BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1477–1485

Statistical analysis of a small set of time-ordered gene expression data using linear splines

M.J.L. de Hoon, S. Imoto and S. Miyano*

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

Received on December 21, 2001; revised on March 18, 2002; accepted on April 17, 2002

ABSTRACT

Motivation: Recently, the temporal response of genes to changes in their environment has been investigated using cDNA microarray technology by measuring the gene expression levels at a small number of time points. Conventional techniques for time series analysis are not suitable for such a short series of time-ordered data. The analysis of gene expression data has therefore usually been limited to a fold-change analysis, instead of a systematic statistical approach.

Methods: We use the maximum likelihood method together with Akaike's Information Criterion to fit linear splines to a small set of time-ordered gene expression data in order to infer statistically meaningful information from the measurements. The significance of measured gene expression data is assessed using Student's *t*-test.

Results: Previous gene expression measurements of the cyanobacterium *Synechocystis* sp. PCC6803 were reanalyzed using linear splines. The temporal response was identified of many genes that had been missed by a fold-change analysis. Based on our statistical analysis, we found that about four gene expression measurements or more are needed at each time point.

Availability: An extension module for Python to calculate linear spline functions is available at http://bonsai.ims.u-tokyo.ac.jp/~mdehoon. This software package (with patent pending) is free of charge for academic use only.

Contact: mdehoon@ims.u-tokyo.ac.jp

INTRODUCTION

In recent years, many cDNA microarray experiments have been performed measuring gene expression levels under different conditions. Measured gene expression data have become widely available in publicly accessible databases, such as the KEGG database (Nakao *et al.*, 1999).

In some of these experiments, the steady-state gene expression levels are measured under several environmental conditions. For instance, the expression levels of the cyanobacterium *Synechocystis* sp. PCC6803 and a mutant have been measured at different temperatures, leading to the identification of the gene Hik33 as a potential cold sensor in this cyanobacterium (Suzuki *et al.*, 2001).

In other experiments, the temporal pattern of gene expression is considered by measuring expression levels at a limited number of points in time. Gene expression levels that vary periodically have, for instance, been measured during the cell cycle of the yeast *Saccharomyces cerevisiae* (Spellman *et al.*, 1998). The expression levels of the same yeast species were measured during the metabolic shift from fermentation to respiration (DeRisi *et al.*, 1997). In this experiment, the environmental conditions were changing slowly over time. Conversely, the gene response to an abruptly changing environment can be measured. As an example, the gene expression levels of the cyanobacterium *Synechocystis* sp. PCC 6803 were measured at several points in time after a sudden shift from low light to high light (Hihara *et al.*, 2001).

In these experiments, gene expression levels are typically measured at a small number of time points. Conventional techniques for time series analysis, such as Fourier analysis or autoregressive or moving-average modeling, are not suitable for such a small number of data points. Instead, the gene expression data are often analyzed by clustering techniques or by considering the relative change in the gene expression level only. Such a fold-change analysis may miss significant changes in gene expression levels, while it may inadvertently attribute significance to measurements dominated by noise. In addition, a simple fold-change analysis may not be able to identify important features in the temporal gene expression response.

Several techniques to analyze gene expression data, such as deriving Boolean or Bayesian networks, have been proposed in the past (Liang *et al.*, 1998; Akutsu *et al.*, 2000; Friedman *et al.*, 2000; Imoto *et al.*, 2002). Whereas describing gene interactions in terms of a regulatory network is very important, deriving a network model requires gene expression data at a large number of time points, which is currently often not yet available. It should be noted that the number of genes is on the order of several thousands, while the gene expression levels are usually

*To whom correspondence should be addressed.

measured at only five or ten points in time.

So far, a systematic method has been lacking to statistically analyze gene expression measurements from a small number of time-ordered data. In this paper, we will outline a strategy based on fitting linear spline functions to time-ordered data using the maximum likelihood method and Akaike's Information Criterion (Akaike, 1973, 1974). The significance of the gene expression measurements is assessed by applying Student's t-test. This allows us to infer information from gene expression measurements while taking the statistical significance of the data into consideration. This kind of analysis should be viewed as a first step towards building gene regulatory networks.

As an example, we reanalyzed the gene expression measurements of the cyanobacterium *Synechocystis* sp. PCC 6803 (Hihara *et al.*, 2001). It is shown that information can be inferred from the measured data that is missed when considering the fold-change only. By repeating our analysis with a subset of the available data, we were able to determine how many measurements are needed at each time point in order to estimate the linear spline function reliably.

METHODS

Student's t-test

Gene expression data are usually given in terms of the base-2 logarithm of the expression ratio, defined as the expression level of a gene relative to its level in some control condition. We would first like to assess if these log-ratios are significantly different from zero. The significance of the measured data can be established by applying Student's t-test for each time point separately. Since multiple comparisons are being made for each gene, the value of the significance level α should be chosen carefully.

We define $\mathrm{H}_0^{(i)}$ as the hypothesis that for a given gene the log-ratio is equal to zero at a given time point t_i, and H_0 as the hypothesis that for a given gene the log-ratios at all time points are equal to zero. If we denote α as the significance level for rejection of hypothesis H_0, and α' as the significance level for rejection of hypothesis $\mathrm{H}_0^{(i)}$, then α' and α are related via

$$1 - \alpha = \left(1 - \alpha'\right)^a, \tag{1}$$

in which a is the number of time points at which the gene expression ratio was measured. Note that by linearizing the right-hand side in α', this equation reduces to Bonferroni's method for adjusting significance levels (see also Anderson and Finn, 1996).

By performing Student's t-test at every time point for each gene, using α' as the significance level, we will find whether $\mathrm{H}_0^{(i)}$ and therefore H_0 should be rejected. If H_0 is not rejected, we can conclude that the gene is not significantly affected by the experimental manipulations, and should therefore not be included in further analyses. If, for a given gene, the null hypothesis H_0 is rejected, we conclude that the gene was significantly affected by the experimental manipulations.

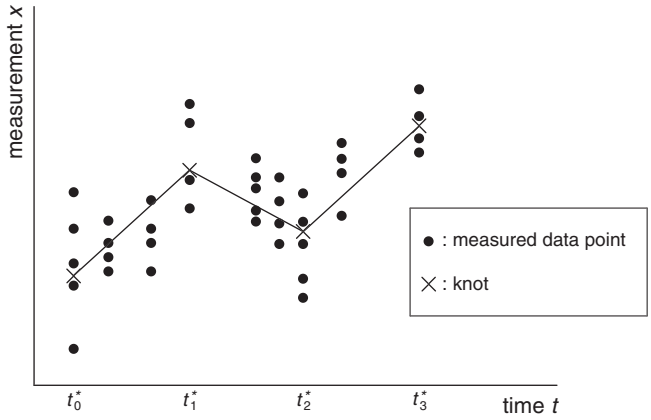

Fig. 1. A conceptual example of a linear spline function fitted to measured data.

Analyzing time-ordered data using linear splines

Next, we analyze the temporal gene expression response for genes that were found to be significantly affected. The measured log-ratios form a small set of time-ordered data, to which we can fit a linear spline function. A linear spline function is a continuous function consisting of piecewise linear functions, which are connected to each other at knots (Friedman and Silverman, 1989; Higuchi, 1999; Higuchi and Ohtani, 2000). Whereas cubic splines are used more commonly, for the small number of data points we are dealing with linear spline functions are more suitable. A conceptual example of a linear spline function with knots is shown in Figure 1.

Consider a set of data points $\left(t_j, x_j\right)$, $j \in \{1, \ldots, n\}$, in which t_j is the time of measurement and x_j is the log-ratio of the measured gene expression level. We wish to fit a nonparametric regression model of the form

$$x_j = g\left(t_j\right) + \epsilon_j \tag{2}$$

to these data, in which g is a linear spline function with knots $(t_0^*, t_1^*, \ldots, t_q^*)$ and ϵ_j, $j \in \{1, \ldots, n\}$, are independent random variables with a normal distribution with zero mean and variance σ^2. In experiments, the log-ratio of the gene expression level is typically observed to follow a normal distribution.

We estimate the linear spline function g using the maximum likelihood method. The probability distribution

of one data point x_j, given t_j, is

$$f\left(x_j|t_j; g, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(x_j - g\left(t_j\right)\right)^2}{2\sigma^2}\right\}. \tag{3}$$

The log-likelihood function for the n data points is then given by

$$L\left(g, \sigma^2\right) = -\frac{n}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(x_j - g\left(t_j\right)\right)^2. \tag{4}$$

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - g\left(t_j\right)\right)^2. \tag{5}$$

The log-likelihood function can then be written in the form

$$L\left(g, \sigma^2 = \hat{\sigma}^2\right) = -\frac{n}{2}\ln\left[2\pi\hat{\sigma}^2\right] - \frac{n}{2}. \tag{6}$$

The maximum likelihood estimate $\hat{g}$ of the linear spline function g can now be found by minimizing $\hat{\sigma}^2$. It can be shown that the minimum value of $\hat{\sigma}^2$ will be achieved if the linear spline function is chosen such that

$$\underline{\underline{A}} \cdot \underline{\hat{g}} = \underline{b}, \tag{7}$$

in which $\underline{\hat{g}} = \left(\hat{g}_0, \hat{g}_1, \ldots, \hat{g}_q\right)^T$ is a vector containing the estimated values $\hat{g}_i$ of the linear spline function at the knots t_i^*, $\underline{\underline{A}}$ is a tridiagonal symmetric matrix given by

$$A_{00} = \sum_{j:\, t_0^* \le t_j < t_1^*}\left(\frac{t_1^* - t_j}{t_1^* - t_0^*}\right)^2; \tag{8}$$

$$A_{i,i} = \sum_{j:\, t_{i-1}^* < t_j \le t_i^*}\left(\frac{t_j - t_{i-1}^*}{t_i^* - t_{i-1}^*}\right)^2 + \sum_{j:\, t_i^* < t_j < t_{i+1}^*}\left(\frac{t_{i+1}^* - t_j}{t_{i+1}^* - t_i^*}\right)^2 \text{ for } 0 < i < q; \tag{9}$$

$$A_{q,q} = \sum_{j:\, t_{q-1}^* < t_j \le t_q^*}\left(\frac{t_j - t_{q-1}^*}{t_q^* - t_{q-1}^*}\right)^2; \tag{10}$$

$$A_{i+1,i} = A_{i,i+1} = \sum_{j:\, t_i^* < t_j < t_{i+1}^*}\frac{\left(t_{i+1}^* - t_j\right)\left(t_j - t_i^*\right)}{\left(t_{i+1}^* - t_i^*\right)^2} \quad \text{for } 0 \le i < q; \tag{11}$$

and $\underline{b}$ is a vector given by

$$b_0 = \sum_{j:\, t_0^* \le t_j < t_1^*}\frac{t_1^* - t_j}{t_1^* - t_0^*}x_j; \tag{12}$$

$$b_i = \sum_{j:\, t_{i-1}^* < t_j \le t_i^*}\frac{t_j - t_{i-1}^*}{t_i^* - t_{i-1}^*}x_j + \sum_{j:\, t_i^* < t_j < t_{i+1}^*}\frac{t_{i+1}^* - t_j}{t_{i+1}^* - t_i^*}x_j \quad \text{for } 0 < i < q; \tag{13}$$

$$b_q = \sum_{j:\, t_{q-1}^* < t_j \le t_q^*}\frac{t_j - t_{q-1}^*}{t_q^* - t_{q-1}^*}x_j. \tag{14}$$

In the case of time-ordered gene expression data, the gene expression ratio is typically calculated with respect to the gene expression level at time zero. By definition, at time zero, the log-ratio will then have a fixed point equal to zero. In general, a fixed point at time zero with a value g_0 can be incorporated in our methodology by modifying Equation (7). It can be shown that in this case the minimum value of $\hat{\sigma}^2$ will be achieved by choosing the linear spline function such that $\hat{g}_0 = g_0$, while $(\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_q)$ are determined from

$$\begin{pmatrix} A_{11} & A_{12} & 0 & \cdots & 0 & 0 \\ A_{21} & A_{22} & A_{23} & \cdots & 0 & 0 \\ 0 & A_{32} & A_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{q-1,q-1} & A_{q-1,q} \\ 0 & 0 & 0 & \cdots & A_{q,q-1} & A_{q,q} \end{pmatrix} \cdot \begin{pmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \\ \vdots \\ \hat{g}_{q-1} \\ \hat{g}_q \end{pmatrix} = \begin{pmatrix} b_1' \\ b_2 \\ b_3 \\ \vdots \\ b_{q-1} \\ b_q \end{pmatrix}, \tag{15}$$

in which $b_1' \equiv b_1 - A_{10}g_0$. For time-ordered gene expression levels given as log-ratios, we will have a fixed point $g_0 = 0$ and therefore $b_1' = b_1$.

The maximum number of knots $q_{\max}$ is equal to the number of time points at which the gene expression levels were measured. The number of possible knot placements increases exponentially with the maximum number of knots as $1 + 2^{q_{\max}-1}$. As an example, Figure 2 shows the possible knot positions for the experiment described below, in which measurements were made at 15 minutes, 1 hour, 6 hours, and 15 hours. The number of knots, q, will be between zero and four, in addition to a fixed knot equal to zero at time zero. For $q = 2$ and $q = 3$, three possibilities exist for the placement of the knots between the linear segments of the linear spline function. These

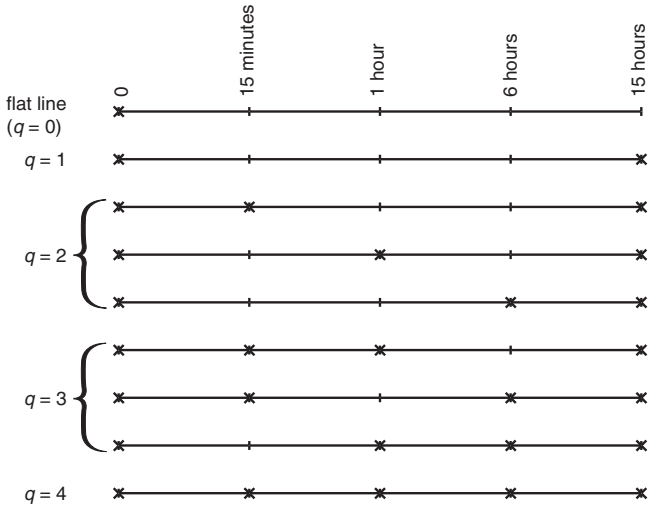

Fig. 2. Possible placement of knots for a time-ordered set of measurements at four time points, in addition to a fixed point at time zero.

are indicated in Figure 2, together with the cases $q = 0$, $q = 1$, and $q = 4$.

The fitted model depends on the number of knots, which can be chosen using Akaike's Information Criterion, known as the *AIC* (Akaike, 1973, 1974)

$$AIC = -2 \cdot \begin{bmatrix} \text{log-likelihood} \\ \text{of the estimated} \\ \text{model} \end{bmatrix} + 2 \cdot \begin{bmatrix} \text{number of} \\ \text{estimated} \\ \text{parameters} \end{bmatrix}, \quad (16)$$

in which the estimated parameters are $\hat{\sigma}^2$ and $(\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_q)$. The *AIC* is based on information theoretic concepts and is nowadays regarded as one of the most reliable methods for statistical model identification, particularly for time series model fitting (Priestley, 1994). Substituting the estimated log-likelihood function from Equation (6), we find

$$AIC = n \ln\left[2\pi\hat{\sigma}^2\right] + n + 2q + 2, \quad (17)$$

in which $\hat{\sigma}^2$ is given by Equation (5) after substitution of the maximum likelihood estimate $\hat{g}$ for the linear spline function g.

For each value of q, we calculate the value of the *AIC* after fitting the linear spline function as described above, and select the value of q that yields the minimum value of the *AIC*. The case $q = 1$ corresponds to linear regression. For the special case $q = 0$, we are effectively fitting a flat line to the data. If we find that for a particular gene, the minimum *AIC* is achieved for the constant function ($q = 0$), then we can conclude that the expression level of that gene was unaffected by the experimental manipulations.

Table 1. The gene expression measurements as a time-ordered set of data

Time point	Number of measurements
$t = 15$ minutes	6
$t = 1$ hour	6
$t = 6$ hours	4
$t = 15$ hours	4

This information can then be used to check the results of Student's t-test.

RESULTS

Student's t-test

We will illustrate the usage of Student's t-test and linear spline functions by reanalyzing the measured gene expression profile of the cyanobacterium sp. PCC 6803 after a sudden exposure to high light (HL); (Hihara *et al.*, 2001). The expression levels of 3079 ORFs were measured at 15 minutes, 1 hour, 6 hours, and 15 hours both for cyanobacteria exposed to HL and cyanobacteria that remained in the low light (LL) condition. Table 1 shows the number of measurements at each time point. Data from the cDNA expression measurements were obtained from the KEGG database (Nakao *et al.*, 1999). It should be mentioned that the data used for the original analysis (Hihara *et al.*, 2001) may not be identical to the raw data submitted to KEGG.

First, the background signal intensities were subtracted from the HL and LL raw data. Consistent with the original fold-change analysis (Hihara *et al.*, 2001), in order to reduce the effects of noise, genes with the 2000 lowest expression levels in either the LL or the HL condition were removed from the data set. The measured expression levels of these genes, being comparable to the background fluorescence levels, were dominated by noise (Hihara *et al.*, 2001). Global normalization was applied to the expression levels of the 912 remaining genes and the ratio of the HL to the LL signal intensities was calculated to find the gene expression ratio with respect to the control (LL) condition. Using global normalization directly without removing genes with low signal levels first may result in noisier gene expression ratios. The log-ratios of the gene expression levels were then found by taking the base-2 logarithm of the gene expression ratios.

In the fold-change analysis, a gene was regarded as being affected by HL if its expression level changed by a factor of two or more, which corresponds to the log-ratio changing by at least unity. The statistical significance of such changes was assessed heuristically by considering the size of the standard deviation of the measurements (Hihara *et al.*, 2001).

Table 2. Student's t-test of gene expression measurements

Significance level	Number of ORFs
$p < 0.0003$	58
$p < 0.001$	90
$p < 0.005$	171
$p < 0.01$	208
$p < 0.05$	362

The results of Student's t-test on the log-ratios of each gene separately are shown in Table 2. At a significance level of $\alpha = 0.005$, 171 genes were found to be significantly affected by the HL condition. Note that we would expect about five type-I errors among these 171 genes. In comparison, 164 ORFs were found to be affected by the HL condition in the fold-change analysis (Hihara *et al.*, 2001). An explicit example of the t-test is shown on our webpage at http://bonsai.ims.u-tokyo.ac.jp/∼mdehoon/publications/.

Table 3 lists the genes that were identified by the t-test at a significance level $\alpha = 0.001$. As indicated in the table, 36 genes of those genes had not been found by the fold-change analysis (Hihara *et al.*, 2001). Conversely, of the 164 ORFs that were identified in the fold-change analysis, six are not significantly affected by HL according to Student's t-test, even at a significance level $\alpha = 0.01$. Note that at this significance level, the t-test analysis yielded 208 ORFs whose expression levels had been significantly affected by HL (Table 2). At a significance level $\alpha = 0.001$, an additional fifteen genes are found not to be affected by HL. These ORFs are listed in Table 4.

Analysis using linear spline functions

In the fold-change analysis, the temporal gene expression patterns were classified into six categories (Hihara *et al.*, 2001), listed in Table 5. This classification was based on the average of the measured gene expression ratios at each time point. Instead, we will fit linear spline functions to the measured gene expression data.

Table 6 shows the knot positions of the linear spline function fitted to the measured log-ratios of each gene. Genes were included only if Student's t-test showed that they were significantly affected by HL, using a significance level $\alpha = 0.001$. None of the gene expression levels was described by a flat line, which is consistent with the results from Student's t-test.

As an example, we again consider the measured expression levels of the gene *cpcG* (*sll1471*). Table 7 shows the calculated *AIC* for the different sets of knot positions. The minimum *AIC* is achieved for knots at 0, 15 minutes, 1 hour, and 15 hours. Figure 3 shows the measured log-ratio for this gene as well as the fitted linear spline function.

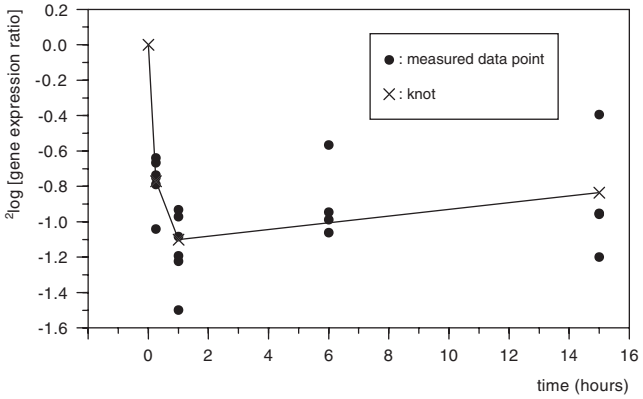

Fig. 3. The measured log-ratio for the gene *cpcG* (*sll1471*), together with the fitted linear spline. This gene was considered to be unaffected by HL in the fold-change analysis (Hihara *et al.*, 2001).

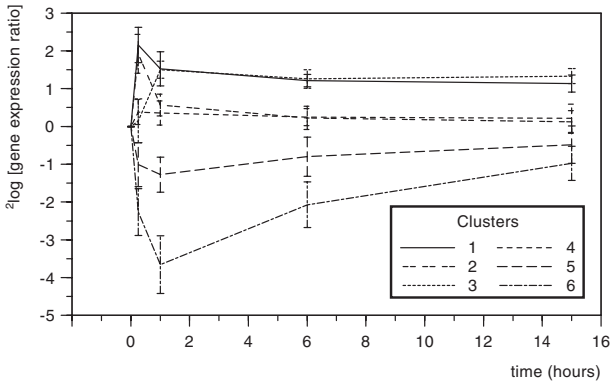

Fig. 4. The log-ratio of each cluster, as determined from the linear spline function fitted to the gene expression data. The error bars are a measure for the within-cluster deviation.

To assess the biological significance of our method, we applied k-means clustering (MacQueen, 1967) to the linear spline functions fitted to the measured log-ratios for the 90 genes considered to be significantly affected. Figure 4 shows the log-ratio of the gene expression level as a function of time for each cluster. The error bars shown at each time point are equal to the standard deviation over all genes in one cluster. The number of clusters was chosen to be six, which was the largest number of clusters without a significant overlap between the clusters.

The clusters that were found are shown in Table 3 above, together with the knot positions of the linear spline function that was fitted to the expression data of each gene. As a measure of the goodness of fit of the linear spline function, the percentage variance explained is shown for each gene. Most clusters contain functionally related genes, as well as some ORFs of unknown function. The functional annotations of the ORFs, as shown in the

Table 3. ORFs identified by Student's t-test at a significance level $\alpha = 0.001$, together with their gene name and biological function, if known. The ORFs were grouped using k-means clustering applied to the linear spline functions fitted to the measured log-ratio for each gene. For an explanation of the notation used in fourth and sixth column, see Table 6 and Table 5, respectively

ORF	*Gene*	Biological function	Knot positions	Percentage variance explained	Result of the fold-change analysis
Cluster 1					
sll0322	*hypF*	Transcriptional regulatory protein HypF	(c)	75	Type 2
sll0927	*metX*	*S*-adenosylmethlonine synthetase	(f)	86	Type 2
sll1028	*ccmK*	Carbon dioxide concentrating mechanism protein CcmK	(f)	85	Type 2
sll1029	*ccmK*	Carbon dioxide concentrating mechanism protein CcmK	(f)	90	Type 2
sll1031	*ccmM*	Carbon dioxide concentrating mechanism protein CcmM	(f)	96	Type 2
slr1963		Unknown	(i)	98	Type 1
slr2075	*groES*	10 kD chaperonin	(g)	96	Type 2
sll0416	*groEL-2*	60 kD chaperonin 2	(i)	99	Type 2
Cluster 2					
sll0170	*dnaK*	DnaK protein	(i)	89	Type 1
sll0430	*htpG*	Heat shock protein	(i)	97	Type 1
sll0521	*ndhG*	NADH dehydrogenase subunit 6	(i)	69	Not identified
slr1280	*ndhK*	NADH dehydrogenase subunit NdhK	(f)	89	Type 1
slr0011	*rbcX*	Unknown function	(f)	75	Type 2
slr1350	*desA*	Fatty acid desaturase	(f)	86	Type 1
slr1604	*ftsH*	Cell division protein FtsH	(f)	95	Type 1
slr1641	*clpB*	ClpB protein	(i)	94	Type 1
sll0814		Unknown	(f)	92	Type 1
slr0476		Unknown	(f)	87	Type 1
slr1687		Unknown	(f)	91	Not identified
Cluster 3					
sll1096	*rps12*	30S ribosomal protein S12	(f)	93	Type 2 (see text)
sll1097	*rps7*	30S ribosomal protein S7	(f)	92	Type 3
sll1260	*rps2*	30S ribosomal protein S2	(f)	89	Not identified
sll1743	*rpl11*	50S ribosomal protein L11	(i)	95	Type 3
sll1745	*rpl10*	50S ribosomal protein L10	(i)	93	Type 3
sll1799	*rpl3*	50S ribosomal protein L3	(f)	95	Type 3
sll1801	*rpl23*	50S ribosomal protein L23	(h)	91	Type 3
sll1802	*rpl2*	50S ribosomal protein L2	(i)	97	Type 3
sll1804	*rps3*	30S ribosomal protein S3	(f)	82	Type 3
sll1807	*rpl24*	50S ribosomal protein L24	(i)	95	Type 3
sll1809	*rps8*	30S ribosomal protein S8	(i)	91	Type 3
sll1810	*rpl6*	50S ribosomal protein L6	(f)	75	Type 3
sll1816	*rps13*	30S ribosomal protein S13	(i)	93	Type 3
Cluster 4					
sll0020	*clpC*	ATP-dependent Clp protease regulatory subunit	(f)	67	Not identified
sll0144	*pyrH* or *smbA*	Uridine monophosphate kinase	(d)	84	Not identified
sll0185		Unknown	(i)	88	Not identified
sll0262	*desD (des6)*	Delta-6 desaturate	(g)	67	Not identified
sll0414		Hypothetical protein	(f)	74	Not identified
sll0519	*ndhA*	NADH dehydrogenase subunit 1	(i)	59	Not identified
sll0587	*pykF*	Pyruvate kinase	(d)	57	Not identified
sll0680	*pstS* or *phoS*	Phosphate-binding periplasmic protein precursor (PBP)	(c)	74	Not identified
sll0854		Unknown	(g)	68	Not identified
sll1002	*ycf22*	Hypothetical protein	(f)	69	Not identified
sll1327	*atpC*	ATP synthase g subunit	(f)	76	Not identified
sll1614	*pma1*	Cation-transporting ATPase (E1-E2 ATPase)	(f)	51	Not identified
slr0208		Unknown	(f)	78	Not identified
slr0452	*ilvD*	Dihydroxyacid dehydratase	(h)	50	Not identified
slr0642		Integral membrane protein	(f)	54	Not identified
slr0839	*hemH*	Ferrochelatase	(c)	54	Not identified
slr1237	*codA*	Cytosine deaminase	(h)	88	Not identified
slr1469	*rnpA*	Protein subunit of ribonuclease P (RNase P)	(d)	63	Not identified

Table 3. *Continued.*

ORF	*Gene*	Biological function	Knot positions	Percentage variance explained	Result of the fold-change analysis
slr1513		Unknown	(f)	92	Not identified
slr1557		Unknown	(f)	58	Not identified
slr1718		Unknown	(c)	35	Not identified
ssr2554		Unknown	(d)	57	Not identified
Cluster 5					
sll0258	*psbV*	Cytochrome c550	(i)	94	Type 4
sll0617	*im30*	Chloroplast membrane-associated 30 kD protein	(h)	78	Not identified
sll0851	*psbC*	Photosystem II CP43 protein	(d)	80	Not identified
sll0901	*pure* or *ade6*	Phosphoribosylaminoimidazole carboxylase	(c)	85	Not identified
sll1091		391 aa (43 kD) bacteriochlorophyll synthase subunit	(c)	92	Type 4
sll1214	*PNIL34* or *AT103*	Phytochrome-regulated gene	(c)	89	Type 4
sll1306		Unknown	(c)	89	Type 4
sll1471	*cpcG*	Phycobilisome rod-core linker polypeptide CpcG	(f)	95	Not identified
sll1626	*lexA*	SOS function regulatory protein	(i)	93	Type 4
sll1694	*hofG* or *hopG*	General secretion pathway protein G	(i)	95	Type 6
sll1712		DNA binding protein HU	(i)	92	Not identified
slr0151		Unknown	(c)	63	Not identified
slr0329	*xylR*	Xylose repressor	(c)	90	Not identified
slr0374		Cell division cycle protein	(f)	91	Type 5
slr1128		Erthyrocyte band 7 integral membrane protein, protein 7.2B, stomatin	(i)	74	Not identified
slr1348	*cysE*	Serine acetyltransferase	(h)	91	Type 6
slr1459	*apcF*	Phycobilisome core component	(c)	94	Type 4
slr1545	*rpoE*	RNA polymerase sigma-E factor	(g)	90	Not identified
slr1793	*talB*	Transaldolase	(h)	90	Not identified
slr1853		Unknown	(c)	74	Not identified
slr1856		Anti-sigma B factor antagonist	(f)	92	Type 4
ssl1533		Unknown	(f)	96	Type 4
ssl3093	*cpcD*	Phycocyanin associated linker protein	(c)	96	Type 4
Cluster 6					
sll1577	*cpcB*	Phycocyanin b subunit	(i)	96	Type 5
sll1578	*cpcA*	Phycocyanin a subunit	(f)	92	Type 5
sll1579	*cpcC*	Phycocyanin associated linker protein	(i)	97	Type 5
sll1580	*cpcC*	Phycocyanin associated linker protein	(i)	97	Type 5
slr2051	*cpcG*	Phycobilisome rod-core linker polypeptide CpcG	(f)	95	Type 4
sll0819	*psaF*	Photosystem I subunit III	(i)	93	Type 5
slr0737	*psaD*	Photosystem I subunit II	(i)	97	Type 5
slr1655	*psaL*	Photosystem I subunit XI	(f)	94	Type 5
slr1834	*psaA*	P700 apoprotein subunit Ia	(h)	97	Type 5
slr1835	*psaB*	P700 apoprotein subunit Ib	(h)	95	Type 5
slr1855		Unknown	(g)	92	Type 4
slr1986	*apcB*	Allophycocyanin b chain	(i)	93	Type 4
slr2067	*apcA*	Allophycocyanin a chain	(i)	95	Type 4

table, were obtained from CyanoBase (Nakamura *et al.*, 1998).

Table 3 also shows the results of the fold-change analysis. Using a factor of two change in the expression level as a criterion to decide whether a gene is significantly affected leads to missing genes in several clusters. For instance, in the second cluster the gene *ndhK* (NADH dehydrogenase subunit NdhK) is present while the related gene *ndhG* (NADH dehydrogenase subunit 6) is missing in the fold-change analysis. Similarly, the gene *rps2* (30S ribosomal protein S2) is missing from the third cluster, containing ribosomal proteins only. None of the genes in cluster 4 were identified in the fold-change analysis.

For the fold-change analysis, better results can be obtained by using the *t*-test to select significantly affected genes. Even then, the gene *rps12* (30S ribosomal protein S12) is misclassified to the first cluster instead of the ribosomal protein cluster 3. This is due to the log-ratios being described by their averages at each time point instead of by their linear spline estimates.

Table 4. ORFs identified in the fold-change analysis (Hihara *et al.*, 2001), although they were not significantly affected by HL according to Student's t-test at a significance level $\alpha = 0.01$ or $\alpha = 0.001$

	$\alpha = 0.01$	$\alpha = 0.001$
Type 1	slr1311, sll1867	slr0228, slr0394, slr1516
Type 2	slr2076	slr0009, slr0012
Type 3	ssl3437, sll1818	sll0533, sll1261, sll1742, sll1800
Type 4	–	sll0427, sll1713, slr0335, slr1295, slr1854
Type 5	–	–
Type 6	sll1688	slr0272

Table 5. Six categories were used in the fold-change analysis to classify the temporal pattern of gene expression (Hihara *et al.*, 2001)

Type 1	Induced within 15 minutes, then decreased
Type 2	Induced continuously at high levels
Type 3	Induced at approximately one hour
Type 4	Repressed within 15 minutes, then increased
Type 5	Repressed continuously at low levels
Type 6	Repressed at approximately one hour

Table 6. Classification of the temporal gene expression response based on the knot positions of the fitted linear spline function

	Knot positions	Number of genes
(a)	Flat line	0
(b)	0, 15 hours	0
(c)	0, 15 minutes, 15 hours	13
(d)	0, 1 hour, 15 hours	5
(e)	0, 6 hours, 15 hours	0
(f)	0, 15 minutes, 1 hour, 15 hours	33
(g)	0, 15 minutes, 6 hours, 15 hours	5
(h)	0, 1 hour, 6 hours, 15 hours	8
(i)	0, 15 minutes, 1 hour, 6 hours, 15 hours	26

Finally, we would like to establish whether the number of measurements at each time point was sufficient to reliably determine the knot positions of the linear spline function. To do so, we repeated the estimation of the linear spline function using subsets of the measured data. We then counted for how many genes the estimated knot positions change if a subset of the data was used instead of the complete set of data. The average and standard deviation of this number for different numbers of data points used is shown in Table 8.

Even if only two data points are removed both at the 15 minutes and the 1 hour time point, and four data points are used at each time point, in 29% of the cases the estimated knot positions change. This suggests that in this experiment, four or more data points are needed at each time point to reliably deduce information from the gene expression measurements.

Table 7. The calculated *AIC* for the different sets of knot positions for the gene *cpcG* (*sll1471*)

Knot positions	*AIC*
Flat line	57.3
0, 15 hours	50.8
0, 15 minutes, 15 hours	7.5
0, 1 hour, 15 hours	19.4
0, 6 hours, 15 hours	48.3
0, 15 minutes, 1 hour, 15 hours	**2.4**
0, 15 minutes, 6 hours, 15 hours	9.5
0, 1 hour, 6 hours, 15 hours	19.9
0, 15 minutes, 1 hour, 6 hours, 15 hours	2.7

Table 8. Reliability of the linear spline fitting procedure as a function of number of data used. Only those genes were considered which were significantly affected by HL according to Student's t-test at a significance level $\alpha = 0.01$

Using six data points at 15 minutes and 1 hour, and four data points at 6 hours and 15 hours	linear spline functions are estimated for 208 genes
Using five data points at 15 minutes and 1 hour, and four data points at 6 hours and 15 hours	41 ± 11 estimated differently
Using four data points at each time point	60 ± 15 estimated differently
Using three data points at each time point	90 ± 16 estimated differently
Using two data points at each time point	120 ± 13 estimated differently

DISCUSSION

We have described a strategy based on the maximum likelihood method to analyze a set of time-ordered measurements. By applying Student's t-test to the measured gene expression data, we first establish which of the measured genes are significantly affected by the experimental manipulation. The expression responses of those genes are then described by fitting a linear spline function. The number of knots to be used for the linear spline function is determined using Akaike's Information Criterion (*AIC*).

There are several advantages to using linear spline functions. First, it allows us more flexibility in describing the measured gene expression compared to using a nominal classification. Also, in order to set up a gene regulatory network, it is important that the gene response as determined from gene expression measurements is available in a numerical form. Finally, the positions of the knots

specify those time points at which the expression of a gene changes markedly, which is important in identifying its biological function.

In the model described above, the knots of the linear spline function are placed at measurement time points only. A more general model can be considered in which knots can be placed at any time (Higuchi, 1999). For instance, in Figure 3 the two knots at 15 minutes and 1 hour can be replaced by one knot suitably placed between those two time points. Notice, however, that by allowing knots to be placed at any time, effectively two free parameters are assigned to each knot, which would complicate the model.

As a next step, the classification of gene expression responses based on the position of the knots can be refined. As an example, subcategories can be created that consider the change in slope of the linear spline function at the knots.

Applying the technique of linear spline functions to measured gene expression data, we identified the temporal expression response pattern of genes that were significantly affected by the experimental manipulations. The response of 36 of those genes was not noticed in earlier fold-change analyses of expression data. Furthermore, it was shown that for six genes the expression level response found in a fold-change analysis were not significant even to the 1% level according to Student's *t*-test.

Gene expression data tend to be noisy and are often plagued by outliers. Whereas Student's *t*-test and maximum likelihood methods described here take the statistical significance of noisy data into account, the issue of outliers needs to be addressed separately. As a simple procedure to remove outliers, we can calculate the mean and standard deviation of the data at each time point, and remove data that deviate more than two standard deviations from the mean.

The number of expression measurements needed at each time point to reliably fit a linear spline function was determined by removing some data points and fitting a linear spline function anew. It was found that if four data points per time point are used, in about 29% of the cases the knot positions will not be estimated reliably. For this experiment, it would therefore have been advisable to make more than four measurements per time point. In general, the number of measurements will depend on the pattern of the gene expression response as well as the magnitude of measurement errors.

REFERENCES

Akaike,H. (1973) Information theory and an extension of the maximum likelihood principle. Research Memorandum No. 46, Institute of Statistical Mathematics, Tokyo (1971). In Petrov,B.N. and Csaki,F. (eds), *2nd International Symposium on Informational Theory*. Akadémiai Kiadó, Budapest, pp. 267–281.

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **AC-19**, 716–723.

Akutsu,T., Miyano,S. and Kuhara,S. (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.

Anderson,T.W. and Finn,J.D. (1996) *The New Statistical Analysis of Data*. Springer, New York.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Friedman,J.H. and Silverman,B.W. (1989) Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*, **31**, 3–39.

Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Higuchi,T. (1999) Automatic identification of the large scale field aligned current systems as an example of knowledge discovery from the large database (in Japanese with English abstract). *Proceedings of the Institute of Statistical Mathematics*, **47**, 291–306.

Higuchi,T. and Ohtani,S. (2000) Automatic identification of a large-scale field-aligned current structures. *J. Geophys. Res.*, **105**, 25305–25315.

Hihara,Y., Kamei,A., Kanehisa,M., Kaplan,A. and Ikeuchi,M. (2001) DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. *The Plant Cell*, **13**, 793–806.

Imoto,S., Goto,T. and Miyano,S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.

Liang,S., Fuhrman,S. and Somogyi,R. (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, **3**, 18–29.

MacQueen,J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, **1**, 281–297.

Nakamura,Y., Kaneko,T., Hirosawa,M., Miyajima,N. and Tabata,S. (1998) CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res.*, **26**, 63–67.

Nakao,M., Bono,H., Kawashima,S., Kamiya,T., Sato,K., Goto,S. and Kanehisa,M. (1999) Genome-scale gene expression analysis and pathway reconstruction in KEGG. In Asai,K., Miyano,S. and Takagi,T. (eds), *Genome Informatics*, Vol. 10, University Academy Press, Tokyo, Japan, pp. 94–103.

Priestley,M.B. (1994) *Spectral Analysis and Time Series*. Academic Press, London.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., 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**, 3273–3297.

Suzuki,I., Kanesaki,Y., Mikami,K., Kanehisa,M. and Murata,N. (2001) Cold-regulated genes under control of the cold sensor Hik33 in *Synechocystis*. *Mol. Microbiol.*, **40**, 235–244.