Fast Convex Optimization for Connectivity Enforcement in Gene Regulatory Network Inference

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

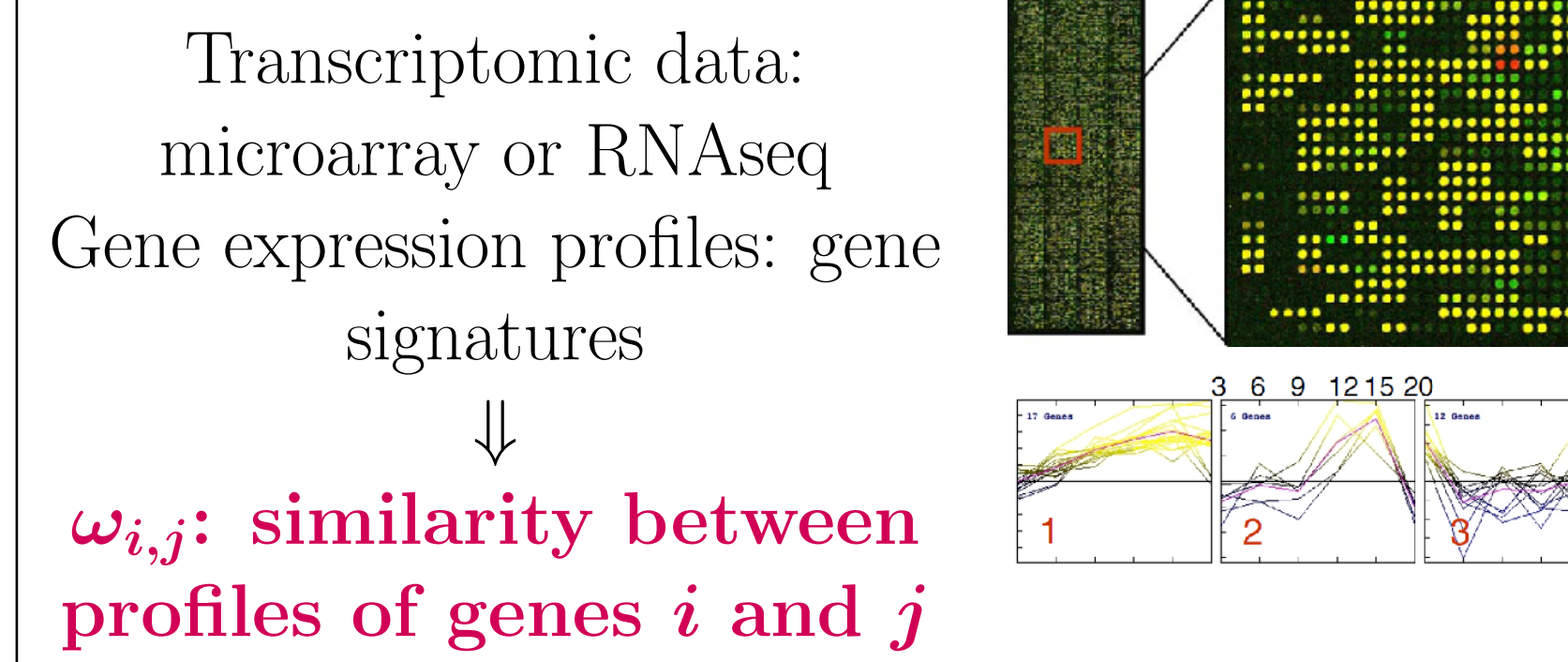## Introduction: reverse biological engineering

What is a Gene Regulatory Network (GRN)?

$\mathcal{V}$: genes

$\begin{cases} \mathcal{T} : \text{regulating genes TFs} \\ \mathcal{V}\backslash\mathcal{T} : \text{regulated genes } \overline{\text{TFs}} \end{cases}$

$\mathcal{E}^*$: regulatory relationships
between genes
$\Downarrow$
$\mathcal{G}^*(\mathcal{V},\mathcal{E}^*)$: inferred GRN

How to obtain $\mathcal{E}^*$?

$\mathcal{G}(\mathcal{V},\mathcal{E})$: fully connected network
$\mathcal{E}$: edges weighted by ω
ω: strength of the links between genes
$\Downarrow$
Thresholding on ω

How to obtain ω?

Transcriptomic data: microarray or RNAseq
Gene expression profiles: gene signatures
$\Downarrow$
$\omega_{i,j}$: similarity between profiles of genes i and j
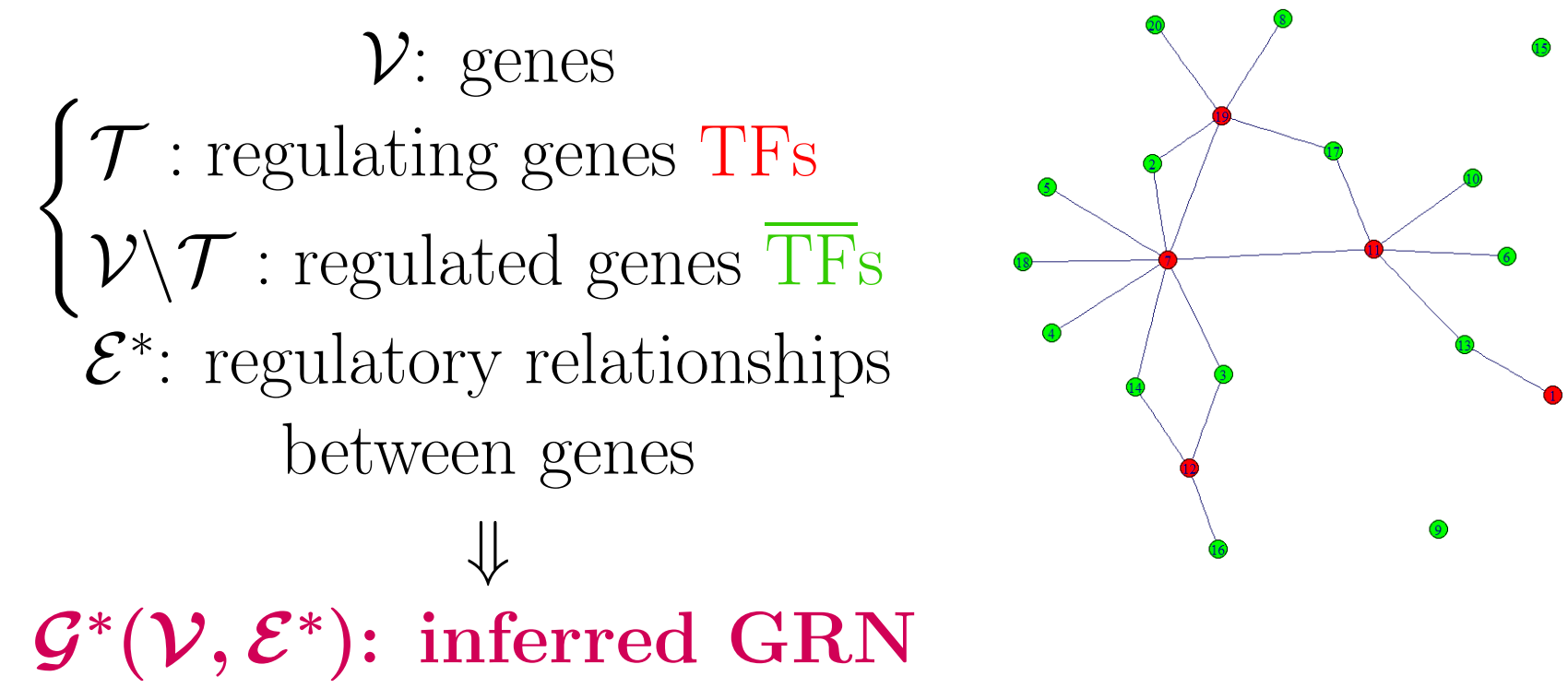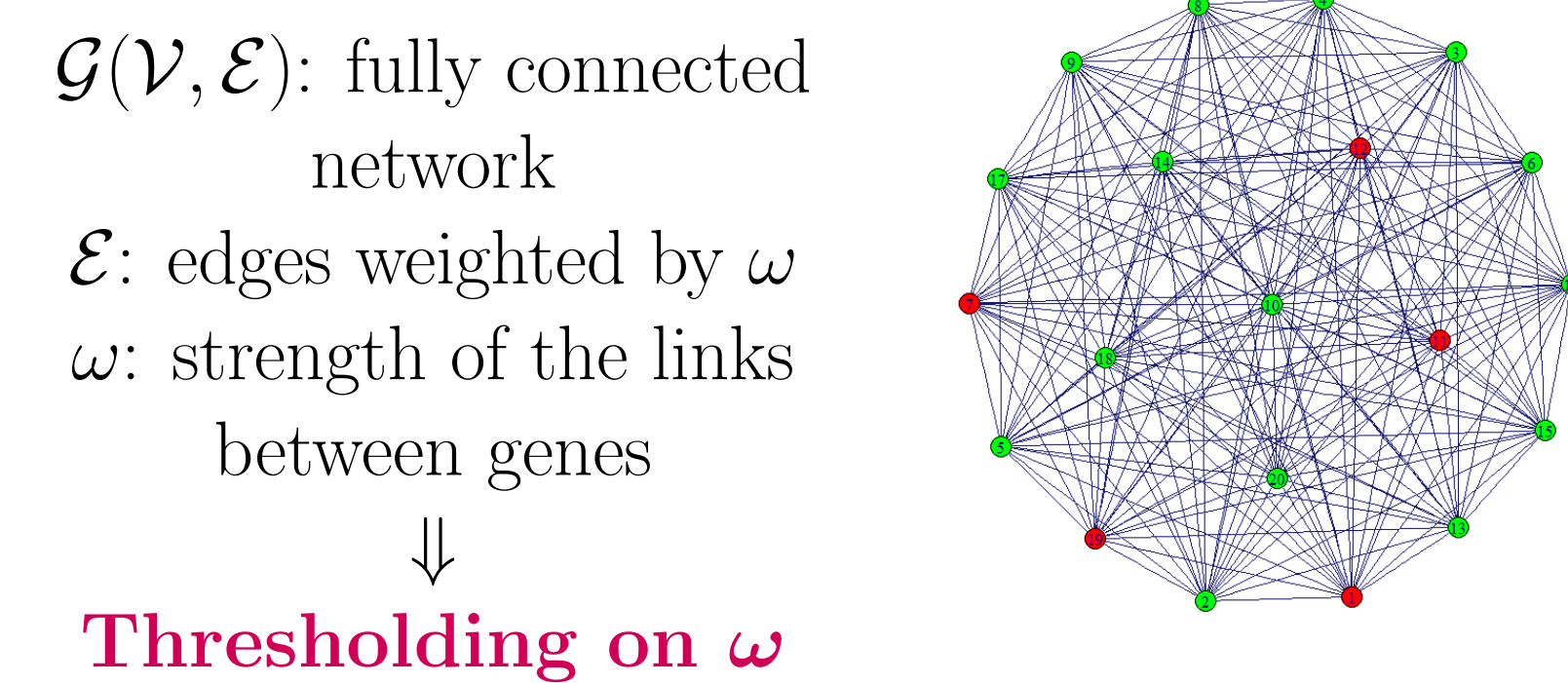
Inferring a GRN: in a complete graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ weighted by ω, find a set of edges $\mathcal{E}^*$ ($\subseteq \mathcal{E}$) reflecting regulatory links between genes
Many inference methods: score-based (e.g. mutual information [2]) or model-based approaches (e.g. gaussian graphical models [4])
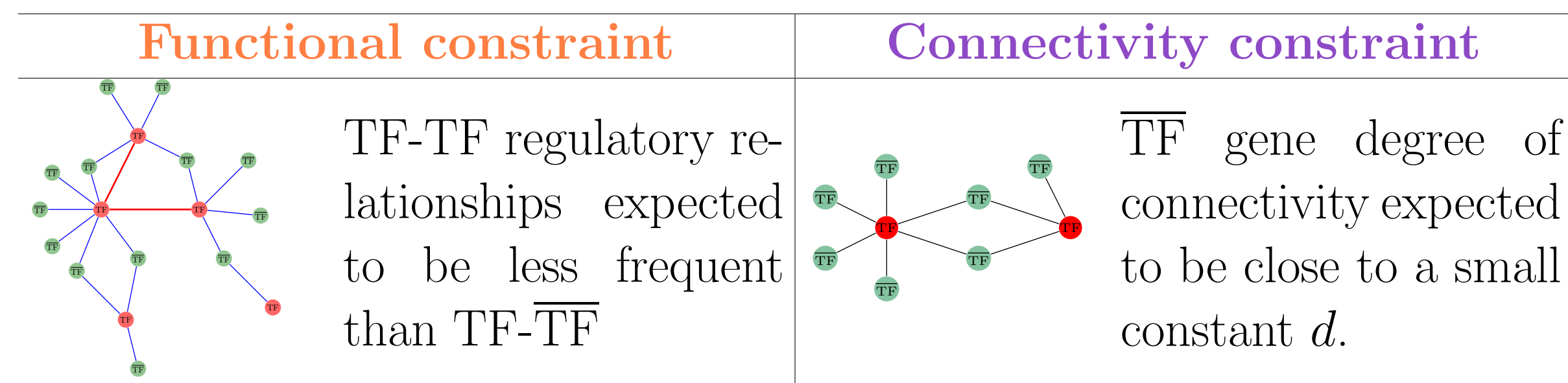
Problem formulation

Thresholding step on ω treated as an edge segmentation problem

Let $x_{i,j}$ be the binary label of the edges $e_{i,j}$ such that $x_{i,j} = \begin{cases} 1 & \text{if } e_{i,j} \in \mathcal{E}^* \\ 0 & \text{otherwise.} \end{cases}$

$\Rightarrow$ **Find the optimal labeling x^* coding for regulatory links**

Biological *a priori*

Functional constraint	Connectivity constraint
TF-TF regulatory relationships expected to be less frequent than TF-$\overline{\text{TF}}$	$\overline{\text{TF}}$ gene degree of connectivity expected to be close to a small constant d.

Proposed cost function

$$\underset{x\in\{0,1\}^n}{\text{minimize}} \quad \underbrace{\sum_{(i,j)\in\mathcal{E}} \omega_{i,j}(1-x_{i,j})}_{\text{Favors strongly weighted edges}} + \underbrace{\sum_{(i,j)\in\mathcal{E}} \lambda_{i,j}x_{i,j}}_{\text{Favors TF-}\overline{\text{TF}}\text{ edge presence}} + \mu \underbrace{\sum_{i\in\mathcal{V}\backslash\mathcal{T}} \Phi\left(\sum_{j\in\mathcal{V}} x_{i,j} - d\right)}_{\text{Enforces } \overline{\text{TF}} \text{ degree to be close to } d}$$

with

- $\lambda_{i,j} \in [0,1]$: parameter depending on the nature (TF or $\overline{\text{TF}}$) of genes i and j
- $\mu \geq 0$: regularization parameter
- $\Phi(.)$: Lipschitz-gradient distance function

Optimization strategy

Relaxation of the integrality constraint on x

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \underbrace{\omega^\top(\mathbb{1}-x) + \lambda^\top x + \mu\,\Phi(\Omega x - d)}_{f_1} + \underbrace{\iota_{[0,1]^n}(x)}_{f_2}$$

- f_1: differentiable and Lipschitz-gradient function
- f_2: convex function

$\Rightarrow$ **Optimization *via* proximal method**

- proximal operator: $\text{prox}_f(x) = \arg\min_y \left(f(y) + \frac{1}{2}\|y-x\|_R^2\right)$
- closed-form for prox_{f_2}: simple projection onto the convex set $[0,1]^n$

Acceleration tricks [1] and algorithm

- **Preconditioning approach**
  Majorization-Minimization principle to f_1 $\Rightarrow$ Quadratic majorant of f_1
- **Block-coordinate approach**
  Additively block separable function f_2 $\Rightarrow$ Optimization of a subset of variables at each iteration

**Algorithm 1: Block-Coordinate Preconditioned Forward-Backward (BC-P-FB) algorithm**

Fix $x_0 \in \mathbb{R}^N$
for $n = 0, 1, \ldots$ **do**
  Select the index $k_n \in \{1, \ldots, p\}$ of a block of variables
  $z_n^{(k_n)} = x_n^{(k_n)} - \gamma_n \mathbf{A}_{k_n}^{-1}\Omega_{k_n}^\top \nabla\Phi(\Omega x_n - d)$
  $x_{n+1}^{(k_n)} = \text{prox}_{\gamma_n^{-1}\mathbf{A}_{k_n} f_2^{(k_n)}}(z_n^{(k_n)})$
  $x_{n+1}^{(k)} = x_n^{(k)}, \quad k \in \{1, \ldots, p\} \setminus \{k_n\}$

Results

- Dataset validation: DREAM4 international challenge
- Comparison to state-of-the-art: CLR [2] and GENIE3 [3]
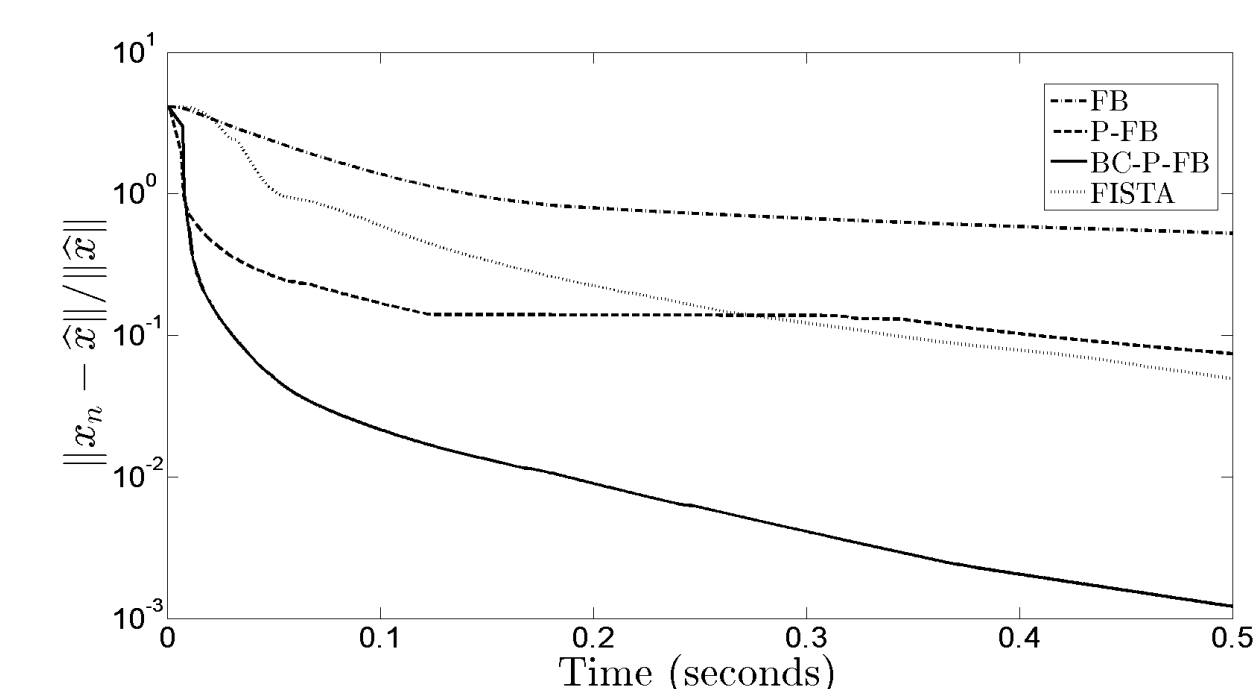- Evaluation criterion: AUPR (Area Under the Precision-Recall curve)

Inference results

	Network index	1	2	3	4	5	Mean
	CLR	0.252	0.275	0.312	0.313	0.317	0.294
AUPR	GENIE3	0.257	0.265	0.317	0.298	0.302	0.293
	BRANE Relax	**0.268**	**0.296**	**0.346**	**0.317**	**0.332**	**0.312**
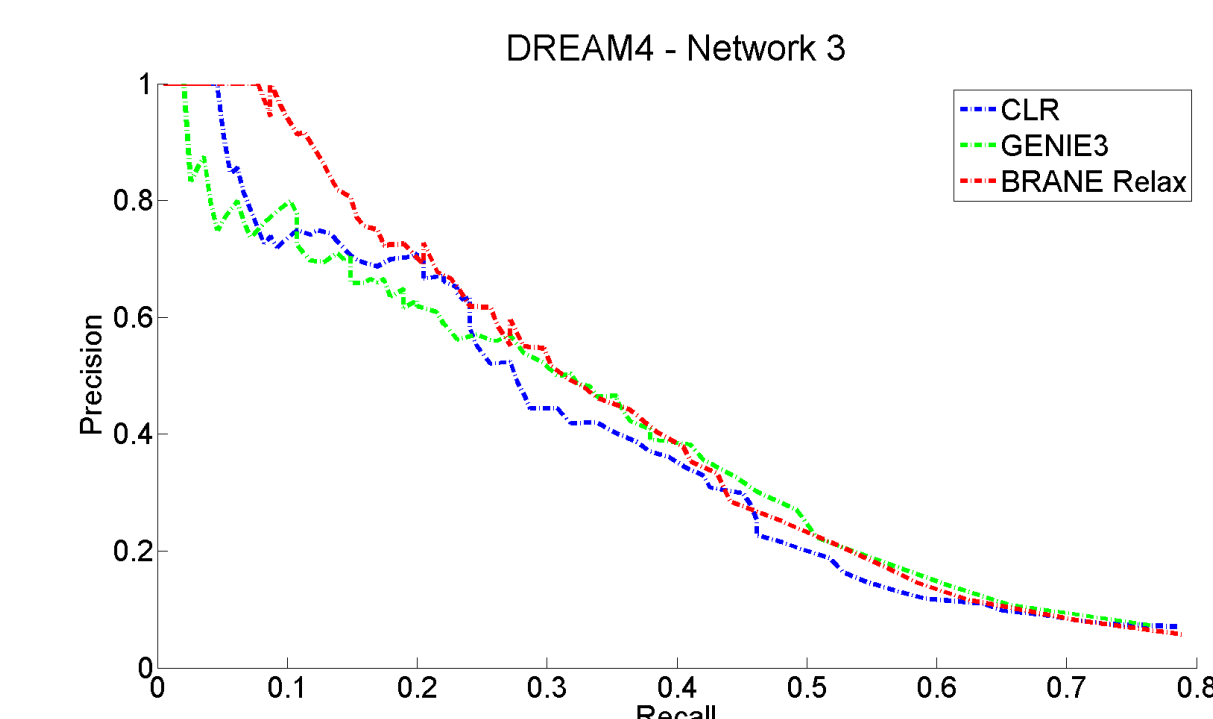
$\Rightarrow$ **Improvement of 6% over the five networks**

Speedup results

Conclusions

- The variational formulation, taking into account biological assumptions translated into structural *a priori*, delivers promising results
- CLR and GENIE3 are outperformed on DREAM4 international challenge
- Existing GRN methods may benefit from our approach, as they take a weighted graph as input

References

[1] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward-backward algorithm. Technical report, 2013.

[2] 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.

[3] 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.

[4] 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.

[5] 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 in May 2015.