Bayesian approaches for reverse engineering of cellular networks: a performance evaluation on simulated data

Ferrazzi F (1), Sebastiani P (3), Kohane IS (2), Ramoni MF (2), Bellazzi R (1)

(1) Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, Italy
(2) Children's Hospital Informatics Program and Harvard Partners Center for Genetics and Genomics, Harvard Medical School, Boston, USA
(3) Department of Biostatistics, Boston University School of Public Health, Boston, USA

Motivation

Time series measurements of gene expression or protein concentrations allow us to model the cell as a dynamic system, whose instantaneous state can be characterized by a set of state variables. Dynamic Bayesian networks (DBNs) are a special class of BNs particularly suited to study dynamic data. In particular, Linear Gaussian Networks allow researchers to avoid information loss associated with discretization and render the learning process computationally tractable even for hundreds of variables. However, it is often argued that linear models cannot capture the complex nonlinear dynamics of cellular systems. For this reason, we here propose a model that uses a linear regression of nonlinear transformations of the parent values. We evaluate both approaches using simulated data produced by a mathematical model of cell cycle control.

Methods

Assuming to have a database of measurements for n genes/proteins in p consecutive time points, it is possible to derive the DBN which encodes the dependencies over time of the random variables representing gene expression or protein concentration values. Supposing that the process under study is first order Markovian and stationary and that no instantaneous relationship between the values of two variables is possible, we need to learn only the transition network between the variables at time t and at time t + 1. To this aim, a probability model and a search strategy must be chosen. Linear Gaussian Networks treat variables as continuous and exploit a linear regression model to describe the conditional mean of a variable with respect to its parents. We here propose also a model in which this conditional mean is instead described as a linear combination of nonlinear functions of the parents. As it is reasonable to assume that expression/protein levels cannot indefinitely grow in proportion to their parent values, we decided to use the hyperbolic tangent function, in order to model a saturated effect of the parent on its child. In accordance with the Bayesian literature, we look for the network with maximum posterior probability given the data: to this aim, we exploit a finite horizon local search and we explore the dependency of each variable on all the variables at the previous time point.

Results

In order to have some insights regarding the suitability of Gaussian networks to describe relationships among genes or proteins, we decided to carry out an experiment with simulated data coming from a model of the budding yeast cell cycle (Chen et al., Mol. Biol. Cell, 2004). The whole model contains 36 differential equations: almost all the 36 variables represent protein concentrations, while the others represent the mass and the timing of cell cycle events. We used the profiles simulated in the case of wild-type cells and sampled values every 5 mins, from time 0 to 100 mins (about one cell cycle length). We analyzed the obtained dataset with the proposed dynamic Bayesian network approach, using both the linear regression model and its modification with nonlinear functions. If we consider as "true parents" of a variable A the other variables that appear in the differential equation describing A's dynamics, we can compare these with the parents found by the DBN algorithm. It is therefore possible to calculate the recall and precision: the recall corresponds to the fraction of "true parents" correctly inferred by the DBN algorithm, while the precision is the fraction of inferred parents that are also "true parents". However, considering only the parent variables in the regulatory network can constitute a very restrictive criterion: the value of a variable A at time t is in fact at the same time influenced by its parent values at time t - 1 and an important determinant for its children values at time t + 1. We therefore decided to repeat the

accuracy calculations comparing the "true parents" of each node with the variables in its Markov blanket, given by the union of its parents, its children and the parents of its children. The recall obtained with the nonlinear function model is always slightly higher and the precision slightly lower than the ones obtained with the linear model. On average both models provide results characterized by a 30% recall and an analogous precision. The inferred networks are generally very parsimonious (each variable has few parents) and statistical analysis showed that the goodness of fit is satisfactory. These preliminary results confirm the suitability of Gaussian networks for a first level, genome-wide analysis of high throughput dynamic data: the models here proposed can indeed infer a synthetic and quite accurate description of the system under study, useful to guide researchers to further, deeper studies.

Contact email: fulvia.ferrazzi@unipv.it