Graph Inference Enhancement with Clustering: *Application to Gene Regulatory Network Reconstruction*

Aurélie Pirayre[1,2], Camille Couprie[1], Laurent Duval[1], and Jean-Christophe Pesquet[2]

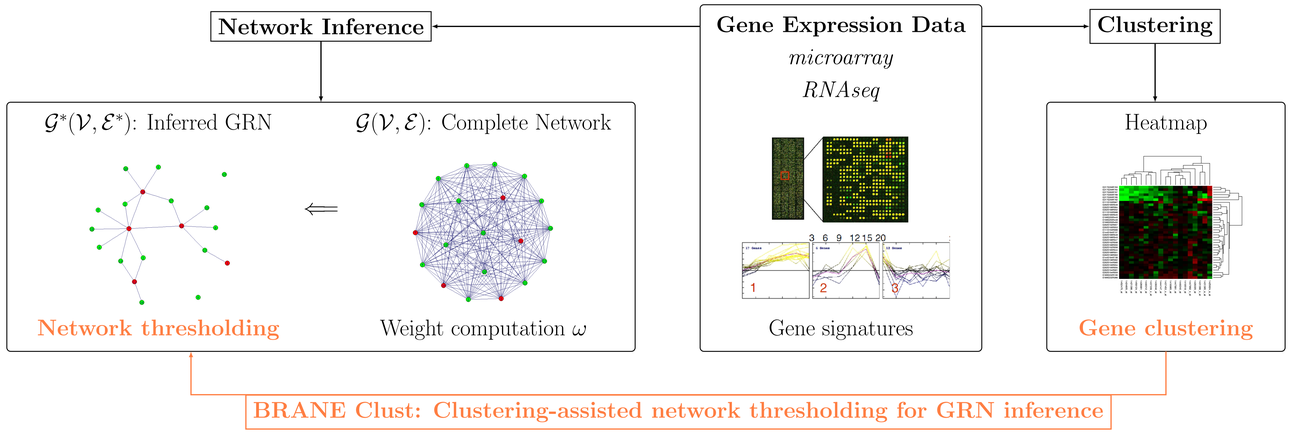
[1] IFP Energies nouvelles, Mecatronics, Computer Science and Applied Mathematics Division, Rueil-Malmaison, France
[2] Université Paris-Est, LIGM, UMR CNRS 8049, France

Introduction

When clustering improves network inference... like Gene Regulatory Network (GRN).

Goal: GRN inference

Complete graph
$\mathcal{G}(\mathcal{V}, \mathcal{E})$
$\Downarrow$
Complete weighted graph
$\mathcal{G}(\mathcal{V}, \mathcal{E}, \omega)$
$\Downarrow$
Find a set of edges $\mathcal{E}^*(\subseteq \mathcal{E})$ reflecting gene regulatory links
$\Downarrow$
Final Gene Regulatory Network
$\mathcal{G}^*(\mathcal{V}, \mathcal{E}^*)$



Classical GRN thresholding

Given $\mathcal{G}(\mathcal{V}, \mathcal{E}, \omega)$, how to find $\mathcal{E}^*$?

- $x_{i,j}$ a binary label of edge $e_{i,j}$: $x_{i,j} = \begin{cases} 1 & \text{if } e_{i,j} \in \mathcal{E}^* \\ 0 & \text{otherwise.} \end{cases}$
- The optimal labeling $\boldsymbol{x}^*$ reflects regulatory links

GRN thresholding: maximization of a cost function F [4]

- $\omega_{i,j}$ a weight between nodes i and j computed from gene expression data
- $\lambda > 0$ a thresholding parameter
- Select strongly weighted edges $e_{i,j} : \omega_{i,j} > \lambda$

$$\underset{x \in \{0,1\}^n}{\text{maximize}} \sum_{\substack{(i,j) \in \mathcal{V}^2 \\ i<j}} \omega_{i,j} x_{i,j} + \lambda(1 - x_{i,j})$$

- Explicit solution:

$$x^*_{i,j} = \mathbb{1}(\omega_{i,j} > \lambda)$$

Clustering-assisted GRN thresholding: use prior on gene clusters

Proposed formulation

We want to:

- favor strongly weighted edges
- favor edges $e_{i,j}$ if nodes i and j belong to the same cluster
- promote a modular structure around T central nodes

$$\underset{x \in \{0,1\}^n}{\text{maximize}} \sum_{(i,j) \in \mathcal{V}^2} \omega_{i,j} x_{i,j} \frac{\beta - \mathbb{1}(y_i \neq y_j)}{\beta} + \lambda(1 - x_{i,j}),$$
$$\text{subject to} \quad y \in C$$

where

$-\ C = \{(z_i)_{1 \leq i \leq G} \in \mathbb{N}^G \mid \forall \tau \in \{1, \dots, T\}, z_{i_\tau} = t_\tau\}$,
$-\ y_i \in \mathbb{N}$ the cluster labeling of the node i,
$-\ t_\tau$ marker linked to the central node τ
$-\ G$ the number of nodes

- At y fixed and x variable, an explicit solution is given by

$$x^*_{i,j} = \begin{cases} 1 & \text{if } \omega_{i,j} > \dfrac{\lambda\beta}{\beta - \mathbb{1}(y_i \neq y_j)} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

- At x fixed and y variable, the problem reduces to

$$\underset{y \in \mathbb{N}^G \cap C}{\text{minimize}} \sum_{\substack{(i,j) \in \mathcal{V}^2 \\ i<j}} \alpha_{i,j} \mathbb{1}(y_i \neq y_j)$$

Optimization strategy

- NP-hard combinatorial problem: multi-label formulation...
  Let $y^{(1)}, \dots, y^{(L)}$ be a set of L vectors, whose components are

$$y_i^{(l)} = \begin{cases} 1 & \text{if } y_i = l \\ 0 & \text{otherwise.} \end{cases}$$

- ... and quadratic relaxation:

$$\underset{\substack{y^{(1)} \in C^{(1)}, \dots, y^{(L)} \in C^{(L)} \\ (y^{(1)}, \dots, y^{(L)}) \in \widehat{D}}}{\text{minimize}} \sum_{l=1}^{L} \left(\sum_{\substack{(i,j) \in \mathcal{V}^2 \\ i<j}} \alpha_{i,j} (y_i^{(l)} - y_j^{(l)})^2 \right) \tag{2}$$

with

$-\ C^{(l)}$ a set of T markers $z_{i_\tau}^{(l)}$ valued by 1 if $t_\tau = l$ and 0 otherwise

$-\ \widehat{D}$ a set coding for the sum-to-one constraint

- Dirichlet problem $\rightarrow$ solution of $L - 1$ systems of linear equations
- Optimal gene cluster labeling given by

$$y_i^* = \underset{l \in \{1, \dots, L\}}{\arg\max}\, y_i^{(l)}.$$

Proposed algorithm

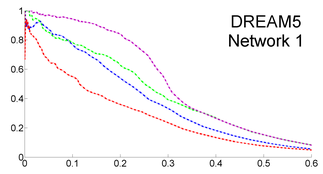
For β and λ fixed

- Compute $\alpha_{i,j}$ weights with $\alpha_{i,j} = \begin{cases} 0 & \text{if } \omega_{i,j} \leq \lambda \\ \omega_{i,j} - \lambda & \text{if } \lambda < \omega_{i,j} \leq \frac{\lambda\beta}{\beta - 1} \\ \frac{\omega_{i,j}}{\beta} & \text{if } \omega_{i,j} > \frac{\lambda\beta}{\beta - 1} \end{cases}$
- Based on $\alpha_{i,j}$ weights, compute the optimal gene cluster labeling $\boldsymbol{y}^*$ given by (2)
- Using $\boldsymbol{y}^*$, compute the optimal edge labeling $\boldsymbol{x}^*$ given by (1)

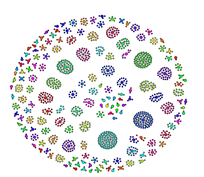
Results

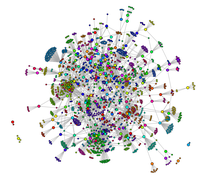
Many inference methods are score-based (e.g. mutual information [1]) or model-based (e.g. Gaussian graphical models [3]).

- Dataset validation: DREAM4 and DREAM5
- Comparison: CLR and GENIE3
- Criterion: Area Under Precision-Recall curve



	DREAM4					DREAM5
Network	1	2	3	4	5	1
GENIE3 [2]	0.239	0.260	0.323	0.301	0.295	0.290
CLR [1]	**0.245**	0.255	0.299	0.298	0.299	0.248
BRANE Clust	0.243	**0.277**	**0.369**	**0.328**	**0.332**	**0.342**





Conclusions

- CLR and GENIE3 results are improved by
  $-\ 10.5\,\%$ and $8.8\,\%$ respectively on DREAM4
  $-\ 38\,\%$ and $19\,\%$ respectively on DREAM5
- A priori on clustering improves network inference
- Existing GRN methods may benefit from BRANE Clust
- Perspective: more sophisticated clustering method.

References

[1] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):54–66, 2007.

[2] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9):1–10, Sep. 2010.

[3] A. Wiesel, Y. C. Eldar, and A. O. Hero III. Covariance estimation in decomposable Gaussian graphical models. *IEEE Trans. Signal Process.*, 58(3):1482–1492, Mar. 2010.

[4] A. Pirayre, C. Couprie, L. Duval, F. Bidard, and J.-C. Pesquet BRANE Cut: Biologically-Related Apriori Network Enhancement with Graph cuts for Gene Regulatory Network Inference. *submitted.* 2015