

A comparative approach for multiple gene network inference using time-series gene expression data

Guillaume Bourque,¹ David Sankoff²

Keywords: gene network inference, time series, differential equations, comparative approach

1 Introduction

Microarrays with their massively parallel capabilities for measuring gene expression have become an attractive tool to reverse-engineer gene regulatory networks [6, 7]. The task is challenging as the genes which are part of such networks are typically hidden within the thousands genes found in the genomes. Experiments remain expensive and, with limited data, the problem of sorting through the combinatorial number of potential networks is most difficult.

When modeling a system consisting of n genes with differential equations, even under the simplest linear model, there are $n(n-1)$ directional effects, n self effects and n constant effects for a total of $n(n+1)$ unknown parameters. Assume the gene expression of these n genes is measured at T time points, then we have a system with $n(n+1)$ unknowns and nT equations. Because in gene expression experiments we typically have that $T \ll n$, the problem is under-determined and extra constraints must be incorporated into the model to unambiguously resolve it. These constraints involve for instance the smoothness of the differential equations [1]. But recently, a different approach has been to favor the simplicity of the overall solution by minimizing the number of non-zero coefficients [4, 9, 10, 11]. Such an approach makes sense biologically since it can be expected that each gene interacts with a limited number of other genes. It will also reduce the complexity of the network and focus on the most important interactions. The fact that very few interactions are necessary to explain the gene expression data has been demonstrated by Holter et al. [2, 3] and Hörnquist et al. [5].

Our system was design for the recovery of gene interactions concurrently in many gene regulatory networks related by a graph or a tree. See Fig. 1 for an example. Suppose we are studying a certain regulatory network in different species of known phylogeny. We can think of the different networks as being related to each other in that way and use this information. Alternatively, we might be interested in the development stages of this network or we could be studying the same system but in different tissues related at a different level. The idea is that, given gene expression data for each species, or each stage of development, or each tissue, we seek to recover each individual network while minimizing a cost based on the differences along the edges of the graph or the tree. We show how this comparative framework allows new insights and facilitates the gene network inference process.

2 Method

The method we propose models the network by a system of linear equations and it limits the number of non-zero coefficients, or regulators, for each gene. The idea is not to force a fixed number of interactions [10], or set a restrictive upper bound, but to favor solutions with few interactions [4, 11]. We use a generalized stepwise multiple linear regression to solve the system of equations.

¹Centre de Recherches Mathématiques, Université de Montréal. E-mail: bourque@crm.umontreal.ca

²Department of Mathematics and Statistics, University of Ottawa.

Our approach not only can incorporate biological knowledge at every stage of the decision process; it can be used for the inference but also for the revision of gene regulatory networks.

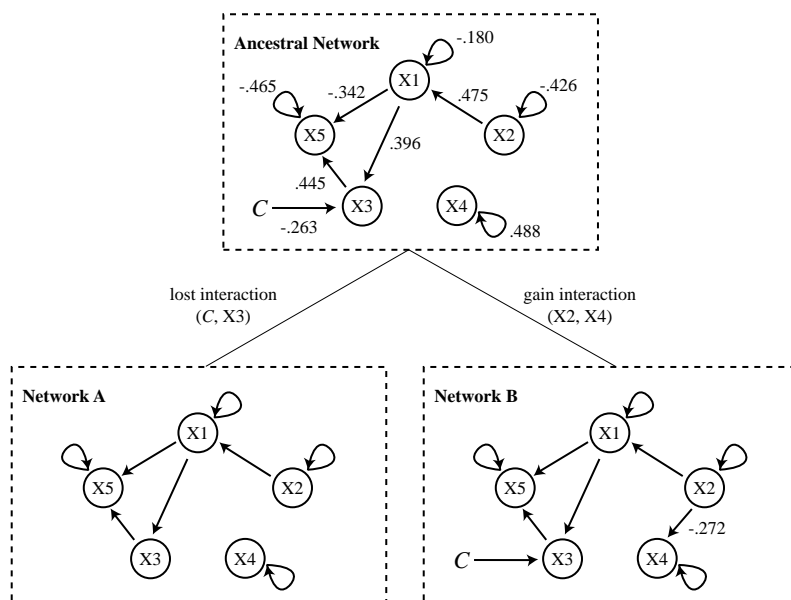


Figure 1: Example with two observable gene networks (A and B) and an unobservable ancestral network which are related by a tree. The networks describe the interactions between five genes (X1, X2, X3, X4 and X5). The interaction coefficients are only displayed in the ancestral network but they are also preserved in network A and B except for the interaction (C, X3), where C stands for a constant effect, which is lost in network A and the interaction (X2,X4) which is added in network B. Typically, gene expression data would be collected for both observable networks and the goal would be to recover all the interaction coefficients.

References

- [1] P. D’Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. *Pac Symp Biocomput*, pages 41–52, 1999.
- [2] N. S. Holter, A. Maritan, M. Cieplak, N. V. Fedoroff, and J. R. Banavar. *PNAS*, 98(4):1693–1698, Feb 2001.
- [3] N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff. *PNAS*, 97(15):8409–8414, Jul 2000.
- [4] M. De Hoon, S. Imoto, and S. Miyano. In S. Lange, K. Satoh, and C. H. Smith, editors, *Fifth International Conference on Discovery Science*, pages 267–274, 2002.
- [5] M. Hornquist, J. Hertz, and M. Wahde. *Biosystems*, 65(2-3):147–156, Mar 2002.
- [6] D. J. Lockhart and E. A. Winzeler. *Nature*, 405(6788):827–836, Jun 2000.
- [7] A. Schulze and J. Downward. *Nat Cell Biol*, 3(8):190–195, Aug 2001.
- [8] E. van Someren, L. Wessels, and M. Reinders. In *Proc. of SPIE*, 2001.
- [9] E. P. van Someren, L. F. A. Wessels, and M. J. T. Reinders. In J. Biemond, editor, *21st Symp. on Information Theory in the Benelux*, pages 215–222, Wassenaar (NL), 2000.
- [10] E. P. van Someren, L. F. A. Wessels, M. J. T. Reinders, and E. Backer. In *Proc. of ICSB*, 2001.
- [11] M. K. S. Yeung, J. Tegner, and J. J. Collins. *PNAS*, 99(9):6163–6168, Apr 2002.