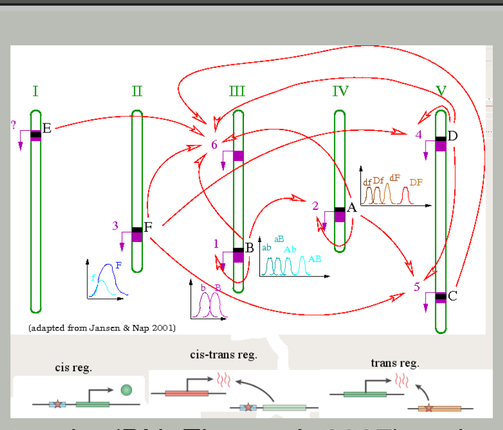
From gene clustering to genetical genomics: analysing or reconstructing biological networks

Matthieu Vignes[1] Jimmy Vandel[1] Nathalie Keussayan[1] Juliette Blanchet[2]
Simon de Givry[1] Brigitte Mangin[1]

[1]BIA Unit - INRA Toulouse, Castanet Tolosan, France
[2]WLF/SLF, Davos, Switzerland

Introduction & Biological issues

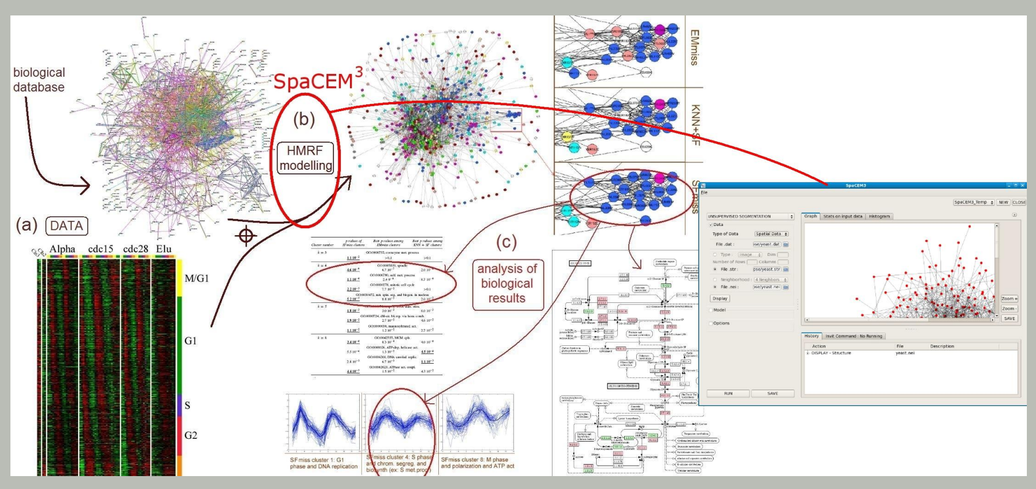
- Goal: Gene Regulatory Network (GRN) inference.
- Means: 2 kinds of information available. (i) Polymorphisms on genes (eQTL), observed in crosses between 2 known strains, as causes for (ii) variation on RNAm levels (snapshot of cell activity).



- Considered approaches: Discrete Bayesian Networks (BN, Zhu et al. 2007) and Structural Equation Modelling (SEM, Liu et al. 2008).
- Results on synthetic data in the context of genetical genomics data.

1 - Gene expression clustering accounting for missing observations in a Markovian setting

- Noisy observations **Y**, some possibly missing (at random), network structure on labels **Z** between the p biological entities.
- Tool: Hidden Markov Random Fields with genuine EM algorithm with mean-field like approximations for model learning. Implemented in SpaCEM[3] available at `http://spacem3.gforge.inria.fr/`.
- Model: $P(Y, Z) = P(Y \mid Z).P_G(Z) = \prod_1^p f(Y_i, \theta_{Z_i}).\exp(-H(Z;\Delta))/W(\Delta)$; model selection: BIC.



- Biological analysis for validation (Blanchet & Vignes 2009).

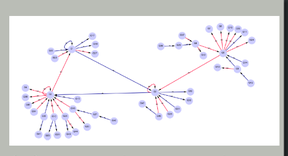
Gene network reconstruction with genetical genomics data

- Estimating weights on edges to infer (partially) the graph.
- Include genetic information in an additional blanket **X** on the graph: Triplet Markov Fields (Blanchet & Forbes 2008)
  $P(Y, X, Z) \propto \exp\left(-\sum_{c\in\mathcal{C}} V_c(X_c, Z_c) - \sum_i \log(f(Y_i \mid \theta_{X_i,Z_i}))\right)$ at present only limited to supervised classification.

2 - Reconstruction of networks combining genetic and genomics data

2.1 Genetical genomics data simulation

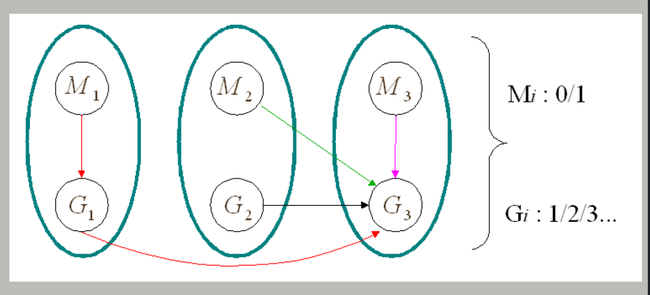
- Choose network with features close to known features of biological networks → `http://www.comp-sys-bio.org/AGN/`.



- Simulate genotype given a RIL population, chromosome(s) size(s), number and distribution of markers → CarthaGène `http://www.inra.fr/mia/T/CarthaGene/`.
- Compute steady state gene expression levels from coupled ODE (mimicking biochemistry) → COmplex PAthway SImulator `http://www.copasi.org`.

2.2 Structure learning of a discrete Bayesian Network

- BN = DAG $\oplus$ $P(V) = \prod_{i=1}^p P(V_i \mid V_{pa(V_i)})$; tested score-based (BIC) algorithms: Greedy Search, K2...
- Fig.: M_i: genotype, G_i: transcript level of gene i.



- eQTL analysis with MCQTL `http://carlit.toulouse.inra.fr/MCQTL/`.

2.3 SEM of genetical genomics data

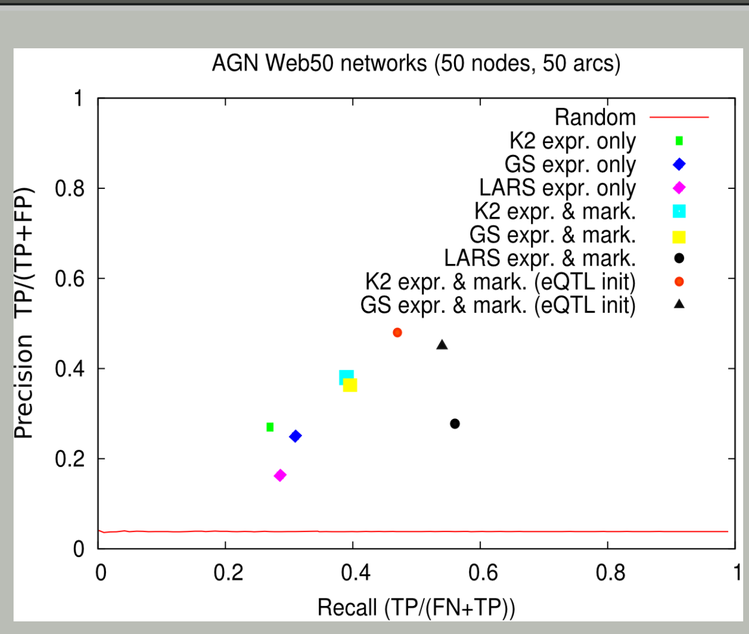
- $Y = Y.B + X.\Theta + \epsilon$, where: Y transcript levels, X genotypes, B_{km} direct effect of expr. k on expr. m and Θ_{jm} direct effect of marker j on expr. m.
- Lasso gene-by-gene regression to estimate paramater $\neq 0$: $\widehat{\begin{bmatrix} B_k \\ \Theta_k \end{bmatrix}} = \arg\min |Y_k - [Y_{\setminus k} X].\begin{bmatrix} B_k \\ \Theta_k \end{bmatrix}|_{L_2} + \lambda|\begin{bmatrix} B_k \\ \Theta_k \end{bmatrix}|_{L_1}$ ($|\widehat{\begin{bmatrix} B_k \\ \Theta_k \end{bmatrix}}|_{L_1} \leq \tau$); LAR algorithm (refs on `http://www-stat.stanford.edu/~tibs/lasso.html`); BIC and Meinshausen criteria to determine the best λ.

2.4 BN vs. SEM – pros and cons

	Continuous data	Comput. time	Model. cycles	Param./likelihood estimation	Non-lin. depend.
BN					
SEM					

(thanks to lars)

2.5 Comparative results



Conclusion and prospects

- Wide range of available methods to integrate genetic and genomics data to infer GRN.
- Gain in using genetic information to infer the network.
- Plausible synthetic genetical genomics dataset simulation, but Gold standard dataset on model organism needed.
- Prospects: (i) fairly assess merits of different approaches and develop algorithms to optimize score devoted to data at hand (ii) analyze data from collaborations (SUNYFUEL, FRAGENOMICS...).

References

R. Jansen and J. Nap, Genetical genomics: the added value from segregation, *Trends Gen.*, 17:388-91 (2001).

J. Zhu et al., Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations, *PLoS Comput. Biol.*, 3:e64 (2007).

J. Blanchet and F. Forbes. Triplet Markov fields for the supervised classification of complex structured data. *IEEE PAMI*, 30:1055-67 (2008).

B. Liu et al., Gene network inference via structural equation modeling in genetical genomics experiments, *Genetics*, 178:1763-76 (2008).

J. Blanchet and M. Vignes, A model-based approach to gene clustering with missing observations reconstruction in a Markov Random Field framework, *J. Comput. Biol.*, 16:475-86 (2009).