Integrating GENomic prediction with GENe regulatory networks to optimize genetic value prediction : biological and statistical challenges

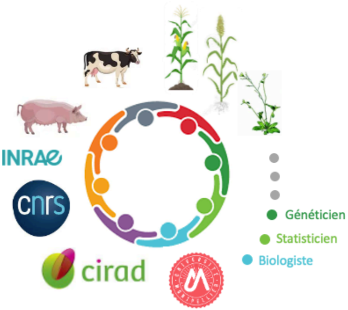
Marie Denis and David Pot

NETBIO seminar, Novembre 15, 2023

cirad
AGRICULTURAL RESEARCH
FOR DEVELOPMENT

RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

IGEN: A 2 days seminar in December 2022



- Geneticists
  - Vincent Segura
  - Maud Fagny
  - Renaud Rincent
  - Stéphane Nicolas
  - Celine Carilier Jacquin
  - Mathilde Causse
  - Emlie Millet
  - Laurence Flori
  - Gabriel Krouk
  - ...

- Biologists
  - Christophe Perin
  - Nancy Terrier
  - Antoine Martin
  - Gabriel Krouk
  - ...

- Statisticians
  - Andrea Rau
  - Sophie Lèbre
  - Mickael Lucas
  - Gabriel Krouk
  - ...

Biological motivations

Need to infer gene expression network/graph for addressing two biological objectives:

1. To gain insights into complex biological mechanisms involved in important processes, such as disease progress or growth

2. To improve prediction of important phenotypes in genetic improvement context

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation

2. Towards a better prediction

3. Conclusion

4. Bibliography

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation
2. Towards a better prediction
3. Conclusion
4. Bibliography

Towards a better understanding (genomic context)

To infer links/connections between genes for identifying biological mechanisms (such as key genes, functional modules, relations between network and a phenotype of interest, etc.)

Example:

How potassium and sodium fertilization impact biological mechanisms involved in response to water deficiency in *Eucalyptus grandis* ?

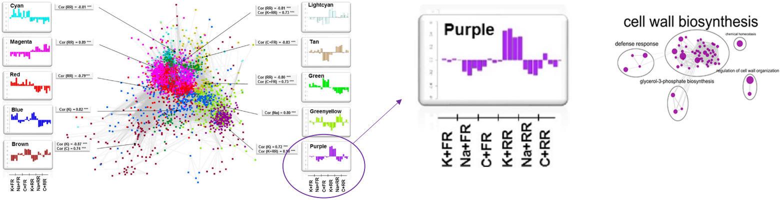


Figure 1: Gene co-expression network (on left), bar plot representing the average gene significance of the genes within the cluster purple (middle), and the associated enrichment map (on right) (Favreau et al., 2019).

Statistical questions

Network inference

How to build co-expression network from gene expression data?

Data:

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,1} & X_{m,2} & \cdots & X_{m,n} \end{pmatrix}$$

with $X_{i,j}$ the expression level of gene j for sample i

We want to infer **network/graph** where:

- Vertices: genes
- Edges: **links** between genes (gene-gene interactions)

Statistical questions

Network inference

What do we mean by links (gene-gene interactions)?

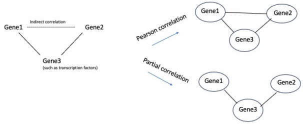
- Does it depend on biological question and/or experimental design?
- Does co-expression network aim at focusing on direct co-expression between genes? (Villa-Vialaneix et al., 2013; Grimes et al., 2019)

In the litterature

- **Pearson-based correlation networks** (relevance networks): marginal relationships between genes. Each pair of genes is considered alone: very dense networks, edges represent marginal connections not direct or causal
- **Partial correlation based networks**: direct relationships between genes. Correlation between two genes corrected for all other genes under investigation

Statistical questions

Network inference



Gene2 and Gene3 correlated but not dependent on each other

Gaussian Graphical Models (GGMs) (Lauritzen, 1996) commonly used to estimate partial correlations

- Improve measurement of direct relations between gene expressions by accounting for the effect of all expression data
- More efficient for grouping together genes with a common function / more consistent to prior biological knowledge (Werhli et al., 2006; Krumsiek et al., 2011; Villa-Vialaneix et al., 2013)

Statistical questions

Network inference: Gaussian Graphical Model

Let X_i be the m-vector (gene expression) of observed data for subject i such that

$$X_i \sim \mathcal{N}_m(\mu, \Sigma),\ \ i = 1, \dots n,$$

with $\mu \in \mathbb{R}^m$ is the mean vector, Σ is the covariance matrix which is a positive semi-definite symmetric matrix, and $\Omega = \Sigma^{-1} \in \mathbb{R}^m \times \mathbb{R}^m$ is the precision matrix.

$\looparrowright$ Conditional independence implied by the form/structure of the precision matrix:

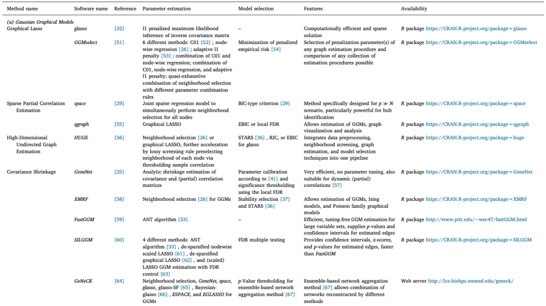
$$\text{Gene j and Gene k are linked } \Leftrightarrow \Omega_{jk} > 0.$$

Problem: When $n < m$, Σ is not full rank $\Rightarrow$ can not be inverted

Statistical questions

Network inference: Gaussian Graphical Model

Various estimation techniques (from the review done by Altenbuchinger et al. (2020)):



Method name	Software name	Reference	Parameter estimation	Model selection	Features	Availability
(a) Gaussian Graphical Models						
Graphical Lasso	*glasso*	[32]	l1 penalized maximum likelihood inference of inverse covariance matrix	–	Computationally efficient and sparse solution	*R* package https://CRAN.R-project.org/package = glasso
	GGMselect	[51]	6 different methods: C01 [52] ; node-wise regression [26] ; adaptive l1 penalty [53] ; combination of C01 and node-wise regression; combination of C01, node-wise regression, and adaptive l1 penalty; quasi-exhaustive combination of neighborhood selection with different parameter combination rules	Minimization of penalized empirical risk [54]	Selection of penalization parameter(s) of any graph estimation procedure and comparison of any collection of estimation procedures possible	*R* package https://CRAN.R-project.org/package = GGMselect
Sparse Partial Correlation Estimation	*space*	[29]	Joint sparse regression model to simultaneously perform neighborhood selection for all nodes	BIC-type criterion [29]	Method specifically designed for $p \gg N$ scenario, particularly powerful for hub identification	*R* package https://CRAN.R-project.org/package = space
	qgraph	[55]	Graphical LASSO	EBIC or local FDR	Allows estimation of GGMs, graph visualization and analysis	*R* package https://CRAN.R-project.org/package = qgraph
High-Dimensional Undirected Graph Estimation	*HUGE*	[56]	Neighborhood selection [26] or graphical LASSO, further acceleration by lossy screening rule preselecting neighborhood of each node via thresholding sample correlation	STARS [36] , RIC, or EBIC for *glasso*	Integrates data preprocessing, neighborhood screening, graph estimation, and model selection techniques into one pipeline	*R* package https://CRAN.R-project.org/package = huge
Covariance Shrinkage	*GeneNet*	[25]	Analytic shrinkage estimation of covariance and (partial) correlation matrices	Parameter calibration according to [41] and significance thresholding using the local FDR	Very efficient, no parameter tuning, also suitable for dynamic (partial) correlations [57]	*R* package https://CRAN.R-project.org/package = GeneNet
	XMRF	[58]	Neighborhood selection [26] for GGMs	Stability selection [37] and STARS [36]	Allows estimation of GGMs, Ising models, and Poisson family graphical models	*R* package https://CRAN.R-project.org/package = XMRF
	FastGGM	[59]	ANT algorithm [33]	–	Efficient, tuning-free GGM estimation for large variable sets, supplies *p*-values and confidence intervals for estimated edges	*R* package http://www.pitt.edu/~wec47/fastGGM.html
	SILGGM	[60]	4 different methods: ANT algorithm [33] , de-sparsified nodewise scaled LASSO [61] , de-sparsified graphical LASSO [62] , and (scaled) LASSO GGM estimation with FDR control [63]	FDR multiple testing	Provides confidence intervals, *z*-scores, and *p*-values for estimated edges, faster than *FastGGM*	*R* package https://CRAN.R-project.org/package = SILGGM
	GeNeCK	[64]	Neighborhood selection, *GeneNet*, *space*, *glasso*, glasso-SF [65] , Bayesian-glasso [66] , *ESPACE*, and *EGLASSO* for GGMs	*p*-Value thresholding for ensemble-based network aggregation method [67]	Ensemble-based network aggregation method [67] allows combination of networks reconstructed by different methods	Web server http://lce.biohpc.swmed.edu/geneck/

- More or less adapted for dealing with high-dimensional data: low to high differences observed
- More or less user friendly

⇒ Need guidelines for choosing the most adapted/to compare them

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation

2. Towards a better prediction

3. Conclusion

4. Bibliography

Statistical questions

Network evaluation

As most of co-expression networks in plants are Pearson-based correlation networks ⇒ Need to compare Pearson-based correlation network and partial correlation based network (Werhli et al., 2006; Krumsiek et al., 2011)

How to compare the inferred networks? How to evaluate their biological relevance ?

- Functional enrichment analysis for testing the biological relevance, detection of key genes, relevance of networks to the phenotype of interest, etc. (Villa-Vialaneix et al., 2013; Lee et al., 2020)
- To compare to a "reference" network (obtained from data base such as STRING protein-protein interactions database)

↪ Which statistical measures?

- Co-expression Differential Network Analysis: to extract the common structure (Grimes et al., 2019; Peterson et al., 2020)

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation
2. Towards a better prediction
3. Conclusion
4. Bibliography

Towards a better prediction

Is ≪ functional understanding ≫ relevant for prediction objectives, if it is the case how we take it into consideration ?

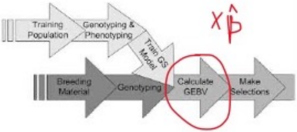
Idea: To use gene expression data or prior knowledge information into GS models

Genomic Selection (GS) model

$$Y = \mu + \underbrace{X\beta}_{GEBV} + \varepsilon$$

with $Y \in n \times 1$ the phenotype of interest, $X \in n \times p$ the marker matrix, $\beta \in p \times 1$ the marker effects, and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 Idn)$. GEBV: Genomic Estimated Breeding Value

↬ Various statistical approaches for estimating marker effects $\hat{\beta}$ (Ridge regression, Bayesian Lasso (BayesB), BayesC, etc)



Towards a better prediction

Idea: To use gene expression data or **prior knowledge information** into GS models

$$Y = \mu + X \underbrace{\beta}_{\text{prior knowledge information}} + \varepsilon$$

Which type of information?

- From previous experimental studies (Co-expression networks, GO terms, GWAS results, selection signature,...) but may be not adequate with data at hand
- From ”physical” knowledge: markers belonging to the same gene

Towards a better prediction

Idea: To use gene expression data or **prior knowledge information** into GS models

$$Y = \mu + X \underbrace{\beta}_{\text{prior knowledge information}} + \varepsilon$$

”Although there are many databases that provide information on biochemical relationships under normal conditions, the available reference networks may be incomplete or inappropriate for the experimental condition or set of subjects under study” (Peterson et al., 2016)

- ↬ Need to use statistical approaches integrating different degrees of fidelity/belief to the prior knowledge (to guard against mis-specification) (Stingo et al., 2010; Kundu et al., 2018; Denis et al., 2022)
- ↬ Need to use statistical approaches providing a trade-off between prior knowledge and computational complexity

Towards a better prediction

Idea: To use gene expression data or **prior knowledge information** into GS models

$$Y = \mu + X \underbrace{\qquad\beta\qquad}_{\text{prior knowledge information}} + \varepsilon$$

How to integrate those information into GS models?

Bayesian framework is a natural framework where prior knowledge may be specified via prior on regression coefficients (Bayesian fused and group Lasso (Kyung et al., 2010), Ising prior (Li and Zhang, 2010))

Example: $\beta \sim \mathcal{N}_p(0, \Sigma)$ with Σ related to structure between variables specified via for instance by undirected graph (Graph Laplacian prior(Liu et al., 2014), Gaussian Markov random field horseshoe prior (Denis and Tadesse, 2023))

Results: Improvement in prediction quality depends on several factors such as quality of information, relevance to the trait considered, etc. (Peterson et al., 2016; Mollandin et al., 2022)

Towards a better prediction

Idea: To use **gene expression data** or prior knowledge information into GS models

$$Y = \mu + \underbrace{X}_{\text{gene expression data}} \beta + \varepsilon$$

GS models may be used BUT questions about the interest of using transciptomic data instead of or in addition of genetic data.
Low gain in using transcriptomic data in prediction/Results vary according to environments Chateigner et al. (2020):

Questions:

- How to predict a phenotype measured at one time point given that gene expressions vary over tissues, time, and environments?
- Do we need to provide more stable information ? Via a common graph structure obtained across multiple co-expression networks?

”The problem of identifying predictors that are both relevant to a response variable of interest and functionally related to one another.”

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation
2. Towards a better prediction
3. Conclusion
4. Bibliography

Conclusion

- Various statistical and biological questions raised...
- But the bibliography is not exhaustive.... there are certainly already responses to our questions....
- But seems interesting for biologists, geneticists, and statisticians

↬ Master student for working on the first part with Bénédicte Favreau (Biologist, Cirad) on Eucalyptus

↬ To continue exchanging on those subjects...

Outline

1. Towards a better understanding
   - Network inference
   - Network evaluation

2. Towards a better prediction

3. Conclusion

4. Bibliography

Altenbuchinger, M., Weihs, A., Quackenbush, J., Grabe, H. J., and Zacharias, H. U. (2020). Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1863(6):194418.

Chateigner, A., Lesage-Descauses, M.-C., Rogier, O., Jorge, V., Leplé, J.-C., Brunaud, V., Roux, C. P.-L., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Sanchez, L., et al. (2020). Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC genomics*, 21(1):1–16.

Denis, M. and Tadesse, M. G. (2023). Graph-structured variable selection with gaussian markov random field horseshoe prior. *HAL*.

Denis, M., Varghese, R. S., Barefoot, M. E., Tadesse, M. G., and Ressom, H. W. (2022). A bayesian two-step integrative procedure incorporating prior knowledge for the identification of mirna-mrnas involved in hepatocellular carcinoma. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 81–86. IEEE.

Favreau, B., Denis, M., Ployet, R., Mounet, F., Peireira da Silva, H., Franceschini, L., Laclau, J.-P., Labate, C., and Carrer, H. (2019). Distinct leaf transcriptomic response of water deficient eucalyptus grandis submitted to potassium and sodium fertilization. *PLoS One*, 14(6):e0218528.

Grimes, T., Potter, S. S., and Datta, S. (2019). Integrating gene regulatory pathways into differential network analysis of gene expression data. *Scientific reports*, 9(1):5479.

Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J. (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC systems biology*, 5:1–16.

Kundu, S., Cheng, Y., Shin, M., Manyam, G., Mallick, B. K., and Baladandayuthapani, V. (2018). Bayesian variable selection with graphical structure learning: Applications in integrative genomics. *PloS one*, 13(7):e0195070.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Lee, J., Shah, M., Ballouz, S., Crow, M., and Gillis, J. (2020). Cococonet: conserved and comparative co-expression across a diverse set of species. *Nucleic acids research*, 48(W1):W566–W571.

Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105(491):1202–1214.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A. C., et al. (2014). Bayesian regularization via graph Laplacian. *Bayesian Analysis*, 9(2):449–474.

Mollandin, F., Gilbert, H., Croiseau, P., and Rau, A. (2022). Accounting for overlapping annotations in genomic prediction models of complex traits. *BMC bioinformatics*, 23(1):1–22.

Peterson, C. B., Osborne, N., Stingo, F. C., Bourgeat, P., Doecke, J. D., and Vannucci, M. (2020). Bayesian modeling of multiple structural connectivity networks during the progression of alzheimer's disease. *Biometrics*, 76(4):1120–1132.

Peterson, C. B., Stingo, F. C., and Vannucci, M. (2016). Joint bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in medicine*, 35(7):1017–1031.

Stingo, F. C., Chen, Y. A., Vannucci, M., Barrier, M., and Mirkes, P. E. (2010). A bayesian graphical modeling approach to microrna regulatory network inference. *The annals of applied statistics*, 4(4):2024.

Villa-Vialaneix, N., Liaubet, L., Laurent, T., Cherel, P., Gamot, A., and SanCristobal, M. (2013). The structure of a gene co-expression network reveals biological functions underlying eqtls. *PloS one*, 8(4):e60045.

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531.