Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions

Ming YUAN and Christina KENDZIORSKI

Among the first microarray experiments were those measuring expression over time, and time course experiments remain common. Most methods to analyze time course data attempt to group genes sharing similar temporal profiles within a single biological condition. However, with time course data in multiple conditions, a main goal is to identify differential expression patterns over time. An intuitive approach to this problem would be to apply at each time point any of the many methods for identifying differentially expressed genes across biological conditions and then somehow combine the results of the repeated marginal analyses. But considering each time point in isolation is inefficient, because it does not use the information contained in the dependence structure of the time course data. This problem is exacerbated in microarray studies, where low sensitivity is a problematic feature of many methods. Furthermore, a gene's expression pattern over time might not be identified by simply combining results from repeated marginal analyses. We propose a hidden Markov modeling approach developed to efficiently identify differentially expressed genes in time course microarray experiments and classify genes based on their temporal expression patterns. Simulation studies demonstrate a substantial increase in sensitivity, with little increase in the false discovery rate, compared with a marginal analysis at each time point. This increase is also observed in data from a case study of the effects of aging on stress response in heart tissue, where a significantly larger number of genes are identified using the proposed approach.

KEY WORDS: Gene expression; Hidden Markov models; Microarrays; Time course.

1. INTRODUCTION

In the mid to late 1990s, advances in DNA microarray technology generated tremendous enthusiasm within the scientific community. Microarrays were referred to as "the first great hope" for providing global views of biological processes (Lander 1999) and were expected to revolutionize genomics. [The Chipping Forecast (1999) summarizes expectations at that time.] The enthusiasm was not misguided. Microarrays are now the most widely used tool in genomics to efficiently measure an organism's gene expression levels.

Among the first microarray experiments were those measuring expression over time (DeRisi, Iyer, and Brown 1997; Chu et al. 1998; Spellman et al. 1998), and time course microarray experiments remain common. In fact, they comprise more than one-third of the experiments catalogued in the Gene Expression Omnibus, the expression database maintained by the National Center for Biotechnology Information (*http://www.ncbi.nlm.nih.gov/geo/*).

A general goal common to many time course experiments is to characterize temporal patterns of gene expression within a single biological condition and group genes by these patterns. Doing so could provide insight into the biological function of genes if one assumes that genes with similar temporal patterns of expression share similar function. To accomplish these tasks, many have used unsupervised learning methods, such as hierarchical clustering (Eisen, Spellman, Brown, and Botstein 1998; Spellman et al. 1998) or k-means clustering (Tavazoie, Hughes, Campbell, Cho, and Church 1999), self-organizing maps (Tamayo et al. 1999), and singular value decomposition (Alter, Brown, and Botstein 2000; Wall, Dyck, and Brettin 2001). Cyclic patterns in particular have been identified using numerical scores based on a Fourier transform followed by correlation to known cyclic genes (Spellman et al. 1998; Whitfield et al. 2002).

Model-based approaches have also been developed. Ramoni, Sebastiani, and Kohane (2002) considered each gene's expression profile as output from an autoregressive (AR) process; genes with the highest posterior probability of being generated by the same AR process are clustered together. The total number of clusters is identified using an iterative procedure that begins with each profile in its own cluster. At each step, chosen profiles are merged into a single cluster if doing so increases a marginal likelihood function. Schliep, Schönhuth, and Steinhoff (2003) addressed similar goals. In their work, partially supervised learning is used to identify an initial set of clusters at each time point, represented by a hidden Markov model (HMM). An iterative procedure then determines the particular assignment of data to clusters that maximizes the joint likelihood of the clustering; cluster number is determined by state splitting and state deletion in HMM "model surgery." Zhao, Prentice, and Breeden (2001) introduced the application of the single-pulse model to identify genes undergoing a transcriptional response to a stimulus. Resulting estimates of the mean time of cycle activation and deactivation provide information on individual transcript profiles and can be used to assess the quality of clusters.

A second, more recent goal of many time course experiments is to collect profiles in multiple biological conditions and identify temporal patterns of differential expression. The previously described approaches consider data within one condition and thus cannot provide information on differential expression among conditions. To address this, at each time point one could apply any of the many methods for identifying differentially expressed genes across biological conditions. (For a review of these methods, see Parmigiani, Garrett, Irizarry, and Zeger 2003.) However, a consideration of each time point in isolation can be inefficient, because it does not use the information contained in the dependence structure of the time course

Ming Yuan is Assistant Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: *myuan@isye.gatech.edu*). Christina Kendziorski is Assistant Professor, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53706 (E-mail: *kendzior@biostat.wisc.edu*). Yuan was supported in part by National Science Foundation (NSF) grant DMS-00-72292. Kendziorski was supported in part by NSF grants R01-ES12752 and P30-CA14520. The authors thank Keith Baggerly, Hongzhe Li, Michael Newton, and Terry Speed for helpful comments made during the Workshop on Microarrays and Proteomics held at the Institute for Mathematics and Its Applications, Minneapolis, Minnesota, September 2003, and also the associate editor and two anonymous referees for comments that greatly improved the manuscript.

Journal of the American Statistical Association
December 2006, Vol. 101, No. 476, Applications and Case Studies
DOI 10.1198/016214505000000394

data. This problem is exacerbated in microarray studies, where low sensitivity is a problematic feature of many methods. In addition, a gene's expression pattern over time might not be identified by simply combining results from repeated marginal analyses.

The method presented here was developed to efficiently identify differentially expressed genes in time course microarray experiments and classify genes based on their temporal expression patterns. It was motivated by an experiment to investigate the transcriptional response to oxidative stress in the heart and how it changes with age (Edwards et al. 2003). 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 increased susceptibility to various forms of stress, little is known about the genetic basis for this change. The details of the experiment are given in Section 5.

In Section 2 we describe the general model and fitting procedure for analyzing time course microarray data. We consider a specific model implementation in Section 3, followed by a simulation study to illustrate and evaluate the approach. We analyze the case study in Section 5. In Section 6 we provide a discussion and outline open questions.

2. GENERAL MODEL

The general data structure and primary questions of the case study described earlier are similar to many time course microarray experiments. There are multiple time points, and for each time point there are microarray measurements from at least two biological conditions. Intensity values are background corrected and normalized to account for known sources of variation, leaving a single summary score of expression for each replicate measurement of each gene at each time in each condition. The primary goals of the study are to identify genes with different levels of expression at each time and to classify genes into temporal expression patterns. The approach discussed here is developed to accomplish these goals.

2.1 Modeling and Inference

Consider K different biological conditions and T time points. Let $\mathbf{x}_t$ be an $m \times n$ matrix of expression values for m genes probed with n arrays at time t. Clearly, $n \geq K$, and the equality holds if and only if there are no replicates within any biological condition. The heart has $K = 2$ biological conditions (young and old), $T = 5$ time points, $m = 12{,}588$ genes, and $n = 6$ arrays at each time point (3 replicates in each biological condition). The full set of observed expression values is then denoted by

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T).$$

With slight abuse of notation, let $\mathbf{x}_g$ denote one row of this matrix containing data for gene g over time, $\mathbf{x}_{gt}$ contains data for gene g at time t, and x_{gtc} consists of data for gene g at time t under condition c. Our interest lies in the relationship among the K latent mean levels of expression for each gene g at each time t denoted by $\mu_{gt1}, \mu_{gt2}, \ldots, \mu_{gtK}$.

Equality and inequality relationships among the means across conditions induce distinct expression patterns, or states. For example, if $K = 2$, as in the heart study, then there are two potential expression states for a given gene: equivalent expression ($\mu_{gt1} = \mu_{gt2}$) and differential expression ($\mu_{gt1} \neq \mu_{gt2}$).

The goal of the experiment that we are concerned with can be restated as questions about these underlying states. In short, for each gene g at each time t, we would like to estimate the probability of each state $[\pi_k(g,t) = P(s_{gt} = k)]$ for $k = 1, 2, \ldots, B_K$, and also estimate the most likely configuration of expression states over time $(s_{g1}, s_{g2}, \ldots, s_{gT})$. Note that the most likely configuration of states need not equal the collection of states that maximize $\pi_k(g,t)$ marginally at each t.

The most natural estimates for s_{gt} are the maximum a posteriori (MAP) estimates, or the estimates obtained by the Bayes rule under 0–1 loss (Berger 1985). Depending on the quantity to be estimated, the MAPs are given by

$$(\widehat{s}_{gt} : g = 1, \ldots, m) = \underset{(s_{gt} : g = 1, \ldots, m)}{\arg\max} P(s_{gt} : g = 1, \ldots, m | \mathbf{X}), \quad t = 1, \ldots, T, \quad (1)$$

and

$$(\widehat{\mathbf{s}}_{g\cdot} : g = 1, \ldots, m) = \underset{(\mathbf{s}_{g\cdot} : g = 1, \ldots, m)}{\arg\max} P(\mathbf{s}_{g\cdot} : g = 1, \ldots, m | \mathbf{X}), \quad (2)$$

where $\mathbf{s}_{g\cdot} = (s_{g1}, s_{g2}, \ldots, s_{gT})$.

To compute the MAPs, we propose a model for the set of expression measurements taken on a gene g. For a fixed time t, we consider $\mathbf{x}_{gt}$ arising from a conditional distribution

$$\mathbf{x}_{gt} | s_{gt} = i \sim f_{it}(\mathbf{x}_{gt}).$$

The time course $\mathbf{x}_g$ is then governed by two interrelated probabilistic mechanisms: the conditional distributions at each time and the process describing the evolution of states over time $(s_{g1}, \ldots, s_{gT})$. Assuming that the expression pattern (or state) process for each gene can be described by a Markov chain, that the observed expression vector can be characterized by distributions conditional on the underlying state process, and that there is conditional independence in the expression data over time, the proposed model is an HMM; an example is shown in Figure 1. Gene subscripts are dropped for convenience.

Computing the MAPs directly is difficult in the context of the HMM model, because the states are not directly observable and parameters π_0, f_{it}, and the transition matrix $\mathbf{A}$ are usually unknown. For example, consider an HMM with just two states and five time points. There are $2^5 = 32$ possible expression pattern vectors, and thus one might consider modeling the expression vectors as a mixture with 32 components. In principle, parameter estimation could be done using EM, which would require maximizing the complete data likelihood. Tremendous computing effort would be required to conduct such a maximization directly. In addition, numerical accuracy would be questionable as the number of components increased.

Fortunately, the Baum–Welch algorithm can be used to estimate $\mathbf{A}$, π_0, and f_{it} (Durbin, Eddy, Krogh, and Mitchison 1998). The Baum–Welch algorithm exploits the Markov structure of HMMs. The algorithm, a version of EM algorithm, estimates $\mathbf{A}$, π_0, and f_{it} by treating the pattern process as missing data. The algorithm iterates between the E-step and the M-step. In the E-step, given the current parameter estimates, an expectation over the missing data is taken; this is followed by an M-step to obtain a new set of estimates (see Durbin et al. 1998 for details).

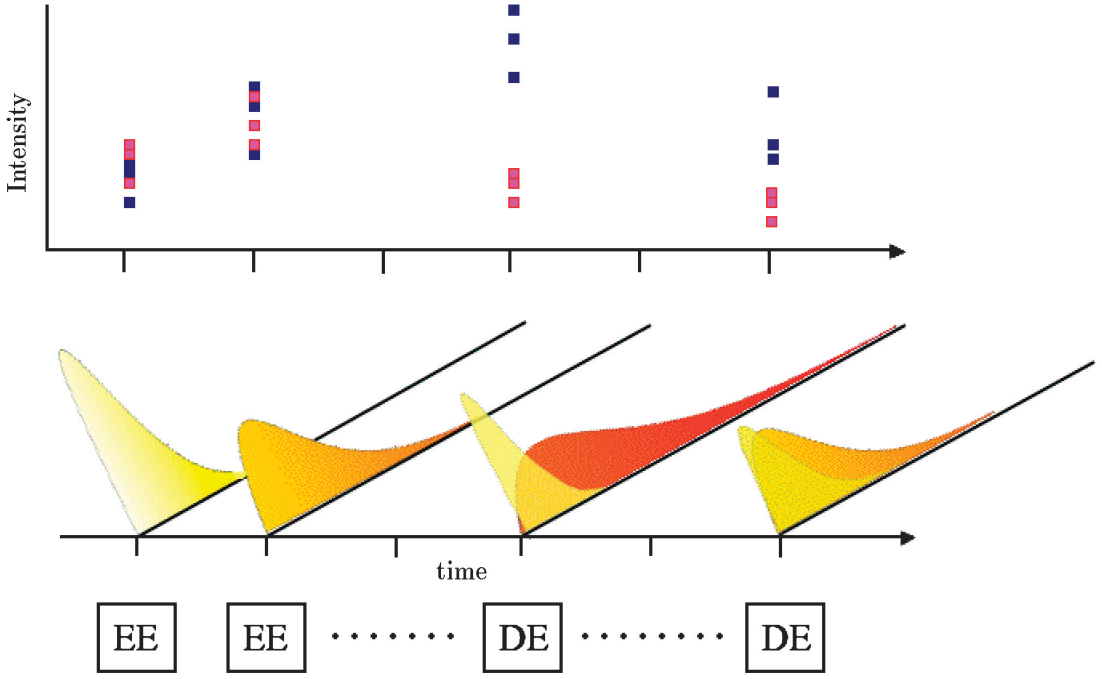

Figure 1. Expression Measurements for a Single Gene g Simulated in Two Biological Conditions (pink and blue, three replicates in each condition) Over Time. The expression patterns, or states, are also shown ($\mu_{gt1} \neq \mu_{gt2}$ and $\mu_{gt1} = \mu_{gt2}$). In practice, the states are not observed. For the HMM model, the unobserved states are expected to change over time according to a Markov process. At time t, the observed output is the expression vector $\boldsymbol{x}_{gt}$.

After obtaining parameter estimates, (1) can be evaluated. In other words, the most probable expression state for each gene at each time can be identified marginally. Because the most likely expression pattern over time might not be the collection of states that are most probable marginally, evaluation of (2) does not follow directly. To compute the most likely paths of expression states for each gene [given by (2)], the Viterbi algorithm (Durbin et al. 1998) can be used. Like the Baum–Welch algorithm, the Viterbi algorithm makes use of the Markov property of the pattern process. Details have been given by Durbin et al. (1998), among many others.

2.2 Extensions

The general HMM approach proposed earlier for expression data in two biological conditions can be extended to other types of measurements in multiple conditions. Implementation in a specific setting requires a number of decisions:

1. Data matrix **X**. In this article we focus on the cases where **X** represents the expression scores. In some applications, however, instead of working on the expression vector itself, some sort of dimension-reduction technique may serve as a preprocessing step (Efron, Tibshirani, Storey, and Tusher 2001; Pan, Lin, and Lee 2003; Allison et al. 2002). A popular choice is a summary statistic y_{gt}, such as the t-statistic or corresponding p value for $x_{gt1}, \ldots, x_{gtK}$. Under these models, the observed random process is $\{y_{gt}\}$ instead of $\mathbf{x}_{gt}$.

2. Expression patterns. For the case of multiple biological conditions, the number of states will be increased. For example, if $K = 3$, then there are five possible states:

State 1: $\mu_{gt1} = \mu_{gt2} = \mu_{gt3}$,
State 2: $\mu_{gt1} \neq \mu_{gt2} = \mu_{gt3}$,
State 3: $\mu_{gt1} = \mu_{gt2} \neq \mu_{gt3}$,
State 4: $\mu_{gt1} = \mu_{gt3} \neq \mu_{gt2}$,

and

State 5: $\mu_{gt1} \neq \mu_{gt2} \neq \mu_{gt3}$.

More generally, the number of states as a function of the number of treatments K is equal to the Bell exponential number of possible set partitions, B_K. Because B_K increases exponentially in K, prior information to narrow down the states worth investigating can be useful (see Kendziorski, Newton, Lan, and Gould 2003). There are situations in which ordered patterns might be of interest. With $K = 2$, one might consider three states: $\mu_{gt1} = \mu_{gt2}$, $\mu_{gt1} > \mu_{gt2}$, and $\mu_{gt1} < \mu_{gt2}$.

3. Homogeneous or nonhomogeneous HMM. A homogeneous HMM (h-HMM) is one in which $A(t)$ does not depend on t. h-HMMs and nonhomogeneous HMMs (nh-HMMs) are useful in different scenarios. Of course, the h-HMM is a special case of the nh-HMM. Thus, to avoid model misspecification, the nh-HMM is recommended unless there is a clear reason to do otherwise. In Section 4 an example is considered where the true data-generating mechanism is an h-HMM but an nh-HMM is specified for the analysis. For that example, there is little loss in efficiency.

4. Specification of the observational model, f. There is theoretically much flexibility in the chosen form for f. Of course, practical constraints exist related to specifying a model that describes the data well and allows for efficient inferences. We illustrate the approach described earlier using a parametric hierarchical model for f.

3. PARAMETRIC EMPIRICAL BAYES MODELS

We described the utility of HMMs applied to time course microarray experiments in multiple biological conditions in

the preceding section. To illustrate the main ideas, here we restrict our attention to a parametric empirical Bayes model introduced by Newton, Kendziorski, Richmond, Blattner, and Tsui (2001) and further developed by Kendziorski et al. (2003). The approach identifies genes differentially expressed among conditions measured at a single time point. For expositional convenience, we consider microarray time course data only in two biological conditions. The approach naturally handles data in more than two conditions.

For gene g at a given time t, $\mathbf{x}_{gt} = (x_{gt1}, \ldots, x_{gtn_1}, x_{gt(n_1+1)}, \ldots, x_{gt(n_1+n_2)})$ denotes n_1 replicated measurements under the first condition and n_2 under the second condition. As discussed before, there are two expression states for this situation. If there is equivalent expression (State 1: EE) between two conditions, then we consider $\mathbf{x}_{gt}$ as $n = n_1 + n_2$ independent samples from $f_{0t}(\cdot|\mu_{gt})$, where μ_{gt} is the common mean; μ_{gt} arises from some genome-wide distribution $G_t(\mu_{gt})$. Consequently, the marginal distribution for $\mathbf{x}_{gt}$ under EE is

$$f_{1t}(\mathbf{x}_{gt}) = \int f_{0t}(\mathbf{x}_{gt}|\mu_{gt})\, dG_t(\mu_{gt}).$$

Alternatively, if there is differential expression (State 2: DE), then $(x_{gt1}, \ldots, x_{gtn_1})$ are n_1 independent samples from $f_{0t}(\cdot|\mu_{gt1})$ and $(x_{gt(n_1+1)}, \ldots, x_{gt(n_1+n_2)})$ are n_2 independent samples from $f_{0t}(\cdot|\mu_{gt2})$, where μ_{gt1} and μ_{gt2} are also from distribution G_t. The distribution for $\mathbf{x}_{gt}$ under DE is then given by

$$f_{2t}(\mathbf{x}_{gt}) = \int f_{0t}\big(x_{gt1}, \ldots, x_{gtn_1}\big|\mu_{gt1}\big)\, dG_t(\mu_{gt1}) \times \int f_{0t}\big(x_{gt(n_1+1)}, \ldots, x_{gt(n_1+n_2)}\big|\mu_{gt2}\big)\, dG_t(\mu_{gt2}).$$

If p_t represents the proportion of DE genes at time t, then the marginal distribution of the data is given by

$$(1 - p_t) f_{1t}(\mathbf{x}_{gt}) + p_t f_{2t}(\mathbf{x}_{gt}).$$

Recall the MAP for gene g at time t: $(\widehat{s_{gt}}) = \arg\max_{(s_{gt})} P(s_{gt}|\mathbf{X})$. In this modeling framework, with just two states, a gene g at time t is classified into State 2 if $P(s_{gt} = 2|\mathbf{X})/P(s_{gt} = 1|\mathbf{X}) > 1$ (according to the Bayes rule under 0–1 loss). If the data at other time points ($\mathbf{x}_{-t}$) are not considered, then

$$\frac{P(s_{gt} = 2|\mathbf{x}_t)}{P(s_{gt} = 1|\mathbf{x}_t)} = \frac{P(s_{gt} = 2) f_{2t}(\mathbf{x}_{gt})}{P(s_{gt} = 1) f_{1t}(\mathbf{x}_{gt})}. \tag{3}$$

But if all of the data are used, then (3) becomes

$$\frac{P(s_{gt} = 2|\mathbf{X})}{P(s_{gt} = 1|\mathbf{X})} = \frac{P(s_{gt} = 2|\mathbf{x}_{-t}) f_{2t}(\mathbf{x}_{gt})}{P(s_{gt} = 1|\mathbf{x}_{-t}) f_{1t}(\mathbf{x}_{gt})}.$$

A closer look at (3) and (4) demonstrates a main advantage of the HMM approach. If $\mathbf{x}_{-t}$ does not provide information on s_{gt}, then $P(s_{gt}|\mathbf{x}_{-t}) = P(s_{gt})$. Consequently, the Markov structure in the pattern process disappears, and the data from different time points are analyzed as if they were independent. Accounting for time dependence can dramatically increase the sensitivity of the marginal inferences. To see this, consider a hypothetical example where the proportion of genes in State 2 at time t is .05 [$P(s_{gt} = 2) = .05$]. Suppose that gene g exhibits only moderate evidence of DE at time t. Then a marginal analysis [by (3)] at time t would not classify g into State 2, because to do so requires that $P(s_{gt} = 2|\mathbf{x}_t)/P(s_{gt} = 1|\mathbf{x}_t) > 1$, which implies that $f_{2t}(\mathbf{x}_{gt})/f_{1t}(\mathbf{x}_{gt})$ must be larger than 19. However, in some cases, by accounting for dependence over time, $P(s_{gt} = 2|\mathbf{x}_{-t})$ will increase. This would happen when, for example, $P(s_{gt} = 2|s_{g,t-1} = 2)$ is large and there is much evidence for gene g to be DE at time $t - 1$. For $P(s_{gt} = 2|\mathbf{x}_{-t}) >= .5$, a gene g is classified into State 2 with much less evidence marginally [$f_{2t}(\mathbf{x}_{gt})/f_{1t}(\mathbf{x}_{gt}) > 1$]. This increase in efficiency is verified numerically in Section 4.

The particular version of the general mixture model considered here is the gamma–gamma (GG) model. In the GG model, f_{0t} is assumed to be a gamma distribution with shape parameter $\alpha_t > 0$ and rate parameter $\lambda_t = \alpha_t/\mu_{gt}$, that is,

$$f_{0t}(z|\mu_{gt}) = \frac{1}{\Gamma(\alpha_t)} \lambda^{\alpha_t} z^{\alpha_t - 1} \exp(-\lambda_t z), \qquad z > 0.$$

Fixing α_t, λ_t is assumed to follow a gamma distribution with shape parameter α_{0t} and rate parameter ν_t. Thus there are three unknown parameters involved $\boldsymbol{\theta}_t = (\alpha_t, \alpha_{0t}, \nu_t)$. For the GG model, explicit forms for f_{1t} and f_{2t} exist (see Kendziorski et al. 2003).

4. SIMULATION STUDY

We carried out a simulation study to investigate the general performance of the proposed approach and to consider the potential loss in efficiency resulting from model misspecification. Datasets were simulated from an h-HMM model with six time points and two biological conditions. The GG mixture model is specified at each time by $\boldsymbol{\theta} = (10, .9, .5)$; transition probabilities are defined as $P(s_t = DE|s_{t-1} = EE) = .1$ for $t > 1$ [$P(s_1 = DE) = .1$]. A total of 100 datasets were simulated for each $k = 1, 2, 3, 4$, where $P(s_t = DE|s_{t-1} = DE) = .1 + .2 \times (k - 1)$, for a total of 400 simulated datasets; each set contains 1,500 genes.

Each simulated dataset was analyzed under three assumptions, summarized in terms of $A(t)$:

I. Independent analysis (IA). $P(s_t = DE|s_{t-1} = DE) = P(s_t = DE|s_{t-1} = EE)$ and there is no dependence over time. This is equivalent to a separate analysis at each time point using the hierarchical GG model.
II. h-HMM. $A(t)$ does not depend on t.
III. nh-HMM. $A(t)$ can depend on t.

Table 1 shows the average number of genes found by each method. When $P(s_t = DE|s_{t-1} = DE) = P(s_t = DE|s_{t-1} = EE) = .1$, there is no dependence over time; as expected, there is little difference among the results of the three methods. As $P(DE|DE)$ increases, both HMM-based methods identify more genes than the IA. In fact, the bigger the difference between $P(s_t = DE|s_{t-1} = DE)$ and $P(s_t = DE|s_{t-1} = EE)$, the greater the number of genes identified.

The increase in sensitivity does not involve a substantial increase in the false discovery rate (FDR), as shown in Figure 2. The left column of Figure 2 gives the FDR for different methods under different settings. Mostly, the difference among different methods is within 1%. Similar patterns can be observed from the specificities shown in the right column. The sensitivities plotted in the middle column, however, show a dra-

Table 1. Homogeneous HMM Simulations: The Average Number of Genes Found by Each Method (average taken over 100 simulations)

P(DE | DE)	*Method*	*Time 1*	*Time 2*	*Time 3*	*Time 4*	*Time 5*	*Time 6*
.1	I	$81.65_{(1.3)}$	$82.33_{(1.2)}$	$82.52_{(1.2)}$	$82.39_{(1.3)}$	$80.01_{(1.3)}$	$82.41_{(1.2)}$
	II	$81.69_{(1.3)}$	$82.18_{(.94)}$	$82.15_{(.95)}$	$82.29_{(.94)}$	$80.50_{(.98)}$	$81.88_{(.87)}$
	III	$81.75_{(1.3)}$	$82.43_{(1.2)}$	$82.79_{(1.2)}$	$82.74_{(1.3)}$	$80.04_{(1.3)}$	$82.50_{(1.2)}$
.3	I	$82.15_{(1.2)}$	$100.8_{(1.3)}$	$106.0_{(1.4)}$	$105.5_{(1.3)}$	$106.3_{(1.5)}$	$105.4_{(1.3)}$
	II	$83.14_{(1.2)}$	$103.2_{(1.0)}$	$108.7_{(1.0)}$	$108.5_{(1.0)}$	$109.0_{(1.2)}$	$106.4_{(1.0)}$
	III	$83.29_{(1.2)}$	$103.2_{(1.3)}$	$109.2_{(1.4)}$	$108.6_{(1.4)}$	$109.7_{(1.5)}$	$106.6_{(1.3)}$
.5	I	$84.05_{(1.4)}$	$120.7_{(1.5)}$	$134.1_{(1.6)}$	$142.6_{(1.6)}$	$144.4_{(1.6)}$	$145.3_{(1.8)}$
	II	$88.12_{(1.4)}$	$133.4_{(1.2)}$	$152.0_{(1.3)}$	$161.7_{(1.3)}$	$163.4_{(1.3)}$	$154.5_{(1.4)}$
	III	$88.27_{(1.4)}$	$133.8_{(1.5)}$	$151.4_{(1.8)}$	$162.3_{(1.7)}$	$163.1_{(1.6)}$	$154.9_{(1.8)}$
.7	I	$82.58_{(1.2)}$	$140.8_{(1.6)}$	$178.5_{(1.8)}$	$197.7_{(1.7)}$	$216.6_{(1.9)}$	$225.6_{(2.2)}$
	II	$91.68_{(1.3)}$	$170.8_{(1.4)}$	$222.9_{(1.5)}$	$252.9_{(1.5)}$	$269.5_{(1.9)}$	$262.9_{(2.0)}$
	III	$91.75_{(1.3)}$	$171.8_{(1.8)}$	$223.1_{(2.0)}$	$252.4_{(2.0)}$	$271.4_{(2.4)}$	$266.8_{(2.9)}$

NOTE: Standard errors are in parentheses.

matic increase using HMM-based methods. The increase of sensitivity can be as large as 15% depending on the transition probabilities and time points. Furthermore, although the true data-generating mechanism is an h-HMM, Figure 2 shows that there is little decrease in sensitivity when using the nh-HMM approach.

5. CASE STUDY

Of interest here is an experiment designed to better understand the genetic basis underlying the relationship between longevity and the ability to resist oxidative stress. Affymetrix MG–U74A arrays were used to measure the expression levels

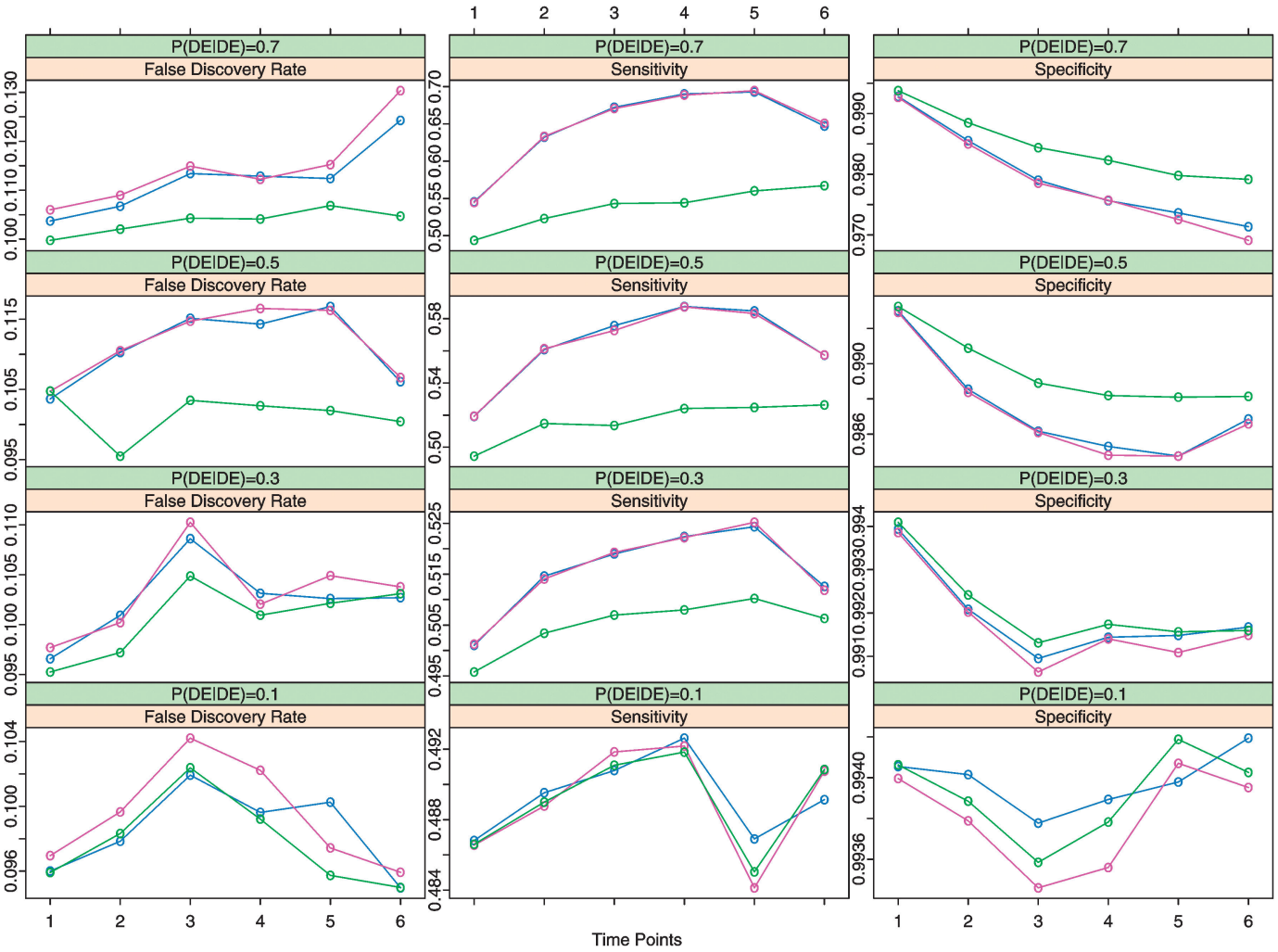

Figure 2. The Average Sensitivity, Specificity, and FDR for the IA (green), h-HMM (blue), and nh-HMM (pink) (separate; h-HMM; nh-HMM). Averages are taken over 100 simulations. The maximal increase in FDR for the HMM approach is less than 2%; most increases are less than 1%. The maximal increase in sensitivity is near 15%. Note that the scales on the y-axis are different for different panels.

of 12,588 genes in the heart tissue of young and old mice at baseline and at four times after stress induction (1, 3, 5, and 7 hours). Three mice were considered for each time and age combination, giving a total of 30 arrays. After data collection, Affymetrix disclosed that approximately 20 % of the genes on the MG–U74A arrays in the heart study were defective. As a result, 2,545 probes were removed from the analysis, leaving 10,043 genes. Details of the data processing have been given by Edwards et al. (2003). The data were normalized across arrays using robust multi-array analysis (RMA) (Irizarry et al. 2003). The dataset was analyzed via the nh-HMM model. All calculations were carried out in R 1.9.1 (R Development Core Team 2004). Expression paths were assessed via the Viterbi algorithm. An analysis assuming no dependence over time (IA) was also done for comparison. The numbers of genes identified with each method for the case study are presented in Table 2.

If there is no strong temporal dependence, then one would expect the sets of DE genes identified by IA at different time points to be quite different. This is certainly not the case here, because most of the genes (732 out of 835) found to be DE at Time 2 are also found to be DE at Time 1. Similar phenomena are observed at the other time points. These observations indicate that compared with an EE gene, a DE gene is more likely to be DE at the next time point.

Table 2. Patterns Identified by Each Method

State	Method	Time 1	Time 2	Time 3	Time 4	Time 5
1: EE	IA	8,023	9,208	9,238	9,415	9,293
	nh-HMM	8,050	8,796	8,829	8,910	8,889
	Both methods	7,869	8,766	8,793	8,894	8,837
2: DE	IA	2,020	835	805	628	750
	nh-HMM	1,993	1,247	1,214	1,133	1,154
	Both methods	1,839	805	769	612	698

The nh-HMM often results in a dramatic increase in the number of genes showing some DE. The example discussed in Section 3 suggests that this is due to the ability of the nh-HMM to identify genes that are consistently DE over time, even if there is little evidence of DE at any given time point.

Figure 3 demonstrates that this is the case. There were 11 genes identified as EE by the IA at each time, but as DE by nh-HMM. Figure 3 shows the expression vectors for these genes. As shown, there is little evidence for DE marginally, but there is consistent evidence over time. In terms of fold change, the nh-HMM approach is finding genes with an average fold

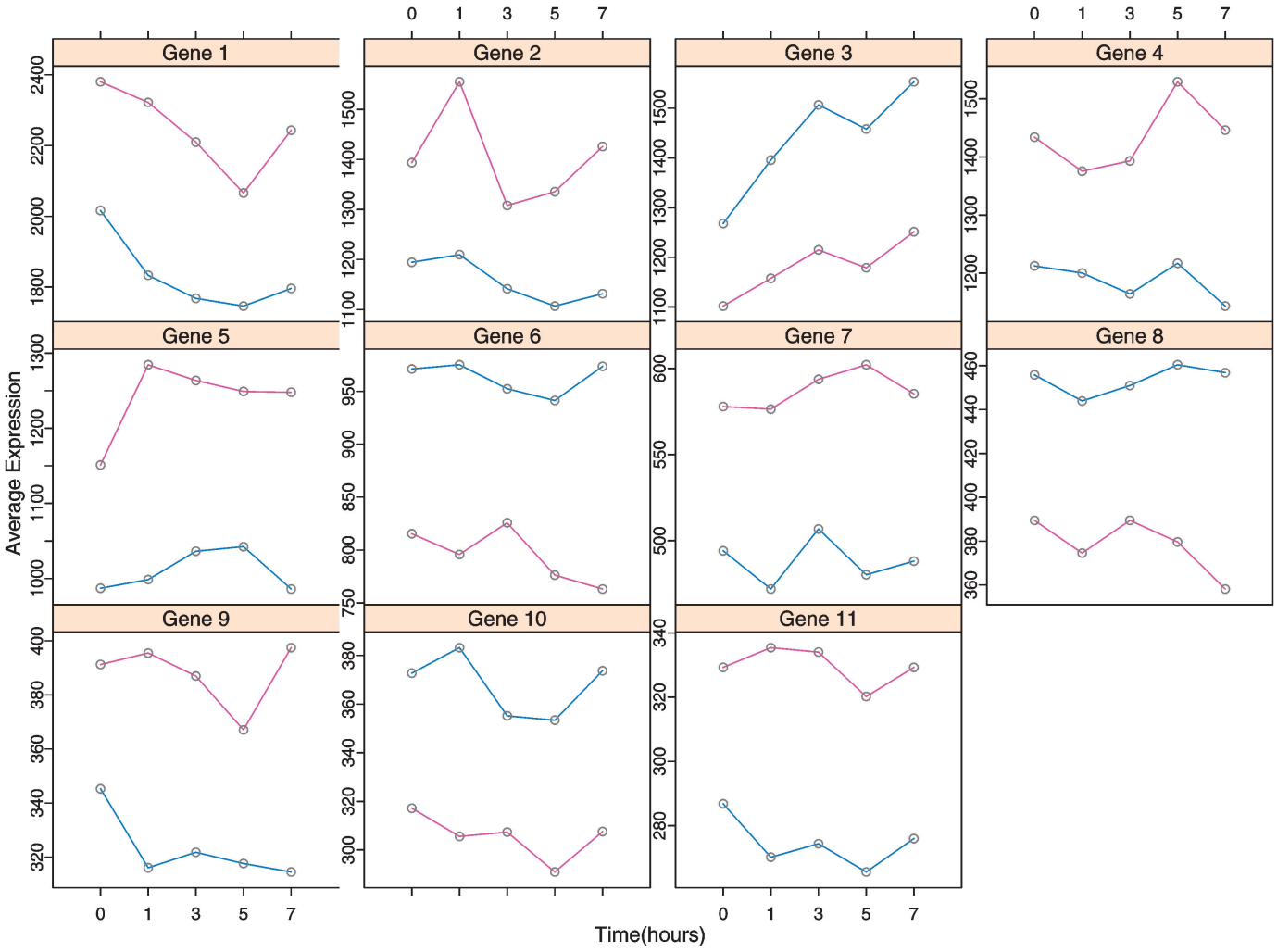

Figure 3. Eleven Genes Identified as EE at All Times via the IA and as DE at All Times via the nh-HMM Approach. The blue lines correspond to the older group; the pink lines to the younger group. Note that the scales on the y-axis are different for different panels.

change difference of .46; marginal analyses at each time are not sensitive enough to identify changes of this magnitude.

Figure 3 was generated as follows:

1. The intensity level data were processed to give one summary score of expression for each gene at each time point in each biological condition. Normalization was done using RMA (Irizarry et al. 2003).
2. A hierarchical GG mixture model was used to describe the data at each time point. HMM assumptions as described in Section 2 were considered appropriate for these data. The GG model assumptions were checked using diagnostics described by Newton and Kendziorski (2003).
3. The Baum–Welch algorithm was used to obtain parameter estimates.
4. For each gene at each time, posterior probabilities of the two possible states were calculated under the IA and nh-HMM models.
5. A total of 11 genes were identified as State 1 (EE) via IA but as State 2 (DE) via nh-HMM for every time point.
6. The expression vectors for these genes were averaged over replicates, and the averages were plotted at each time.

In addition to classifying genes into states, the posterior probabilities can be used to identify particular expression patterns over time. For example, investigators in this study are also interested in identifying genes showing equivalent expression at the earlier time points but differential expression later in the experiment. Viterbi paths corresponding to this pattern were identified; the expression profiles for these 25 genes are shown in Figure 4.

6. DISCUSSION

Microarray experiments that collect expression profiles over time in multiple biological conditions are becoming increasingly common. Many methods to date analyze time course data within conditions and attempt to cluster genes with similar profiles over time. As a result, these methods do not apply to the problems of identifying genes DE over time and classifying genes based on their DE patterns. To do this, one could apply at each time point any of the methods for identifying DE genes across multiple conditions and combine results across time after the marginal analyses. As we have shown here, this approach is not efficient. An alternative approach would be to slightly modify the ANOVA methods for microarrays proposed by Kerr and Churchill (2001) or Wolfinger et al. (2001) and identify genes with significant condition-by-time interactions.

One might expect such an analysis to suffer from low power because of few replicates and stringent adjustments for multiple tests. Park, Yi, and Lee (2003) showed that this is the case in a study of rat cortical stem cells in two biological conditions over time. They performed an ANOVA on 3,840 genes and found that none had a significant condition-by-time interaction. To address this, these authors proposed a two-stage approach in which the first stage removes the effect of time and the second stage is used to identify DE genes. In particular, for their dataset, after initial identification of no significant genes based on interaction coefficients, they fit a reduced ANOVA model with the interaction term removed; p values for the condition effects were then calculated in two ways. The first way, which identified 53 genes with significant group effects at the 5% level, was to assume normality of the test statistics and use a Bonferroni correction; the second way involved obtaining residuals from a model with group effect only, calculating t-statistics after permutations of the residuals, and using the t-statistics to determine adjusted p values by the method of Westfall and Young (1993). This second approach identified 90 genes at a 5% significance level. To obtain some idea about each gene's temporal expression profile, the 53 genes identified after a Bonferroni correction were then clustered using K-means.

Although a standard ANOVA approach is intuitive, there are a number of questions that it does not address. Time dependence is not considered explicitly (i.e., identical results would be obtained if the columns were reordered); there is no information indicating which time points contribute most to a gene being identified as DE across conditions; and the cluster analysis provides no quantitative information on temporal patterns of differential expression.

The HMM approach presented here addresses these questions directly. In particular, the unobserved expression patterns over time are assumed to follow a Markov process, with intensity values taken from some distribution conditional on the expression pattern state. The posterior probability of each expression pattern (DE or EE for two conditions; multiple patterns for more than two conditions) is reported at each time for every gene. These posterior probabilities, specific to gene and time, prove very useful for identifying genes that are in a particular pattern at each time. The Viterbi algorithm is used to identify the most likely temporal expression path, and a posterior probability associated with each path is reported. As we have shown, this posterior probability can be useful in organizing genes into groups and provides a quantitative way to evaluate a gene's membership in any given group. Another strength of the proposed approach is its ability to handle both EE and DE. In practice, often a gene is classified as EE if it fails some test of DE. This is not correct, of course, because lack of evidence for DE does not necessarily imply EE. For this HMM approach, the posterior probability of EE can be used to better quantify the uncertainty in classifying a gene as EE.

A comparison with marginal analyses repeated each time has shown that the HMM approach substantially increases the number of genes identified as DE. Simulations suggest that this increase is due almost completely to an increase in sensitivity, because there is very little change in the FDR. In the marginal analyses, the reported FDRs are near 10%, whereas in the HMM approach, the FDRs are around 11–12%. (Recall that the Bayes rule was used to classify a gene as DE; control of FDR was not targeted.) If desired, adjusting the threshold to target a specific FDR can be done (Genovese and Wasserman 2003; Storey and Tibshirani 2003; Newton, Noueiry, Sarkar, and Ahlquist 2004). For example, the expected posterior FDR associated with a list of size N at time t is $(1/N)\sum_{l=1}^{N} 1 - P(s_{lt} = DE|\mathbf{X})$. One could simply choose the largest number of genes for which the FDR is below some prespecified level. When error rates other than FDR are of interest, thresholds can be determined by considering appropriate loss functions that quantify an investigator's tolerance for both false positives and false negatives.

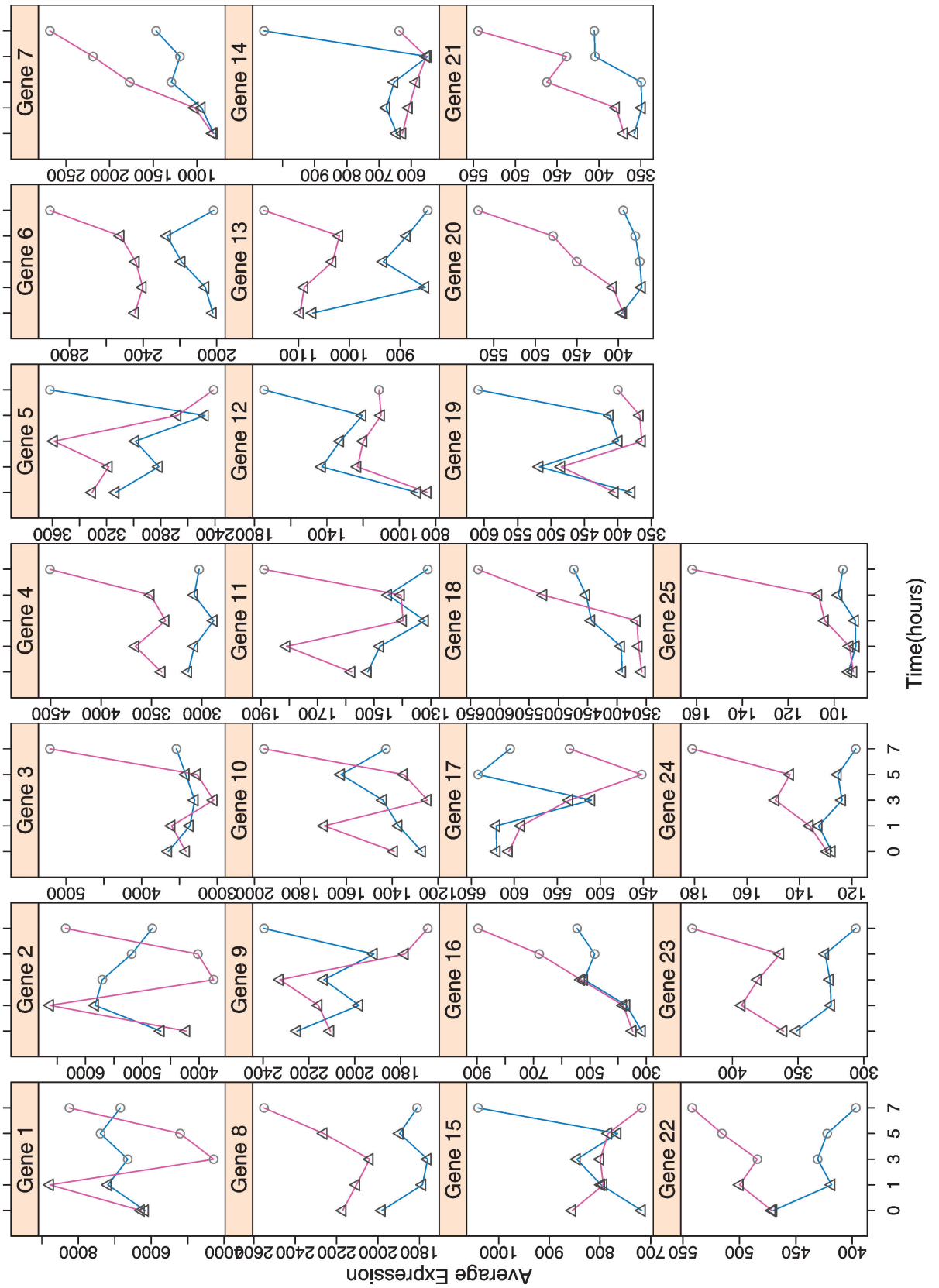

Figure 4. A Total of 25 Genes Identified as EE at Earlier Time Points and as DE at Later Time Points via the nh-HMM Approach. The blue lines correspond to the older group; the pink lines, to the younger group. Triangles indicate the time points at which the genes are identified as EE; circles indicate the time points at which the genes are identified as DE. Note that the scales on the y-axis are different for different panels.

Table 3. Simulation With Gene-Specific Transition Probabilities: The Average Number of Genes Found, the FDR, and the Sensitivity and Specificity of Each Method (average taken over 100 simulations)

P(DE|DE)	Method	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Number of DE genes	I	$79.45_{(1.4)}$	$138.4_{(1.1)}$	$183.5_{(2.2)}$	$207.0_{(1.2)}$	$219.9_{(1.7)}$	$225.9_{(1.5)}$
	II	$89.05_{(1.4)}$	$158.7_{(1.0)}$	$215.1_{(1.6)}$	$244.4_{(1.2)}$	$257.3_{(1.2)}$	$265.3_{(1.7)}$
	III	$89.10_{(1.4)}$	$158.2_{(1.3)}$	$217.2_{(2.3)}$	$245.2_{(1.5)}$	$258.5_{(1.4)}$	$261.9_{(2.2)}$
FDR	I	.1050	.0942	.1125	.0996	.1057	.1098
	II	.1148	.0922	.0984	.0884	.0865	.1283
	III	.1163	.0943	.1045	.0911	.0891	.1237
Sensitivity	I	.4711	.5234	.5459	.5597	.5589	.5625
	II	.5231	.6016	.6524	.6694	.6684	.6467
	III	.5225	.5979	.6531	.6692	.6695	.6410
Specificity	I	.9936	.9897	.9823	.9823	.9796	.9783
	II	.9923	.9884	.9823	.9814	.9805	.9701
	III	.9921	.9881	.9807	.9808	.9798	.9713

NOTE: Standard errors (SEs) are shown in parentheses for number of DE genes; for FDR, sensitivity, and specificity, SEs $< .003, .006, .0007$.

The approach can be extended to account for different orders, different parametric assumptions, and more biologically relevant transition matrices. Here we have here considered Markov chains of order 1 because for this case study, there are relatively few time points, and HMMs of higher order were not necessary. In some applications, first-order HMMs might not be sufficient and techniques presented by Durbin et al. (1998) may prove useful. For illustration purposes, we have also restricted our attention to the GG model. Model diagnostics have been discussed in detail by Newton and Kendziorski (2003), and we recommend checking the parametric assumptions on a case-by-case basis.

An extension that we have not yet considered extensively is to allow distinct probability transition matrices for individual genes or clusters of genes. This would allow one to account for processes evolving at different time scales and also possibly allow for the incorporation of gene groups known to have similar expression patterns. Under consideration are possible approaches for identifying such groups of genes and incorporating this into our analyses. One possibility is to cluster genes and assume that genes in each cluster follow the same transition matrix. To investigate how the method proposed here performs if $A(t)$ does vary across genes, we simulated 100 datasets in a similar fashion as described in Section 4. Instead of fixing the transition matrix for all genes, we simulated gene-specific transition probabilities from a beta distribution with shape parameters 7 and 3 such that the mean is .7 if a gene is differentially expressed at the previous time point and beta$(1, 9)$ otherwise. Table 3 reports the number of identified differentially expressed genes, the FDR, and the specificity and sensitivity averaged over 100 datasets. The results are similar to those given in Table 1. The proposed approach continues to show a substantial increase in sensitivity with very little increase in the FDR. Further work in this area is underway.

The proposed HMM approach should prove useful in a number of studies collecting gene expression profiles in multiple biological conditions over time. We have illustrated the approach using a specific parametric model with assumptions that can be checked. However, because the general approach makes few assumptions, there is much flexibility regarding alternative models that could be considered within this HMM framework.

[Received December 2003. Revised November 2004.]

REFERENCES

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Kee, C., Prolla, T. A., and Weindruch, R. (2002), "A Mixture Model Approach for the Analysis of Microarray Gene Expression Data," *Computational Statistics & Data Analysis*, 39, 1–20.

Alter, O., Brown, P., and Botstein, D. (2000), "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proceedings of the National Academy of Sciences*, 97, 10101–10106.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Chipping Forecast (1999), *Nature Genetics Supplement*, 21, 1–60.

Chu, S., DeRisi, J. L., Eisen, M., Mullholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998), "The Transcriptional Program of Sporulation in Budding Yeast," *Science*, 282, 699–705.

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.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, London, U.K.: Cambridge University Press.

Edwards, M. G., Sarkar, D., Klopp, R., Morrow, J. D., Weindruch, R., and Prolla, T. A. (2003), "Age-Related Impairment of the Transcriptional Response to Oxidative Stress in the Mouse Heart," *Physiological Genomics*, 13, 119–127.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 456, 1151–1160.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868.

Genovese, C., and Wasserman, L. (2003), "Bayesian and Frequentist Multiple Testing," in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, eds. M. J. Bayarri, A. P. Dawid, J. O. Berger, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, pp. 145–162.

Irizarry, R., Hobbs, B., Collins, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), "Exploration, Normalization, and Summaries of High-Density Oligonucleotide Array Probe Level Data," *Biostatistics*, 4, 249–264.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003), "On Parametric Empirical Bayes Methods for Comparing Multiple Groups Using Replicated Gene Expression Profiles," *Statistics in Medicine*, 22, 3899–3914.

Kerr, M. K., and Churchill, G. A. (2001), "Statistical Design and Analysis of Gene Expression Microarray Data," *Genetical Research*, 77, 123–128.

Lander, E. S. (1999), "Array of Hope," *Nature Genetics Supplement*, 21, 3–4.

Newton, M. A., and Kendziorski, C. M. (2003), "Parametric Empirical Bayes Methods for Microarrays," in *The Analysis of Gene Expression Data: Methods and Software*, eds. G. Parmigiani, E. S. Garrett, R. Irizarry, and S. L. Zeger, New York: Springer-Verlag, pp. 254–271.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression With a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176.

Pan, W., Lin, J., and Lee, C. T. (2003), "A Mixture Model Approach to Detecting Differentially Expressed Genes With Microarray Data," *Functional & Integrative Genomics*, 3, 117–124.
Park, T., Yi, S. G., and Lee, S. (2003), "Statistical Tests for Identifying Differentially Expressed Genes in Time-Course Microarray Experiments," *Bioinformatics*, 19, 694–703.
Parmigiani, G., Garrett, E. S., Irizarry, R., and Zeger, S. L. (eds.) (2003), *The Analysis of Gene Expression Data: Methods and Software*, New York: Springer-Verlag.
R Development Core Team (2004), "R: *A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.
Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002), "Cluster Analysis of Gene Expression Dynamics," *Proceedings of the National Academy of Sciences*, 99, 9121–9126.
Schliep, A., Schönhuth, A., and Steinhoff, C. (2003), "Using Hidden Markov Models to Analyze Gene Expression Time Course Data," *Bioinformatics*, 19, 255–263.
Spellman, P. T., Sherlock, G., Zhang, M., 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," *Molecular Biology of the Cell*, 9, 3273–3297.
Storey, J., and Tibshirani, R. (2003), "Statistical Significance for Genome-Wide Studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.
Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999), "Interpreting Patterns of Gene Expression With Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences*, 96, 2907–2912.
Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999), "Systematic Determination of Genetic Network Architecture," *Nature Genetics*, 22, 281–285.
Wall, M. E., Dyck, P. A., and Brettin, T. S. (2001), "Singular Value Decomposition Analysis of Microarray Data," *Bioinformatics*, 17, 566–568.
Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., and Botstein, D. (2002), "Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors," *Molecular Biology of the Cell*, 13, 1977–2000.
Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing*, New York: Wiley.
Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. (2001), "Assessing Gene Significance From cDNA Microarray Expression Data via Mixed Models," *Journal of Computational Biology*, 8, 625–637.
Zhao, L. P., Prentice, R., and Breeden, L. (2001), "Statistical Modeling of Large Microarray Data Sets to Identify Stimulus-Response Profiles," *Proceedings of the National Academy of Sciences*, 98, 5631–5636.

Discussion

Hongzhe Li and Fangxin Hong

We would like to congratulate Yuan and Kendzioski (YK for short) on their interesting work on using the hidden Markov model (HMM) to analyze microarray time course (MTC) gene expression data. Because many important biological systems or processes are dynamic systems, it is important to study the gene expression patterns over time in a genomic scale to capture the dynamic behavior of gene expression. DNA microarray technologies make it possible to monitor changes in gene expression levels over time during these biological processes. There are many interesting statistical problems related to the analysis of MTC gene expression data, including identification of genes with certain expression patterns over time, identification of periodically regulated genes, clustering of MTC gene expression data, and investigation of genetic networks using MTC data. YK considered the important problem of identifying genes that are temporally differentially expressed (TDE) between two or more MTC experiments. They clearly demonstrated that such MTC studies can potentially identify more genes that are differentially expressed than considering gene expression levels at one single time point. Their method therefore has many potential practical applications.

Our comments focus on the dependency structures of MTC gene expression data and the assumptions made by YK in developing their HMM model. We also provide an alternative functional hierarchical model for identifying TDE genes.

1. DEPENDENCY STRUCTURE OF MTC GENE EXPRESSION DATA

What makes MTC gene expression data unique is the dependency structure of the gene expression data measured over time. Different study designs can induce different dependency structures. The real data examples considered in YK are all from cross-sectional designs, where gene expression levels are measured from cells of different subjects. For such a design, a simple model for the log gene expression measurement y_{jikt} can be written as

$$\begin{aligned} y_{jikt} &= f_{ji}(t) + \epsilon_{jikt}, \\ & j = 1, \ldots, n \text{ genes}, \\ & i = 1, 2 \text{ groups}, \\ & k = 1, \ldots, K \text{ replications}, \\ & t = t_1, \ldots, t_T \text{ time points}, \end{aligned} \tag{1}$$

where $f_{ji}(t)$ is the true gene expression level at time t and ϵ_{jikt} is the noise. In this model, the dependency of the gene expression measurements over time is modeled by the gene- and experiment-specific mean function $f_{ji}(t)$. The error terms ϵ_{jikt} are usually assumed to be independent. However, for cDNA microarray data from reference designs, the error terms might be dependent because the data are all measured relative to a pool of common mRNAs. In addition, the error variances can increase as time elapses; this is especially likely when cells are initially synchronized.

Hongzhe Li is Professor of Biostatistics, Department of Biostatistics and Clinical Epidemiology, University of Pennsylvania School of Medicine, 423 Guardian Drive, 920 Blockley Hall, Philadelphia, PA 19104-6021 (E-mail: *hli@cceb.upenn.edu*). His research is supported by National Institutes of Health grant R01 ES009911. Fangxin Hong is Bioinformatics Scientist, Salk Institute, La Jolla, CA 92037.

DOI 10.1198/016214506000000915