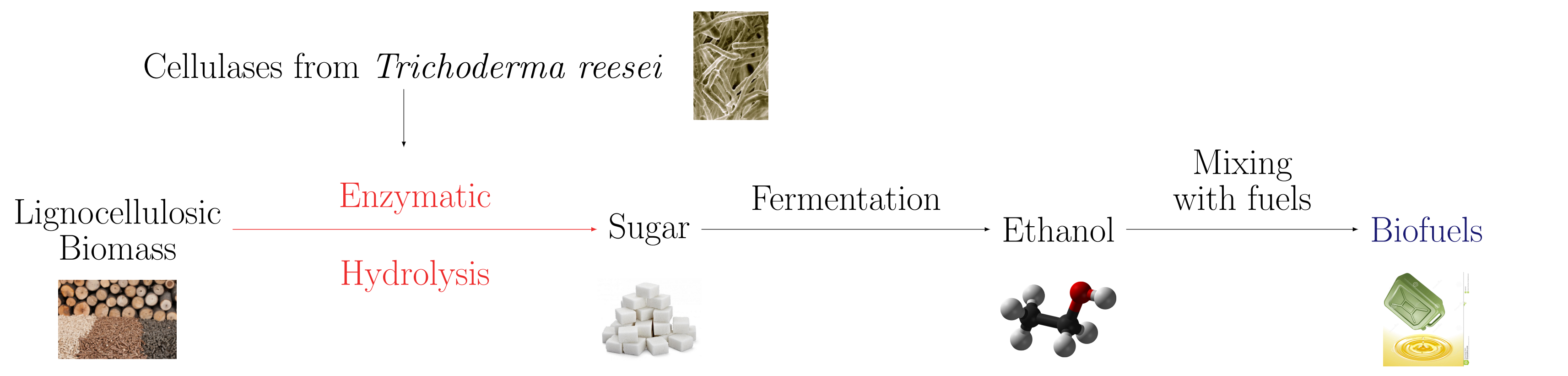
Discrete vs Continuous Optimization for Gene Regulatory Network Inference

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


Cellulases from *Trichoderma reesei*

Lignocellulosic Biomass → Enzymatic Hydrolysis → Sugar → Fermentation → Ethanol → Mixing with fuels → Biofuels

Energetic context

Improving the production efficiency of the second generation biofuels by optimizing the enzymatic hydrolysis phase

Biological context

Genetic target identification in *Trichoderma reesei* to improve the cellulase production, involved in the biofuel production process

Mathematical context

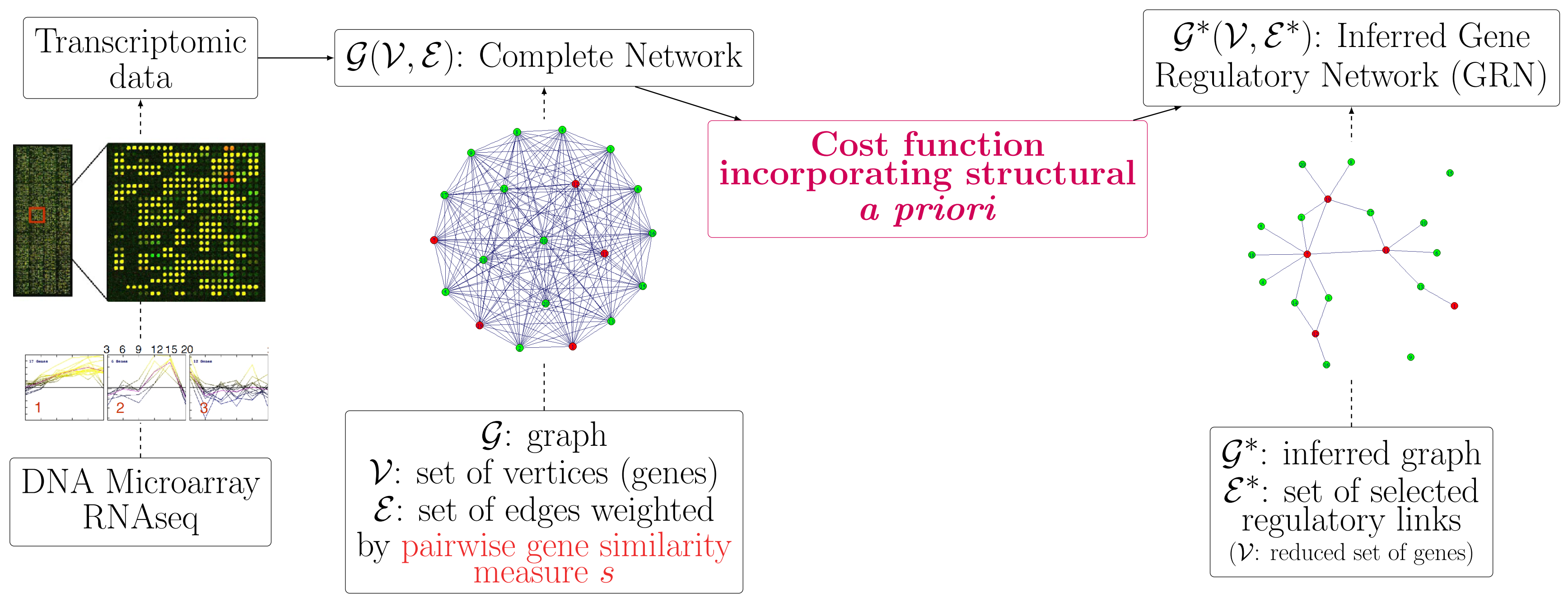
Novel mathematical models based on graph optimization to infer Gene Regulatory Networks (GRNs) and identify new target genes

GRNs: powerful tools to visualize gene interaction relationships from high-throughput data
Difficult problem: thousands of genes expressed in only few conditions/replicates

Very active community with DREAM challenges and many inference methods:
Relevance Network, ARACNE, SIMoNe, NARROMI, CLR, GENIE3...

Global strategy

Inferring a GRN: recovering interactions between transcription factors and their target genes i.e. in a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, find a set of edges $\mathcal{E}^*(\subseteq \mathcal{E})$ reflecting regulatory links


Transcriptomic data

DNA Microarray
RNAseq

$\mathcal{G}(\mathcal{V},\mathcal{E})$: Complete Network

$\mathcal{G}$: graph
$\mathcal{V}$: set of vertices (genes)
$\mathcal{E}$: set of edges weighted by pairwise gene similarity measure s

Cost function incorporating structural *a priori*

$\mathcal{G}^*(\mathcal{V},\mathcal{E}^*)$: Inferred Gene Regulatory Network (GRN)

$\mathcal{G}^*$: inferred graph
$\mathcal{E}^*$: set of selected regulatory links
($\mathcal{V}$: reduced set of genes)

GRN inference problem treated as a segmentation problem

- Let $x_{i,j}$ be the binary label of the edges $e_{i,j}$ such that

$$x_{i,j} = \begin{cases} 1 & \text{if } e_{i,j} \in \mathcal{E}^* \\ 0 & \text{otherwise.} \end{cases}$$

- Inference problem re-expressed as cost function minimization → optimal labeling $\mathbf{x}^*$ signaling the edge presence (or absence) in the inferred graph $\mathcal{G}^*(\mathcal{V},\mathcal{E}^*)$

How to define a biologically sound cost function ?

Proposed cost function

Generic Cost function

$$\underset{\mathbf{x}\in\{0,1\}^n}{\text{minimize}} \underbrace{\sum_{(i,j)\in\mathcal{E}} s_{i,j}\Psi(x_{i,j}-1)}_{\text{Disfavors the deletion of strongly weighted edges}} + \underbrace{\sum_{(i,j)\in\mathcal{E}} \lambda_{i,j}\Psi(x_{i,j})}_{\text{Favors the selection of edges linked to a transcription factor (TF)}} \underbrace{+\mu\Phi((x_{i',j'})_{(i',j')\in\mathcal{N}_{i,j}})}_{\textbf{Structural } \textbf{\textit{a priori}}}$$

where

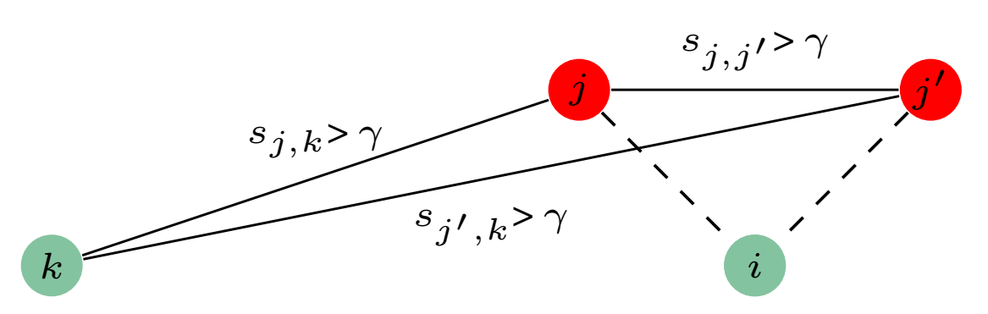
- $s_{i,j} \in [0,1]$ is a similarity weight between the expression profiles of genes i and j
- $\lambda_{i,j} \in [0,1]$ a parameter depending on the nature (regulator or not) of genes i and j
- $\mu \geq 0$ a regularization parameter
- $\mathcal{N}_{i,j}$ a local neighborhood of the edge $e_{i,j}$

$\mathcal{T} \subset \mathcal{V}$: a set of transcription factors (TFs)

Structural *a priori*

Co-regulation property

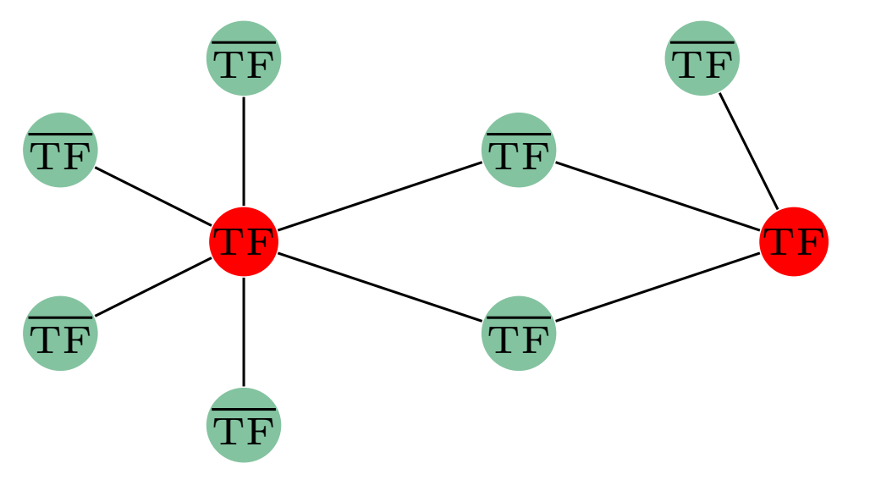
- Assuming that a gene k is co-regulated by two TFs (j,j'), then $\forall i \in \mathcal{V}\backslash\mathcal{T}$ the inferences of $e_{i,j}$ and $e_{i,j'}$ are coupled



- $\Phi(x_{i,j}) = \sum_{\substack{i\in\mathcal{V}\backslash\mathcal{T}\\(j,j')\in\mathcal{T}^2}} \alpha_{i,j,j'}|x_{i,j}-x_{i,j'}|$

Connectivity constraint

- The degree of connectivity of non transcription factors ($\overline{\text{TF}}$s) is enforced to be close to a constant number d



- $\Phi(x_{i,j}) = \sum_{i\in\mathcal{V}\backslash\mathcal{T}} \left(\sum_{j\in\mathcal{V}} x_{i,j} - d\right)^2$

Optimization strategy

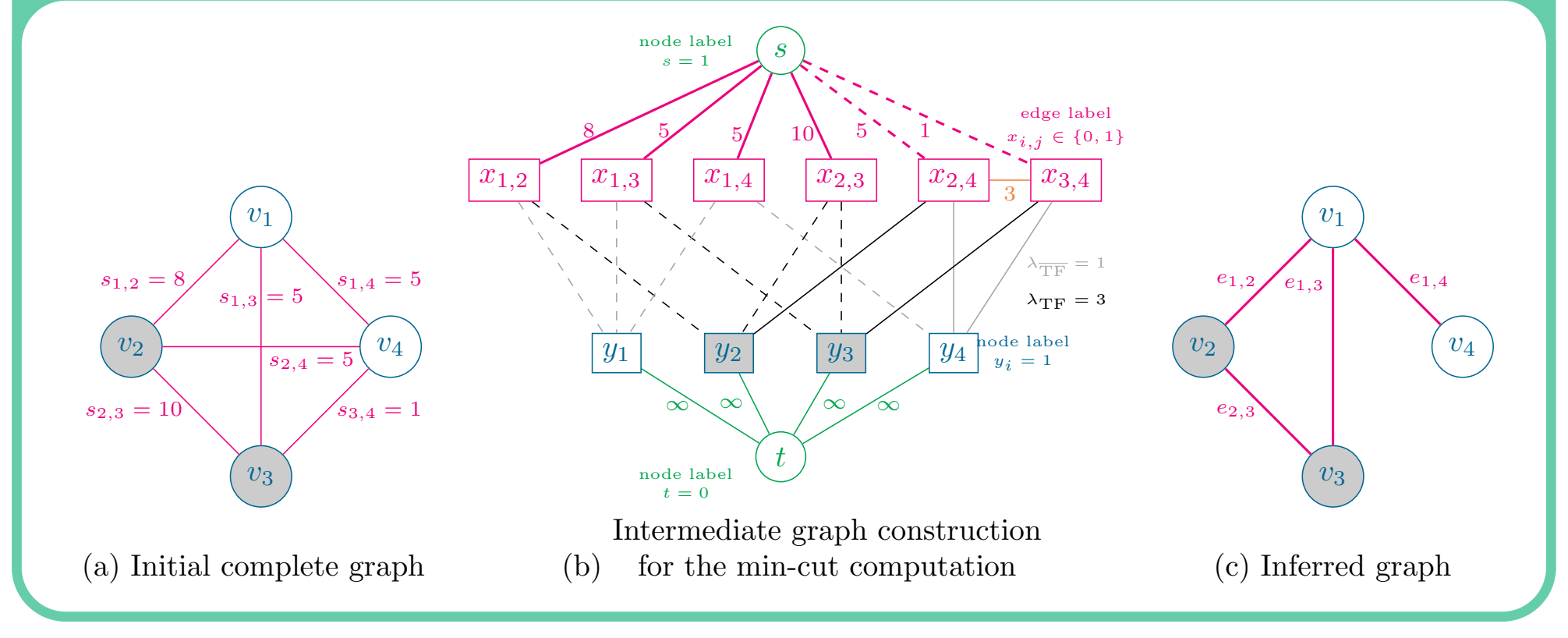
Objective: Design appropriate algorithms to compute the optimal labeling $\mathbf{x}^*$

- **BRANE Cut: Discrete Optimization *via* Maximal Flow algorithm** [4]

$$\underset{\mathbf{x}\in\{0,1\}^n}{\text{minimize}} \underbrace{\sum_{(i,j)\in\mathcal{E}} s_{i,j}|x_{i,j}-1| + \sum_{(i,j)\in\mathcal{E}} \lambda_{i,j}|x_{i,j}| + \mu \sum_{\substack{i\in\mathcal{V}\backslash\mathcal{T}\\(j,j')\in\mathcal{T}^2}} \alpha_{i,j,j'}|x_{i,j}-x_{i,j'}|}_{f}$$

- f: Sub-modular function
- Minimal Cut - Maximal Flow duality
- Maximal Flow algorithm applied to an appropriate flow network $\mathcal{G}_f$

Algorithm 1: Maximal Flow algorithm



(a) Initial complete graph (b) Intermediate graph construction for the min-cut computation (c) Inferred graph

- **BRANE Relax** [6]**: Continuous Optimization *via* Proximal methods** [1]

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \underbrace{\sum_{(i,j)\in\mathcal{E}} s_{i,j}(1-x_{i,j}) + \sum_{(i,j)\in\mathcal{E}} \lambda_{i,j}x_{i,j} + \mu \sum_{i\in\mathcal{V}\backslash\mathcal{T}} \left(\sum_{j\in\mathcal{V}} x_{i,j} - d\right)^2}_{f_1} + \underbrace{\iota_{[0,1]^n}(x)}_{f_2}$$

- f_1: differentiable function with β-Lipschitz gradient
- f_2: convex function (relaxation)
- Solved by Forward-Backward algorithm using Preconditioning and Block-Coordinate improvement strategies

Algorithm 2: 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
    $\boldsymbol{z}_n^{(k_n)} = \boldsymbol{x}_n^{(k_n)} - \gamma_n \mathbf{A}_{k_n}^{-1}\boldsymbol{\Omega}_{k_n}^\top \nabla\Phi(\boldsymbol{\Omega}\boldsymbol{x}_n - \mathbf{d})$
    $\boldsymbol{x}_{n+1}^{(k_n)} = \text{prox}_{\gamma_n^{-1}\mathbf{A}_{k_n}, f_1^{(k_n)}}(\boldsymbol{z}_n^{(k_n)})$
    $\boldsymbol{x}_{n+1}^{(k)} = \boldsymbol{x}_n^{(k)}, \quad k \in \{1,\ldots,p\} \setminus \{k_n\}$
```

Results

Comparison, on the DREAM4 in silico multifactorial challenge dataset [5] containing five networks, to two state-of-the-art methods:

- Information-theoretic score-based: CLR [2]
- Model-based: GENIE3 [3]

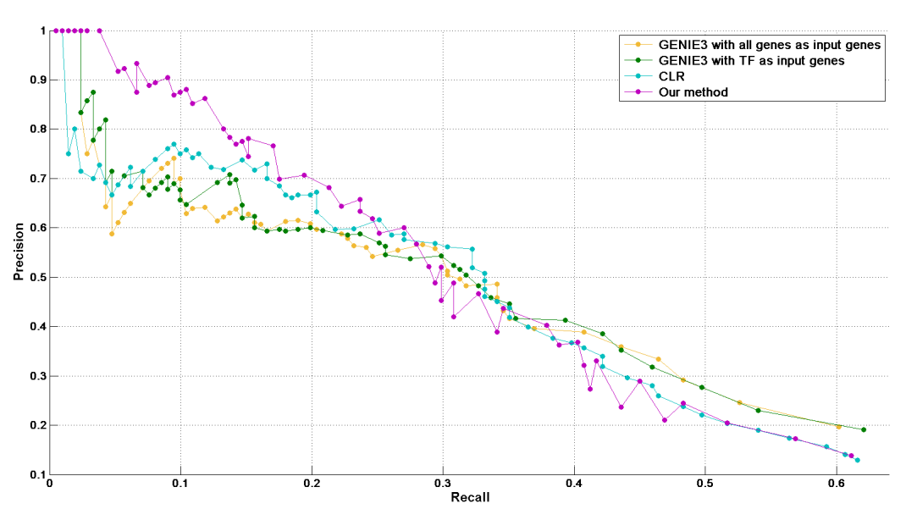
The evaluation is performed by computing:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad \text{and} \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$

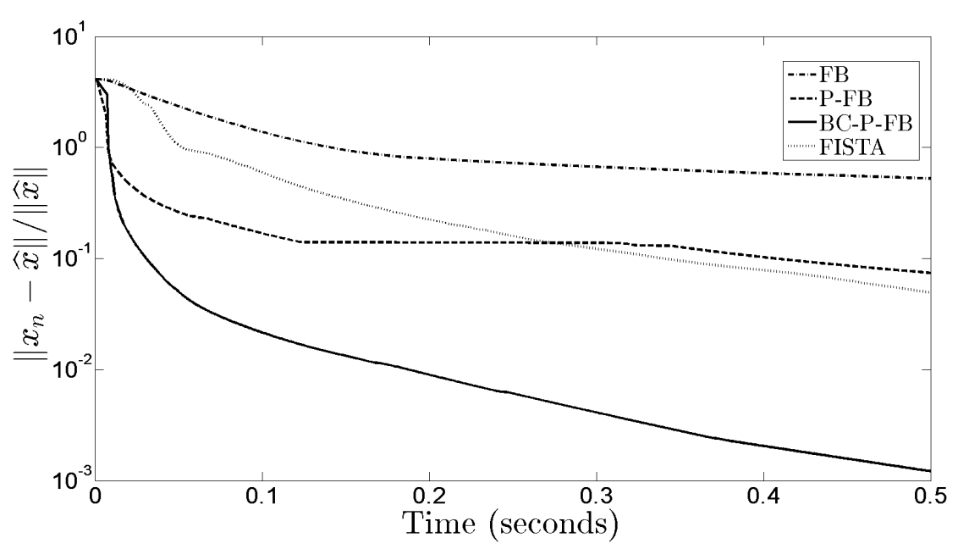
where TP: True Positive, FP: False Positive and FN: False Negative.

Results are given in terms of AUPR: Area Under the Precision-Recall curve.

Network	1	2	3	4	5
GENIE3	0.239	0.260	0.316	0.301	0.295
CLR	0.249	0.258	0.294	0.296	0.299
BRANE Cut	**0.256**	0.261	0.317	**0.317**	**0.316**
BRANE Relax	0.246	**0.264**	**0.321**	**0.317**	**0.317**



(a) Precision-Recall (PR) curves for various GRN inference method: CLR, GENIE3 and BRANE Cut



(b) Comparison of the convergence speed for various algorithms: FB, Preconditioned-FB, BlockCoordinate-P-FB and FISTA for BRANE Relax formulation

Conclusion

- Two variational formulations of the inference problem, taking into account structural *a priori*, deliver promising results
- On this tested dataset, CLR and GENIE3 are outperformed
- The continuous approach allows us to interpret the result as a confidence score of the edge presence
- 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. http://www.optimization-online.org/DBHTML/2013/12/4178.html.

[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.*, 5:8, 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):e12776, Sep. 2010.

[4] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE Trans. Patt. Anal. Mach. Int.*, 26:65–81, 2004.

[5] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Nat. Acad. Sci. U.S.A.*, 107(14):6286–6291, Apr. 2010.

[6] A. Pirayre, C. Couprie, L. Duval, and J.-C. Pesquet. Fast convex optimization for connectivity enforcement in gene regulatory network inference. To be presented in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2015.