Reconstructing Gene Networks from Microarray Time-Series Data via Granger Causality*

Qiang Luo, Xu Liu, and Dongyun Yi

Department of Mathematics and Systems Science
National University of Defense Technology
Changsha, Hunan 410073, China
dongyun.yi@gmail.com

Abstract. Reconstructing gene network structure from Microarray time-series data is a basic problem in Systems Biology. In gene regulation networks, the time delays and the combination effects which are not considered by most existent models are key factors to understand the genetic regulatory networks. To address these problems, this paper proposed a fast algorithm to learn initial network structures for gene networks from time-series data by employing the Granger causality model to analyze the time delays and the combination effects for gene regulation. The simulation results on a synthetic network and the ethylene pathway in *Arabidopsis* show that the proposed algorithm is a promise tool for learning network structures from time-series data.

Keywords: partial Granger causality, gene regulatory networks, time series data, projection pursuit.

1 Introduction

Gene networks controlling how genes are up and down regulated in response to signals play a key role in the life phenomena [1], such as development, metabolizability, adaptability, immunity, etc. Earlier, the genetic networks are investigated by constructing the mathematical model of few genes, and the characteristics of the biology networks are analyzed through simulation [2]. These approaches work well in the small scale networks. Nowadays, the invention and application of high-throughput technologies make it possible to study the genes in the genome scale enabling the quantitative understanding of large gene networks [3]. For analysis in the genome scale, it is more suitable to reconstruct the network models of the genetic regulatory networks from data [4]. Recently, as more and more biology databases available, reverse engineering cellular networks has become a hot issue in biology, computer science, as well as mathematics, and the results of these researches are fruitful [5, 6, 7]. Methods for gene network reconstruction have

* This work is partially supported by National Basic Research Program of China (No. 2005CB321800) , and the Graduate Innovation Foundation of National University of Defense Technology (No. B060203).

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 196–209, 2009.

been proposed on the basis of statistical analysis such as Boolean models [8], differential equation models [9], Beyesian networks [10], and so on.

In recent years, learning the structures of Bayesian networks from massive data to reconstruct the gene networks from Microarray data attracted many scholars' attention [11, 12, 13, 14, 15]. Actually, the data used by Bayesian networks is produced by the perturbation experiments which usually knock out a gene and study the downstream effects, i.e., these data reflect the stationary status of the gene networks response to a stimulus. However, the use of perturbation experiments is limited due to technical and biological reasons [16]. Time-series expression data which imply a number of regulatory interactions have been widely used to study biological systems in many different species [17]. To model the time-series Microarray data, the dynamic Bayesian network (DBN) which has been shown to be appropriate for representing complex stochastic non-linear relationships among multiple random variables has been employed [18].

The initial structure of gene networks plays a key role in the present structure learning algorithms [19,20,21,22] for DBN, since the learning algorithms usually start from prior network structures which are constructed by expert according to the background knowledge, and perform heuristic searches in the space of directed acyclic graphs to improve the network structure in the light of the information contained in data. Unfortunately, the complexity of the prior network structure avoids structure missing, but exponentially increases the computation complexity of the learning algorithm. Considering the combination effects in the regulation, one has to do the permutation test $m2^{m-1}$ times for the gene network with m genes. Besides, the published learning algorithms of DBN are almost based on the first order Markov assumption of the variables, but many researches on the practical data sets challenge this assumption [23, 24].

To address these problems, we proposed a learning algorithm for DBN from Microarray time-series data without the first order Markov assumption. This algorithm consists of three steps: first, the Granger causality model are employed for pairs of genes with their time sequence expression data to build an initial network; second, the false regulations in the initial network are deleted by partial-Granger causality; third, if necessary, it depends on the practical data, we could apply the partial-Granger causality to further filter out the false regulation between genes by computing the combination effects of the conditional candidate genes. The application results of the proposed algorithm on both the simulation data and the practical data show that it is a promising method of dynamical network structure learning.

2 Method

Consider a gene network with n genes, denoted by $\mathbf{G} = (G_1, G_2, \cdots, G_n)$ and the expression of the genes are jointly stationary. The expression time series with length T of the genes in this network are available, $g_i(t) \in R^+$ $(i = 1, 2, \cdots, n, t = 1, 2, \cdots, T)$ is the stochastic time process for each gene. The main aim of this paper is to reconstruct the structures of the gene networks from the Microarray

time-series data. Since the most concern of this study is the relationship between changes of genes, we can assume that $EG_i(t) = 0$ for all i.

2.1 Granger Causality

Assume that each process $G_i(t)$ $(i = 1, 2, \cdots, n)$ admits an autoregressive representation

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{(i)} G_i(t-p) + \varepsilon_i(t),\ \mathrm{var}(\varepsilon_{\mathrm{i}}(\mathrm{t})) = \sigma_{\varepsilon_{\mathrm{i}}}^2. \tag{1}$$

Jointly, they are represented as

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{(i|j)} G_i(t-p) + \sum_{q=1}^{\infty} b_q^{(i|j)} G_j(t-q) + \varepsilon_{i|j}(t),\ \mathrm{var}(\varepsilon_{\mathrm{i|j}}(\mathrm{t})) = \sigma_{\varepsilon_{\mathrm{i|j}}}^2. \tag{2}$$

The intensity of causal influence of G_j on G_i can be measured by

$$F_{G_j \to G_i} = \frac{\sigma_{\varepsilon_i}^2}{\sigma_{\varepsilon_{i|j}}^2} - 1. \tag{3}$$

If G_i and G_j are independent, then b_q are uniformly zero and $\sigma_{\varepsilon_i} = \sigma_{\varepsilon_{i|j}}$; otherwise, the expression of G_j will be helpful for the prediction of G_i, i.e., $\sigma_{\varepsilon_i} > \sigma_{\varepsilon_{i|j}}$. Hence, it is clear that $F_{G_j \to G_i} = 0$ when there is no causal influence from G_j to G_i, and the greater the value of $F_{G_j \to G_i}$ is the more strong the causal influence will be. With the expression data, if the orders in this model (2) are determined by some criterion (e.g., Akaike Information Criterion, AIC) as P_i and Q_j for G_i and G_j, respectively, the variances can be estimated as follows:

$$\hat{\sigma}_{\varepsilon_i} = \frac{1}{T - 2P_i} \sum_{i=P_i+1}^{T} \hat{\varepsilon}_i^2, \tag{4}$$

$$\hat{\sigma}_{\varepsilon_{i|j}} = \frac{1}{T - 2P_i - 2Q_j} \sum_{i=P_i+Q_j+1}^{T} \hat{\varepsilon}_{i|j}^2, \tag{5}$$

where $\hat{\varepsilon}_i$ and $\hat{\varepsilon}_{i|j}$ are the residuals of models (1) and (2), respectively. Then, we have

$$\hat{F}_{G_j \to G_i} = \frac{\hat{\sigma}_{\varepsilon_i}}{\hat{\sigma}_{\varepsilon_{i|j}}} - 1 \sim F(2Q_j, T - 2P_i - 2Q_j). \tag{6}$$

Given a significance level F_1, an F test of the null hypothesis that G_j does not have causality influence on G_i. Now, the dynamic network structure can be given below:

$$M^{(1)} = \left((m_{ij}^{(1)}) \right), \tag{7}$$

where

$$m_{ij}^{(1)} = \begin{cases} 1, \hat{F}_{G_j \to G_i} > F_1; \\ 0, \text{otherwise.} \end{cases} \tag{8}$$

2.2 Partial-Granger Causality

The conditional independence confuses pairwise algorithms, for instance, gene Y is correlated with gene Z when $X \to Y$ and $X \to Z$, i.e., given X, Y and Z are conditional independence. To filter out the fake edges from the learning structure, the partial-Granger causality (P-G-C) is induced.

When $m_{ij}^{(1)} = 1$ $(i \neq j, i, j = 1, 2, \cdots, n)$, the conditional independence between G_i and G_j given G_k will be examined for each $k = 1, 2, \cdots, n$ and $k \neq i, k \neq j$. Consider the following models

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{(i|k)} G_i(t-p) + \sum_{q=1}^{\infty} c_q^{(i|k)} G_k(t-q) + \varepsilon_{i|k}(t),\ \mathrm{var}(\varepsilon_{i|k}(t)) = \sigma^2_{\varepsilon_{i|k}}, \quad (9)$$

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{(i|j,k)} G_i(t-p) + \sum_{r=1}^{\infty} b_r^{(i|j,k)} G_j(t-r) + \sum_{q=1}^{\infty} c_q^{(i|j,k)} G_k(t-q) + \varepsilon_{i|j,k}(t), \quad (10)$$

where

$$\mathrm{var}(\varepsilon_{i|j,k}(t)) = \sigma^2_{\varepsilon_{i|j,k}}.$$

Similarly, we assume that the noise terms in these models are white noise, and the null hypothesis that G_j is conditional independence with G_i given G_k, i.e., $b_r^{(i|j,k)}$ in (10) are uniformly zero, could be tested with the following statistic:

$$F_{G_j \to G_i|G_k} = \frac{\sigma^2_{\varepsilon_{i|k}}}{\sigma^2_{\varepsilon_{i|k,j}}} - 1, \quad (11)$$

and its estimation with the observation data is similar to (6):

$$\hat{F}_{G_j \to G_i|G_k} \sim F(2R_{j|k}, T - 2P_{i|k} - 2Q_k - 2R_{j|k}), \quad (12)$$

where $P_{i|k}$, $R_{j|k}$ and Q_k are orders given by AIC for G_i, G_j and G_k, respectively. Given the significant level F_2, the F test can be performed to filter out the fake correlations in $M^{(1)}$ to get the initial dynamical network structure $M^{(2)}$ as follows:

$$m_{ij}^{(2)} = \begin{cases} 0, & \text{for } m_{ij}^{(1)} = 1 \text{ and } \hat{F}_{G_j \to G_i|G_k} \leq F_2; \\ m_{ij}^{(1)}, & \text{otherwise.} \end{cases} \quad (13)$$

In gene networks, the number of conditional genes is $n-2$ for each regulation pair indicated by $M^{(1)}$, i.e., we need to compute the P-G-C $n^2(n-2)$ times in the worst case. Besides, the combination effect of the conditional genes can not be exclude in this one by one version of P-G-C. Let the set of the conditional genes denoted by $\mathbf{G}(i,j) = \mathbf{G} \backslash \{G_i, G_j\}$, and the causality model can be given below:

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{i|\mathbf{G}(i,j)} G_i(t-p) + \sum_{k \neq i,j} \sum_{q=1}^{\infty} c_q^{(k)} G_k(t-q) + \varepsilon_{i|\mathbf{G}(i,j)}(t), \quad (14)$$

$$G_i(t) = \sum_{p=1}^{\infty} a_p^{(i|\mathbf{X}(i,j),j)} G_i(t-p) + \sum_{r=1}^{\infty} b_r G_j(t-r) + \sum_{k\neq i,j} \sum_{q=1}^{\infty} c_q^{(k,j)} G_k(t-q) + \varepsilon_{i|\mathbf{G}(i,j),j}(t). \tag{15}$$

Generally, it is not practical to perform F test for this model, since the number of parameters in this model is most likely to be much bigger than the sample size of the data, i.e., the parameters can not be well estimated with limited sample size. Next, let's give a fast algorithm to compute the multivariate partial Granger causality in a projection pursuit manner.

For $G_j \to G_i$, let $G_{k_s} \in \mathbf{G}(i,j)(s = 1, \cdots, n-2)$, then define

$$H(t) = I - B^T(t)(B^T(t)B(t))^{(-1)}B(t), \tag{16}$$

where

$$B(t) = (G_{k_1}(t), G_{k_2}(t), \cdots, G_{k_{n-2}}(t)), t = 1, 2, \cdots, T.$$

Then, the effects of the conditional variables could be excluded from variable pairs of (G_i, G_j) by

$$G_i'(t) = H(t)G_i(t), \tag{17}$$

$$G_j'(t) = H(t)G_j(t), \tag{18}$$

and the partial-Granger causality model can be defined as follows:

$$G_i'(t) = \sum_{p=1}^{\infty} a_p^{(i)} G_i'(t-p) + \varepsilon_i(t), \tag{19}$$

$$G_i'(t) = \sum_{p=1}^{\infty} a_p^{(i|j)} G_i'(t-p) + \sum_{q=1}^{\infty} b_q G_j'(t-q) + \varepsilon_{i|j}(t). \tag{20}$$

Performing the F test with the statistic

$$\hat{F}_{G_j \to G_i|\mathbf{G}(i,j)} \sim F(2Q_j, T - 2P_i - 2Q_j), \tag{21}$$

the network structure $M^{(2')} = (m_{ij}^{2'})$ is established by

$$m_{ij}^{(2')} = \begin{cases} 0, & \text{for } m_{ij}^{(1)} = 1 \text{ and } \hat{F}_{G_j \to G_i|\mathbf{G}(i,j)} < F_{2'}; \\ m_{ij}^{(1)}, & \text{otherwise.} \end{cases} \tag{22}$$

Now the multivariate Granger causality needs to be computed for a pair of regulation genes only once. The steps of the main algorithm of this paper are described as follows:

MAIN ALGORITHM

Step 1. Data Centralizing: G_i is replaced by $G_i - \frac{1}{T}\sum_{t=1}^{T} G_i(t)$ for $i = 1, 2, \cdots, n$;

Step 2. G-C analysis: The network $M^{(1)}$ is given by calculating the statistics $\hat{F}_{G_j \to G_i}$ as (6);

Step 3. P-G-C analysis: Univariate P-G-C analysis is performed to build network structure $M^{(2)}$ according to (13);

Step 3′. P-G-C analysis: Multivariate P-G-C analysis is carried out to exclude false relationships in the network structure (22).

In this algorithm, you can choose to use Step 3 or Step 3′ in your application: when the prior knowledge about the conditional gene of a pair of regulation genes is available, Step 3 needs to be run only for the candidate conditional gene; If no prior information is available or the combination effects is noticeable, Step 3′ is preferred.

3 Experimental Results

3.1 Synthetic Dynamical Networks

We first test our algorithm of structure learning for the synthetic network of 5 genes consisting 5 regulations [25]. As presented on Fig. 1(a), the directed edges represent the regulations between genes which can also be formulated by a dynamical system:

$$
\begin{aligned}
x_1(t) &= 0.95\sqrt{2}x_1(t-1) - 0.9025x_1(t-2) + \varepsilon_1(t),\\
x_2(t) &= 0.5x_1(t-2) + \varepsilon_2(t),\\
x_3(t) &= -0.4x_1(t-3) + \varepsilon_3(t),\\
x_4(t) &= -0.5x_1(t-1) + 0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \varepsilon_4(t),\\
x_5(t) &= -0.25\sqrt{2}x_4(t-1) + 0.25\sqrt{2}x_5(t-1) + \varepsilon_5(t).
\end{aligned}
$$

Without loss of generality, we may set $\varepsilon_1 \sim N(0, 0.6), \varepsilon_2 \sim N(0, 0.5), \varepsilon_3 \sim N(0, 0.3), \varepsilon_4 \sim N(0, 0.3), \varepsilon_5 \sim N(0, 0.6)$ and, for simplicity, we assume that $\forall i \neq j, Cov(\varepsilon_i, \varepsilon_j) = 0$.

As described in last section, our algorithm learns the network structure from data by many steps, and thereby the learning results given by Step 2 and Step 3′ are portrayed in Fig. 1(b) and Fig. 1(c), respectively. Fig. 1(b) shows many fake relationships, such as $gene1 \to gene5$, an indirect casual interaction generated by $gene1 \to gene4$ together with $gene4 \to gene5$. From Fig. 1(c), we can see that except for the edge $gene4 \to gene5$ the edges of the original synthetic network have all be reconstructed from the simulation data, meanwhile no false edge have been learned from the data. The result presented in Fig. 1 is one of the results given by the proposed algorithm with threshold values 0.9 and 0.9 with no repeat, and the length of the time series, i.e., the sample size, generated by the synthetic network is 500. Clearly, the sample size and the threshold values

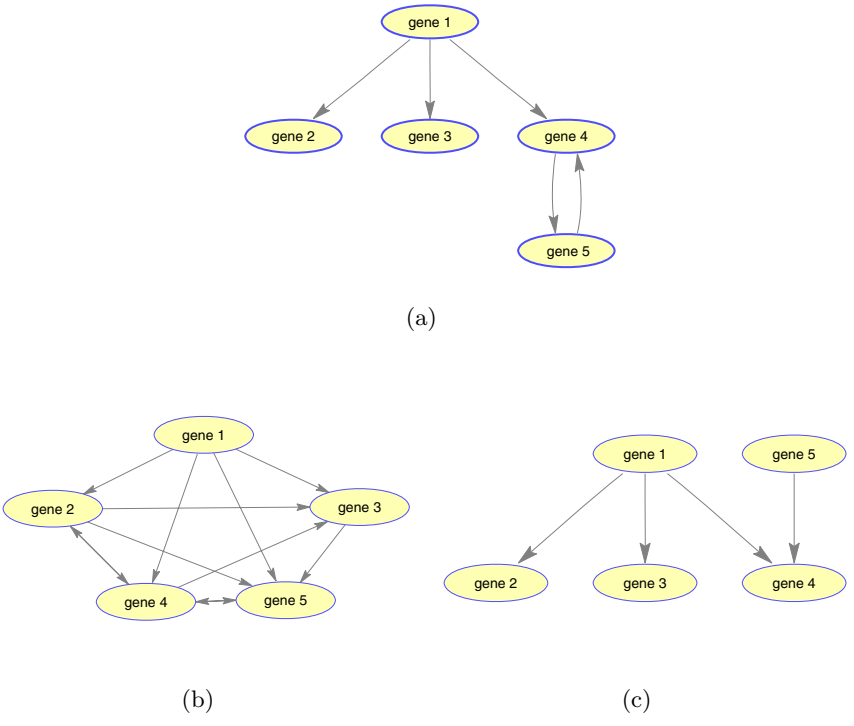


Fig. 1. The numerical experimental results of our algorithm on the synthetic network. (a) the original synthetic network; (b) the learned structure by G-C analysis; (c) the learned structure by Step 3′

of the F-tests, which are key factors for the performance of our algorithm, need further investigation.

Practically, the sample size are often limited, so the performance of our algorithm, as a function of sample size is also instructive. The sensitivity (SE) and the specificity (SP) have been employed to quantitatively compare the performances of our algorithm over different sample sizes. The SE of an approach to learn the gene regulations from expression data is a measure of the probability to detect the regulation by this approach but does not say whether any candidate regulation is truly a transcriptional regulation, and it can be computed by the number of true positives (TP) and the number of true regulations in gene network (NG) as follows

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{NG}} \times 100\%. \tag{23}$$

The SP measures the accuracy of a given approach, and can be estimated from the percentage of the predicted regulation of this approach that are present in the reference gene network by the following equation:

$$\mathrm{SP} = \frac{\mathrm{TP}}{\mathrm{NL}} \times 100\%, \tag{24}$$

where NL stands for the number of edges in the learned network structure. For a given sample size, the data are iteratively generated from the synthetic network, and the network structure is learned from these data by our algorithm with some threshold values. A total of 50 repeated for each experimental setting are conducted and the averages of the sensitivity and the specificity are computed. The numerical experimental results are shown in Fig. 2 by our algorithm with different threshold values vary from 70 to 700. Clearly, no matter which threshold values are chosen, the SE and the SP totally increase with the sample size.

Note that the results in Fig. 2 also show that the proposed algorithms with different threshold values exhibit different performances. Therefore, the two threshold values, in which one for the F-test in G-C analysis and the other for the F-test in P-G-C analysis, deserve to be discussed in details. In this paper, we set the two threshold to be the same value. As the threshold values specify the confidence level of the F-test in the algorithm, the higher value of the threshold will get higher specificity and lower sensitivity. For a given sample size 90, Fig. 3(a) shows the box plots for the performance comparison between the proposed methods with different threshold values ranging from 0.5 to 0.98. The p-values of the one-way analysis of variance are both 0 for the recall percentage and the precision percentage given by the proposed algorithms with different threshold values, which means the effect of threshold values on the performance of our algorithm is statistically significant. Fig. 3(b) illustrates the average performance level variation against the threshold value and the threshold value somewhere between 0.7 and 0.8 reaches the balance between sensitivity and specificity.

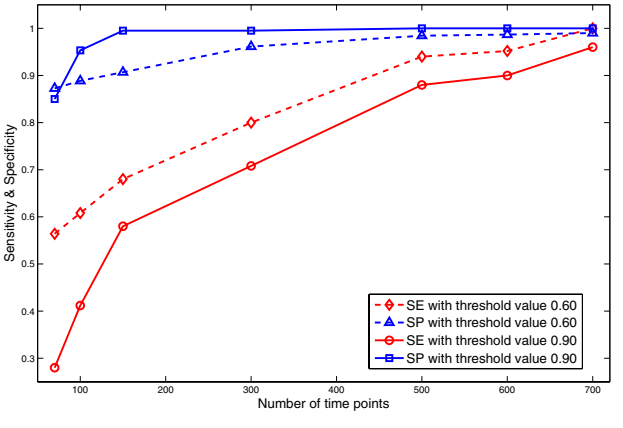
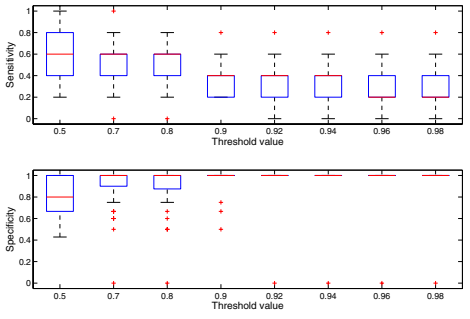
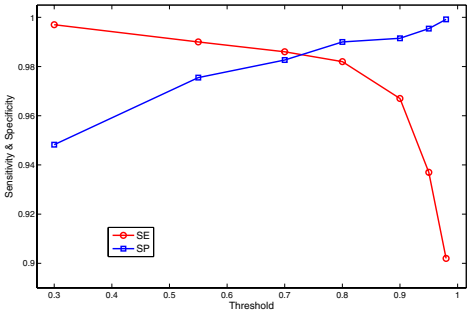


Fig. 2. Performance comparison of our methods on the data sets with different sample sizes



(a) Box plots for SE and SP



(b) Average performances against threshold values

Fig. 3. Performance comparison of our methods with different threshold values

3.2 Genetic Regulatory Networks

In practice, the gene regulatory networks or the signaling pathways are much more complex than the synthetic one. For example, the actually interactions between genes are expected to be nonlinear and noisy instead of the linear interactions in the synthetic network. Therefore, the results given by the proposed algorithm with gene expression data only provide initial dynamical network structure of the genetic regulatory networks for further analysis, such as Bayesian networks. Here, in this paper, we applied our algorithm on the Microarray time-series data of 7 genes (ETR1, ETR2, ERS1, ERS2, EIN4, CTR1 and MPK6) related to the detection of ethylene stimulus provided by the functional analysis of regulatory genes involved in *Arabidopsis* leaf senescence at Warwick HRI, which have 22 time points in each time series and 16 replicates. Table. 1 lists the gene catma_id [1]. Fig. 4 illustrates the results obtained by the proposed algorithm. Since the space is limited, the description details of these genes can be

[1] The aim of the Complete *Arabidopsis* Transcriptome MicroArray (CATMA) project (http://www.catma.org/) was the design and production of high quality Gene-specific Sequence Tags (GSTs) covering most *Arabidopsis* genes.

Table 1. Gene name and catma_id discussed in this paper

Gene	Name	Catma_id
ETR1	At1g66340	CATMA1a55610
ETR2	At3g23150	CATMA3a23140
ERS1	At2g40940	CATMA2a39280
ERS2	At1g04310	CATMA1a03150
EIN4	At3g04580	CATMA3a03560
CTR1	At5g03730	CATMA5a02913
MPK6	At2g43790	CATMA2a42185

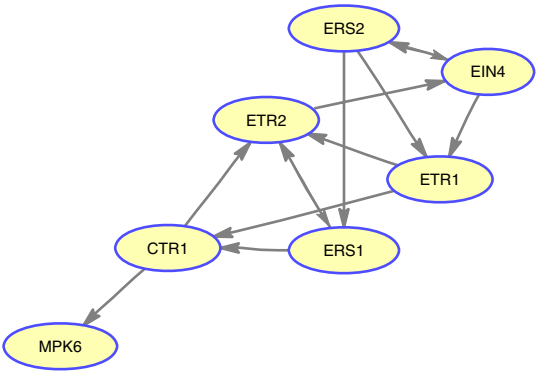


Fig. 4. Learned network structure by the proposed algorithm

found on the web site of the Complete *Arabidopsis* Transcriptome MicroArray (CATMA) project (http://www.catma.org/).

During the past decade, the reference plant *Arabidopsis* have been well studied and the ethylene is a gaseous plant hormone involved in many life process [26]. We've already known that the ethylene is perceived by a family of five membrane-associated receptor (ETR1, ETR2, ERS1, ERS2 and EIN4) in *Arabidopsis*, and the ethylene binding leads to functional inactivation of the receptors which are negative regulators of ethylene responses. In the presence of ethylene, CTR1 which is another negative regulator of the pathway loses its ability to repress the downstream genes. That's the early events of ethylene perception and signaling. However, the detailed network structure of these genes needs to be further investigated, and a MAPK pathway involving MPK6 in *Arabidopsis* has recently been proposed in operating downstream of CTR1 as a positive regulator in the pathway. From the result given by the proposed algorithm, we can see that except for EIN4 and ERS2 the other ethylene receptors are all have directed relationship with CTR1. The existence of the directed edge from CTR1 to MPK6 supports that the MAPK cascade is involved in this pathway.

The coefficients in the partial Granger casuality model have been listed in Table 2. We set the maximum order for delay to be 5 and the AIC is used to select the optimal order. On the diagonal, the coefficients are for the autoregulatory

Table 2. Coefficients given by partial Granger Casuality model

	ETR1	ETR2	ERS1	ERS2	EIN4	CTR1	MPK6
ETR1	0.2479	-0.0638				0.4304	
		-1.0446				-1.1091	
ETR2		0.15	0.7018		-0.0798		
			-0.4192		-0.1611		
ERS1		-0.7355	0.1045			-0.0224	
		1.1327				0.6591	
ERS2	-0.1398		0.1535	0.0239	0.1793		
	0.2747		0.2808				
EIN4	0.1822			-0.087	-0.0056		
	-0.2256			-0.4767			
CTR1		-0.4751				0.1063	-0.251
		0.4915					0.2538
MPK6							0.243
							-0.477

equation (19); and the left coefficients are for the partial Granger casuality equation (20). For example, the first entry in this table means the autoregulatory equation for ETR1 is

$$G'_{ETR1}(t) = 0.2479G'_{ETR1}(t-1) + \varepsilon. \tag{25}$$

We may say it is the positive autoregulation that works behind the expression behavior of ETR1. However, for MPK6, there is a second order delay, which makes some confusion:

$$G'_{MPK6}(t) = 0.243G'_{MPK6}(t-1) - 0.477G'_{MPK6}(t-2) + \varepsilon. \tag{26}$$

This can be explained biologically that MPK6 not only has a directly positive autoregulation but also has some negative regulation which might be works in some negative loop instead of the directly regulation. Therefore, there will be a delay in this negative effect, which means that some genes other than the 7 genes we studied here also have great effects on the regulation of MPK6, i.e., more genes need to be included in this network for further understanding.

In Table 2, the second entry is the coefficients of the partial Granger model which describes how ETR1 regulates ETR2 excluding the influence of the other 5 genes. And it can be written as follows:

$$G'_{ETR2}(t) = -0.0638G'_{ETR2}(t-1) - 1.0446G'_{ETR1}(t-1) + \varepsilon. \tag{27}$$

Similarly, the above equation may help us to understand the negative regulation from ETR1 to ETR2, which indicates that the receptors are not function independently. As listed in Table 2, the autoregulation for ETR2 is a positive one, but in the partial Granger model the first coefficient in (27) is -0.0638, which represents a negative autoregulation. It can be explained as follows: the autoregulation coefficients on the diagonal of the result table describe regulations that

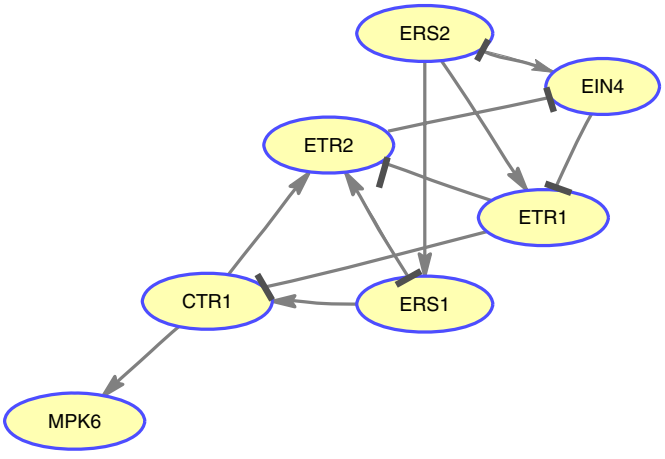


Fig. 5. Reconstructed gene network for the ethylene response pathway in *Arabidopsis*

include all genes influences. Instead of the effects contributed to by the whole gene set, the partial Granger causality model excludes the regulation effect on ETR2 given by the genes other than ETR1. That's to say, the autoregulation can not be determined at the current stage. Therefore, the combination effects must be considered during the reconstruction of genetic regulatory networks, and partial Granger causality model provides us a promise tool to address this problem. At last, the result shown on Fig. 4 can be redrawn with more details about the negative or positive regulations between the genes, which is shown in Fig. 5.

4 Conclusion

Reconstructing gene regulatory networks from Microarray time-series data has been a big challenge in systems biology due to the complex regulatory relationships among genes and the limit of available data. Many algorithms have been developed to deal with this reverse engineering problem, and most of them need to start from the initial network structures which affect the sensitivity and specificity of these algorithms. In this paper, an algorithm for learning structure from expression time-series data has been proposed by applying the partial Granger causality model. Instead of the first order Markov assumption, the proposed algorithm considers the multiple delays of the regulation effects by optimally determining the orders in the model with AIC. The present learning algorithms for Bayesian networks are embarrassed by the exponential computation complexity. However, in our algorithm, the combination effects of the candidate conditional genes need to be computed only once for each pair of regulation genes in the projection pursuit manner. Despite the linear assumptions inherent in the Granger causality models, the results reported above indicate that it is a promise tool for reconstructing gene networks. Therefore, one of the further investigation direction of this algorithm is its nonlinear extension.

References

1. Boone, C., Bussey, H., Andrews, B.J.: Exploring Genetic Interactions and Networks with Yeast. Nature Review Genetics 8, 437–449 (2007)
2. Huang, S.: Gene Expression Profiling, Genetic Networks, and Cellular States: An Integrating Concept for Tumori-genesis and Drug discovery. Journal of Molecular Medicine 77, 469–480 (1999)
3. Kitano, H.: Systems Biology: a brief overview. Science 295, 1662–1664 (2002)
4. Smolen, P., Baxter, D.A., Byrne, J.H.: Modeling Transcriptional Control in Gene Networks — Methods Recent Results, and Future Directions. Bulletin of Mathematical Biology 62, 247–292 (2000)
5. Jong, H.D.: Modeling and Simulation of Genetic Regulatory System: A Literature Review. Journal of Computational Biology 9(1), 67–163 (2002)
6. Werhli, A.V., Grzegorczyk, M., Husmeier, D.: Comparative Evaluation of Reverse Engineering Gene Regulatory Networks with Relevance Networks, Graphical Gaussian Models and Bayesian Networks. Bioinformatics 22, 2523–2531 (2006)
7. Schlitt, T., Brazma, A.: Current Approaches to Gene Regulatory Network Modelling. BMC Bioinformatics, 8(suppl. 6), S9 (2007)
8. Kauffman, S., Peterson, C., Samuelsson, B., Troein, C.: Random Boolean Network Models and the Yeast Transcriptional Network. Proc. Natl. Acad. Sci. USA 100, 14796–14799 (2003)
9. Chen, K.C., Wang, T.Y., Tseng, H.H., Huang, C.Y.F., Kao, C.Y.: A Stochastic Differential Equation Model for Quantifying Transcriptional Regulatory Network In Saccharomyces Cerevisiae. Bioinformatics 21(12), 2883–2890 (2005)
10. Friedman, N., Nachman, I., Pe'er, D.: Learning Bayesian Network Structure from Massive Datasets: the Sparse Candidate Algorithm. In: Laskey, K.B., Prade, H. (eds.) UAI 1999. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 206–215. Morgan Kaufmann, Stockholm (1999)
11. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian Networks to Analyze Expression Data. In: Proceeding of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB), Tokyo, Japan, pp. 127–135 (2000)
12. Pe'er, D., Regev, A., Elidan, G., Friedman, N.: Inferring Subnetworks from Perturbed Expression Profiles. Bioinformatics 17 (suppl. 1), S215–S224 (2001)
13. Friedman, N.: Inferring Cellular Networks Using Probabilistic Graphical Models. Science 303(5659), 799–807 (2004)
14. Sachs, K., Perez, O., Pe'er, D.: Casual Protein-Signaling Networks Derived from Multi-Parameter Single-Cell Data. Science 308(5721), 523–529 (2005)
15. Gevaert, O., Smet, F.D., Timmerman, D., Moreau, Y., Moor, B.D.: Predicting the Prognosis of Breast Cancer by Integrating Clinical and Microarray Data with Bayesian Networks. Bioinformatics 22, e184–e190 (2006)
16. Tong, A., Evangelista, M., Parsons, A.B., et al.: Systematic Genetic Analysis with Ordered Arrays of Yeast Deletion Mutants. Science 294, 2364–2368 (2001)
17. Ernst, J., Nau, G.J., Bar-Joseph, Z.: Clustering Short Time Series Gene Expression Data. Bioinformatics 21(suppl. 1), I159–I168 (2005)
18. Murphy, K., Mian, S.: Modelling Gene Expression Data Using Dynamic Bayesian Networks. Technical report, Computer Science Division, University of California, Berkeley, CA (1999)
19. Perrin, B.E., Ralaivola, L., Mazurie, A., et al.: Gene Networks Inference Using Dynamic Baysian Networks. Bioinformatics 19(suppl. 2), ii138–ii148 (2003)

20. Kim, S., Imoto, S., Miyano, S.: Dynamic Bayesian Network and Nonparametric Regression for Nonlinear Modeling of Gene Networks from Time Series Gene Expression Data. Biosystems 75, 57–65 (2004)
21. Zou, M., Conzen, S.D.: A New Dynamic Bayesian Network Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data. Bioinformatics 21(1), 71–79 (2005)
22. Dojer, N., Gambin, A., Mizera, A., Wilczyski, B., Tiuryn, J.: Applying Dynamic Bayesian Networks to Perturbed Gene Expression Data. BMC Bioinformatics 7, 249 (2006)
23. Li, X., Rao, S., Jiang, W., Li, C., Yun, X.: Discovery of Time-Delayed Gene Regulatory Networks Based on Temporal Gene Expression Profiling. BMC Bioinformatics 7, 26 (2006)
24. Shi, Y., Mitchell, T., Bar-Joseph, Z.: Inferring Pairwise Regulatory Relationships from Multiple Time Series Datasets. Bioinformatics 23(6), 755–763 (2007)
25. Schelter, B., Winterhalder, M., Timmer, J.: Handbook of Time Series Analysis: Recent Theoretical Developments. Wiley-VCH, Weinheim (2006)
26. Li, H., Guo, H.: Molecular basis of the ethylene signaling and response pathway in Arabidopsis. Journal of Plant Growth Regulation 26(2), 106–117 (2007)