Sparse and Stable Reconstruction of Genetic Regulatory Networks Using Time Series Gene Expression Data

Roozbeh Manshaei
Ryerson University
ECE Department, Eric Palin Hall,
87 Gerrard Street East, Toronto, Canada
roozbeh.manshaei@ryerson.ca

Matthew Kyan
Ryerson University
ECE Department, Eric Palin Hall,
87 Gerrard Street East, Toronto, Canada
mkyan@ee.ryerson.ca

ABSTRACT
Gene regulatory networks represent the regulatory and physical interactions between genes of an organism. In this application, we are presented with a set of time series gene expression data, from which an unknown topology describing the regulatory interactions between genes must be inferred. To this end, we formulate an algorithm for reconstructing a genetic regulatory network to explain time series data obtained from genetic experiments. Our algorithm minimizes the trade-off between of the sparsity of gene interactions in the inferred network and the best model accuracy, where stability and prior knowledge are considered as constraints. Our algorithm is applied to time series gene expression data from yeast cell-cycle regulation, and results show improved reconstruction. The convex nature of the proposed model makes it suitable for application to large-scale networks.

Categories and Subject Descriptors
G.1.6 **[Optimization]**: Convex Programming; I.5 **[Pattern Recognition]**; I.5.4 **[Applications]:** Signal Processing

General Terms
Theory and application

Keywords
Genetic network reconstruction, optimization, sparsity, stability.

1. INTRODUCTION
The ability to measure the expression of genes on a genome-wide scale has garnered recent attention from biologists for deciphering the dynamics of gene interaction. A significant role in the progress of systems biology is played by engineering methods [1], in particular, model-based optimization. A key question pertains to which criteria (objective functions) should be optimized in genetic networks [2]. Optimization has been used to understand optimal biological circuit design, biochemical networks, and in inference of bio-molecular networks, such as transcriptional regulatory networks and gene regulatory networks (GRNs) [3].

In order to better scale, we present an algorithm for reconstructing the smallest GRN using genetic experimental data, using a weighted convex relaxation, which converts an inherently concave problem into one that is convex. Using prior knowledge, we can infer whether one gene affects another gene or not, or whether this effect is positive (activation) or negative (inhibition). In addition, stability and sparsity of genetic networks is considered in the model formulation. This allows for the application of convex optimization whereby linear and other constraints are considered to achieve both a best fit on the genetic data while satisfying a priori knowledge on gene interconnections.

2. RECONSTRUCTION ALGORITHM
2.1 Ordinary Differential Equations Model
The processes of transcription and translation in a gene regulatory network (GRN) consisting of n genes can be modeled as the following dynamic system:

$$\begin{cases} \dot{g} = Cg + Sd \\ d = f(g) \end{cases} \quad (1)$$

where: $g = [g_1, g_2, \dots, g_n]^T \in R^n$ represents the concentration of mRNAs over n genes; $C = diag\,[-c_1, -c_2, \dots, -c_n] \in R^{n\times n}$ with c_i as the degradation rate of gene i; $d = [d_1, d_2, \dots, d_n]^T \in R^n$ represents the reaction rates; and $S \in R^{n\times n}$ is the reaction topology of the biological network. Assuming the reaction rate d is a linear combination of g and $E \in R^{n\times n}$ is the coefficient matrix ($d = Eg$). Then: $\dot{g} = Cg + SEg$.

A standard discretization method like zero-order hold is used in this case on m observation points for a sampling time Δt, thus:

$$g(k+1) = \left(e^{C\Delta t} + \left(e^{C\Delta t} - I\right)C^{-1}SE\right) * g(k) = W * g(k) \quad (2)$$

where SE describes the structure of gene regulatory network: $se_{ij} > 0$, $se_{ij} = 0$ and $se_{ij} < 0$ if gene j activates gene i, does not regulate gene i, and represses gene i respectively. To reconstruct the gene regulatory network, we need to determine the sign of elements in SE: akin to estimating the elements of matrix W.

2.2 Problem Formulation
The proposed method is based on convex optimization, for which there are a set of efficient techniques for resolution [4]. Key observations in modeling a gene regulatory network that we consider in our formulation (as constraints), include:

Error: Let $G \in R^{n\times m}$ be gene expression levels of n genes at m time points. Our problem is divided into $m-1$ sub-problems (unique time transitions), each solved individually. Let $G^{(p)}$ and $G^{(p+1)}$ $(p = 1, \dots, m-1)$ be the p^{th} and $(p+1)^{th}$ columns of G, respectively. Assuming noise, thus we can define:

$$G^{(p+1)} - W^{(p\to p+1)} * G^{(p)} = \varepsilon^{(p)}, \quad (p = 1, \dots, m-1) \quad (3)$$

wherein we try to minimize $\varepsilon^{(p)} \in R^n$ as a function of $W^{(p\to p+1)} \in R^{n\times n}$, while obtaining a minimal model for $W^{(p\to p+1)}$ and satisfying any a priori constraints that might be imposed on $W^{(p\to p+1)}$. We then establish the error level that a model can attain. For instance, using the total squared error as the error criterion: $\left\|G^{(p+1)} - W^{(p\to p+1)} * G^{(p)}\right\|_2 < \varepsilon^{(p)}$.

Sparsity: The requirement that $W^{(p\to p+1)}$ be sparse is related to biological networks being sparse in nature. We achieve this using a cardinality function (number of non-zero matrix elements):

$$card(x) = \sum_{i=1}^{n} I(x_i), \qquad I(z) = \begin{cases} 1, & z \neq 0 \\ 0, & Otherwise \end{cases} \quad (4)$$

The problem of reconstructing the sparsest W that satisfies the error constraint leads to the constrained optimization problem:

$$\begin{cases} minimize & \lambda\, card\left(W^{(p\to p+1)}\right) + (1-\lambda)\varepsilon^{(p)} \\ subject\ to & \left\|G^{(p+1)} - W^{(p\to p+1)} * G^{(p)}\right\|_2 < \varepsilon^{(p)}, \\ & \varepsilon^{(p)} > 0 \end{cases} \quad (5)$$

BCB'13, September 22 - 25, 2013, Washington, DC, USA
ACM 978-1-4503-2434-2/13/09."

The problem data are the matrix G and the parameter $0 < \lambda < 1$ while the matrix $W^{(p\to p+1)}$ and fitting error $\varepsilon^{(p)}$ are variables. λ is used to control the trade-off between sparsity ($card(W^{(p\to p+1)})$), and best fitting ($\varepsilon^{(p)}$).

Weighted Relaxation: The sparsity function ($card(W^{(p\to p+1)})$) in the objective function is concave. Methods to solve [5] are very slow, and do not scale well to medium and large networks. Thus, convex relaxation of the cardinality cost function is employed by replacing sparsity with the weighted l_1 − norm in (6), where for small and positive δ, b_{ij}s are chosen as per (7):

$$card\left(W^{(p\to p+1)}\right) = \sum_{i,j=1}^{n} b_{ij}|w_{ij}| \quad (6)$$

$$b_{ij} = \left(\frac{\delta}{\delta+|w_{ij}|}\right) \quad (7)$$

Consequently, we can eliminate any weak genetic interactions in the final inferred $W^{(p\to p+1)}$:

$$\begin{cases} minimize \quad \lambda \sum_{i,j=1}^{n} b_{ij}|w_{ij}| + (1-\lambda)\varepsilon^{(p)} \\ subject\ to \ \left\|G^{(p+1)} - W^{(p\to p+1)} * G^{(p)}\right\|_2 < \varepsilon^{(p)}, \varepsilon^{(p)} > 0 \end{cases} \quad (8)$$

Prior knowledge: if available, can be encoded as a sign matrix $H = (h_{ij}) \in \{0, +, -, ?\}^{n\times n}$, for which positive, negative and no interactions between any two genes are signed as (+), (-) and (0) respectively, while a (?) represents no prior knowledge. We consider prior knowledge in (8) with added linear constraints:

$$W^{(p\to p+1)} \in H \Leftrightarrow \begin{cases} w_{ij} > 0 \quad if \quad h_{ij} = + \\ w_{ij} < 0 \quad if \quad h_{ij} = - \\ w_{ij} = 0 \quad if \quad h_{ij} = 0 \\ w_{ij} \in R \quad if \quad h_{ij} = \{+ \ or \ - \ or \ 0\} = ? \end{cases} \quad (9)$$

Stability: behavior of GRNs has important biological implication. According to [6] for discrete models, $W^{(p\to p+1)}$ is stable if (10) is satisfied, ultimately reducing to our proposed formulation (11).

$$\sum_{i\neq j}|w_{ij}| \leq |w_{ii}|\ ,\ i = 1, \dots, n \quad (10)$$

$$\begin{cases} min \quad \lambda\left(\sum_{i,j=1}^{n} b_{ij}|w_{ij}|\right) + (1-\lambda)\varepsilon^{(p)} \\ s.t. \quad \left\|G^{(p+1)} - W^{(p\to p+1)} * G^{(p)}\right\|_2 \leq \varepsilon^{(p)} \ , \varepsilon^{(p)} \geq 0 \\ \quad\quad W^{(p\to p+1)} \in H,\ \sum_{j=1}^{n}|w_{ij}| \leq 0\ ,\ i = 1, \dots, n \end{cases} \quad (11)$$

3. RESULTS AND DISCUSSION

In order to evaluate our algorithm, we obtained optimal Ws based on yeast (*Saccharomyces cerevisiae*) cell cycle microarray time series data sets. We have focused on twelve yeast genes playing key roles in the control of cell cycle from the Yeast Proteome Database [7]. Our algorithm is used as an identification method to find all possible genetic interaction networks that fit the data for the set of twelve genes. To test the capability of our algorithm, we used the algorithm 1 for 18+24 times to extract gene regulatory network (GRN) structures from inferred W's, and evaluate against GRNs extracted from KEGG database [8]. Our algorithm was implemented in MATLAB using the CVX toolbox for convex optimization problems [5] and run on an Intel Core i7, 3.40 GHz processor with 8 GB RAM.

Algorithm 1: GRN Reconstruction Algorithm

Control parameter $0 < \lambda < 0.2$
Initialize b_{ij}'s = 1 and w_{ij}'s = 0 i, j = 1,2,..,n
$\delta = 0.1, \gamma = 10^{-3}$
While $\left\|W_{new}^{(p\to p+1)} - W_{old}^{(p\to p+1)}\right\|_2 \geq \gamma$
  Solve the linear program (11)
  Update b_{ij}'s by (7)
End

Table 1: Comparison of the proposed algorithm with other methods using statistical criteria.

	Sensitivity	Precision	F-Score
DBN [9]	12.1%	80%	21%
VBEM [10]	15.2%	62.5%	23.5%
PF subjected to LASSO [11]	21.2%	70%	32.5%
Proposed	27%	90%	41.5%

We compared the reconstructed GRNs against four algorithms for the same dataset. The identified networks from our proposed algorithm have the best performance among other three methods, demonstrating suitable matches with pathways reported in KEGG. Indeed, we observe that our algorithm is capable to extract 9 true connections out of 33 (validated as ground truth from the literature), versus 4,5, and 7 true connections by DBN [9], VBEM [10], and PF subjected to LASSO [11] respectively (Table 1).

4. CONCLUSIONS

In this paper, we described a novel algorithm based on convex optimization for gene regulatory network reconstruction. We considered the problem of identifying an optimal model that best explains genetic data. We relaxed the cardinality function by employing its weighted $l-1$ approximation and extended our formulation to explain a priori knowledge on the network structure, as well as stability constraints. The convex nature of our algorithm leads to a solution that can handle large-scale reconstruction problems, and was successfully validated using yeast cell cycle data.

5. REFERENCES

[1] Kremling A. et al: **Systems biology - an engineering perspective**. *Journal of biotechnology*, 129(2), 329-351, 2007.

[2] Nielsen J.: **Principles of optimal metabolic network operation**. *Molecular Systems Biology*, 3:2, 2007.

[3] Thomas R. et al: **A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data**. *BMC Bioinformatics*, 2007.

[4] Boyd S., et al: **Convex optimization,** Cambridge University Press, 2004.

[5] Boyd S.: **l1-norm methods for convex cardinality problems**, Lecture Notes for EE364b, 2007.

[6] G. Golub, et al: **Matrix Computations**, *Johns Hopkins Press, 2nd ed.,* 1989.

[7] Costanzo M.C., et al.: **The Yeast Proteome Database and Caenorhabditis elegans Proteome Database**, *Nucleic Acids Research*, 28(1):73-76, 2000.

[8] Kanehisa M., et al: **KEGG for integration and interpretation of large-scale molecular datasets**, *Nucleic Acids Research*, 40:D109-D114, 2012.

[9] Kim S., et al: **Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data**, *Biosystems*, 75(1–3):57-65, 2004.

[10] Tienda-Luna I. M., et al: **Sensitivity and Specificity of Inferring Genetic Regulatory Interactions with the VBEM Algorithm**, *IADIS International Journal on Computer Science and Information Systems*, 4(1):54-63, 2009.

[11] Noor A, et al: **Inferring Gene Regulatory Networks via Nonlinear State-Space Models and Exploiting Sparsity**, *IEEE/ACM Trans Comput. Biol. Bioinform.*, 9(4):1203-1211, 2012.