GENE NETWORK RECONSTRUCTION WITH LATENT FACTORS BASED ON INDEPENDENT COMPONENT ANALYSIS

W.F. Wang*, Y.Y. Fang**, C.C, Huang*, C.M. Chen*

* Inst. of Biomed. Engin., National Taiwan University, No.1, Sec. 4, Roosevelt Rd., Taipei, Taiwan
** Dept of Electrical Engin., National Central University, No.300, Jhongda Rd., Jhongli, Taiwan

ming@lotus.mc.ntu.edu.tw

Abstract: Latent factors, though not observable from the experiments, are generally of existence and potentially play influential roles on the gene expressions of interest. While various approaches have been proposed to estimate the genetic network from a set of microarray data, the influence of the latent factors has not been taken into account by most of previous reconstruction algorithms. As a consequence, the gene-gene interactions may be over- or under-estimated, even with a correct network topology. To account for the effects of latent factors, a new gene network reconstruction algorithm based on ICA is proposed in this paper. The expression level of each gene is assumed to be a linear function of latent factors and observable gene expressions. The latent factors are extracted from the observed gene expressions by using independent component analysis, the cost function of which is the kurtosis of the latent signals regularized by the mean squared errors between the predicted and observed gene expressions. AIC is used as the cost function to determine the number of latent factors and the network topology. The optimal solutions are sought by fast fix-point method. To evaluate the performance, the proposed algorithm has been validated by using simulated time-course data.

Introduction

Reconstruction of genetic regulatory networks has emerged as one of the most exciting and challenging task in postgenomic biology. With the large-scale genetic expression data produced by microarray technologies, various approaches have been proposed to reconstruct the genetic regulatory networks, such as Boolean networks [1-2], graph models [3], Bayesian networks [4-6], additive regulation models [7-10], and so on.

As one of the first approaches to modeling genetic regulation, Boolean networks [1-2] simplify the gene expression into two states, i.e., *on* or *off*, and attempt to describe the logic of gene interactions. The advantages of the Boolean networks consist in their noise immunity and weak statistical requirement on the amount of expression data. However, Boolean networks have been criticized for the unrealistic modeling of continuous gene expressions into binary states. Even though Boolean networks may be generalized to logical networks with multiple states [11], the limited ability of logical networks in describing real gene interactions remain as an intrinsic problem.

Instead of shooting for a quantitative characterization of gene interactions, graph models [3] aim to build a directed graph or a digraph to describe the gene relations. Compared to Boolean networks, graph models offer an even simpler description of the gene interactions. More specifically, graph models reveal only structural information of genetic networks rather than how genes interact with each other. Basically, graph models share similar pros and cons with Boolean networks.

Bayesian networks attempt to capture the causal relations among genes by encoding the conditional joint probability distribution of gene expressions. The Bayesian network is an attractive approach to modeling genetic regulatory network because of its solid basis in statistics, which enables it to handle the stochastic aspects of noisy microarray measurements in a natural manner. Moreover, it can be used even with incomplete data. However, Bayesian networks suffer a major drawback that they cannot deal with feedback loops generally existing in genetic regulation.

Additive regulation models refer to the genetic network models using weighted sums as the core, as suggested by D'haeseleer [12] to unify those similar ideas with different names. Some typical examples of additive regulation models are linear model [7,10], weighted matrices with nonlinear dose-response curve [8] and recurrent neural networks [9]. Additive regulation models represent a parametric approach to quantify the dynamic behavior of genetic interactions, e.g., dependencies between genes at different time instances. Theoretically, additive regulation models stand at a better position to characterize genetic interactions more realistic than Boolean networks and graph models for their continuous nature of data representation. Nevertheless, the limited data provided by microarray measurements have been generally insufficient to derive the parameters involved in the additive regulation models,

ISSN: 1727-1983

a problem known as curse of dimensionality. Although the dimensionality problem may be alleviated by reducing the number of parameters, e.g., clustering gene expression profiles [13-14], using Boolean networks [15] to define the structure of gene networks, etc., the number of parameters may still remain too many to be estimated with a high confidence.

While each class of genetic networks has been shown to be informative for understanding underlying gene relations, most of them share a common potential deficiency in elucidating genetic interactions, namely, the overfitting problem. The overfitting problem arises from the fact that the microarray data provide only the gene expression levels, while genetic interactions may involve participants other than mRNAs, such as proteins. Moreover, most genetic regulatory networks reconstructed in previous studies considered only a subset of genes. As a result, the unobserved factors may be incorrectly embedded in the overfitted gene networks containing only the interested genes.

Materials and Methods

Gene expression is regulated by complex interactions which may be co-effect of many underlying factors. In the recent literatures, although relation networks between genes can be estimated by many algorithms developed to reconstruct the gene network topology, most take gene and gene interactions into account simply without considering the underlying latent factors which can not be observed from large-scale microarray data such as temperature, enzyme and glucose concentration , etc.

Independent component analysis (ICA) is an algorithm widely used to extract the latent features underlying the observed signals. The observed signals in ICA are modeled as a linear mixture of unobservable factors, each of which is called an independent component (IC). The latent variables are assumed as nongaussian and mutually independent. To reconstruct a genetic regulatory network involving relations within genes and interactions between genes and latent factors, linear regression is used to estimate the part of causal relation among genes and the effects of underlying factors are drawn by ICA. It is assumed that gene-gene interaction is time-independent.

Links extracted by partial correlation are used to estimate the regulatory relation among genes. By incorporating the capability of ICA in extracting latent factors, the proposed gene reconstruction algorithm is formulated as a constrained optimization problem defined as

$$\min \quad J = -\sum_{w=1}^{n}[\frac{1}{T}\sum_{t=1}^{T} f(s_i(t))^4 - 3(\frac{1}{T}\sum_{t=1}^{T} f(s_i(t))^2)^2] \qquad (1)$$

$$\text{subject to } \Lambda = \sum_{t=2}^{T} \left\| (g(t) - \mathbf{A}\mathbf{s}(t) - \mathbf{B}g(t-1)) \right\|^2 < \varepsilon$$

where J is the negative of the absolute value of kurtosis, which is a measure of nongaussianity of ICs, Λ the sum of square errors between the observed and predicted gene expressions, s_i the ith latent factor (IC), A mixing matrix of latent factors, B weighting matrix of gene-gene interaction, T the number of observed data for each gene, and ν a Gaussian random variable.

The constrained optimization problem is converted to an unconstrained optimization problem by using a Lagrange multiplier and solved by fast fix-point iterative algorithm used to find the optimal solution more efficiently than the gradient method. The unknowns to be solved include the latent factors s, mixing matrix A and interaction matrix B.

Network topology determination is the most important and challenging step throughout the gene network reconstruction task. Links connecting genes are determined by partial correlation. de la Fuente [18] suggested that partial correlation be used to gain the meaningful correlation between two variables. Different from correlation, partial correlation determines the association between two variables when other variables are constrained. It is the correlation between the residuals of the interested variables as the common factors are adjusted. For instance, the real association between variables i and j, which are both correlated to variable k, is the Pearson correlation between the residuals of linear regressions of i and j when k is conditioned, respectively [18,19].

Accurate ICs estimation depends on the reliable relationships among genes. On the other hand, incorrect IC estimation would result in erroneous genetic network. To overcome this issue, an Expectation-Maximization algorithm is proposed to derive the ICs and network topology iteratively. In the E-step, given model parameters, including the network topology, mixing matrix **A** and weighting matrix **B**, the ICs are sought by solving the constrained optimization problem posed in Eq. (1) using the fix-point method. In the M-step, given the latent factors, i.e., the ICs, the model parameters are determined by optimizing the Akaike information criterion (AIC) of the reconstructed gene network based on partial correlation and linear regression.

Results

To demonstrate the performance of the proposed gene network reconstruction algorithm, Fig. 1 illustrates a gene network with 6 observable genes and 2 latent factors. The number on the directed link from gene g_i to gene g_j indicates the influence of gene *i* on gene *j*, which corresponds to $\mathbf{B}(j,i)$. Similarly, the number on the directed link from latent factor F_i to gene g_j indicates the influence of latent factor *i* on gene *j*, which corresponds to $\mathbf{A}(j,i)$.

ISSN: 1727-1983

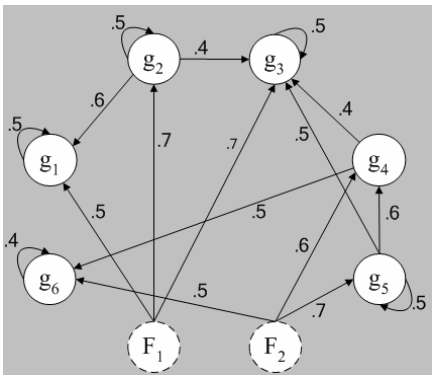

Figure 1: The simulation model for performance evaluation of the proposed gene network reconstruction algorithm, in which $g_1 \sim g_6$ are six observable gene expressions and F_1 and F_2 are two latent factors.

In the first experiment, we show that given the true network topology, the proposed algorithm is able to derive the latent factors, mixing matrix **A** and weighting matrix **B** very close to the true setups. For example, the time-course data of these two latent factors estimated by the proposed algorithm and those of the true latent factors are plotted in Fig. 2, which are denoted by solid curves and curves marked by 'x', respectively. It is clear that both estimated latent factors reasonably resemble the true ones.

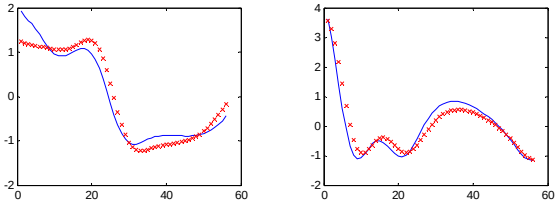

Figure 2: Two true (marked by x) and estimated (solid curve) latent factors in a simulation, given the true gene-gene network topology.

Table 1: The estimated and true mixing matrices A and interaction matrices B by a linear regression method.

	Interaction matrix **B**						Mixing matrix **A**	
	g_1	g_2	g_3	g_4	g_5	g_6	s_1	s_2
g_1	.58 (.5)	.65 (.6)					0 (.5)	0 (0)
g_2		.80 (.5)	.30 (.4)				0 (.7)	0 (0)
g_3			.81 (.5)	-1.02(.4)	1.48(.5)		0 (.7)	0 (0)
g_4					.95 (.6)		0 (0)	0 (.6)
g_5					.91 (.5)		0 (0)	0 (.7)
g_6				1.39 (.5)		-.07(.4)	0 (0)	0 (.5)

Table 1 shows the mixing matrix **A** and weighting matrix **B** estimated by a conventional linear regression method. The left part of Table 1 shows the weighting matrix **B** estimated by linear regression. As contrast, Table 2 lists the estimated and true mixing matrics A and interaction matrics B. The true values are given in the parentheses. In fact, the root mean squared error between the observed and predicted gene expressions is only 0.03. Compared to Fig.1, the estimated weight on each link is quite close to the true one.

Table 2: The estimated and true mixing matrices A and interaction matrices B by the proposed approach.

	Interaction matrix **B**						Mixing matrix **A**	
	g_1	g_2	g_3	g_4	g_5	g_6	s_1	s_2
g_1	.51 (.5)	.55 (.6)					.56 (.5)	-.13 (0)
g_2		.41 (.5)	.38 (.4)				.78 (.7)	-.18 (0)
g_3			.43 (.5)	.54 (.4)	.42 (.5)		.78 (.7)	-.18 (0)
g_4					.61 (.6)		.07 (0)	.52 (.6)
g_5					.51 (.5)		.08 (0)	.60 (.7)
g_6				.52 (.5)		.4 (.4)	.06 (0)	.43 (.5)

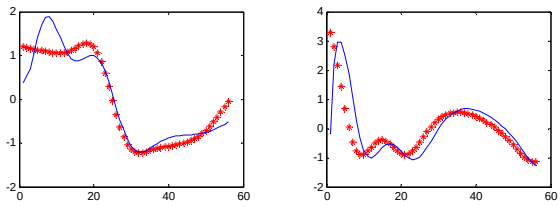

Figure 3: Two true (marked by x) and estimated (solid curve) latent factors without given true network.

Table 3: The estimated and true mixing matrices A and interaction matrices B by the proposed approach without prior information on the true network.

	Interaction matrix **B**						Mixing matrix **A**	
	g_1	g_2	g_3	g_4	g_5	g_6	s_1	s_2
g_1	.46 (.5)	.33 (.6)					1.31(.5)	0 (0)
g_2	.01(0)	-.01(.5)	.47 (.4)				1.86(.7)	0 (0)
g_3		-.5(0)	.49 (.5)	.97 (.4)	0 (.5)		1.96(.7)	0 (0)
g_4				-2.11(0)	4.66(.6)	.12(0)	0 (0)	1.85(.6)
g_5				-1.56(0)	4.67(.5)		0 (0)	2.33(.7)
g_6				-1.26(.5)	3.39(0)	.5 (.4)	0 (0)	1.54(.5)

In the second experiment, we assume no prior information on network topology and attempt to reconstruct the simulated network using the proposed algorithm. The preliminary results show that the true positive rate (i.e., the ratio of the number of identified true links to the number of total true links) can be as

ISSN: 1727-1983

high as 91% and the accuracy can be higher than 81% . The mixing matrix **A** and weighting matrix **B** of the second experiment are summarized in Table 3. Fig. 3 shows the comparison between true and extracted latent factors.

Discussion

It is clear that even with a correct network topology, the estimated weight of each link is quite different from the true one using the linear regression method. More seriously, some links even have wrong signs for the estimated weights. It means that gene g_i might be originally an activator of gene g_i, but concluded as a repressor instead. The accuracy of the proposed algorithm is higher than 80%. However, some of the derived link weights are quite different from the real ones.

Conclusions

A new gene network reconstruction algorithm accounting for the effect of latent factors is proposed. It has the advantage over Bayesian network that no model assumption is needed for the gene-gene interaction.

References

[1] LIANG, S., FUHRMAN, S., and SOMOGYI, R. (1998): 'REVEAL, A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures.', *Pacific Symposium on Biocomputing*, **3**, pp.18-29

[2] AKUTSU, T., MIYANO, S., and KUHARA, S. (1999): 'Identification of Genetic Networks from A Small Number of Gene Expression Patterns under The Boolean Network Model.', *Pacific Symposium on Biocomputing*, **4**, pp.17-28

[3] WAGNER, A. (2002): 'Estimating Coarse Gene Network Structure from Large-Scale Perturbation Data.', *Genome Research*, **12**, pp.309-15

[4] SPELLMAN, P., SHERLOCK, G., ZHANG, M., IYER, V., ANDERS, K., EISEN, M., BROWN, P., BOTSTEIN, D., and FUTCHER, B. (1998) 'Comprehensive Identification of Cell Cycle-Regulated Genes of The Yeast *Saccharomyces Cerevisiae* by Microarray Hybridization.', *Mol. Biol. Cell*, **9**, pp.3273-97

[5] FRIEDMAN, N. LINIAL, M., NACHMAN, I., and PE'ER, D. (2000): 'Using Bayesian Networks to Analyze Expression Data.', *J. Comput. Biol*., **7**, pp. 601-20

[6] PE'ER, D., REGEV., A., ELIDAN, G., and FRIEDMAN, N. (2001): 'Inferring Subnetworks from Perturbed Expression Profiles.', *Bioinformatics*, **17(S1)**, pp. S215-24

[7] D'HAESELLER, P., WEN, X., FUHRMAN, S., and SOMOGYI, R. (1999): 'Linear Modeling of Mrna Expression Levels During CNS Development and Injury.', *Pacific Symposium on Biocomputing*, **4**, pp.41-52

[8] WEAVER, D. C., WORKMAN, C. T., and STORMO, G. D. (1999): 'Modeling Regulatory Networks with Weight Matrices.', *Pacific Symposium on Biocomputing*, **4**, pp.112-23

[9] WAHDE, M., and HERTZ, J. (2000): 'Coarse-Grained Reverse Engineering of Genetic Regulatory Networks.', *Biosystems*, **55**, pp.129-36

[10] TEGNER, J., YEUNG, M. K. S., HASTY, J., and COLLINS, J. H. (2003): 'Reverse Engineering Gene Networks: Integrating Genetic Perturbations with Dynamic Modeling.', *PNAS*, **100(10)**, pp.5944-9

[11] MENDOZA, L., and THIEFFRY, D., and ALVAREZ-BUYLLA, E. R. (1999): 'Genetic Control of Flower Morphogenesis in *Arabidopsis Thaliana*: A Logic Analysis.', *Bioinformatics*, **15(7-8)**, pp.593-606

[12] D'HAESELLER, P. (2000): 'Reconstructing Gene Networks from Large Scale Gene Expression Data.' PhD Dissertation, p. 9, University of New Mexico, Albuquerque, New Mexico, USA

[13] van SOMEREn, E. T., WESSELS, L. F. A., and Reinders, M. J. T. (2000): 'Linear Modeling Genetic Networks from Experimental Data.', *Proceedings of the Eighth International Conference on Intelligent System for Molecular Biology*, 355-66

[14] MJOLSNESS, E., MANN, T., CASTANO, R., and WOLD, B. (2000): 'From Coexpression to Coregulation: An Approach to Inferring Transcriptional Regulation among Gene Classes from Large-Scale Expression Data.', *Advances in Neural Information Processing Systems*, **12**, pp.928-34

[15] MAKI, Y., TOMINAGA, D., OKAMOTO, M., WATANABE, S., and EGUCHI, Y. (2001), 'Development of A System for The Inference of Large-Scale Genetic Networks.', *Pacific Symposium on Biocomputing*, **6**, pp.446-58

[16] HYVARINEN, A. et al. (2001), '*Independent Component Analysis*.', (J. Wiley and Sons, New York)

[17] SCHÄFER, J. and STRIMMER, K. (2005): 'An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks.', *Bioinformatics*, **21**, pp. 754 - 64

[18] de la FUENTE, A., , BING, A., HOESCHELE, I., and MENDES, P. (2004): 'Discovery of Meaningful Associations in Genomic Data Using Partial Correlation Coefficients.', *Bioinformatics*, **20**, pp. 3565 - 74

[19] ROSNER, B. (2000):'Regression and Correlation Methods', in DUXBURY (5): 'Fundamental of biostatistics', (Duxbury, CA), pp. 495-96