Gene network reverse engineering: comparison of algorithms

ID - 127

Bansal Mukesh[1,2], Belcastro Vincenzo[1,3], Ambesi Alberto[4], Di Bernardo Diego[1,2]
[1]Telethon Institute of Genetics and Medicine, Via P.Castellino 111, Naples, Italy
[2]European School of Molecular Medicine, Naples, Italy
[3]University of Naples 'Federico II', Naples, Italy
[4]Joint Centers for Systems Biology Columbia University, New York, NY 10032

Motivation

Inferring, or 'reverse-engineering', gene networks can be defined as the process of identifying gene interactions from experimental data through computational analysis [1]. Gene expression data from microarrays are typically used for this purpose. We show that reverse-engineering algorithms are indeed able to correctly infer regulatory interactions among genes, at least when one performs perturbation experiments complying with the algorithm requirements [2]. These algorithms are superior to classic clustering algorithms for the purpose of finding regulatory interactions among genes, and, although further improvements are needed, have reached a discrete performance for being practically useful.

Methods

Mathematical Models We tested and compared different reverse-engineering algorithms for which ready-to-use softwares are available and that has been tested on experimental data sets. We also compared them to classic clustering algorithm. Tool we used for Bayesian inference is BANJO [3], for Information-Theoretic approach we used ARACNe [4] and NIR [5] for Ordinary differential equation model. For clustering we used the classical hierarchical clustering.

Data Sets To test the performance of mentioned algorithms, we generated an in-silico (*) data coming a linear model described by

$X(t) = A X(t) + B U(t)$

where X(t) is represents the measured concentrations of mRNAs at time t following the perturbation. X represents the measured rate of change of X at time t. U represents external perturbations to the rate of accumulation of X at time t, B represents the effect of the perturbation on the genes, and A, the connectivity matrix, is an NxN matrix of coefficients describing the regulatory interactions between the species in X. We generated the data set of various size (10, 100 and 1000 genes), different types (time series and steady state), different conditions (single and multiple gene perturbation), and added different noise to the system. We tested the performance of all models on all data sets and also compared their performance by varying the number of expression used.

Results

In-silico analysis gives reliable guidelines on algorithms performance in line with the results obtained on real datasets: ARACNE (Information theoretic approach) performs well for steady-state data and can be applied also when few experiments are available, as compared to the number of genes. BANJO (Bayesian Networks) is very accurate, but with a very low sensitivity, on steady-state data when more than 100 different perturbation experiments are available, independently of the number of genes. NIR works very well for steady-state data, also when few experiments are available, but requires knowledge on the genes that have been perturbed directly in each perturbation experiment. We found all these algorithms to be superior to classic clustering algorithms for the purpose of finding regulatory interactions among genes A general consideration is that the nature of experiments performed in order to challenge the cells and measure gene expression profiles can make the task of inference easier. From our results, local perturbation experiments, i.e. Single gene over-expression or knock-down, seem to be much more informative than global perturbation experiments, i.e. over-expressing tens of genes simultaneously or submitting the cells to a strong shock.

Email: bansal@tigem.it