Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data of Bacillus Subtilis Using Differential Equations

M.J.L. de Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, S. Miyano

Pacific Symposium on Biocomputing 8:17-28(2003)

INFERRING GENE REGULATORY NETWORKS FROM TIME-ORDERED GENE EXPRESSION DATA OF BACILLUS SUBTILIS USING DIFFERENTIAL EQUATIONS

MICHIEL J.L. DE HOON[1], SEIYA IMOTO[1], KAZUO KOBAYASHI[2],
NAOTAKE OGASAWARA[2], SATORU MIYANO[1]
[1]*Human Genome Center, Institute of Medical Science, University of Tokyo*
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2]*Graduate School of Biological Science, Nara Institute of Science and Technology,*
8916-5 Takayama, Ikoma, Nara 630-0101, Japan

We describe a new method to infer a gene regulatory network, in terms of a linear system of differential equations, from time course gene expression data. As biologically the gene regulatory network is known to be sparse, we expect most coefficients in such a linear system of differential equations to be zero. In previously proposed methods, the number of nonzero coefficients in the system was limited based on ad hoc assumptions. Instead, we propose to infer the degree of sparseness of the gene regulatory network from the data, where we use Akaike's Information Criterion to determine which coefficients are nonzero. We apply our method to MMGE time course data of *Bacillus subtilis*.

1 Introduction

The recently developed cDNA microarray technology allows gene expression levels to be measured for the whole genome at the same time. While the amount of available gene expression data has been increasing rapidly, the required mathematical techniques to analyze such data is still in development. Particularly, deriving a gene regulatory network from gene expression data has proven to be difficult.

In time-ordered gene expression measurements, the temporal pattern of gene expression is investigated by measuring the gene expression levels at a small number of points in time. Periodically varying gene expression levels have for instance been measured during the cell cycle of the yeast *Saccharomyces cerevisiae*.[1] The gene response to a slowly changing environment has been measured during the diauxic shift of the same yeast.[2] Other experiments consider the temporal gene expression pattern due to an abrupt change in the environment of the organism. As an example, the gene expression response was measured of the cyanobacterium *Synechocystis* sp. PCC 6803 after to sudden shift in the intensity of external light.[3,4]

Several methods have been proposed to infer gene interrelations from expression data. In cluster analysis,[2,5,6] genes are grouped together based on the similarity between their gene expression profiles. Inferring Boolean or

Bayesian networks from measured gene expression data has been proposed previously,[7,8,9,10,11] as well as modeling gene expression data using an arbitrary system of differential equations.[12] To reliably infer such an arbitrary system of differential equations, however, a long series of time-ordered gene expression data would be needed, which currently is often not yet available.

Instead, we will construct a linear system of differential equations from gene expression data. This approach maintains the advantages of quantitativeness and causality inherent in differential equations, while being simple enough to be computationally tractable.

Previously, modeling biological data with linear differential equations was considered theoretically by Chen.[13] In this model, both the mRNA and the protein concentrations were described by a system of linear differential equations. Such a system can be described as

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{x}\left(t\right)=\underline{\underline{\Lambda}}\cdot\underline{x}\left(t\right), \tag{1}$$

in which the vector $\underline{x}\left(t\right)$ contains the mRNA and protein concentrations as a function of time, and the matrix $\underline{\underline{\Lambda}}$ is constant with units of $[\mathrm{second}]^{-1}$. This equation can be considered as a generalization of the Boolean network model, in which the number of levels is infinite instead of binary.

In cDNA microarray experiments, usually only the gene expression levels are determined by measuring the corresponding mRNA concentrations, while the protein concentration is unknown. We therefore focus on a system of differential equations describing gene interactions only. A matrix element Λ_{ij} then represents the effect of gene j on gene i, $\left[\Lambda_{ij}\right]^{-1}$ being the reaction time.

To infer the coefficients in the system of differential equations from measured data, it was previously suggested[13] to discretize the system of differential equations, substitute the measured mRNA and protein concentrations, and solve the resulting linear system of equations to find the coefficients Λ_{ij} in the system of linear differential equations. The system of equations is usually underdetermined. Using the additional requirement that the gene regulatory network should be sparse, Chen showed that the model can be constructed in $O\left(m^{h+1}\right)$ time, where m is the number of genes and h is the number of nonzero coefficients allowed for each differential equation in the system.[13]

The parameter h is chosen ad hoc, which has two unexpected consequences. As each row in the matrix $\underline{\underline{\Lambda}}$ will have exactly h nonzero elements, every gene or protein in the network has h parent genes or proteins, and consequently no genes or proteins can exist at the top of a network. Secondly, every gene will inevitably be a member of a feedback loop. While feedback loops are likely to exist in gene regulatory networks, their existence should be determined from

the measured data instead of created artificially.

Bayesian networks, on the other hand, do not allow the existence of loops. Bayesian networks rely on the joint probability distribution of the estimated network to be decomposable in a product of conditional probability distributions. This decomposition is possible only in the absence of loops. We further note that Bayesian networks tend to contain many parameters, and therefore need a large amount of data for a reliable estimation.

We therefore aim to find a method that allows the existence of loops in the network, but does not require their presence. Using Eq. 1, we construct a sparse matrix by limiting the number of nonzero coefficients that may appear in the system. Instead of choosing this number ad hoc, we estimate which coefficients in the interaction matrix are zero from the data by using Akaike's Information Criterion (AIC), allowing the number of gene regulatory pathways to be different for each gene.

Our method can be applied to find a network between individual genes, as well as a regulatory network between clusters of genes. As an example, we infer a gene regulatory network between clusters of genes using time course data of *Bacillus subtilis*. Clusters are created using the k-means clustering algorithm. The biological function of the clusters can be determined from the functional categories of the genes belonging to each cluster.

2 Method

We consider a regulatory network between m genes in terms of a linear system of differential equations (Eq. 1), where the vector $\underline{x}\,(t)$ contains the expression ratios of the m genes at time t. This system of differential equations can be solved as

$$\underline{x}\,(t) = \exp\left(\underline{\underline{\Lambda}} t\right) \cdot \underline{x}_0 \ , \tag{2}$$

in which $\underline{x}_0$ contains the gene expression ratios at time zero. In this equation, the matrix exponential is defined in terms of a Taylor expansion as[14]

$$\exp\left(\underline{\underline{A}}\right) \equiv \sum_{i=0}^{\infty} \frac{1}{i!} \underline{\underline{A}}^i \ . \tag{3}$$

As Eq. 2 depends nonlinearly on $\underline{\underline{\Lambda}}$, it will be difficult to solve for $\underline{\underline{\Lambda}}$ in terms of the measured data $\underline{x}\,(t)$. An approximate solution can be found by replacing the differential equation (Eq. 1) by a difference equation:

$$\frac{\Delta \underline{x}}{\Delta t} = \underline{\underline{\Lambda}} \cdot \underline{x} \ , \tag{4}$$

or

$$\underline{x}\,(t+\Delta t)-\underline{x}\,(t)=\Delta t\cdot\underline{\underline{\Lambda}}\cdot\underline{x}\,(t)\ , \tag{5}$$

which is of the form considered by Chen.[13] To be able to statistically determine the sparseness of matrix $\underline{\underline{\Lambda}}$, we explicitly add an error $\underline{\varepsilon}\,(t)$, which will invariably be present in the data:

$$\underline{x}\,(t+\Delta t)-\underline{x}\,(t)=\Delta t\cdot\underline{\underline{\Lambda}}\cdot\underline{x}\,(t)+\underline{\varepsilon}\,(t)\ . \tag{6}$$

By using this equation, we effectively describe a gene regulatory network in terms of a multidimensional linear Markov model.

We assume that the error has a normal distribution independent of time:

$$f\left(\underline{\varepsilon}\,(t)\,;\sigma^2\right)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m\exp\left\{-\frac{\underline{\varepsilon}\,(t)^{\mathrm{T}}\cdot\underline{\varepsilon}\,(t)}{2\sigma^2}\right\}\ , \tag{7}$$

with a standard deviation σ equal for all genes at all times. The log-likelihood function for a series of time-ordered measurements $\underline{x}_i$ at times t_i, $i\in\{1,\ldots,n\}$ at n time points is then

$$L\left(\underline{\underline{\Lambda}},\sigma^2\right)=-\frac{nm}{2}\ln\left[2\pi\sigma^2\right]-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\hat{\underline{\varepsilon}}_i^{\mathrm{T}}\cdot\hat{\underline{\varepsilon}}_i\ , \tag{8}$$

in which

$$\hat{\underline{\varepsilon}}_i=\underline{x}_i-\underline{x}_{i-1}-(t_i-t_{i-1})\cdot\underline{\underline{\Lambda}}\cdot\underline{x}_{i-1} \tag{9}$$

is the measurement error at time t_i estimated from the measured data.

The maximum likelihood estimate of the variance σ^2 can be found by maximizing the log-likelihood function with respect to σ^2. This yields

$$\hat{\sigma}^2=\frac{1}{nm}\sum_{i=1}^{n}\hat{\underline{\varepsilon}}_i^{\mathrm{T}}\cdot\hat{\underline{\varepsilon}}_i\ . \tag{10}$$

Substituting this into the log-likelihood function (Eq. 8) yields

$$L\left(\underline{\underline{\Lambda}},\sigma^2=\hat{\sigma}^2\right)=-\frac{nm}{2}\ln\left[2\pi\hat{\sigma}^2\right]-\frac{nm}{2}. \tag{11}$$

To find the maximum likelihood estimate $\hat{\underline{\underline{\Lambda}}}$ of the matrix $\underline{\underline{\Lambda}}$, we use Eq. 9 to write the total squared error $\hat{\sigma}^2$ as

$$\begin{aligned}\hat{\sigma}^2=\frac{1}{nm}\sum_{i=1}^{n}\quad&\Big[\left(\underline{x}_i^{\mathrm{T}}-\underline{x}_{i-1}^{\mathrm{T}}\right)\cdot\left(\underline{x}_i-\underline{x}_{i-1}\right)+(t_i-t_{i-1})^2\,\underline{x}_{i-1}^{\mathrm{T}}\cdot\underline{\underline{\Lambda}}^{\mathrm{T}}\cdot\underline{\underline{\Lambda}}\cdot\underline{x}_{i-1}\\&-2\left(\underline{x}_i^{\mathrm{T}}-(t_i-t_{i-1})\,\underline{x}_{i-1}^{\mathrm{T}}\right)\cdot\underline{\underline{\Lambda}}\cdot\underline{x}_{i-1}\Big]\ ,\end{aligned} \tag{12}$$

and take the derivative with respect to $\underline{\underline{\Lambda}}$. We find a linear equation in $\underline{\underline{\Lambda}}$:

$$\hat{\underline{\underline{\Lambda}}} = \underline{\underline{B}} \cdot \underline{\underline{A}}^{-1}, \tag{13}$$

in which the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are defined as

$$\underline{\underline{A}} \equiv \sum_{i=1}^{n} \left[(t_i - t_{i-1})^2 \cdot \underline{x}_{i-1} \cdot \underline{x}_{i-1}^{\mathrm{T}} \right] ; \tag{14}$$

$$\underline{\underline{B}} \equiv \sum_{i=1}^{n} \left[(t_i - t_{i-1}) \cdot \left(\underline{x}_i - \underline{x}_{i-1} \right) \cdot \underline{x}_{i-1}^{\mathrm{T}} \right] . \tag{15}$$

In the absence of errors, the estimated matrix $\hat{\underline{\underline{\Lambda}}}$ is equal to the true matrix $\underline{\underline{\Lambda}}$. We know from biology that the gene regulatory network and therefore $\underline{\underline{\Lambda}}$ is sparse. However, all of the elements in the estimated matrix $\hat{\underline{\underline{\Lambda}}}$ may be nonzero due to the presence of noise, even if the corresponding elements in the true matrix $\underline{\underline{\Lambda}}$ are zero. We may decide to set a matrix element equal to zero if the resulting increase in the total squared error, as given by Eq. 12, is small.

Formally, we would use Akaike's Information Criterion[15,16]

$$\mathrm{AIC} = -2 \cdot \left[\begin{array}{c} \text{log-likelihood of the} \\ \text{estimated model} \end{array} \right] + 2 \cdot \left[\begin{array}{c} \text{number of estimated} \\ \text{parameters} \end{array} \right] \tag{16}$$

to decide which matrix elements should be set equal to zero. The AIC avoids overfitting of a model to data by comparing the total error in the estimated model to the number of parameters that was used in the model. The model with the lowest AIC is considered to be optimal. The AIC is based on information theory and is widely used for statistical model identification, especially for time series model fitting.[17]

We use a mask $\underline{\underline{M}}$ to set matrix elements of $\hat{\underline{\underline{\Lambda}}}$ equal to zero:

$$\hat{\underline{\underline{\Lambda}}}' = \underline{\underline{M}} \circ \hat{\underline{\underline{\Lambda}}}, \tag{17}$$

where $\circ$ denotes the Hadamard (element-wise) product,[14] and the mask $\underline{\underline{M}}$ is a matrix whose elements are either one or zero. The corresponding total squared error $\hat{\sigma}^2$ can be found by replacing $\hat{\underline{\underline{\Lambda}}}$ by $\hat{\underline{\underline{\Lambda}}}'$ in Eq. 12. The total squared error, given the mask $\underline{\underline{M}}$, can be minimized by solving the set of equations

$$\begin{aligned} &\text{if } M_{ij} = 1: \quad \left[\hat{\underline{\underline{\Lambda}}}' \cdot \underline{\underline{A}} \right]_{ij} = B_{ij}; \\ &\text{if } M_{ij} = 0: \qquad \hat{\Lambda}'_{ij} = 0; \end{aligned} \tag{18}$$

yielding the maximum likelihood estimate $\underline{\underline{\hat{\Lambda}}}'$. In this equation, $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are determined from Eqs. 14 and 15 using the measured gene expression levels $\underline{x}_i$.

We then calculate the AIC corresponding to $\underline{\underline{M}}$ by substituting the estimated log-likelihood function from Eq. 11 into Eq. 16:

$$\mathrm{AIC} = nm \ln\left[2\pi\hat{\sigma}^2\right] + nm + 2 \cdot \left(1 + \left[\begin{array}{c}\text{sum of the mask} \\ \text{elements } M_{ij}\end{array}\right]\right) , \qquad (19)$$

the estimated parameters being $\hat{\sigma}^2$ and the elements of the matrix $\underline{\underline{\hat{\Lambda}}}$ that we allow to be nonzero. From this equation, we see that while the squared error decreases, the AIC may increase as the number of nonzero elements increases. A gene regulatory network may now be inferred from gene expression data by finding the mask $\underline{\underline{M}}$ that yields the lowest value for the AIC.

For any but the most trivial cases, the number of possible masks $\underline{\underline{M}}$ is extremely large, making an exhaustive search to find the optimal mask infeasible. Instead, we propose a greedy search. Initially, we choose a mask at random, with an equal probability of zero or one for each mask element. We attempt to reduce the AIC by changing each of the mask elements M_{ij}. This process is continued until we find a final mask, for which no further reduction in the AIC can be achieved. We repeat this algorithm many times starting from different (random) initial masks, and choose the final mask $\underline{\underline{M}}$ that has the smallest corresponding AIC. If this optimal mask is found in several tens of trials, we assume that no better masks exist.

3 Results

We will demonstrate our technique of finding a gene regulatory network using gene expression data that were recently measured in an MMGE gene expression experiment of *Bacillus subtilis*.[18] MMGE is a synthetic minimal medium containing glucose and glutamine as carbon and nitrogen sources. In this medium, the expression of genes required for biosynthesis of small molecules, such as amino acids, is induced. The expression levels of 4320 ORFs were measured at eight time points at one hour intervals in this experiment, making two measurements at each time point.

3.1 Data preprocessing

To reduce the effect of measurement noise present in the data, the expression levels of each gene were compared to the measured background level. Genes with an average gene expression level lower than the average background level in either the red or the green channel were removed from the analysis.

Global normalization was then applied to the 3823 remaining genes, and the base-2 logarithms of the gene expression ratios were calculated. Since we are only interested in genes with appreciably changing expression levels during the experiment, we applied a statistical test to the measured log-ratios to determine if they are significantly different from zero. Usually, the t-test would be performed at every time point to determine which log-ratios are significantly different from zero. However, a t-test would be unreliable in this experiment, as there are only two measurements at each time point. We therefore devised a statistical test incorporating the measurements at all eight time points.

Under the null hypothesis, we assume that a gene was not affected by the experimental manipulation. The measured log-ratios at different time points are then equivalent. We further assume that the log-ratios have a normal distribution with zero mean. The standard deviation is then estimated from all $8 \times 2 = 16$ measurements:

$$\hat{\sigma}_{j|\mathrm{H}_0} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1,2} \left(x_{ji}\left[k\right]\right)^2}, \tag{20}$$

in which $x_{ji}\left[k\right]$ denotes the data value of measurement k at time point i for gene j. At each time point, we calculate the average log-ratio as

$$\bar{x}_{ji} = \frac{1}{2} \sum_{k=1,2} x_{ji}\left[k\right]. \tag{21}$$

Under the null hypothesis, $\bar{x}_{j\cdot}$ (the average of two gene expression log-ratios at a time point) is a random variable with a normal distribution with zero mean and an estimated standard deviation $\hat{\sigma}_{j|\mathrm{H}_0}/\sqrt{2}$. The joint probability for $\bar{x}_{j\cdot}$ to be larger in absolute value than the measured values $\bar{x}_{ji}$ is then

$$\begin{aligned} P = \prod_{i=1}^{n} P_i &= \prod_{i=1}^{n} p\left(|\bar{x}_{j\cdot}| > |\bar{x}_{ji}|\right) \\ &= \prod_{i=1}^{n} \left[1 - \mathrm{erf}\left(\frac{|\bar{x}_{ji}|}{\hat{\sigma}_{j|\mathrm{H}_0}/\sqrt{2}}\right)\right], \end{aligned} \tag{22}$$

in which erf is the error function. For a single factor P_i in this product, we would normally choose a significance level α, and reject the null hypothesis if $P_i < \alpha$. Accordingly, we adopt the criterion that $P < \alpha^n$ for rejection of the null hypothesis. This allows us to determine whether the expression levels of a gene change significantly during the experiment by making use of all the available data for that gene.

We chose a significance level $\alpha = 0.00025$ such that the expected number of false positives ($0.00025 \times 3823 = 1$) is acceptable. By applying this criterion to the 3823 genes, we found that 684 genes were significantly affected.

3.2 Clustering

The 684 genes were subsequently clustered into five groups using k-means clustering. The Euclidean distance was used to measure the distance between genes, while the centroid of a cluster was defined by the median over all genes in the cluster. The number of clusters was chosen such that a significant overlap was avoided. The k-means algorithm was repeated 1,000,000 times starting from different random initial clusterings. The optimal solution was found 81 times. The full clustering result is available at http://bonsai.ims.u-tokyo.ac.jp/∼mdehoon/publications/Subtilis/clusters.html.

In order to determine the biological function of the clusters that were created, we considered the functional category in the SubtiList database[19,20] for all genes in each cluster. Table 1 lists the main functional categories for the five clusters that were formed.

Figure 1 shows the log-ratio of the gene expression as a function of time for each cluster. While the expression levels of clusters I, II, and V change considerably during the time course, clusters II and III have fairly constant expression levels. Cluster IV in particular can be considered as a catchall cluster, to which genes are assigned that do not fit well in the other clusters.

3.3 Network construction

From the measured log-ratios of those twelve genes, we constructed the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ and calculated the matrix $\hat{\underline{\underline{\Lambda}}}$. The process of calculating a mask $\underline{\underline{M}}$, starting from a random initial mask, was repeated 1000 times. The optimal solution was found 55 times. It is therefore unlikely that there are other masks with a lower AIC. Note that the total number of possible masks is $2^{25} = 33,554,432$.

The network that was found is shown in Figure 2. The number of parents of a cluster in the network varies between zero and five. Clusters III and IV appear as the top of the network, while clusters I, II, and V are connected in a loop. Note that this network can neither be generated by the previously proposed method,[13] nor by a Bayesian network model.

The two strongest interactions in the network are the positive and negative effect of cluster IV on cluster V and cluster II respectively. This suggests that the opposite behavior of the gene expression levels of cluster II and V are caused by cluster IV, instead of a direct interaction between clusters II and V.

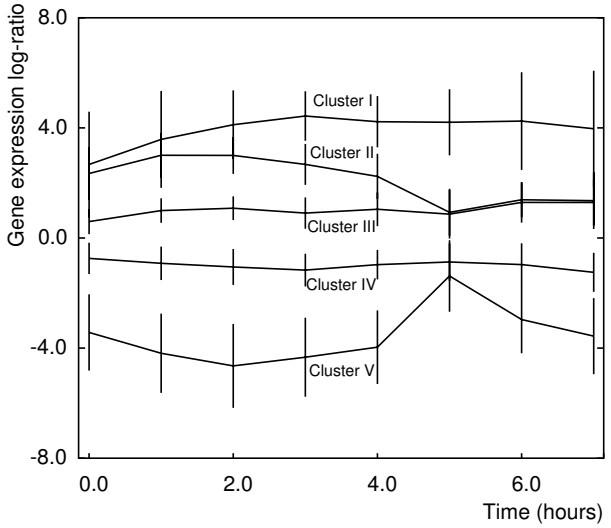

Figure 1: The log-ratio of the gene expression as a function of time for each cluster, as determined from the measured gene expression data.

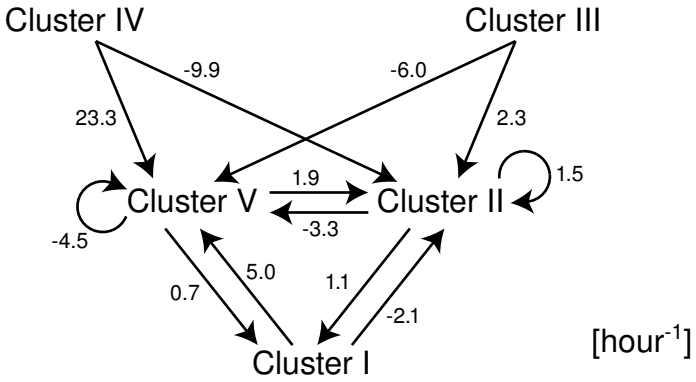

Figure 2: The network between the five gene clusters, as determined from the MMGE time-course data. The values show how strongly one gene cluster affects another gene cluster, as given by the corresponding elements in the interaction matrix $\hat{\underline{\underline{\Lambda}}}'$. In effect, this matrix represents how rapidly gene expression levels respond to each other. As an example, a change in the gene expression level of Cluster I would cause the expression level of Cluster V to change considerably within $1/(5.0 \text{ hour}^{-1}) = 12$ minutes, if the expression levels of Clusters II, III, and IV are unchanged.

Table 1: The main functional categories for the five clusters created using k-means clustering. The functional categories refer to the SubtiList database at Institut Pasteur.

Cluster	Number of genes	Main functional categories
I	42	2.2: 11 genes; 1.1: 9 genes
II	62	1.2: 15 genes; 2.2: 12 genes
III	187	5.1: 30 genes; 6.0: 23 genes; 1.2: 22 genes
IV	343	5.1: 40 genes; 5.2: 39 genes; 1.2: 33 genes
V	50	1.2: 15 genes; 2.1.1: 15 genes

Functional categories

1.1:	Cell wall.
1.2:	Transport/binding proteins and lipoproteins.
2.1.1:	Metabolism of carbohydrates and related molecules — Specific pathways.
2.2:	Metabolism of amino acids and related molecules.
5.1:	Similar to unknown proteins from *Bacillus subtilis*.
5.2:	Similar to unknown proteins from other organisms.
6.0:	No similarity.

4 Discussion

We have shown a method to infer a gene regulatory network in the form of a linear system of differential equations from measured gene expression data. Due to the limited number of time points at which measurements are typically made, finding a gene regulatory network is usually an underdetermined problem. Since biologically the resulting gene regulatory network is expected to be sparse, we set some of the matrix entries equal to zero, and infer a network using only the nonzero entries. The number of nonzero entries, and thus the sparseness of the network, was determined from the data using Akaike's Information Criterion without using any ad hoc parameters.

Describing a gene network in terms of differential equations has three advantages. First, the set of differential equations describes causal relations between genes: a coefficient Λ_{ij} of the coefficient matrix determines the effect of gene j on gene i. Second, it describes gene interactions in an explicitly numerical form. Third, because of the large amount of information present in a system of differential equations, other network forms can easily be derived from it. In addition, we can link the inferred network to other analysis or visualization tools, such as *Genomic Object Net*[22].

In previously described methods, either loops cannot be found (such as

in Bayesian network models) or the method artificially generates loops in the network. While the method proposed here allows loops to be present in the network, their existence is not required. Loops are found only if warranted by the data. When inferring a regulatory network between gene clusters using time-course data of *Bacillus subtilis* in an MMGE medium, we found that some of the clusters were part of a loop, while others were not.

If the number of genes m is equal to or larger than the number of experiments n, the matrix $\underline{\underline{A}}$ in Eq. 18 is singular. The problem is then underdetermined, and an interaction matrix $\hat{\underline{\underline{\Lambda}}}$ can be found with zero total error $\hat{\sigma}^2$ and an AIC of $-\infty$. This breakdown of our proposed method can be avoided by applying it to a sufficiently small number of genes or gene clusters, or by limiting the number of parents in the network.

References

1. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization" *Mol. Biol. Cell* **9** (1998) 3273–3297.
2. J.L. DeRisi, V.R. Iyer, and P.O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale" *Science* **278** (1997) 680–686.
3. Y. Hihara, A. Kamei, M. Kanehisa, A. Kaplan, and M. Ikeuchi, "DNA microarray analysis of cyanobacterial gene expression during acclimation to high light" *The Plant Cell* **13** (2001) 793–806.
4. M.J.L. de Hoon, S. Imoto, and S. Miyano, "Statistical analysis of a small set of time-ordered gene expression data using linear splines" *Bioinformatics*, in press.
5. M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns" *Proc. Natl. Acad. Sci. USA* **95** (1998) 14863–14868.
6. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation" *Proc. Natl. Acad. Sci. USA* **96** (1999) 2907–2912.
7. S. Liang, S. Fuhrman, and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inference of genetic network architectures" *Proc. Pac. Symp. on Biocomputing* **3** (1998) 18–29.
8. T. Akutsu, S. Miyano, and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways" *Bioinformatics* **16** (2000)

727–734.
9. N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data" *J. Comp. Biol.* **7** (2000) 601–620.
10. S. Imoto, T. Goto, and S. Miyano, "Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression" *Proc. Pac. Symp. on Biocomputing* **7** (2002) 175–186.
11. S. Imoto, S.-Y. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano, "Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network" *Proc. IEEE Computer Society Bioinformatics Conference* (2002) 219–227.
12. E. Sakamoto and H. Iba, "Evolutionary inference of a biological network as differential equations by genetic programming" *Genome Informatics* **12** (2001) 276–277.
13. T. Chen, H.L. He, and G.M. Church, "Modeling gene expression with differential equations" *Proc. Pac. Symp. on Biocomputing* **4** (1999) 29–40.
14. R.A. Horn and C.R. Johnson, *Matrix Analysis.* Cambridge University Press, Cambridge, UK (1999).
15. H. Akaike, "Information theory and an extension of the maximum likelihood principle" Research Memorandum No. 46, Institute of Statistical Mathematics, Tokyo (1971). In B.N. Petrov and F. Csaki (editors), *2nd Int. Symp. on Inf. Theory.* Akadémiai Kiadó, Budapest (1973) 267–281.
16. H. Akaike, "A new look at the statistical model identification" *IEEE Trans. Automat. Contr.* **AC-19** (1974) 716–723.
17. M.B. Priestley, *Spectral Analysis and Time Series.* Academic Press, London (1994).
18. Microbial Advanced Database Organization (Micado). http://www-mig.versailles.inra.fr/bdsi/Micado/.
19. I. Moszer, P. Glaser, and A. Danchin, "SubtiList: a relational database for the Bacillus subtilis genome" *Microbiology* **141** (1995) 261–268.
20. I. Moszer, "The complete genome of Bacillus subtilis: From sequence annotation to data management and analysis" *FEBS Letters* **430** (1998) 28–36
21. T.W. Anderson and J.D. Finn, *The New Statistical Analysis of Data.* Springer Verlag, New York (1996).
22. H. Matsuno, A. Doi, Y. Hirata, and S. Miyano, "XML documentation of biopathways and their simulation in Genomic Object Net" *Genome Informatics* **12** (2001) 54–62. *Genomic Object Net* is available at http://www.GenomicObject.net.