

# A Genetic Algorithm for Inferring Time Delays in Gene Regulatory Networks

Fang-Xiang Wu<sup>1</sup>, Anthony J. Kusalik<sup>1,2</sup>, and Wen-Jun Zhang<sup>1,3</sup>,

Department of <sup>1</sup>Biomedical Engineering, <sup>2</sup>Computer Science, and <sup>3</sup>Mechanical Engineering,  
University of Saskatchewan, Saskatoon, SK, S7N 5A9, CANADA  
faw341@mail.usask.ca; zhanc@engr.usask.ca; kusalik@cs.usask.ca

## Abstract

Recently we proposed a state-space model with time delays for gene regulatory networks. Although the system can be uniquely determined under some assumptions, the solution space is still too large to use an exhaustive search method to find the optimal solution. This work employs Boolean variables to capture the existence of the discrete time delays of the regulatory relationships among the internal variables, and proposes a genetic algorithm (GA) to determine the optimal Boolean variables (the optimal solution) and to further infer gene regulatory networks with time delays. Computational experiments performed on a real gene expression dataset show that GA is effective at finding the optimal solution. Not only does the regulatory network with time delay obtained from the dataset possess the expected properties of a real one, but the approach also improves the prediction accuracy by 72%, compared to gene regulatory network without time delays.

## 1. Introduction

Analysis of gene expression data reveals a considerable number of time delayed interactions suggesting that time delay is ubiquitous in gene regulation. Although many models have been proposed for gene regulatory networks [1,2,3], most of them are underestimated from the current volume of gene expression datasets even if time delays are not accounted for. In other work, we propose a state-space model with time delays for gene regulatory networks [4]. The model views genes as observation variables which are regulated by a number of internal variables. The number internal variables and their expressions are estimated from observation data (i.e. gene expression data) using the Bayesian information criterion (BIC) [5] and probabilistic principle component analysis (PPAC) [6]. In order to uniquely determine all state transition matrices,  $p^2(\tau_{\max} + 1)$  equations are needed, where  $p$  is the number of internal variables and  $\tau_{\max}$  is the

maximum number of the discrete time delays accounted for. As there are  $mp$  expression values of internal variable, only  $mp$  equations are available. This implies that the system parameters can be estimated only if  $m > p(\tau_{\max} + 1)$ , where  $m$  is the number of time points in gene expression dataset. We considered this case with  $\tau_{\max} = 1$  [4]. In reality, for many gene expression data, the inequality  $m > p(\tau_{\max} + 1)$  does not come true even if  $\tau_{\max} = 1$ , and implying the system is underestimated. Although the system can be uniquely determined under the assumption that each regulatory interaction only has a single time delay, the solution space is still too large to search for the optimal one using an exhaustive search method.

The present work employs Boolean variables to capture the existence of the discrete time delays in the regulatory relationships among the internal variables, and proposes a genetic algorithm (GA) to determine the optimal Boolean variables (corresponding to the optimal solution) and to further infer gene regulatory networks with time delays. Computational experiments have been performed using a real gene expression dataset to investigate the performance of the presented methods.

## 2. Overview of Methods

The state-space model with time delays can be described mathematically by

$$\begin{cases} \mathbf{z}(t+1) = \sum_{\tau=0}^{\tau_{\max}} \mathbf{B}_{\tau} \circ \mathbf{A}_{\tau} \cdot \mathbf{z}(t-\tau) + \mathbf{n}_1(t) \\ \mathbf{x}(t) = \mathbf{C} \cdot \mathbf{z}(t) + \mathbf{n}_2(t) \end{cases} \quad (1)$$

where the symbol “ $\circ$ ” denotes the Hadamard (element-wise) multiplication of two matrices [7]. The matrices  $\mathbf{B}_{\tau} = [b_{ij\tau}]_{p \times p}$  ( $\tau = 0, \dots, \tau_{\max}$ ) are Boolean matrices, which capture time-delayed regulatory relationships,  $b_{ij\tau} = 1$  if internal variable  $j$  regulates internal variable  $i$  with time delay  $\tau$  and  $b_{ij\tau} = 0$  otherwise, and

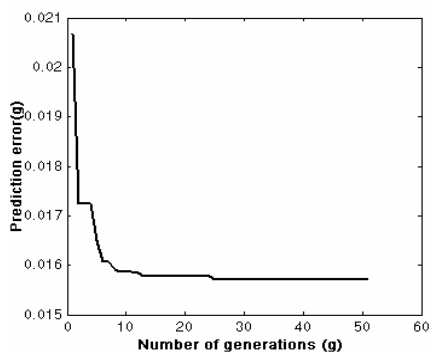
$$\sum_{\tau=0}^{\tau_{\max}} b_{ij\tau} = 1 \quad (i, j = 1, \dots, p). \quad (2)$$

The meanings of other symbols in (1) are the same as those in Eq. (1) of Ref [4].

The task of parameter identification in model (1) is the estimation of all elements in matrices  $\mathbf{A}_\tau = [a_{ij\tau}]_{p \times p}$  ( $\tau = 0, \dots, \tau_{\max}$ ) and  $\mathbf{C} = [c_{ik}]_{n \times p}$  such that both the system error and the observation error are minimized. The constructing of model (1) from microarray gene expression data may be divided into two phases. Phase one employs PCCA and BIC to estimate the number of internal variables and their expression from gene expression data, and establishes the observation equation (the lower one in (1)), subject to minimizing the observation error with BIC. Phase two employs GA to find optimal Boolean matrices  $\mathbf{B}_\tau$  and the multiple regression method to determine matrices  $\mathbf{A}_\tau$ , subject to minimizing the prediction error.

The solution space for  $\mathbf{B}_\tau$  ( $\tau = 0, \dots, \tau_{\max}$ ) consists of all Boolean matrix sets  $\{\mathbf{B}_0, \dots, \mathbf{B}_{\tau_{\max}}\}$  satisfying (2). The GA encodes an individual (a solution) by a binary string consisting of all elements of  $\{\mathbf{B}_0, \dots, \mathbf{B}_{\tau_{\max}}\}$  in a given order, and includes a selection operator, a crossover operator and a mutation operator. The GA considers the prediction error as the fitness value of an individual. The selection operator determines if an individual stays in the present population according to the normal distribution  $N(\mu, \sigma^2)$ , where  $\mu$  and  $\sigma^2$  are the mean and the variance, respectively, estimated from fitness values of all individuals in the previous population. The crossover operator adopts single-point crossover methods for simplicity. The mutation operator mutates each position on a string which is randomly selected with a mutation probability  $P_m$ .

### 3. Experiment Study and Results



**Figure 1.** Plot of prediction error with respect to the number of generations

To investigate the proposed method, it is applied to the gene expression dataset (BAC) from Laub at al. [8],

which consists of the expression profiles of 1590 genes over 11 equally-spaced time points with no missing data. The dataset is available at <http://caulobacter.stanford.edu/CellCycle>. The expression profile of each gene is normalized to have length 1 and then expression values on each microarray to have a mean of 0 and a standard deviate of 1.

Using PCCA and BIC, the number of internal variables (five) and their expression are estimated from the dataset. In this experiment, we consider  $\tau_{\max} = 1$  for the sake of simplicity. In this case, it is sufficient to code only Boolean matrix  $\mathbf{B}_0$  by a binary string. Let the size of the population  $N = 30$  and the mutation probability  $P_m = 0.02$ . The GA converges in 25 generations according to Figure 1.

The results show that the GA is effective at finding the optimal solution. Furthermore, compared to gene regulatory networks without time delays, gene regulatory network with time delays may improve the prediction accuracy by 72%. According to the optimal  $\mathbf{B}_\tau$  ( $\tau = 0, 1$ ), there are 14 (of 25) regulatory relationships with time delays. The 10 eigenvalues of the system are  $\{0.7061 \pm 0.7673i, -0.5284 \pm 0.5904i, -0.2564 \pm 0.1324i, 0.2862 \pm 0.0514i, 0.8332 \pm 0.2225i\}$ , all of which except for the first pair are inside the unit circle. The modulus of the first pair of eigenvalues is 1.0427 which is very close to the unit circle. From difference equation theory, the regulatory system obtained from the dataset not only is almost stable, but also is periodic at the stable states. These are the expected properties of a real gene regulatory network.

### Reference

- [1] Liang, S., *et al.* "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures" *PSB* **3**: 18-29, (1998).
- [2] Chen, T., He, H. L., and Church, G. M. "Modeling Gene Expression with Differential Equations" *PSB* **4**: 29-40, (1999).
- [3] D'haeseleer, P., *et al.* "Linear Modeling of mRNA Expression Levels During CNS Development and Injury" *PSB* **4**: 41-52, (1999).
- [4] Wu, F.X., Zhang, W.J., and Kusalik, A.J. "State-Space model with time delays for gene regulatory networks" in this volume.
- [5] Schwarz, G. "Estimating the dimension of a model" *Annals of Statistics* **6**: 461-464, (1978).
- [6] Tipping, M. E. and Bishop C. M. "Probabilistic principal component analysis" *Journal of the Royal Statistical Society, Series B* **61**: 611-622, (1999).
- [7] Schott J. R. "Matrix Analysis for Statistics" New York: John Wiley & Sons, Inc., (1997).
- [8] Laub, M. T. "Global analysis of the genetic network controlling a bacterial cell cycle" *Science* **290**: 2144-2148, (2000).