Bayesian inference of transcriptional networks from gene temporal expression patterns

ID - 232

Di Camillo Barbara[1], Toffolo Gianna[1], Cobelli Claudio[1]
[1]University of Padova, Information Engineering Department

Motivation

A main objective of System Biology is to infer on transcriptional and protein-protein regulatory interactions from gene and protein dynamic expression data, a learning process known as reverse engineering. At the moment, high-throughput technologies to measure protein expression as microarrays do for RNA are not available, therefore microarray data are used as a proxy of protein activity. This approximation and the heuristic solutions adopted to overcome the analytical and computational problems that rise from data dimensionality (thousands of genes measured in tens of conditions), leads to a great number of false negative and false positive interactions. The use of a priori knowledge on DNA-protein and protein-protein interactions can help to overcome these limitations. A natural reverse engineering approach to integrate a priori knowledge in the learning process is represented by Bayesian networks, which however, are efficient if applied to small systems of few nodes (the genes) with a large number of observations (the measurements), which is not the case of microarrays studies. Recently, module networks have been introduced in the literature to model sets of genes with similar profiles as a single variable, thus diminishing the number of nodes in the network and augmenting the number of observations for each node (each gene in the module contributes to observations for the node). At each iteration first the module network is identified based on given sets of genes characterized by the same temporal pattern, second the temporal patterns are updated based on the network results. However, this process does not use a priori knowledge, neither for network reconstruction or pattern definition.

Methods

Here we concentrate on transcription regulation and present a method to integrate a priori knowledge with microarray data in a module network approach that searches for the main temporal pattern of expression and for sets of transcription factors determining the observed dynamics. The method iterates a Bayesian network search with a pattern search as in module networks, but integrates a priori knowledge in both steps. The main computational steps of the method are described in the following.

1) The potential transcriptional regulators of the genes under analysis are identified by data base mining. 2) A contingency matrix is constructed with all the genes on the rows and only regulators on the column. The element (i, j) in the matrix is equal to 1 means if j is a known regulator of i; is equal to 0 if there is no evidence that j regulates i; is greater than 0 and lower than 1 if there is some evidence that j regulates i (this value is assigned e.g. using a probability score proportional to the probability to observe characteristic motifs of transcription factor j in the promoter region of gene i).

3) The main temporal patterns are searched separately for regulatory genes and for all the other genes, using a recently proposed model-based clustering that uses expectation maximization, iteratively performing a temporal pattern search (E step) and a gene-specific parameter identification step (M step). The method is based on a linear model whose parameters can be identified using weighted least squares, thus it explicitly accounts for the measurement error, does not require the user to fix the number of clusters and is not computationally demanding.

4) Bayesian networks model is applied to solve the networks, where each node represents a temporal pattern and only regulator patterns are allowed to have outcoming edges (i.e. to regulate transcription). A priori knowledge is codified using the contingency table identified at step 2.

5) The network identified at step 4) defines a regulator activity table organized as the contingency table at step 2, but with possibly different values. This table is used to re-define the temporal patterns. Namely, groups of regulated genes are identified as sets of genes sharing the same regulators and for each of these groups the main temporal patterns are searched; groups of regulatory genes are identified as sets of genes sharing the same regulated genes and for each of these groups the main temporal patterns are searched; 6) Steps 4 and 5 are re-iterated until the identified network does not change.

Results

The method was implemented in R and tested on 100 synthetic data-sets of gene expression data generated from networks of known topology, simulating different degrees of a priori knowledge in terms of percentage of known regulators. Results indicate that the use of temporal patterns allows diminishing the number of false negatives; whereas the use of a priori knowledge allows diminishing the number of false positives.

Email: dicamill@dei.unipd.it