# Discrete vs Continuous Optimization for Gene Regulatory Network Inference

Aurélie Pirayre*§, Camille Couprie*, Laurent Duval* and Jean-Christophe Pesquet§
* IFP Energies nouvelles, Direction Mécatronique et Numérique, Rueil-Malmaison, France
§ Univ. Paris-Est, LIGM, UMR CNRS 8049, France

***Abstract*—With the advent of high-throughput biological techniques, arose the need to handle gene expression data. Inferring Gene Regulatory Networks (GRNs) from this kind of data is especially useful for sketching transcriptional regulatory pathways and helps to understand phenotype variations. Given all pairwise gene similarity information, we formulate GRN inference as an energy minimization problem to determine the presence of edges in the final graph. Taking into account expected patterns in the graph structure, biological *a priori* are incorporated into the variational formulation. Different priors lead to different mathematical properties of the cost function, for which various optimization strategies can be applied. Experimental results show a performance improvement (in terms of Area Under the Precision-Recall curve) and/or computation time compared with state-of-the-art methods.**

## I. Introduction

One way of improving biological knowledge is to handle and analyze "omics" data generated by high-throughput techniques. Focusing on the context of transcriptomic data, the identification of genes involved in phenotypic variations is currently performed thanks to Gene Regulatory Network (GRN) analysis such as gene clustering [1]. A GRN is a graph containing gene regulatory pathways for a given living organism. It is obtained from gene expression signals: for each gene, the signal corresponds to the gene expression level in different conditions (physico-chemical or temporal conditions, culture medium or mutated strains). Then, inferring a GRN aims at selecting, among all plausible links, a subset of regulatory links reflecting actual regulatory relationships between genes. Unfortunately, recovering useful information from this collection of signals remains a difficult task due to the small number of observations (number of conditions) compared with the number of genes. In this work, we develop a novel variational approach for taking into account expected graph patterns according to some biological *a priori*. Translating such biological assumptions into an appropriate cost function, we thus formulate the GRN inference problem as an optimization one.

## II. Models

A complete gene network may be viewed as a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \cdots, v_g\}$ is a set of vertices (corresponding to the genes), $g$ is the number of genes, and $\mathcal{E} = \{e_1, \cdots, e_n\}$ a set of edges (corresponding to plausible gene interactions), the number of edges being $n = g(g-1)$. Inferring a GRN $\mathcal{G}^*$ from $\mathcal{G}$ aims at selecting an optimal subset of edges $\mathcal{E}^* \subset \mathcal{E}$ reflecting actual regulatory relationships between genes. This selection problem may be formulated by defining a cost function to minimize where the variables $x_{i,j}$ correspond to edge labels for $e_{i,j}$ such that $x_{i,j} = 1$ if the edge $e_{i,j}$ is in the final graph and 0 otherwise.

Weighting all possible pairwise gene relationships by the similarity $s_{i,j}$ between gene expression profiles for gene $i$ and $j$ and assuming that a reliable list of putative transcription factors (i.e. regulator genes), denoted by $\mathcal{T}$, is available, we define some biological and structural *a priori* which may be incorporated into our cost function based on two rationales: i) the larger the edge weights $s_{i,j}$, the more favorable the edge selection, ii) links involving a regulator gene are favored *via* the parameters $(\lambda_{i,j})_{(i,j)\in\mathcal{E}}$. Two variational priors were used according to the choice of function $\Phi$, leading to the following general criterion form to minimize:

$$\sum_{(i,j)\in\mathcal{E}} s_{i,j}(1 - x_{i,j}) + \sum_{(i,j)\in\mathcal{E}} \lambda_{i,j} x_{i,j} + \mu\Phi((x_{i',j'})_{(i',j')\in\mathcal{N}_{i,j}}), \quad (1)$$

where $\mu$ is a regularization parameter and, for every $(i,j) \in \mathcal{E}$, $\mathcal{N}_{i,j}$ denotes some local neighborhood of edge $e_{i,j}$. Depending on the prior used, mathematical properties of cost function (1) are changed and suitable optimization strategies have to be devised.

- Keeping the degree of regulated genes close to a constant number $d$ is enforced by choosing $\Phi$ as a composition of a linear averaging operator with a norm. A relaxation of the binary constraint on the vector $\mathbf{x}$ of edge labels is then necessary to minimize (1) efficiently by using recent convex optimization methods [2].
- Enforcing a co-regulation property (i.e. favoring a similar label for $x_{i,j}$ and $x_{i,j'}$ when genes $j$ and $j'$ are likely to act together) *via* a total variation like function $\Phi$ makes the criterion submodular. Thus, a discrete optimization process can be carried out, such as a maximum flow algorithm, in order to obtain an optimal labeling [3].

## III. Results

We performed comparisons of our approach with two state-of-the-art methods: Context Likelihood Relatedness (CLR) [4] and GENIE3 [5]. The performance was evaluated in terms of Area Under the Precision-Recall curves (AUPR), where the precision reflects the proportion of correctly inferred edges compared to the total number of inferred edges, while the recall indicates the proportion of correctly inferred edges with respect to the edges corresponding to the gold standard. Results obtained on the synthetic data from the DREAM4 multifactorial challenge are quite promising. In term of AUPR, our method outperforms both CLR and GENIE3 approaches while having a low computational complexity.

## References

[1] D. Zhu and A. O. Hero, "Network constrained clustering for gene microarray data," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 5, Philadelphia, PA, USA, Mar. 18-23, 2005, pp. 765–768.
[2] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "A block coordinate variable metric forward-backward algorithm," Tech. Rep., 2013.
[3] L. R. Ford, Jr. and D. R. Fulkerson, "Maximal flow through a network," *Canad. J. Math.*, vol. 8, no. 0, pp. 399–404, Jan. 1956.
[4] 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.*, vol. 5, p. 8, 2007.
[5] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, "Inferring regulatory networks from expression data using tree-based methods," *PLoS One*, vol. 5, no. 9, p. e12776, Sep. 2010.