Flexible temporal expression profile modelling using the Gaussian process

Ming Yuan*

School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, GA 30332, USA

Abstract

Time course gene expression experiments have proved valuable in a variety of biological studies [e.g., Chuang, Y., Chen, Y., Gadisetti, V., et al., 2002. Gene expression after treatment with hydrogen peroxide, menadione or *t*-butyl hydroperoxide in breast cancer cells. Cancer Res. 62, 6246–6254; Edwards, M.G., Sarkar, D., Klopp, R., Morrow, J.D., Weindruch, R., Prolla, T.A., 2003. Age-related impairment of the transcriptional response to oxidative stress in the mouse heart. Physiol Genomics 13, 119–127]. A general goal common to many of these time course experiments is to identify genes that exhibit different temporal expression profiles across multiple biological conditions. Such experiments are, however, often hampered by the lack of data analytical tools. Taking advantage of the great flexibility of Gaussian processes, we propose a statistical framework for modelling time course gene expression data. It can be applied to both long and short time series and also allows for multiple differential expression patterns. The method can identify a gene's temporal differential expression pattern as well as estimate the expression trajectory. The utility of the method is illustrated on both simulations and an experiment concerning the relationship between longevity and the ability to resist oxidative stress.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Temporal gene expression; Differential expression; Gaussian process

1. Introduction

Time course gene expression experiments monitor simultaneously individual markers in multiple samples over time, and hence allow for not only a snapshot of the activity in the cell, but also the temporal relationships among different genes. Such experiments have proved to be a valuable tool in a variety of biological studies (Chuang et al., 2002; Edwards et al., 2003). A general goal common to many of these time course experiments is to collect gene expression time series in multiple biological conditions such as different cancer tumor types or different treatments, and identify genes that exhibit different temporal expression profiles across multiple biological conditions.

In contrast to the rich literature on how to analyze gene expression data under a single time point (Parmigiani et al., 2003), few methods are available to identify temporally differentially expressed genes. Early statistical approaches are

* Tel.: +1 4048942359; fax: +1 4048942301.
 E-mail address: myuan@isye.gatech.edu.

largely based on linear regression or two-way analysis of variance (ANOVA) (Guo et al., 2003; Xu et al., 2002). These methods often suffered from the problem of low sensitivity due to the limited number of replicates in most time course experiments (Park et al., 2003).

Yuan and Kendziorski (2006) pointed out that more power could be gained if the temporal dependence is appropriately taken into account. They proposed a hidden Markov modelling (HMM) framework to efficiently identify differentially expressed genes at each time point and classify genes based on their temporal expression patterns. HMM is applicable for comparing two or more biological conditions and can be applied to both short and long time series. By directly targeting the relationship among expression profiles under different conditions, the method not only identifies genes that have different expression profiles across conditions, but also pinpoints when such difference in expression occurs.

In many time course gene expression studies, the expression trajectory under each condition is also of great interest. This is especially true when one wants to infer the expression profile between sampling time points. In such cases, one may want to model the temporal expression profile under individual conditions. Hong and Li (2004) proposed to model the profile as a linear combination of several B-spline basis functions. Genes that have different trajectories across different conditions are identified through the inference on the linear coefficients. The model fitting and empirical Bayes inferences are carried out by a Monte Carlo EM algorithm. Alternatively, Storey et al. (2004) modelled the gene specific expression profile using splines and a random scalar is introduced to specify the condition dependent variation. The inference on a gene's differential expression pattern is based on hypothesis testing of the variance component.

Clearly, HMM and methods based on modelling individual expression profiles each has their own appeal depending on the experiment setup and goal. HMM provides flexible inference tools for the gene expression level at the sampling points but does not provide information on the expression profile between time points. The inferences that profile-modelling based methods can make are more limited, but they yield better description of the expression trajectory under each condition.

Both of the existing profile-modelling based approaches, however, have their limits when it comes to more than two conditions. Hong and Li's approach can only be applied to two conditions. Although Storey et al.'s approach can handle more than two conditions, it can only classify genes into two patterns depending on whether it is equivalently expressed across all conditions or not.

To elaborate on this, consider the aging experiment from Edwards et al. (2003). The experiment was designed to better understand the genetic basis underlying the relationship between longevity and the ability to resist oxidative stress as we shall discuss in detail later. After stress induction, the investigators monitored the gene expression level for young, middle-aged and old mice at five different time points. There are no natural ways of applying Hong and Li's approach to compare the three age groups. Storey et al.'s approach can tell us which genes are not equivalently expressed across all three groups. But it cannot provide information on whether or not the differential expression occurs only for one group. Furthermore, the validity of both existing profile-modelling methods is questionable with such a small number of time points.

Similar to Hong and Li (2004) and Storey et al. (2004), the method proposed in this paper is based on modelling the expression profiles and is useful when expression level between time points is of interest. Taking advantage of the flexibility of Gaussian processes, the approach proposed here overcomes the aforementioned problems of the existing approaches. It can be applied to both long and short time series and also allows for multiple differential expression patterns. The main objective of time course experiment is to make inference regarding functions that represent temporal expression profiles. In contrast to the traditional approaches that model the functions in a parametric form, the main idea of our modelling strategy is to view it as a sample from the space of functions. Any inference regarding the profile then takes place directly in function space.

The rest of the paper is organized as follows. In Section 2, we described a Gaussian process approach to model the gene expression profiles. Inferences based on the proposed model are discussed in Section 3. The utility of the proposed method is illustrated by both simulations and a real case study in Sections 4 and 5, respectively. We conclude with some discussions in Section 6.

2. Gaussian process model

We assume that some preprocessing technique has been applied to adequately normalize the data so that the measurements reflect the true underlying gene expression. The general data structure of many time course microarray

experiments is as follows. There are multiple biological conditions; and for each condition, there are expression measurements taken over a set of time points. Often times, replicate measurements are obtained under different biological conditions at each time point.

Consider gene g under condition c. Suppose thats n_{gcm} replicate expression intensities for gene g are taken at possibly condition-dependent time point t_{cm}, where $m = 1, \ldots, M$. For brevity, we shall omit the subscript of n_{gcm} if no confusion occurs. Denote by $\mathbf{x}_{gcm} = (x_{gcm1}, \ldots, x_{gcmn})$ the log transformed intensity measurements. These measurements are noisy observations of the true underlying temporal gene expression profile, denoted by $\mu_{gc}(t)$, sampled at discrete time points $t = t_{c1}, \ldots, t_{cM}$. Specifically, the observational model could be described as

$$x_{gcmr} = \mu_{gc}(t_{cm}) + \varepsilon_{gcmr}, \tag{1}$$

where $\varepsilon_{gc} = (\varepsilon_{gcmr} : r = 1, \ldots, n_{gcm}, m = 1, \ldots, M) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ represents the measurement error where $\boldsymbol{\Sigma}$ is a $(\sum_m n_{gcm}) \times (\sum_m n_{gcm})$ positive definite matrix. The actual experiment design determines the modelling of variance–covariance matrix $\boldsymbol{\Sigma}$. In many studies, replicate measurements are taken for different subjects. In these applications, it is reasonable to model the measurement errors as independent variables, i.e., $\boldsymbol{\Sigma} = \sigma_0^2 I$. In some other studies, however, expression measurements are taken from the same subjects repeatedly. To account for the structure of repeated measurements, one needs to model the correlation among the measurement errors. Detailed discussion on this aspect is beyond the scope of this paper and the interested readers are referred to Heagerty et al. (2004) for common modelling strategies of $\boldsymbol{\Sigma}$.

The main challenge is how to model the true gene expression profile described by the function $\mu_{gc}(\cdot)$. Hong and Li (2004), among others, propose to model it as a linear combination of a finite set of B-spline basis functions. The rationale behind this modelling approach is the assumption that the temporal process evolves possibly nonlinearly but smoothly and this smoothness is governed by the number of basis functions used in modelling μ_{gc}. How well the profile can be approximated heavily depends on the number of basis functions and their respective locations. Unfortunately, the selection of basis functions, most of the time, can only be done on a case-to-case basis. A flexible alternative to the B-spline approach is to view $\mu_{gc}(\cdot)$ as a realization of a Gaussian process.

Although Gaussian process is routinely used in time series and spatial statistics, it may be less familiar to readers whose main interests are in gene expression analysis. For completeness, we first review some necessary concepts of Gaussian processes and covariance functions before proceeding. Interested readers are referred to Cressie (1993) or Stein (1999) for more details.

The idea of Gaussian process modelling is, without parametrizing a function, to view it as a sample from the space of functions. A Gaussian process defines a distribution over functions. It can be thought of as the generalization of a multivariate normal distribution over a finite vector space to a function space of infinite dimension. Different from parametric approaches such as the one used by Hong and Li (2004) where inferences about a function is made via the inference on the linear coefficients, any inference regarding the function takes place directly in function space with Gaussian process modelling.

A Gaussian process $Z(t)$ is a stochastic process whose finite dimensional distribution is multivariate normal for every n and every collection $\{Z(t_1), Z(t_2), \ldots, Z(t_n)\}$. Similar to the multivariate normal distribution, Gaussian processes are specified by their mean function $E(Z(t))$ and covariance function

$$\mathrm{cov}(Z(t_i), Z(t_j)) = K(t_i, t_j). \tag{2}$$

Just as a covariance matrix must be positive definite, a covariance function must also be positive definite in the sense that for every $n > 0$, $t_1, \ldots, t_n$ and $z_1, \ldots, z_n$

$$\sum_{i,j=1}^{n} z_i z_j K(t_i, t_j) \geqslant 0, \tag{3}$$

where the equality holds if and only if $z_1 = \cdots = z_n = 0$. For a stationary Gaussian process, $K(t_1, t_2)$ is a function of $t_1 - t_2$. We write $K(t_1, t_2) = K(t_1 - t_2)$. One of the most commonly used covariance functions is the Gaussian

covariance function

$$K(t_1, t_2) = \tau^2 \exp\left(-\frac{(t_1 - t_2)^2}{2\kappa^2}\right).$$ (4)

Gaussian processes with such covariance function are infinitely mean square differentiable and therefore especially appropriate for modelling functions that are known a priori to be smooth.

Now consider the true expression profile $\mu_{gc}(\cdot)$ for gene g under condition c. We view $\mu_{gc}(\cdot)$ as a realization of a Gaussian process with mean function $\mu_0(\cdot)$ and covariance function $K(\cdot, \cdot)$. We further model the mean function $\mu_0(\cdot)$ as a linear combination of a set of known basis functions $\{\phi_v(\cdot)\}$:

$$\mu_0(t) = \sum_{v=1}^{d} \alpha_v \phi_v(t).$$ (5)

Common choices of basis functions $\{\phi_v(\cdot)\}$ include prespecified B-spline basis functions or low order polynomials.

The primary goal of the time course experiment with multiple conditions is to make inferences concerning relationships among profiles from different conditions, denoted by $\mu_{g1}, \ldots, \mu_{gC}$, based on the observed expression measurements. Any two profiles μ_{gc_1} and μ_{gc_2} can either be equivalently expressed, i.e., $\mu_{gc_1} = \mu_{gc_2}$, or differentially expressed, i.e., $\mu_{gc_1} \neq \mu_{gc_2}$. This type of comparison between profiles can be naturally extended to more than two conditions. For example, in the aging experiment mentioned earlier, there are three conditions: aged, middle-aged and young. Correspondingly, there are three expression profiles $\mu_{g,\text{aged}}$, $\mu_{g,\text{middle}}$ and $\mu_{g,\text{young}}$, and the potential expression patterns include

$$H_1 : \mu_{g,\text{aged}} = \mu_{g,\text{middle}} = \mu_{g,\text{young}},$$ (6)

$$H_2 : \mu_{g,\text{aged}} \neq \mu_{g,\text{middle}} = \mu_{g,\text{young}},$$ (7)

$$H_3 : \mu_{g,\text{middle}} \neq \mu_{g,\text{aged}} = \mu_{g,\text{young}},$$ (8)

$$H_4 : \mu_{g,\text{young}} \neq \mu_{g,\text{aged}} = \mu_{g,\text{middle}},$$ (9)

$$H_5 : \mu_{g,\text{aged}} \neq \mu_{g,\text{middle}} \neq \mu_{g,\text{young}}.$$ (10)

More generally, the number of all possible patterns as a function of the number of conditions C is equal to the Bell exponential number of possible set partitions. There are also other types of hypotheses that might be of interest for time course microarray experiments such as the ones considered in Peddada et al. (2003). We chose the current set of differential expression patterns because it is more natural for comparing multiple conditions (Kendziorski et al., 2003; Yuan and Kendziorski, 2006). It is worth pointing out that the number of distinct patterns increases exponentially as the number of conditions increases. Therefore, in practice, prior knowledge narrowing down plausible patterns is very useful in the case of many conditions. For a gene equivalently expressed under conditions c_1 and c_2, $\mu_{gc_1} = \mu_{gc_2}$ is one realization of a Gaussian process; and for a differentially expressed gene where $\mu_{gc_1} \neq \mu_{gc_2}$, we assume that both μ_{gc_1} and μ_{gc_2} are two independent realizations of the same Gaussian process.

3. Inferences

There are two most important questions in time course experiments. First, one wants to identify a gene's differential expression pattern. Second, we are also interested in estimating the temporal expression profile. The first goal can be achieved by gauging the posterior probabilities for the possible patterns. The second task can be fulfilled by exploring the posterior distribution in the function space. To fix ideas, in the following, we will assume that $\mu_0(\cdot) = \alpha_0$ is a constant function and $K(\cdot, \cdot)$ is a Gaussian covariance function known up to parameters τ^2 and κ^2. The discussion should, however, be easily extended to more general setup.

3.1. Expression pattern

Denote by K the number of distinct patterns of expression under investigation. For pattern H_k, let $r(k)$ be the number of distinct temporal profiles among $\mu_{g1}, \ldots, \mu_{gC}$. Conditions $1, \ldots, C$ can be divided into $r(k)$ exclusive subsets $S_i, i = 1, \ldots, r(k)$, with conditions from S_i sharing the same profile. For example, in the case of two conditions, for pattern $\mu_{gc_1} = \mu_{gc_2}$, $r(1) = 1$ and $S_1 = \{c_1, c_2\}$; for pattern $\mu_{gc_1} \neq \mu_{gc_2}$, $r(2) = 2$ and $S_1 = \{c_1\}$, $S_2 = \{c_2\}$. In the aging experiment, $K = 5$ as described in the previous section and

$$r(1) = 1, \quad S_1 = \{\text{aged, middle-aged, young}\},$$

$$r(2) = 2, \quad S_1 = \{\text{aged}\}, \quad S_2 = \{\text{middle-aged, young}\},$$

$$r(3) = 2, \quad S_1 = \{\text{middle-aged}\}, \quad S_2 = \{\text{aged, young}\},$$

$$r(4) = 2, \quad S_1 = \{\text{young}\}, \quad S_2 = \{\text{aged, middle-aged}\},$$

$$r(5) = 3, \quad S_1 = \{\text{aged}\}, \quad S_2 = \{\text{middle-aged}\}, \quad S_3 = \{\text{young}\}.$$

For a gene g with pattern H_k, the marginal distribution of $\mathbf{x}_g$ can be expressed as $f_k(\mathbf{x}_g) = \prod_{i=1}^{r(k)} f(\mathbf{x}_{gS_i})$ where $\mathbf{x}_g$ is the vector of all measurements taken for gene g, $\mathbf{x}_{gS_i}$ is the vector of all measurements taken under conditions within S_i for gene g, and $f(\mathbf{x}_{gS_i})$ is its marginal distribution. $f_k(\mathbf{x}_g)$ can be derived in closed-form and therefore allows for fast computation. Details are provided in the appendix. Let π_k be the prior probabilities for the hypotheses $H_k, k = 1, \ldots, K$. Applying the Bayes Theorem, the posterior probabilities for the hypothesis are

$$P(H_k|\mathbf{x}_g) = \frac{\pi_k f_k(\mathbf{x}_g)}{\pi_1 f_1(\mathbf{x}_g) + \cdots + \pi_K f_K(\mathbf{x}_g)}. \tag{11}$$

Once the posterior probabilities defined in (11) are obtained, inferences can be made based on these quantities. For example, under 0–1 loss, we shall assign gene g to an expression pattern with the highest posterior probability (Berger, 1985). Certainly, in practice, other thresholds might also be used to give more conservative lists of potential differentially expressed genes. A natural question is how to measure the effectiveness of a cutoff probability ζ. The false discovery rate (FDR) introduced by Benjamini and Hochberg (1995) is a common criterion in the multiple testing setup. In the current context, when claiming that a gene has pattern H_k, it can be interpreted as P (a gene has pattern other than $H_k|$ the posterior probability that it has pattern $H_k > \zeta$). Simple mathematical derivation leads to the following estimate of the false discovery rate (Newton et al., 2004):

$$\widehat{FDR}_k = \frac{\sum_{g:P(H_k|\mathbf{x}_g) > \zeta}(1 - P(H_k|\mathbf{x}_g))}{\text{card}\{g : P(H_k|\mathbf{x}_g) > \zeta\}}. \tag{12}$$

Using (12), we can estimate FDR for a specific cutoff ζ, i.e. $\zeta = 0.5$. Alternatively, for a given FDR level, i.e. $FDR = 0.05$, we can also identify a cutoff ζ which leads to the most powerful list of genes with FDR controlled at the given level.

3.2. Expression profile

In order to recover the expression trajectory, we should be able to make inferences about $\mu_{gc}(t_0)$ for any time point t_0. Clearly, the estimating procedure of $\mu_{gc}(t_0)$ depends on the expression pattern. To avoid propagating the pattern identification error in estimating $\mu_{gc}(t_0)$, it is often of interest to estimate it using only the observations obtained under condition c, $\mathbf{x}_{gc}$. Let μ_{gc} be an $n_{gc} \times 1$ vector representing the true expression levels corresponding to $\mathbf{x}_{gc}$. By the definition of Gaussian process and (1), $\left(\mu_{gc}(t_0), \mathbf{x}'_{gc}\right)'$ follows a multivariate normal distribution:

$$\begin{pmatrix} \mu_{gc}(t_0) \\ \mathbf{x}_{gc} \end{pmatrix} \sim \mathcal{N}\left(\alpha_0 \mathbf{1}_{n_{gc}+1}, \begin{pmatrix} \tau^2 & s_{t_0} \\ s'_{t_0} & V + \Sigma \end{pmatrix}\right), \tag{13}$$

where $s_{t_0} = \mathrm{cov}\left(\mu_{gc}\left(t_0\right), \mu_{gc}\right)$ and $V = \mathrm{var}\left(\mu_{gc}\right)$. Therefore,

$$\mu_{gc}\left(t_0\right) | \mathbf{x}_{gc} \sim \mathscr{N}\left(\alpha_0 + s_{t_0}(V + \Sigma)^{-1}\left(\mathbf{x}_{gc} - \alpha_0 \mathbf{1}_{n_{gc}}\right), \tau^2 - s_{t_0}(V + \Sigma)^{-1} s'_{t_0}\right). \tag{14}$$

From (14), a natural estimate of $\mu_{gc}\left(t_0\right)$ is its conditional mean:

$$\widehat{\mu}_{gc}\left(t_0\right) = \alpha_0 + s_{t_0}(V + \Sigma)^{-1}\left(\mathbf{x}_{gc} - \alpha_0 \mathbf{1}_{n_{gc}}\right). \tag{15}$$

Furthermore, one could also construct the so-called Bayesian confidence interval (Wahba, 1990) for $\mu_{gc}\left(t_0\right)$ using (14). In particular, the $1 - \zeta$ Bayesian confidence interval for $\mu_{gc}\left(t_0\right)$ is

$$\left(\widehat{\mu}_{gc}\left(t_0\right) \pm z_{\zeta/2}\sqrt{\tau^2 - s_{t_0}(V + \Sigma)^{-1} s'_{t_0}}\right), \tag{16}$$

where $z_{\zeta/2}$ is the $\zeta/2$ critical value of the standard normal distribution.

3.3. Model fitting

In the discussion above, we assume that parameters such as $\sigma_0^2, \tau^2, \kappa^2, \pi_1, \ldots, \pi_K$ and α's are all known a priori. This is certainly not the case in practice. Here, we introduce an EM algorithm to estimate the parameters in an empirical Bayes fashion. The EM algorithm is based on the concept of incomplete data. In the case of our Gaussian process model, the expression patterns J can be treated as the missing data. The complete log-likelihood is then

$$\log f(x, J) = \sum_g \log\left(f_{J_g}\left(\mathbf{x}_g\right)\right). \tag{17}$$

The EM algorithm proceeds by iterating between the so-called E-step and M-step. In the E-step, we compute the expectation of the complete log-likelihood (17) conditional on the observed expressions $\mathbf{x}_g$ and the current estimate of unknown parameters. The resultant quantity is the so-called Q-function. In the M-step, the estimates of the unknown parameters are updated by maximizing the Q-function with respect to the unknown parameters.

4. Simulation

To illustrate the utility of the proposed method, we first apply the method on a simulated dataset. The data were simulated from the Gaussian process model as follows:

(i) Each gene is randomly assigned to a pattern so that 72% of the genes have pattern H_1, and 7% of the genes have each of the rest four patterns.
(ii) Depending on its expression pattern, condition-dependent profiles μ_{gc} are generated from a Gaussian process with mean function 5.27 and Gaussian covariance function with $\tau = 1.08$ and $\kappa = 4$ (hours).
(iii) Three expression measurements under each combination of three conditions and five time points, $0, 1, 3, 5, 7$ h, are generated with $\sigma_0 = 0.28$.

The parameters used here are chosen so that the simulated data share similar characteristics as the aging experiment. For each dataset, 5000 genes were generated and their differential expression patterns are determined using Bayes rule after model fitting. Table 1 summarizes the performance in identifying expression patterns averaged over 100 simulated datasets.

Fig. 1 demonstrates how to predict a gene's temporal expression profile using our Gaussian process model. The circles represent simulated expression measurements of a typical gene under a single condition. The solid black line is

Table 1
Simulation result

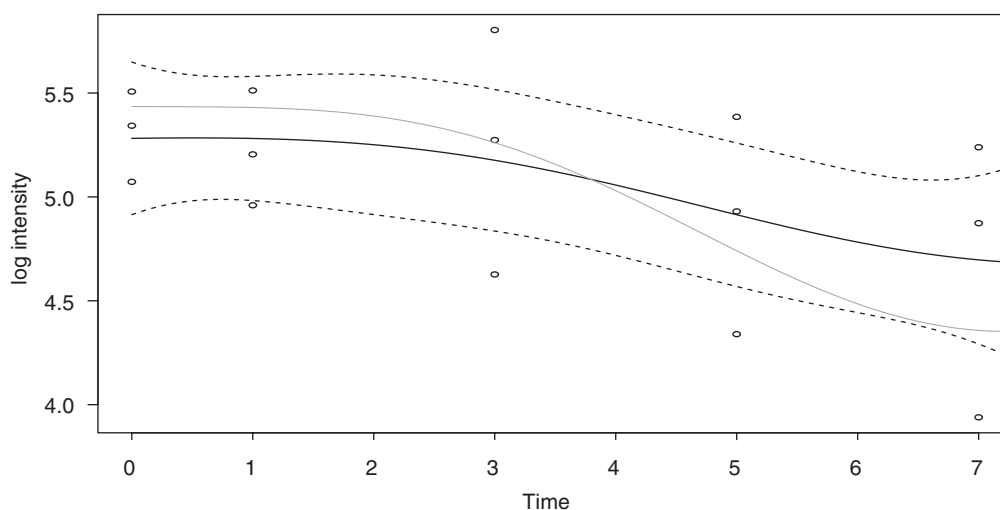| Identified pattern	Truth				
	H_1	H_2	H_3	H_4	H_5
H_1	3599.39	0.21	0.24	0.16	0.00
H_2	16.47	332.96	0.03	0.08	0.46
H_3	16.84	0.12	332.38	0.10	0.56
H_4	17.23	0.11	0.04	332.09	0.53
H_5	1.33	14.13	14.40	13.93	306.21

Fig. 1. True and estimated temporal expression profile.

the estimate obtained using (15) and the broken lines are its 99% Bayesian confidence bands as described by (16). The gray line is the true expression profile simulated.

5. Real example

The experiment reported by Edwards et al. (2003) was done to investigate the transcriptional response to oxidative stress in the heart and how it changes with age. The question is of interest for a number of reasons, a main one being evidence relating longevity with the ability to resist oxidative stress. Although it is well known that age confers varied susceptibility to various forms of stress, little is known about the genetic basis for this change. Affymetrix MG-U74A arrays were used to measure the expression levels of 12,588 genes in the heart tissue of young, middle-age, and old mice at baseline and at four times following stress induction (1, 3, 5, and 7 h). Three mice were considered for each time and age combination to give a total of 45 arrays. Following data collection, Affymetrix disclosed that approximately 20% of the genes on the MG-U74A arrays were defective. As a result, 2545 probes were removed from the analysis leaving 10,043 genes. Details of the data processing and normalization are given in Edwards et al. (2003). In short, all Affymetrix image files were processed using GeneChip Analysis Suite 5.0 software to give a Signal score for each gene. The data were normalized across arrays using the Global Scaling method implemented in that software.

Bayes rule classifies 7396 genes to H_1; 369 genes to H_2; 731 to H_3; 1467 to H_4, and 80 to H_5. Fig. 2 depicts the expression measurements and estimated expression profiles for a sample of 15 genes. The black circles, red triangles and green pluses represent the expression measurements taken for aged, middle-aged and young age group, respectively. The solid lines are the estimated expression profile and the broken lines stand for the 99% Bayesian confidence bands.
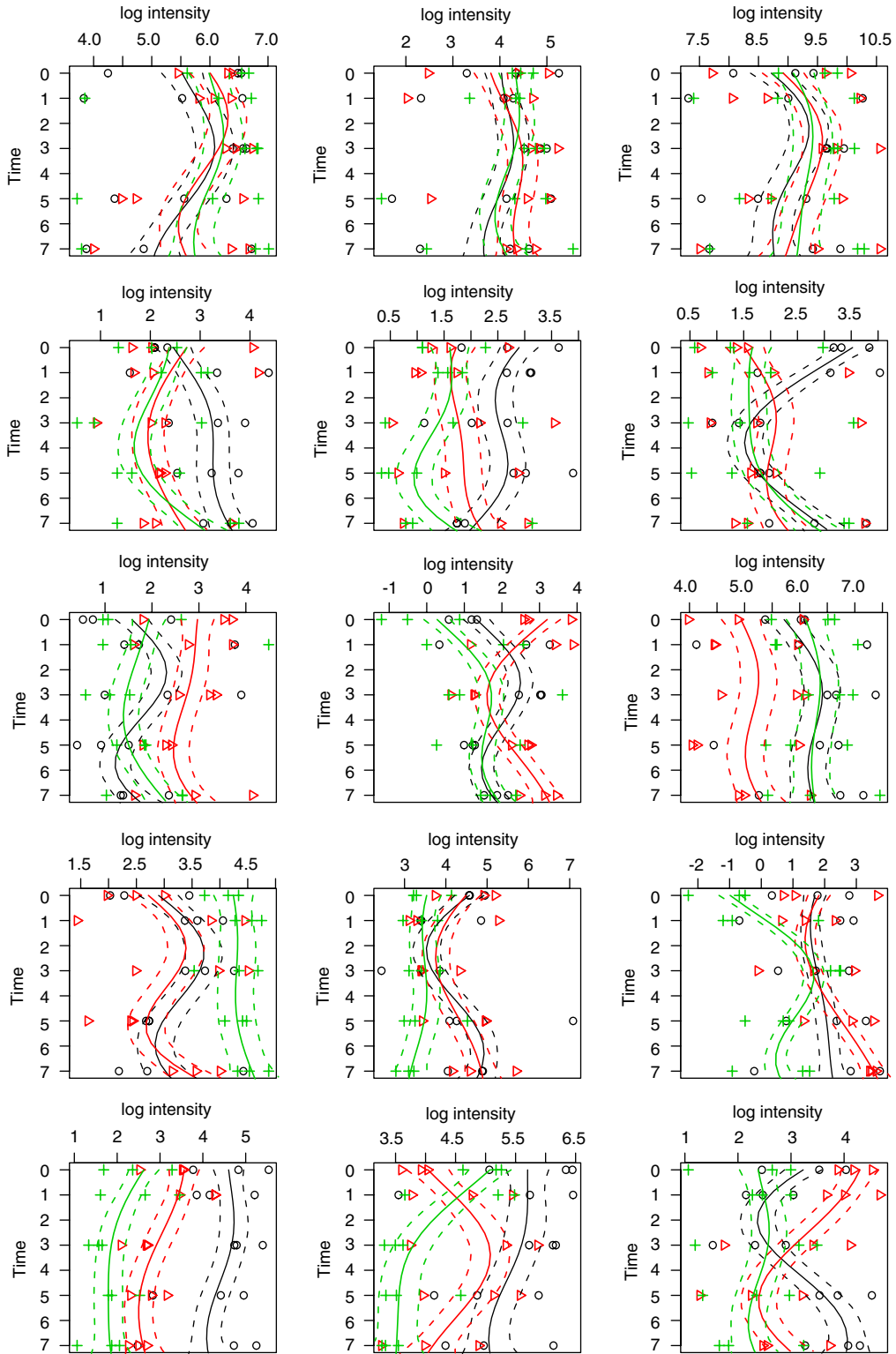
Fig. 2. Genes with different patterns.

Table 2
A sample of genes known to be related to oxidative stress

Symbol	Description	Identified pattern
Atp5l	ATP synthase, H+ transporting, mitochondrial F0 complex subunit, g	Young DE (H_4)
Atp5a1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1	Middle-aged DE (H_3)
Atp5c1	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1	Middle-aged DE (H_3)
Atp5g2	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2	Aged DE (H_2)
Atp6v1a1	ATPase, H+ transporting, V1 subunit A, isoform 1	Middle-aged DE (H_3)
Atp6v1d	ATPase, H+ transporting, V1 subunit D	Young DE (H_4)
Cox6a2	Cytochrome *c* oxidase, subunit VI a, polypeptide 2	Young DE (H_4)
Cox6c	Cytochrome *c* oxidase, subunit VIc& cytochrome-*c* oxidase activity	Middle-aged DE (H_3)
Cox7c	Cytochrome *c* oxidase, subunit VIIc	All DE (H_5)
Epx	Eosinophil peroxidase	Young DE (H_4)
Mpo	Myeloperoxidase	Middle-aged DE (H_3)
MGI:1914434	Genes associated with retinoid-IFN-induced mortality 19	Young DE (H_4)
Ndufb3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 3	Young DE (H_4)
Ndufb8	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 8	Middle-aged DE(H_3)
Tcirg1	T-cell, immune regulator 1	Middle-aged DE (H_3)
Ucp1	Uncoupling protein 1 (mitochondrial, proton carrier)	Young DE (H_4)

The three genes from the first column are identified as H_1, and as indicated by the plot, the three estimated expression profiles are very similar. The second to fourth column each has three genes classified to pattern H_2, H_3 and H_4, respectively, where one age group shows different expression profile from the other two. The fifth column corresponds to pattern H_5. These genes have three different expression profiles under different conditions. Such plot not only helps us determine a gene's expression pattern but also visualizes a gene expression trajectory under different conditions.

Table 2 lists 16 genes that are identified to be differentially expressed under at least one condition. They are all known to be responsive to oxidative stress or involved in the oxidative phosphorylation pathway.

6. Conclusion

Two of the most important tasks in time course gene expression experiments under multiple conditions are identifying temporal differential expression pattern and reconstructing the expression trajectory. Although important, no existing method can satisfactorily address both tasks. In this paper, we proposed a Gaussian process approach which can be used to identify genes' temporal differential expression patterns as well as recover the temporal expression profile. The method extends the empirical Bayes approach developed by Newton et al. (2001) and Kendziorski et al. (2003) to the comparison among multiple temporal expression profiles. The utility of the proposed method was demonstrated on both simulated and case study data.

The main objective in the analysis of time course experiments is the expression profile as a function of time. Gaussian process provides a more flexible modelling approach for functions than traditional methods. In this paper, we introduce such a powerful tool to the modelling of time course gene expression data. Gaussian process based time course experiment modelling is certainly valuable in applications beyond comparing multiple conditions as we focused on in this paper. For example, ongoing research on applying such idea in clustering time course data is encouraging.

7. Appendix—Marginal likelihood function

To compute the marginal likelihood of $\mathbf{x}_g$ under pattern H_k, it suffices to compute the marginal likelihood for $\mathbf{x}_{g S_i}$. Without loss of generality, we assume that S_i contains only one condition c in the following. If S_i contains more than one element, similar calculation can be carried out by pooling the measurements taken under conditions in S_i as if they were taken under a single "condition".

For brevity, we will also assume that $\Sigma = \sigma_0^2 I$. Derivation for more general Σ can proceed similarly. Write $\tau^2 = K(0)$, $\tau^2 V = \left(K\left(t_{ci}, t_{cj}\right) \right)$, $\mu_{gc} = \left(\mu_{gc}\left(t_{c1}\right), \ldots, \mu_{gc}\left(t_{cM}\right) \right)'$, $\mathbf{z}_{gc} = \mathbf{x}_{gc} - \alpha_0$ and $\mathbf{w}_{gc} = \mu_{gc} - \alpha_0 \mathbf{1}_M$. Under the proposed

Gaussian process model, the marginal distribution of $\mathbf{x}_{gc}$ can be derived:

$$\int f\left(\mathbf{x}_{gc}|\mu_{gc}\left(t_{c1}\right),\ldots,\mu_{gc}\left(t_{cM}\right)\right)f\left(\mu_{gc}\left(t_{c1}\right),\ldots,\mu_{gc}\left(t_{cM}\right)\right)$$

$$\times\ \mathrm{d}\mu_{gc}\left(t_{c1}\right)\ldots\mathrm{d}\mu_{gc}\left(t_{cM}\right)$$

$$=\int\left(\prod_{m=1}^{M}f\left(\mathbf{x}_{gcm}|\mu_{gc}\left(t_{cm}\right)\right)\right)f\left(\mu_{gc}\left(t_{c1}\right),\ldots,\mu_{gc}\left(t_{cM}\right)\right)$$

$$\times\ \mathrm{d}\mu_{gc}\left(t_{c1}\right)\ldots\mathrm{d}\mu_{gc}\left(t_{cM}\right)$$

$$=\int\left(1/\sqrt{2\pi}\right)^{n_{gc}}\left(1/\sigma_0^2\right)^{n_{gc}/2}\exp\left(-\frac{\sum_m\sum_r\left(z_{gcmr}-w_{gcm}\right)^2}{2\sigma_0^2}\right)$$

$$\times\left(1/\sqrt{2\pi}\right)^{M}\left(1/\tau^2\right)^{M/2}\left(\det(V)\right)^{-1/2}\exp\left(-\mathbf{w}_{gc}'V^{-1}\mathbf{w}_{gc}/2\tau^2\right)$$

$$\times\ \mathrm{d}w_{gc1}\ldots\mathrm{d}w_{gcM}$$

$$=\left(1/\sqrt{2\pi\sigma_0^2}\right)^{n_{gc}}\exp\left(-\frac{\sum_m\left(\sum_r z_{gcmr}^2-n_{gcm}\bar{z}_{gcm\cdot}^2\right)}{2\sigma_0^2}\right)\left(\sqrt{2\pi}\right)^{M}\left(\sigma_0^2\right)^{M/2}$$

$$\times\ (\det(D))^{-1/2}\int\mathcal{N}\left(\bar{z}_{gcm\cdot}|\mathbf{w}_{gc},\sigma_0^2 D^{-1}\right)\mathcal{N}\left(\mathbf{w}_{gc}|\mathbf{0},\tau^2 V\right)\mathrm{d}w_{gc1}\ldots\mathrm{d}w_{gcM}$$

$$=\left(1/\sqrt{2\pi\sigma_0^2}\right)^{n_{gc}}\exp\left(-\frac{\sum_m\left(\sum_r z_{gcmr}^2-n_{gcm}\bar{z}_{gcm\cdot}^2\right)}{2\sigma_0^2}\right)\left(\sqrt{2\pi}\right)^{M}\left(\sigma_0^2\right)^{M/2}$$

$$\times\ (\det(D))^{-1/2}\mathcal{N}\left(\bar{z}_{gcm\cdot}|\mathbf{0},\sigma_0^2 D^{-1}+\tau^2 V\right)$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^{n_{gc}}\exp\left(-\frac{\sum_m\left(\sum_r z_{gcmr}^2-n_{gcm}\bar{z}_{gcm}^2\right)}{2\sigma_0^2}\right)$$

$$\times\left(\det\left(I+\lambda D^{1/2}V D^{1/2}\right)\right)^{-1/2}$$

$$\times\exp\left(-\frac{\bar{z}_{gc\cdot}'D^{1/2}\left(I+\lambda D^{1/2}V D^{1/2}\right)^{-1}D^{1/2}\bar{z}_{gc\cdot}}{2\sigma_0^2}\right),\tag{18}$$

where $n_{gc}=\sum_{m=1}^{M}n_{gcm}$, $D=\mathrm{diag}\left(n_{gc1},\ldots,n_{gcM}\right)$, $\bar{z}_{gcm}=\sum_{r=1}^{n_{gcm}}z_{gcmr}/n_{gcm}$, $\bar{z}_{gc\cdot}=\left(\bar{z}_{gc1},\ldots,\bar{z}_{gcM}\right)'$ and $\lambda=\tau^2/\sigma_0^2$.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57, 289–300.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Chuang, Y., Chen, Y., Gadisetti, V., et al., 2002. Gene expression after treatment with hydrogen peroxide, menadione, or *t*-butyl hydroperoxide in breast cancer cells. Cancer Res. 62, 6246–6254.

Cressie, N., 1993. Statistics for Spatial Data. Wiley, London.

Edwards, M.G., Sarkar, D., Klopp, R., Morrow, J.D., Weindruch, R., Prolla, T.A., 2003. Age-related impairment of the transcriptional response to oxidative stress in the mouse heart. Physiol Genomics 13, 119–127.

Guo, X., Qi, H., Verfaillie, C., Pan, W., 2003. Statistical significance analysis of longitudinal gene expression data. Bioinformatics 19, 1628–1635.

Heagerty, P., Liang, K., Zeger, S., Diggle, P., 2004. Analysis of Longitudinal Data. second ed. Springer, New York.

Hong, F., Li, H., 2004. Function empirical Bayes methods for identify genes with different time course expression profiles. Technical Report, Center for Bioinformatics and Molecular Biostatistics, University of California.

Kendziorski, C.M., Newton, M.A., Lan, H., Gould, M.N., 2003. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statist. Medicine 22, 3899–3914.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., Tsui, K.W., 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J. Comput. Biol. 8, 37–52.

Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P., 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5, 155–176.

Park, T., Yi, S., Lee, S., Lee, S., Yoo, D., Ahn, J., Lee, Y., 2003. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 19, 694–703.

Parmigiani, G., Garrett, E.S., Irizarry, R., Zeger, S.L., 2003. The Analysis of Gene Expression Data: Methods and Software. Springer, New York.

Peddada, S., Lobenhofer, E., Li, L., Alshari, C., Weinberg, C., Umbach, D., 2003. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. Bioinformatics 19, 834–841.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Storey, J., Leek, J., Xiao, W., Dai, J., Davis, R., 2004. A significant method for time course microarray experiments applied to two human studies. Technical Report, Department of Biostatistics, University of Washington.

Wahba, G., 1990. Spline Models for Observational Data. SIAM, Philadelphia.

Xu, X., Olson, J., Zhao, L., 2002. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. Human Molecular Genetics 11, 1977–1985.

Yuan, M., Kendziorski, C., 2006. Hidden Markov models for microarray time course data in multiple biological conditions (with discussion), J. Amer. Statist. Assoc., to appear.