

Validation of gene regulatory network models inferred from time-course gene expression data at arbitrary time intervals

Michiel J.L. de Hoon, Sascha Ott, Seiya Imoto, Satoru Miyano
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
{mdehoon,ott,imoto,miyano}@ims.u-tokyo.ac.jp

keywords: Gene regulatory networks, dynamic Bayesian networks, transcription factor prediction

Introduction

Dynamic Bayesian networks have been used to infer gene regulatory networks from time-course gene expression data [1, 2]. In a dynamic Bayesian network, the gene expression levels at each time point are described by the conditional probability given the gene expression levels \underline{x} at the previous time point. This can be written as the mapping $\underline{x} \rightarrow g^{\text{DBN}}(\underline{x}) + \text{error}$, requiring equal time intervals between gene expression measurements.

A generalization of the dynamic Bayesian network model that allows unequal time intervals can be derived using a differential equation model [3] by including noise explicitly in the model [4], leading to a system of stochastic differential equations. Here, the rate of change of the gene expression levels at time t is a function of the expression levels at that time point:

$$\frac{d}{dt}\underline{x}(t) = \underline{g}(\underline{x}(t)), \text{ approximated by the difference equation } \frac{\underline{x}_{i+1} - \underline{x}_i}{t_{i+1} - t_i} = \underline{g}(\underline{x}_i), \quad (1)$$

where the function \underline{g} effectively describes the gene regulatory network, and the time intervals $t_{i+1} - t_i$ are allowed to be unequal. To describe the presence of noise in the measured expression data, we add a noise term explicitly:

$$\frac{\underline{x}_{i+1} - \underline{x}_i}{t_{i+1} - t_i} = \underline{g}(\underline{x}_i) + \underline{\dot{\sigma}} \cdot \underline{\varepsilon}_i, \quad (2)$$

where $\underline{\varepsilon}$ a random process with unit variance, and $\underline{\dot{\sigma}}$ (with units of $[\text{hour}]^{-1}$) is a diagonal matrix whose entries on the diagonal represents the magnitude of the noise term for each gene separately. If all time intervals $t_{i+1} - t_i$ are equal, this reduces to a dynamic Bayesian network. In this paper, we approximate gene regulation by a linear model $\underline{g}(\underline{x}_i) = \underline{\underline{A}} \cdot \underline{x}_i$, where $\underline{\underline{A}}$ is a sparse $m \times m$ matrix.

Given a set of measured gene expression data, we can calculate the log-likelihood function for a specific network model, which describes how well the model fits the measured data. The maximum likelihood estimate of the gene regulatory network is the model that maximizes the log-likelihood function, which corresponds to a least-squares estimate of the network. As the number of possible network geometries is much larger than the amount of measured data, in general it is possible to find a network that fits the measured gene expression data perfectly. Such a network, however, most likely overfits the data, and may not be representative of reality. Instead, we can use a constrained maximum likelihood method, in which the maximum likelihood function is maximized under the condition that the number of parent genes of each gene in the network is fixed to some small number h , chosen ad hoc [1, 3]. Alternatively, the number of parents for each gene in the graph can be estimated using information criteria [2, 4, 5], such as Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) [6, 7]:

$$\begin{aligned} \text{AIC} &= -2 \cdot \{\text{log-likelihood}\} + 2 \cdot \{\text{number of estimated parameters in the model}\}; \\ \text{BIC} &= -2 \cdot \{\text{log-likelihood}\} + \ln(n) \cdot \{\text{number of estimated parameters in the model}\}; \end{aligned}$$

where the number of estimated parameters in the model depends on the number of parent genes in the network. These criteria provide a balance between the agreement of the data to the model and the number of parameters needed to reach that level of agreement. To find the optimal network geometry, we use an exhaustive search, which is feasible if the number of genes or the number of candidate parent genes are not exceedingly large.

Evaluation using synthetic data

In order to evaluate the predictive power of our network prediction method, we designed artificial networks and used them to generate simulated expression data. We then inferred our model from the simulated data and compared the predicted network to the true network to find the number of prediction errors. Whereas for the BIC the prediction accuracy improves as the number of measurements increases, the number of prediction errors remains large even for a large number of measurements if the AIC is used. In addition, we found that a single series of measurements of 20 time points or so is too short for an accurate prediction of gene regulation.

Evaluation using measured data

Previously inferred Bayesian network models of gene regulation of *Saccharomyces cerevisiae* [2, 5, 8] and *E. coli* [1] were assessed heuristically. Here, we predict the sigma (transcription) factor of known genes in *Bacillus subtilis* from measured gene expression data, and calculate the statistical significance of the number of correct predictions. Table 1 shows the time-course experiments used in these calculations, together with their measurement time points.

We consider the six sigma factors σ^D , σ^H , σ^F , σ^E , σ^X , σ^W , whose *cy3* and *cy5* expression levels were well above the noise level. We average the expression ratios of the genes in each operon in the regulons of these transcription factors, and use the constrained maximum likelihood method with $h = 1$ to fit our model to the gene expression data and predict the transcription factor of each operon. The significance level of our predictions is assessed by calculating the corresponding p -value, using a purely random prediction as the null hypothesis.

Table 2 shows the frequency that each sigma factor was estimated correctly, as well as the corresponding p -value. Whereas for most sigma factors the estimation procedure does not produce statistically significant results, for the σ^W and σ^X transcription factors we find a significance level of $p = 7.8 \times 10^{-12}$ and $p = 0.062$ respectively.

The difference between σ^D , σ^H , σ^F , σ^E and σ^W , σ^X may lie in the biological mechanism of gene regulation. For example, gene regulation may depend on signaling pathways in which transcription factors that are already present are activated, for example by protein phosphorylation events, without necessarily a change in the concentration of the mRNA molecules encoding those transcription factors. As current models of gene regulation focus on mRNA concentrations only, such effects would not be captured. To improve our understanding of gene regulation, the challenge will be to include such mechanisms in dynamical models in some suitable way.

Transcription factor prediction based on gene expression data provides an alternative in cases where sequence information does not yield a conclusive result. The genes *sigY* and *sigV* have candidate promoter sequences that

Experiment	Measurement time points in minutes
Cold shock	0, 5, 10, 30, 60, 120
Competence	0, 60, 120, 180, 240, 300, 360
Glucose, glutamine added during sporulation	0, 60, 120, 180, 240, 300
Increased aminoacid availability	0, 30, 60, 120, 210, 300, 420, 540
Phosphate, glucose starvation	0, 60, 120, 180, 240, 300, 360, 420
Phosphate limitation	0, 60, 120, 180, 240, 300
Salt stress	0, 5, 10, 30, 60
Sporulation	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 360, 390, 420, 450, 480, 510, 540

Table 1. The time points at which expression measurements were made for the eight time-course experiments of *Bacillus subtilis*. The gene expression levels were measured twice at each time point.

Sigma factor	Number of operons	Number of correct predictions	p -value
σ^D	16	1	0.95
σ^H	11	2	0.57
σ^F	4	1	0.52
σ^E	22	4	0.51
σ^X	6	3	0.062
σ^W	21	18	7.8×10^{-12}

Table 2. Results of the transcription factor prediction, together with the corresponding p -value.

Gene	Predicted transcription factors and corresponding time scale
<i>sigY</i>	σ^X (26 minutes); σ^D (-58 minutes); σ^Y (-61 minutes)
<i>sigV</i>	σ^V (-15 minutes); σ^W (28 minutes)

Table 3. The predicted transcription factors for the genes *sigY* and *sigV*, choosing from the six candidate sigma factors. Negative values indicate inhibitory effects.

match the consensus sequence CGTC, which may be recognized by σ^W and σ^X [9]. Table 3 shows the transcription factors predictions using gene expression data. Whereas *sigY* is predicted to be regulated by σ^X , for *sigV* the predicted sigma factor is σ^W . Both *sigY* and *sigV* are included in their own set of transcription factors. This may be a true autoregulatory effect, or an artifact due to mRNA degradation in the cell. Note that *sigY* and *sigV* are found to be regulated by their respective σ factor with very similar time scales (calculated as $1/\Lambda_{ij}$) of 26 and 28 minutes respectively. The transcription factor prediction of *sigY* and *sigV* demonstrates that a more accurate inference of gene regulatory networks can be achieved by combining gene expression data with sequence information, as suggested by Tamada [10].

References

- [1] I. M. Ong, J. D. Glasner, and D. Page. Modelling regulatory pathways in *E. coli* from time series expression profiles. In *Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB 2002)*, pages 241–248, 2002.
- [2] S. Kim, S. Imoto, and S. Miyano. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In *International Workshop on Computational Methods in Systems Biology (CMSB2003)*, Springer Verlag Lecture Notes in Computer Science, volume 2602, pages 104–113, 2003.
- [3] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 4, pages 29–40, 1999.
- [4] M. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 8, pages 17–28, 2003.
- [5] S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In *Proc. Pac. Symp. on Biocomputing*, volume 7, pages 175–186, 2002.
- [6] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Research Memorandum No. 46, Institute of Statistical Mathematics, Tokyo*. In Petrov, B.N., Csaki, F. (eds.), *2nd Int. Symp. on Inf. Theory, Akadémiai Kiadó, Budapest (1973)*, pages 267–281, 1971.
- [7] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [8] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.
- [9] A. L. Sonenshein, J. A. Hoch, and R. Losick. *Bacillus subtilis and its closest relatives: from genes to cells*. ASM Press, Washington, DC, 2001.
- [10] Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara, and S. Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. In *Proceedings of the Second European Conference on Computational Biology*, 2003.