoire des interactions plantes micro-organismes (LIPM) – Institut National de la Recherche Agronomique, Centre National de la Recherche Scientifique : UMR2594 – Chemin de Borde-Rouge - BP 27 31326 CASTANET TOLOSAN CEDEX, France

Résumé

One of the central targets of Systems Biology is to decipher the complex behavior of a living cell in its environment. A gene regulatory network is a simplified representation of the gene-level interactions. Network inference methods are powerful tools to understand such complex biological processes [1]. However, it could be difficult to identify an algorithm adapted to a specific experimental dataset. Artificial datasets could be used to test different algorithms of inference and select the most accurate one. But, available artificial datasets are more like ideal datasets and consequently quite different from measured datasets. In our case, we didn't know if classical network inference algorithms (like bayesian network, mixed model, penalised regression, random forests, or a combination of them [2]) will work correctly on our dataset. We also didn't find an artificial dataset with the same properties as our experimental data. This is why, we decided to create our own artificial dataset. We present here the characteristics of our experimental dataset and the strategy we elaborate to create an artificial dataset with the same properties as our experimental dataset.
We work on domesticated sunflower (Helianthus annuus), a highly resistant crop plant to drought. The sequencing of the genome of the XRQ line of sunflower, had been published last year [3]. In the context of climate changes it's interesting to understand how sunflower resists to drought at the molecular level and how this resistant interacts with the phenomenon of heterosis when new varieties are created.

To answer this question, two transcriptomic experiments were performed. The goal of the first experiment was to select genes involved in response to drought and in the heterosis phenomenon. This experiment was performed on a hydric-control environment. Eight different parental genotypes of sunflower (4 males and 4 females) and their 16 hybrids were cultivated. The 8 parental genotypes are homozygous for all genes, their hybrids could be homozygous or heterozygous depending on the locus. Sunflowers were cultivated in two hydric conditions: (i) in drought condition and (ii) with sufficient water level. The expression levels of all genes were measured in both conditions and for all genotypes via RNA-sequencing. From those

*Intervenant
†Auteur correspondant: simon.de-givry@inra.fr

transcriptomics measurements we selected 180 genes responding to drought, heterosis and in interaction between drought and heterosis. Because we are focusing on gene regulation we chose transcription factors (detected by iTAK [4] and plantTFCat [5]) to compose the main chunk of our dataset.

The goal of the second experiment was to collect enough data, to performed a gene regulatory network inference, on the 180 selected genes. The experiment was conducted on a field with 435 hybrids created from 72 homozygous parental genotypes (36 females and 36 males) including genotypes from the first experiment. The expressions of the 180 selected genes were measured by qPCR (Fluidigm technology) for all the hybrids. For all parental genotypes, single-nucleotide-polymorphisms (SNPs) were detected against the reference genome of sunflower (XRQ line).

The experimental dataset contains the expression of 180 genes on 435 hybrids. The collected data are not independent as they come from different parental genotypes and their hybrids. We don't know the effect of this dependency between the genotypes on network inference algorithms. To measure the impact of a non-independent dataset on existing network inference methods, we have created artificial datasets to test the accuracy of the methods.

First step is to create an artificial network. We decided to collect informations about interactions between our 180 genes of interest in different public databases. As the sequencing of the sunflower genome is recent, really few informations are available on databases for this plant. For this reason, we decided to collect interactions between the homologous genes of our selected genes on the plant model Arabidopsis thaliana. For 7 sunflower genes, no homologous genes were found. The homologous genes are the nodes of our artificial network. We collected interactions from 3 databases (i) AtPID, a database specific to A. thaliana containing interactions between proteins [6], (ii) AtRegNet specific to A. thaliana containing regulations between transcription factors and target genes [7], and (iii) PlantRegMap a plant database containing regulations between transcription factors and other genes [8]. The three databases contain links found in the literature, or resulting from experiments (as Chip-seq experiments). The third database also contains predicted regulations via detection of binding motifs, on the promoter of target genes, recognized by specific transcription factors. We selected in these databases only directed links, corresponding to expression regulations, involving two genes from our selection. We collected 364 regulations (36 in AtPID, 16 in AtRegNet, and 312 in PlantRegMap), 62% of these regulations were predicted regulations. Those 364 regulations form the edges of our artificial network. The type of regulation (activation or repression of the expression) is known for only two regulations. In the database AtRegNet where the nature of regulations are described, 64% were activation of the expression and 36% were repression of the expression. We decided to randomly associate each edge of our network to a particular type of regulation with a probability of 64% to be an activation of the expression, the rest being a repression of the expression.

In our experimental dataset, each parental genotype has a list of SNPs, associated to a score either 0 if it is like in XRQ-line or 1 if different. It is easy to deduce the SNPs of the hybrids by combining locus-per-locus the SNPs of their parents. In order to study the effect of genetic polymorphism on gene expressions, we created new virtual hybrid genotypes associated to DNA variants on each measured gene. To simplify this analysis, we considered one variant per gene. To be closer to our biological variety we created this DNA variants based on SNPs of the experimental data. For each gene of interest, we collected the SNPs present in their genomic sequence and their promoter region for each parental genotype. Using a K-medoid clustering with a Manhattan distance on the SNP data, genotypes were classified in two groups. The group of genotypes with SNPs close to the SNP values of XRQ-line has a DNA variant score of 0, and the other group of genotypes has a score of 1, for this gene. For hybrids, the score of DNA variant on each gene is equal to the mean score of their parents. It can take 3 values: 0 or 1 if the hybrid is homozygous for this gene, or 0.5 if it is heterozygous. We now had a collection of hybrids, with known DNA variations on our genes of interest.

The third step is to produce artificial measures of expression for the selected genes. The data simulator SysGenSIM simulates steady state gene expressions for different genotypes using ordinary differential equations [9]. The simulation is based on a gene network topology and DNA variant for each gene. In their model, each gene has only one DNA variant. The DNA variant of a gene has either a cis-effect (meaning it influences the rate of transcription of the gene) or a trans-effect (meaning it modifies the efficiency of the gene regulation activity). The equation describing the accumulation of a gene transcript for a given genotype is composed of two parts. The first part of the equation describes the rate of expression of the gene, and the second part describes the rate of degradation of the transcript. The expression rate is modulated by the effect of the DNA variant of the gene and the expression of the regulators of this gene in the network. The DNA variant of the regulators have also an impact on the efficiency of the regulation. For the moment, SysGenSIM only works on recombinant inbred lines (RIL). We slightly modified the simulator to use our heterogenous hybrids, and mimetized the allelic dominance caused by the heterosis phenomenon. In case of genes with heterozygous DNA variant, the DNA variant effect is randomly chosen, with a probability of 0.8 to be an additive effect of the DNA variant effect of both parents and a probability of 0.2 to be a dominant effect of the DNA variant effect of one parent. With the modified version of SysGenSIM we can produce artificial gene expression data for the artificial gene network and hybrids we previously generated.

To adjust the different parameters of SysGenSIM, to produce a dataset as close as possible to our real data, we estimated the part of the variance explained by the genotypes (also called heritability) in the produced dataset and in the real dataset. This heritability is calculated via a mixed model [10].

By choosing at random the type of regulation (activation or repression), the DNA effect (cis- or trans-effect), and the allelic dominance effect for heterosis we produced different simulated gene expression datasets for our 180 genes and 435 genotypes. For each dataset, a particular gene regulation network with the same topology is associated. As a consequence, we can now test different methods of network inference and test the accuracy of these methods by comparing networks produced by the algorithms to the true network. Network inference methods with the best results will be used on the experimental dataset to answer our biological question.

In conclusion, we have developed a strategy to create an artificial dataset of gene expression measurements. The aim of this dataset is to test and select network inference methods adapted to a non-independent dataset for understanding the response to drought and heterosis phenomenon of sunflowers. The strategy is constituted of the following 4 steps, that could be adapted for other biological experiments and other types of data :

(i) Construction of an artificial network based on real biological information available for the same biological process on a close organism ;

(ii) Creation of artificial hybrid genotypes based on genomic information available for the real hybrids used in the experiment ;

(iii) Selection and adaptation of a data simulator emulating the same type of experiment that the one we performed, with steady state measurements on different genotypes ;

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

For each step it's important to use real biological information to in the end obtain an artificial dataset with biological properties close to properties of the real one. Doing like this, we hope the probability that networks inference methods perform the same in simulated as in real data is really high.

Banf & Rhee (2017). Computational inference of gene regulatory networks: Approaches, limitations and opportunities. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1860(1), 41–52.

Vignes et al. (2014). Gene Network Inference, chapter A Panel of Learning Methods for the Reconstruction of Gene Regulatory Networks in a Systems Genetics Contex. Springer.

Badouin et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature, 546(7656), 148–152.

Zheng et al. (2016). iTAK: A Program for Genome-wide Prediction and Classification of Plant Transcription Factors, Transcriptional Regulators, and Protein Kinases. Molecular Plant 9, 1667–1670

Dai et al. (2013). PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. BMC Bioinformatics, 14:321.

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.

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.

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.

Pinna et al. (2011). Simulating systems genetics data with SysGenSIM. Bioinformatics, 27(17), 2459–2462.

Mangin et al. (2017). Genomic Prediction of Sunflower Hybrids Oil Content. Frontiers in Plant Science, 8, 1–12.

Mots-Clés: gene regulatory network, artificial dataset, sunflower