Alak Team

Minor effect of inhibitors on the time-course protein abundances

Seyed-Mohammad-Hadi Daneshmand[1], Ali Sharifi-Zarchi[2,3], OmidAskariSichani[1], Mahdi Jalili[1]

1. Computer Engineering Department, Sharif University of Technology, Tehran, Iran
2. Bioinformatics Department, Institute of Biophysics & Biochemistry, University of Tehran, Tehran, Iran
3. Department of Stem cells and Developmental Biology at the Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, ACECR, Tehran, Iran

Summary Sentence

In comparison with several Bayesian graphical models, our surprisingly simpler method that eliminates the inhibitor effects demonstrated significantly improved results.

Introduction

In silico reconstruction of the signalling networks from perturbation data is a critical and still challenging task for better understanding of the biological processes. Different methods have been used to approach this problem that include linear[1], Monte Carlo[2] and dynamic Bayesian networks[3]. Here we applied three distinct formulations of the dynamical Bayesian graphical model to reconstruct the time-course abundances of a particular set of proteins, using the HPN-DREAM dataset for training. While more complex models have more frequent parameters that cannot be estimated due to the limited size of the training dataset, simpler models exhibited significantly better performances. We also developed a simple *Inhibitor Independent* model that could overcome the other methods by ignoring small effect of the inhibitors.

Method

We used the whole discrete time series of the main dataset, with no particular preprocessing. Missing data were omitted, and the models were trained with the available data. For the replicates we kept only the average value. The protein-protein interaction and signaling networks of the selected proteins were extracted from two databases (string-db and Ingenuity IPA) that were only used for the 1st sub-challenge. In the 2nd sub-challenge, only the correct targets of the inhibitors were incorporated as an external data.

Let $\mathbb{S}$ be the set of all inhibitors. For learning and validating different algorithms, we created the following framework: The same process was executed for each inhibitor x separately. First, the time series data of three different cell lines (BT20, MCF7 and UACC812) treated with inhibitors of S- $\{x\}$ was used to train each model. Next using the data of the same cell lines treated with x was used to measure the accuracy of the results generated by each model, by Mean Squared Error (MSE) and also Pearson Correlation Coefficient (PCC).

First we developed our *in silico* network reconstruction method based on the dynamical Bayesian graphical models, since they are widely used in modelling causal interactions[4]. Consider $Z_t = (z_t^1, z_t^2, ..., z_t^n)$ to be the abundances of proteins at time t, while n is the number of proteins. The general equation of the dynamical Bayesian graphical models is[4]:

$$P(\mathrm{Z}_t \mid Z_{t-1}) = \prod_{i=1}^{n} P(z_t^i \mid Pa(z_t^i)), \quad (1)$$

Where $Pa(z_t^i)$, the parents of z_t^i, is the set of all proteins of the dynamical network that can affect on z_t^i between the time t - 1 to t. Distinct formulations of this equation result in different statistical models that are listed below.

Vector Auto-regressive. This model is based on the following formulation:

$$P(\mathrm{Z}_t \mid \mathrm{Z}_{t-1}) \sim N(\mathrm{W}\, Z_{t-1}, \Sigma), \quad (2)$$

where $N(.)$ denotes the normal distribution. In this model, the matrix $W_{n\times n}$ represents the causal interaction network. To eliminate the effects of one protein into the others (i.e. where the cell line is treated with that protein inhibitor), we set the corresponding row of W to zero.

Sparse Vector Auto-regressive. In the sparse approach, we assumed that the interaction matrix W is sparse (many elements are zero). This assumption leads to the simpler model: the absolute values of the elements in W are exponentially distributed with constant mean λ.

$$\begin{aligned} P(\mathrm{Z}_t \mid \mathrm{Z}_{t-1}; \mathrm{W}) &\sim N(\mathrm{W}\, Z_{t-1}, \Sigma) \\ \mid w_{ij} \mid &\sim \exp(\lambda). \end{aligned} \tag{3}$$

Kalman Filter. This is more complicated model but with less parameters to be estimated. A vector of hidden variables H_t of length m represents the dynamics of the time series at time t. To reduce the number of parameters, we assumed that the hidden variable are less frequent than the proteins ($m<n$). The following equations represent the parent-child relationship:

$$\begin{aligned} P(H_t \mid H_{t-1}) &\sim \mathrm{N}(\mathrm{A}\, H_{t-1}, \Sigma_1) \\ P(Z_t \mid H_t) &\sim \mathrm{N}(\mathrm{B}\, H_t, \Sigma_2). \end{aligned} \tag{4}$$

Causal interaction among the proteins can be obtained through the matrices $A_{m\times m}$ and $B_{\mathrm{n}\times\mathrm{m}}$. More precisely, the matrix $W_{n\times n} = BAB^{-1}$ denotes the causal relationships.

When applied to the real data, we found the simpler methods are exhibiting better results. Hence we developed a surprisingly simple method (*Inhibitor Independent*) that for each protein p, time t, cell line l, inhibitor i and stimuli s, predicts the protein abundance matrix as $A_{p,t,l,i,s} = \frac{1}{|S|}\sum_{j\in S} A_{p,t,l,j,s}$. The accuracies of different methods are provided below.

Method	Vector auto-regressive	Kalman Filter	Sparse Regression	Inhibitor Independent
MSE	29.40	28.90	5.07	1.47
PCC	0.53	0.41	0.82	0.90

Discussion

By ignoring presence of inhibitors, our Inhibitor Independent method could surprisingly exhibit the best results for the part A of 2nd sub-challenge among all the competitors, suggesting the minor effects of the inhibitors in the global network structure and time-course protein abundances. Our results also suggest the mean values of time series can be used as a suitable prior for other Bayesian methods.

References

1. Knapp, B. & Kaderali, L. Reconstruction of Cellular Signal Transduction Networks Using Perturbation Assays and Linear Programming. *PloS one***8**, e69220 (2013).
2. Bender, C. *et al.* Inferring signalling networks from longitudinal data using sampling based approaches in the R-package'ddepn'. *BMC bioinformatics***12**, 291 (2011).
3. Hill, S.M. *et al.* Bayesian inference of signaling network topology in a cancer cell line. *Bioinformatics***28**, 2804-2810 (2012).
4. Murphy, K.P. *Machine learning: a probabilistic perspective*. (The MIT Press, 2012).
5. Bishop, C.M. & Nasrabadi, N.M. *Pattern recognition and machine learning*, Vol. 1. (springer New York, 2006).

Authors Statement

S.D., O.A. and A.Sh. conceived the project. S.D. and O.A. developed the statistical methods and analysed the data. A.Sh. provided the bioinformatics insights into the problem. M.J. led the group.