Inferring gene regulatory networks from time expression profiles

Mukesh Bansal and Diego di Bernardo
Telethon Institute of Genetics and Medicine, Via Pietro Castellino 111, Naples, Italy
bansal,dibernardo@tigem.it

Keywords: Gene network, gene-expression, computational biology

Introduction

Recent developments in large-scale genomic technologies, such as DNA microarrays and mass spectroscopy have made the analysis of gene networks more feasible. However, it is not obvious how the data acquired through such method can be assembled into unambiguous and predictive models of these networks. In a recent study our group developed an algorithm (Network Identification by multiple regression – NIR) that used a series of steady state RNA expression measurements, following transcriptional perturbations, to construct a model of a 9 gene network that is a part of larger SOS network in E.Coli[1]. Though the NIR method proved highly effective in inferring small microbial gene networks, its practical utility is limited because it requires: (i) prior knowledge of which genes are involved in the network of interest; (ii) the perturbation of all the genes in the network via the construction of appropriate episomal plasmids; (iii) the measurement of gene expressions at steady state (i.e., constant physiological conditions after the perturbation). This experimental setup is unpractical for large networks, it is not easily applied to higher organisms, and, most importantly, it is not applicable if there is no prior knowledge of the genes belonging to the network.

Here we are proposing a new algorithm that can infer the network of gene-gene interactions to which a gene of interest belongs and identify its direct targets, using the perturbation of only one of the genes in the network. To this end, we need to measure gene expression profiles at multiple time points following perturbation of only the known gene, or genes, and without the need of the steady-state assumption.

Algorithm

Here we are modeling the network as a system of ordinary differential equations describing a stable linear system that is completely observable and controllable of the form:

$$\dot{x} = \mathbf{A} \cdot x + u + \varepsilon \qquad (1)$$

where x is a Nx1 vector representing the concentrations of RNAs, u is a Nx1 vector representing an external perturbation to the rate of accumulation of x , ε is an Nx1 vector representing measurement and biological noise, that we will assume to be white Gaussian noise, and **A**, the network model, is an NxN matrix of coefficients describing the regulatory interactions between the species in x. We assume that for each of the N genes, we measure the gene expression profile at M time points following the perturbation, with M<<N.

To solve equation (1) we need the number of time points to be M≥N. To artificially increase the number of experimental time points, we tried two different methods: (1) interpolating the time series to double the data points by using cubic polynomial 'spline' interpolation; and (2) extrapolating the time series by predicting future time points with an auto-regressive model (AR) trained on the experimental time points. We then used all the data (collected from experiment and generated ones) to solve equation 1 by using the Least Square method.

To understand what is the best kind of perturbation experiment one can perform to recover the gene network, we used two different types of perturbations to the gene of interest. Input 1 simulates the overexpression of the gene of interest by using, for example, addition of tetracycline in a tetracycline-inducible expression system. Input 2 simulates the overexpression of the gene, followed by its downregulation (addition and removal of tetracycline).

Results and Conclusion

We simulated the expression data for a sparse network of 8 genes with a maximum connectivity of 4 connections per genes, following the perturbation to only one gene (the gene of interest). The simulated data consisted of 8 gene expression profiles measured at 9 time points, sampled at regular interval. We added noise with different strengths to the simulated data. We tested the performance of the algorithm in 100 different sparse networks of 8 genes. We did this for both kind of inputs and for both the methods described above. To assess how well the inferred model described the real network, we counted the

number of genes whose upregulation or downregulation at steady state, following the perturbation of the gene interest, can be correctly simulated by the model. In addition, the model can be used to identify the genes that are direct targets of the perturbed gene by looking at the connections in the recovered network matrix A.
From the simulations we found that extrapolation works better then interpolation for all noise levels in inferring the correct model of the network. (Fig 1). Also we find that an experiment in which the gene of interest is overexpressed and then downregulated yields more information on the regulatory network surrounding that gene of interest, than overexpression alone.
The result shows that even for high noise levels we can correctly identify about 60% of the direct targets of the perturbed gene.

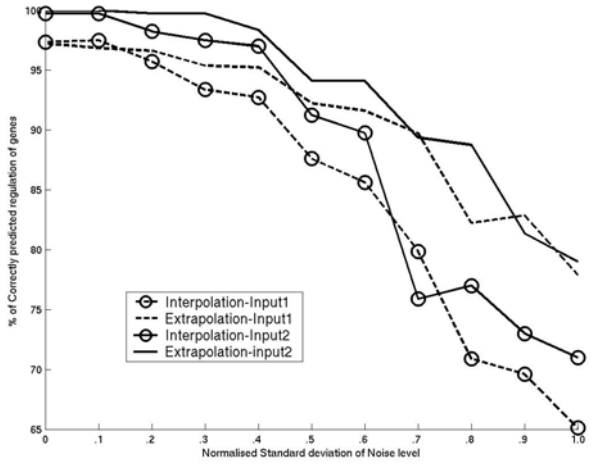

Fig.1 Percentage of correctly predicted regulation of genes for a various level of noise for different method and different inputs.

Future work
In future we will apply dimension reduction techniques in conjunction with the algorithm described above to be able to infer large gene networks containing hundreds of genes.

References

[1] Gardner TS, di Bernardo D, Lorenz D and Collins JJ. (2003) *Inferring gene networks and identifying compound mode of action via expression profiling.* Science 301, 102-5.

[2] Ljung L. (1999) *System identification: Theory for the User.* (Prentice Hall, Upeer Saddle River, NJ).

[3] Van Overschee P and De Moor, B. (1996) *Subspace Identification for Linear Systems*. (Kluwer Academic Publishers, Boston).