PROFILING TIME COURSE EXPRESSION OF VIRUS GENES—AN ILLUSTRATION OF BAYESIAN INFERENCE UNDER SHAPE RESTRICTIONS[1]

BY LI-CHU CHIEN[2], I-SHOU CHANG[2], SHIH SHENG JIANG[2], PRAMOD K. GUPTA, CHI-CHUNG WEN, YUH-JENN WU AND CHAO A. HSIUNG

National Health Research Institutes, National Health Research Institutes, National Health Research Institutes, National Health Research Institutes, Tamkang University, Chung Yuan Christian University and National Health Research Institutes

There have been several studies of the genome-wide temporal transcriptional program of viruses, based on microarray experiments, which are generally useful in the construction of gene regulation network. It seems that biological interpretations in these studies are directly based on the normalized data and some crude statistics, which provide rough estimates of limited features of the profile and may incur biases. This paper introduces a hierarchical Bayesian shape restricted regression method for making inference on the time course expression of virus genes. Estimates of many salient features of the expression profile like onset time, inflection point, maximum value, time to maximum value, area under curve, etc. can be obtained immediately by this method. Applying this method to a baculovirus microarray time course expression data set, we indicate that many biological questions can be formulated quantitatively and we are able to offer insights into the baculovirus biology.

1. Introduction.

1.1. *Transcription program of virus.* With a custom made baculovirus DNA microarray, Jiang et al. (2006) investigated the temporal transcription program of one of the best characterized baculoviruses, Ac*M*NPV, in its host lepidopteran Sf21 cells. They uncovered sequential viral gene expression patterns, which are possibly regulated by different mechanisms during different phases of infection, compared the transcription profile of a mutant virus with that of the wild type, and suggested that the array strategy taken in the study points to a very productive direction for constructing a baculovirus gene regulation network.

The experiments of Jiang et al. (2006) are briefly summarized as follows. They use single color cDNA microarray experiments with external controls for data normalization. Each chip has exactly four spots for each of the 156 open reading frames, referred to as genes henceforth, of baculovirus; total RNA samples of baculovirus genes were taken at several different time points during the 72 hours following infection; the sample for each time point is hybridized to a single chip. The normalized time course expression data are shown to be in good agreement with those obtained by the real-time PCR method for five randomly chosen genes; the data for each gene used in the study of temporal transcription is based solely on the normalized expression levels at these time points and on the crude estimates of its onset time and the time that its expression attains its maximum.

A rough idea regarding virus gene expression is that genes of a virus have their time course expression level being zero initially, then increasing after a while and finally decreasing; because viruses do not have their own machinery for gene transcription, their genes start to express only after getting into cells, and cells may eventually malfunction when infected. It is of interest and feasible to make use of this idea to profile the time course expression of each virus gene, based on microarray data, to estimate salient features of the profile like onset time, time to maximum value, maximum value, area under the profile curve, etc. and to test the shape hypotheses on the profile curve like unimodality on certain time intervals. It is hoped that this approach to gene expression analysis of viruses would eventually provide a sound basis for the study of the temporal transcription program of viruses.

The purpose of this paper is to propose a Bayesian shape restricted regression model based on the above property of a virus, illustrate this model by profiling the time course expression of genes of baculovirus, and indicate that this approach does provide more insights into baculovirus, compared with the crude statistics used in Jiang et al. (2006). Among others, a prominent example in this regard is that this new approach seems to support the widely accepted conjecture that structural genes of the virus may have a larger amount of total expression level, which is hard to examine by the method in Jiang et al. (2006).

This method is illustrated on the dataset for the baculovirus Bac-PH-EGFP in Jiang et al. (2006). With 16 time points, this dataset seems to hold a promising opportunity to capture the main features of the transcription profile. We note that the other two datasets in Jiang et al. (2006) have only 6 time points and 5 of them are in the initial two hours post infection and it is hard to infer some of the main features of the profile based on them.

Because microarray experiments offer feasible approaches to the studies of the genome-wide temporal transcriptional program of viruses, which are generally useful in the construction of gene regulation network, there have been many genome-wide expression studies of virus genes. See, for example, Yang et al. (2002), Iwanaga et al. (2004), Duplessis et al. (2005), van Munster et al. (2006), Majtan et al. (2007), Smith (2007) and references therein; they considered different

viruses and/or different host cells. It seems that all the biological interpretations in these studies are directly based on the normalized data and crude statistics, which seem to provide only naive estimates of limited features of the profile, and there are some discrepancies reported in the literature; see, for example, Smith (2007). It is of great interests to compare the transcriptional studies based on different but related strains of viruses and/or different and related host cells so as to build a gene regulation network. We note that comprehensive comparisons depend on comprehensive and rigorous time course expression profiling of genes in each study. The focus of this paper is the latter.

1.2. *Statistical modeling strategy.* Preliminary examination of the Bac-PH-EGFP data suggests that two of the 156 genes seem to have their expression levels being zero finally as well as initially and the rest of the 154 genes being zero only initially, probably because no data were taken at time point beyond 72 hours and the life cycle of baculovirus is longer than 72 hours, according to Friesen and Miller (2001). To make the presentation concise, we limit our attention to these 154 genes in this paper; the other two genes can be studied similarly.

Let $\mathcal{A}$ denote the set of all smooth functions on [0, 1] that are zero initially, start to increase after a while, and stay positive onward. The task of profiling the time course expression level of virus genes will be considered a shape restricted regression problem with the regression function belonging to $\mathcal{A}$. Let $g = 1, 2, \ldots, 154$ index the 154 genes of the baculovirus. For $g = 1, \ldots, 154$, we assume that, given F_g in $\mathcal{A}$,

$$(1.1) \qquad\qquad Y_{jkg} = F_g(X_k) + \varepsilon_{jkg}.$$

Here $\{X_k \mid k = 0, \ldots, K\}$ are constant design points in [0, 1], $\{Y_{jkg} \mid j = 1, \ldots, m_k, k = 0, \ldots, K, g = 1, \ldots, 154\}$ are response variables, and for every $j = 1, \ldots, m_k, k = 0, \ldots, K, g = 1, \ldots, 154$, ε_{jkg} are independent normal errors with mean μ_g and variance

$$(1.2) \qquad\qquad \sigma_{kg}^2 = \sigma_g^2 \big(F_g(X_k) + \mu_g\big)^{\xi_g}$$

for some $\xi_g = 0, 1$ or 2.

In this paper X_k represents a time point at which the mRNA sample is taken for microarray experiments; Y_{jkg} is the expression level, in terms of fluorescent intensity, obtained at the jth spot of the gth gene for the sample taken at time point X_k. More specifically, in our data, let [0, 1] denote the time period of 72 hours, then $K = 15$, $m_k = 4$, $(X_0, X_1, \ldots, X_{15}) = (0, 1/216, 1/108, 1/72, 1/36, 1/24, 1/12, 1/8, 1/6, 5/24, 1/4, 1/3, 5/12, 2/3, 5/6, 1)$.

The variance structure in (1.2) is a simple way to take into consideration the observation that for single color cDNA microarray experiments, larger intensities often incur larger variances when considering replicates. The reason for not assuming ε_{jkg} having zero mean is that there are always background intensities due

to nonspecific hybridization and, hence, $E(Y_{jkg})$ may not be zero even when the expression level $F_g(X_k)$ is zero.

We now explain that Bernstein polynomials can be used to study the above shape restricted regression model. For integers $0 \leq i \leq n$, let $\varphi_{i,n}(t) = C_i^n t^i (1 - t)^{n-i}$, where $C_i^n = n!/(i!(n-i)!)$. The set $\{\varphi_{i,n} \mid i = 0, \ldots, n\}$ is called the Bernstein basis for polynomials of order up to n. Let $\mathcal{B} = [0, 1] \times \bigcup_{n=3}^{\infty} (\{n\} \times \mathbb{R}^{n-1})$. Define $\mathbf{F} : \mathcal{B} \times [0, 1] \longrightarrow \mathbb{R}^1$ by

$$(1.3) \qquad \mathbf{F}(c, n, b_{2,n}, \ldots, b_{n,n}; t) = \sum_{i=2}^{n} b_{i,n} \varphi_{i,n} \left(\frac{t - c}{1 - c} \right) I_{(c,1]}(t),$$

where $(c, n, b_{2,n}, \ldots, b_{n,n}) \in \mathcal{B}$ and $t \in [0, 1]$. We also denote (1.3) by $F_{c,b_n}(t)$ if $b_n = (b_{2,n}, \ldots, b_{n,n})$. We will see in Section 2 that $F_{c,b_n}(\cdot)$ is a member of $\mathcal{A}$ if $0 \leq \min_{l=2,\ldots,n} b_{l,n} < \max_{l=2,\ldots,n} b_{l,n}$, and every member of $\mathcal{A}$ can be approximated by $F_{c,b_n}(\cdot)$ satisfying these restrictions on b_n. This observation suggests that, by means of (1.3), Bernstein polynomials form a useful tool to introduce priors on $\mathcal{A}$ for a Bayesian analysis.

We will consider Bayesian hierarchical models based on (1.3). With priors on a space of smooth functions satisfying certain shape restrictions and parameters in the priors based on crude estimates from data, our approach has the advantage of utilizing prior knowledge from biology; with 154 correlated and possibly similar profiles to study, hierarchical regression models take advantage of the possibility of data driven shrinkage-type estimates.

We note that Bayesian shape restricted inference with priors introduced by Bernstein polynomials was studied by Chang et al. (2005), which provides a smooth estimate of an increasing failure rate based on right censored data, and by Chang et al. (2007), which compares the Bernstein polynomial method with the density-regression method [Dette, Neumeyer and Pilz (2006)] in estimating an isotonic regression function and a convex regression function. It was also shown there that these Bayesian estimates perform favorably, in addition to the facts that these priors easily take into consideration geometric information, select only smooth functions, can have large support, and can be easily specified. We note that Petrone (1999) made use of these nice properties in her study of random Bernstein polynomials and for sampling the posterior distribution, proposed algorithms that regards the construction of the Bernstein–Dirichlet prior as a histogram smoothing.

The present paper indicates that the expression profiles of virus genes can also be efficiently studied by random Bernstein polynomials, making use of the shape restrictions described above. We will estimate salient features of the profile like onset time, inflection point, maximum value, time to maximum value, area under the profile, etc., utilizing the fact that the derivative of a polynomial has a closed form. We will also test the hypothesis on the shape of the time course expression profile; for example, we will examine whether it is unimodal on the region $[0, \tau]$ for some $\tau < 1$. In fact, by calculating both the posterior probability and the prior

probability that it is unimodal on $[0, \tau]$, we offer an assessment of the strength of the evidence in favor of the hypothesis. We note that this direct approach to hypothesis testing is markedly different from the frequentist p-value approach, as discussed in Kass and Raftery (1995) and Lavine and Schervish (1999), for example.

There is a large literature on shape restricted inference since Hildreth (1954) and Brunk (1955). Most of them treat isotonic and concave regressions from the frequentist viewpoint. Readers are referred to Gijbels (2003) for an excellent review and to Dette, Neumeyer and Pilz (2006) for some of the more recent developments. For the Bayesian approach, there are the works of Lavine and Mockus (1995), Dunson (2005) and Chang et al. (2007), among others. This paper illustrates the use of the Bernstein polynomial in investigating the strength of the evidence provided by the data in favor of the hypothesis on the shape of the regression function, in addition to its use in estimation.

This paper is organized as follows. Section 2 presents the Bernstein polynomial geometry and the hierarchical regression model. Algorithms for Bayesian inference are given in the Appendix. Section 3 illustrates the method by simultaneously analyzing all the data for these genes and indicates that this method does bring insights into baculovirus biology. Section 4 concludes with a brief discussion.

2. Bayesian inference.

2.1. *Bernstein polynomial geometry.* Let $F_{c,a}(t) = \sum_{i=0}^{n} a_i \varphi_{i,n}(\frac{t-c}{1-c}) I_{(c,1]}(t)$, where $a = (a_0, \ldots, a_n)$. Proposition 1 provides a sufficient condition on a under which $F_{c,a}$ is in $\mathcal{A}$. Proposition 2 complements Proposition 1 and provides Bernstein–Weierstrass type approximations for functions in $\mathcal{A}$. In this paper derivatives at 0 and 1 are meant to be one-sided. All the proofs of the propositions in this paper are omitted, because they are similar to those in Chang et al. (2005) and Chang et al. (2007).

PROPOSITION 1.   *Let $n \geq 3$ and $c \in [0, 1)$. If $0 = a_0 = a_1 \leq \min_{l=2,\ldots,n} a_l < \max_{l=2,\ldots,n} a_l$, then $F_{c,a}$ is continuously differentiable, constantly 0 on $[0, c]$, and larger than 0 on $(c, 1)$.*

Let $I_n = \{F_{c,a} \mid c \in [0, 1), a = (a_0, \ldots, a_n)$ satisfying $0 = a_0 = a_1 \leq \min_{l=2,\ldots,n} a_l < \max_{l=2,\ldots,n} a_l\}$. For two continuously differentiable functions f and $\tilde{f}$, define $e(f, \tilde{f}) = \|f - \tilde{f}\|_\infty + \|f' - \tilde{f}'\|_\infty$, where f' denotes the derivative of f, and $\| \cdot \|_\infty$ is the sup-norm for functions on $[0, 1]$. Then we have the following:

PROPOSITION 2.   *Let $\mathcal{D} = \bigcup_{n=3}^{\infty} I_n$. Then $\mathcal{D}$ is dense in $\mathcal{A}$, under e.*

2.2. *Bayesian regression model.*

(i) *Hierarchical prior*

For each $g = 1, \ldots, 154$, we will introduce probabilities π_g on $\mathcal{A}$ as follows. We first describe the framework and then the specific priors to be used. Let $\pi_{1,g}$ be a probability density function on $[0, 1]$, meant to be the prior on the onset time c of gene g; $\pi_{2,g}$ be a probability mass function on the set of positive integers $\{3, 4, \ldots\}$; for each n, $\pi_{3,g}(\cdot|n)$ be a probability density function on $\mathbb{R}^{n-1}$ of b_n. The probability density/mass functions $\pi_{1,g}$, $\pi_{2,g}$ and $\pi_{3,g}$ jointly define a probability $\tilde{\pi}_g$ on $\mathcal{B}$ by the product $\pi_{1,g}(c) \times \pi_{2,g}(n) \times \pi_{3,g}(b_n|n)$; this in turn defines a probability measure on $\mathcal{A}$ by (1.3). Let $\pi_{4,g}$ be a probability density on $\mathbb{R}^1$ for μ_g, the mean of ε_{jkg}. Then $\pi_g = \tilde{\pi}_g \times \pi_{4,g}$ is the prior density we will use on $\mathcal{B} \times \mathbb{R}^1$.

We now describe the strategies to specify $\pi_{1,g}$, $\pi_{2,g}$, $\pi_{3,g}$ and $\pi_{4,g}$. Because our preliminary studies based on a single gene suggest that the posterior distributions of several features do not vary much with the prior order of the Bernstein polynomial so long as it is not too small, we take $\pi_{2,g}$ to have probability 1 for $n = 15$, which has the advantage of lessening the computational burden. The priors $\pi_{1,g}$, $\pi_{3,g}$ and $\pi_{4,g}$ are defined in the following by crude estimates based on all the 154 genes.

For each $g = 1, \ldots, 154$, let $\overline{Y}_{(0)g} \leq \overline{Y}_{(1)g} \leq \cdots \leq \overline{Y}_{(15)g}$ be the order statistics for $\{\overline{Y}_{0g}, \overline{Y}_{1g}, \ldots, \overline{Y}_{15g}\}$, where $\overline{Y}_{kg} = \sum_{j=1}^{4} Y_{jkg}/4$. The prior $\pi_{4,g}$ is the uniform distribution on $[0, 2\overline{Y}_{0g}]$.

We now define $\pi_{1,g}$ for onset time. Let $\tilde{k}(g)$ be the integer such that $\overline{Y}_{\tilde{k}(g)g} = \overline{Y}_{(15)g}$; let $k(g) = \max\{k \mid k = 0, 1, \ldots, \tilde{k}(g) \text{ satisfying } \overline{Y}_{kg} \leq 2\overline{Y}_{0g}\} + 1$. Let $\tilde{X}_g = X_{k(g)}$ and $\hat{X}_g$ equals $X_{(k(g)+\tilde{k}(g))/2}$ if $(k(g) + \tilde{k}(g))/2$ is even, and equals $X_{(k(g)+\tilde{k}(g)+1)/2}$ otherwise. Let α_1 and α_2 be chosen so that the beta distribution $Beta(\alpha_1, \alpha_2)$ has mean $\sum_{g=1}^{154}(\tilde{X}_g/\hat{X}_g)/154$ and variance

$$\left[\left(\max_{g=1,\ldots,154}\{\tilde{X}_g/\hat{X}_g\} - \min_{g=1,\ldots,154}\{\tilde{X}_g/\hat{X}_g\}\right)/4\right]^2.$$

Let $\phi_{11} = \alpha_1 - 0.5$, $\phi_{12} = \alpha_1 + 0.5$, $\phi_{21} = \alpha_2 - 0.5$ and $\phi_{22} = \alpha_2 + 0.5$. We note that for the present dataset, $\alpha_1 = 2.7771$ and $\alpha_2 = 2.4481$, thus, $\phi_{11} = 2.2771$, $\phi_{12} = 3.2771$, $\phi_{21} = 1.9481$ and $\phi_{22} = 2.9481$.

Let ϕ_1 and ϕ_2 be two random variables having distributions respectively *Uniform*(ϕ_{11}, ϕ_{12}) and *Uniform*(ϕ_{21}, ϕ_{22}). Let $U_1, \ldots, U_{154}$ be a random sample of size 154 such that the conditional distribution of U_g given ϕ_1 and ϕ_2 is *Beta*(ϕ_1, ϕ_2) for each $g = 1, \ldots, 154$. We assume that conditional on ϕ_1 and ϕ_2, the prior density $\pi_{1,g}$ of the onset time of gene g is the probability density function of $\hat{X}_g \times U_g$. In particular, we assume the onset time is in the interval $[0, \hat{X}_g]$; this assumption results from examining the data closely.

We next define $\pi_{3,g}(\cdot|n)$, which takes into consideration the range of the observed expression levels and is motivated by the propositions in Section 2.1.

Let $Y_{j[k']g} = Y_{jkg}$, if $\overline{Y}_{(k')g} = \overline{Y}_{kg}$. Denote by $Y_{(1[k])g} \leq Y_{(2[k])g} \leq Y_{(3[k])g} \leq Y_{(4[k])g}$ the order statistics of $\{Y_{1[k]g}, Y_{2[k]g}, Y_{3[k]g}, Y_{4[k]g}\}$. Let ϕ_3 and ϕ_4 be two random variables having distributions respectively $Uniform(\phi_{31}, \phi_{32})$ and $Uniform(\phi_{41}, \phi_{42})$, where $\phi_{31}, \phi_{32}, \phi_{41}$ and ϕ_{42} are constants to be assigned later. Let $V_{2,g}, \ldots, V_{15,g}$ be a random sample such that the conditional distribution of each $V_{i,g}$ given ϕ_3 and ϕ_4 is $Beta(\phi_3, \phi_4)$. We assume that conditional on ϕ_3 and ϕ_4, the prior density function $\pi_{3,g}(\cdot|n)$ of the coefficients $b_{n,g} = (b_{2,n,g}, \ldots, b_{n,n,g})$ is the joint probability density function of $2Y_{(4[15])g} \bullet (V_{2,g}, \ldots, V_{n,g})$. In the present study, $\phi_{31} = \phi_{41} = 0.5$ and $\phi_{32} = \phi_{42} = 1.5$, which give a large support of the prior. Let $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)$, which are the hyperparameters.

Thus, under the assumption that $(c_g, n, b_{n,g}, \mu_g) \in \mathcal{B} \times \mathbb{R}^1$ are conditionally independent given ϕ, the posterior density ν of all the parameters and hyperparameters, given the data, is proportional to

$$(2.1) \qquad \left\{ \prod_{g=1}^{154} \prod_{k=0}^{K} \prod_{j=1}^{m_k} \tilde{g}_{kg}(Y_{jkg} - F_{c_g, b_{n,g}}(X_k)) \pi_g(c_g, n, b_{n,g}, \mu_g | \phi) \right\} \times \psi(\phi),$$

where $\tilde{g}_{kg}$ is the normal density of ε_{jkg} specified in (1.2) and $\psi(\phi) = \prod_{i=1}^{4}(\phi_{i2} - \phi_{i1})^{-1}$ is the joint hyperprior density function.

(ii) *Sampling the posterior distributions*

Based on the hierarchical model, we use a Metropolis-within-Gibbs algorithm to generate the posterior distributions for inference; details of the algorithm are in the Appendix. The software is written in Matlab, which is available from the author upon request. The variance σ_{kg}^2 in (1.2) to be used in the algorithm is decided as follows. Let $\tilde{\sigma}_{kg}^2 = \sum_{j=1}^{4}(Y_{jkg} - \overline{Y}_{kg})^2/3$ and $\hat{\xi}_g \in \{0, 1, 2\}$ be the number that minimizes $L(\xi_g) = \sum_{k=0}^{15}(Q_{kg} - \overline{Q}_g)^2/15$ with $Q_{kg} = \tilde{\sigma}_{kg}^2/\overline{Y}_{kg}^{\xi_g}$ and $\overline{Q}_g = \sum_{k=0}^{15} Q_{kg}/16$ for $\xi_g = 0, 1$ and 2. With $x^{(t)}$ denoting the current state of the Markov chain and $\hat{\mu}_g$ the background noise in the current state $x^{(t)}$, we use

$$\hat{\sigma}_{kg}^2 = \hat{\sigma}_g^2 (\hat{F}_g(X_k) + \hat{\mu}_g)^{\hat{\xi}_g}$$

for the σ_{kg}^2 in (1.2) when updating $x^{(t+1)}$, where $\hat{\sigma}_g^2 = \sum_{k=0}^{15}(\tilde{\sigma}_{kg}^2/\overline{Y}_{kg}^{\hat{\xi}_g})/16$ and $\hat{F}_g$ is the F_g determined by $x^{(t)}$.

We run 5 MCMC chains with initial values chosen randomly from the hyperpriors and the priors of each gene g, and monitor convergence by the Gelman–Rubin statistic $\hat{R}$, following the suggestion in Gelman and Rubin (1992) and Gelman et al. (2004), pages 294–297. For each of the 154 genes, the Gelman–Rubin statistics $\hat{R}$ is calculated for six estimands of interest, which are onset time (Ton), time to maximum (Tmax), maximum (Max), time at which the slope is the highest (Tslope), the highest slope (Slope) and the area under the curve on [0, 1]. Each of the five chains is run with 126,000,000 MCMC iterations and with a burn-in period of 12,600,000 iterations, in which almost all the $\hat{R}$ are less then 1.1. The

56,700 updates, collected by taking one for every 10,000 updates in the last 90% of updates of these 5 sequences, are considered the sample from the posterior distribution, which form the basis for inference.

(iii) *Numerical performance*

To evaluate the numerical performance of the above hierarchical Bayesian method, we studied a similar, but not hierarchical, Bayesian method for the analysis of the time course expression of a single virus gene. This nonhierarchical Bayesian method, modeling the expression profile also by Bernstein polynomials, is more flexible in the sense that it allows nontrivial prior probability on the order of the Bernstein polynomial and is amenable to simulation studies. In fact, the simulation studies indicate its excellent numerical performance. Details of this method and the simulation studies are in the supplementary article [Chang et al. (2008)]. We will evaluate the performance of the hierarchical Bayesian method by comparing it with that of the nonhierarchical Bayesian method, in the context of analyzing our baculovirus expression data. The genes that we chose to conduct this evaluation are selected by the criterion described in the following paragraph; this choice serves also the purpose of comparing the results from our hierarchical Bayesian method and that in Jiang et al. (2006), in addition to evaluating the numerical performance of our method.

For each gene, we consider the differences between the times obtained from the hierarchical Bayesian method and those in Jiang et al. (2006). Figure 1 gives
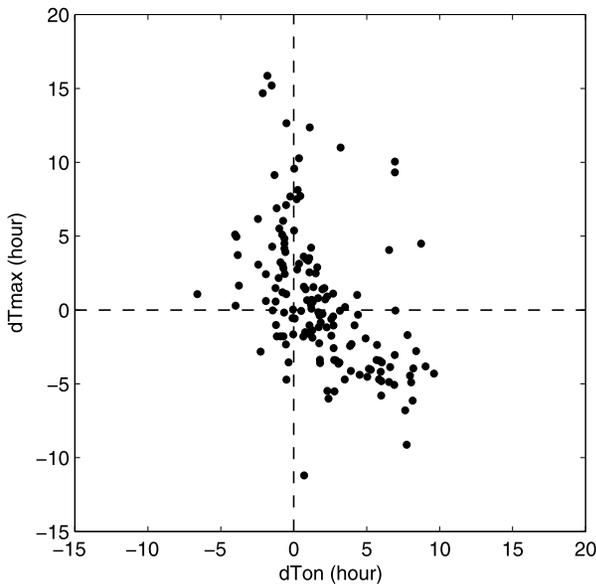

FIG. 1. *The differences between the Ton (Tmax) based on the hierarchical Bayesian method and the crude estimate. The first (second) coordinate of a dot is the onset time (time to maximum) of a gene obtained from hierarchical Bayesian method minus that of the same gene using naive method.*

TABLE 1
Estimates of the onset time based on the naive estimate, hierarchical
Bayesian method and Bayesian method

ID (Name)	Jiang et al. (2006)	Hierarchical Bayesian		Bayesian	
	Estimate	Mean	Stdv	Mean	Stdv
ID 130 (*p10*)	0.0697	0.0724	0.0080	0.0756	0.0043
ID 143 (*pe38*)	0.0335	0.0293	0.0110	0.0211	0.0119
ID 145 (*pk-1*)	0.1785	0.1931	0.0025	0.1930	0.0025
ID 146 (*pk-2*)	0.0552	0.0349	0.0091	0.0292	0.0104
ID 152 (*v-cath*)	0.1374	0.1836	0.0072	0.1802	0.0095

a rough idea of the differences. The first (second) coordinate of a dot in Figure 1 is the onset time (time to maximum) of a gene obtained from the hierarchical Bayesian method minus that of the same gene using the naive method. A gene is selected if either its difference in onset times or that in times to maximum is larger than 10 hours; we note that a difference of this size may cause concerns in biological interpretation. There are in total five such genes and their differences in onset times are not as large as their differences in the time to maximum; we carry out time course expression for these five genes separately by the nonhierarchical Bayesian method. The onset times and the times to maximum of these five genes are shown in Table 1 and Table 2 respectively. The first column of Table 1 gives the ID and the name of these genes; column 2 gives the onset times from Jiang et al. (2006); column 3 gives the means and standard deviations (Stdv) of the posterior distributions of the onset times from the hierarchical Bayesian method; column 4 gives those from the nonhierarchical Bayesian method. The entries in Table 2 bear similar meanings as those in Table 1. It is clear from these tables that the

TABLE 2
Estimates of the time to maximum based on the naive estimate,
hierarchical Bayesian method and Bayesian method

ID (Name)	Jiang et al. (2006)	Hierarchical Bayesian		Bayesian	
	Estimate	Mean	Stdv	Mean	Stdv
ID 130 (*p10*)	0.7343	0.5855	0.0079	0.5859	0.0068
ID 143 (*pe38*)	0.2185	0.3536	0.0248	0.3479	0.0230
ID 145 (*pk-1*)	0.3515	0.5293	0.0046	0.5285	0.0051
ID 146 (*pk-2*)	0.2127	0.4163	0.0179	0.4171	0.0166
ID 152 (*v-cath*)	0.3564	0.4990	0.0082	0.4924	0.0101

results from the hierarchical Bayesian method and those from the nonhierarchical Bayesian method are in quite good agreement. This suggests that the hierarchical Bayesian method seems to produce reliable results in the study of baculovirus gene expression.

We note that one of the genes, *ph*, was knocked out and we included it in the hierarchical Bayesian analysis as a way to see if our method is capable of identifying it. Indeed, it does; it has its time course expression profile much lower than all the others; details are omitted. We also note that we compared other features of several genes obtained from the hierarchical Bayesian method and those from the nonhierarchical Bayesian method and find them in very good agreement. To shorten the paper, we do not report the comparison.

One referee raised the question of whether our procedure automatically identifies genes having different shapes like the two singled out by initially examining the data. Indeed, based on the posterior distributions, we get these two genes identified by performing posterior predictive checking, as described in Gelman (2003) and Gelman, Meng and Stern (1996).

3. Applications to the baculovirus data. Based on the samples from the posterior distribution obtained in Section 2, this section carries out a genome-wide expression analysis of the baculovirus and compares the results with those in Jiang et al. (2006). It seems that the method of this paper reveals more insights into virus biology than the naive method and in case the results from this paper and those in Jiang et al. (2006) are significantly different, it is more often than not that the results from this paper are in better agreement with biology. Since one of the genes, *ph*, was knocked out, the analysis in Jiang et al. (2006) was based on 155 genes and the following studies regard the expression of the 153 genes.

3.1. *Times to maximum.* According to Table 2, the differences in times to maximum for 5 genes are larger than ten hours. Except for the gene *p10*, our method gives larger times to maximum. The following comments seem to suggest that the times to maximum from the current approach allow better or equally sensible interpretation, based on their gene product function.

pe38 encodes a transcription factor important for virulence of the baculovirus [Milks et al. (2003)]. It was shown that it expresses from the immediate early phase throughout the late phase [Knebel-Morsdorf et al. (1996)]. Larger time to maximum might reflect this fact more satisfactorily.

pk-1 is a component of Ac*M*NPV very late gene transcription complex [Mishra, Chadha and Das (2008)]. Reilly and Guarino (1994) indicated that the transcription of *pk-1* peaks in the very late stage of the infection cycle. Larger time to maximum seems more consistent with these observations. Although there is no report on the transcription time of *pk-2*, we tend to think that it is similar to *pk-1* and hence transcribes also in the late stage of the infection cycle.

v-cath encodes a papain type cysteine proteinase with cathapsin L-like property. Its proteinase activity is required for the breakdown of host tissues during the later stages of virus infection/pathogenesis [Hill, Kuzio and Faulkner (1995)]. Larger time to maximum better reflects the needs for its protein expression during this stage, when the host has been exhausted completely and the virus can be spread to other hosts most efficiently.

For the well-known late gene *p10*, although the hierarchical Bayesian method gives a smaller time to maximum than that in Jiang et al. (2006), we note that this smaller time to maximum is still the third largest among all the times to maximum of the 153 genes and hence seems to cause less concern.

3.2. *Time course expression analysis.* To illustrate the use of our method, we now present, in Table 3, the features of the expression profile of the gene *v-cath*, which is one of the genes selected to evaluate the numerical performance of our method. Figure 2 presents the data and the posterior mode of its time course expression. Most of these features can not be reliably obtained by the naive method. This illustration also helps to appreciate that the data have substantial contribution in the inference on these features of *v-cath*. Table 3a reports the posterior probability and the prior probability that the parameter represents a unimodal curve on the interval $[0, \tau]$ for $\tau = 0.6667, 0.8333, 1.0000$; the last two rows give respectively the ratio of the posterior probability to the prior probability and the Bayes factor. Table 3a presents strong evidence, provided by the data, in favor of the unimodality of the time course profile. The posterior probability and the prior probability that the parameter represents a curve that is increasing before reaching its global

TABLE 3
Data analysis for the gene v-cath

Table 3a. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) of being unimodal on $[0, \tau]$.

$[0, \tau]$	$[0, 0.6667]$	$[0, 0.8333]$	$[0, 1.0000]$
Po	1.0000	0.3280	0.0280
Pr	0.4158	0.2658	0.0972
Po/Pr	2.4050	1.2340	0.2881
Bf	∞	1.3482	0.2676

Table 3b. Posterior probability (Po), prior probability (Pr), the ratio of Po to Pr, and the Bayes factor (Bf) that it is increasing before reaching its global maximum.

Po	1.0000
Pr	0.3719
Po/Pr	2.6889
Bf	∞

Table 3c. The Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend of the mode of the posterior density ν in (2.1) is given in the third column in the table. The sample mean, sample Stdv and support of the posterior probability distribution and the prior probability distribution of these features are respectively given in the fourth, fifth and sixth column.

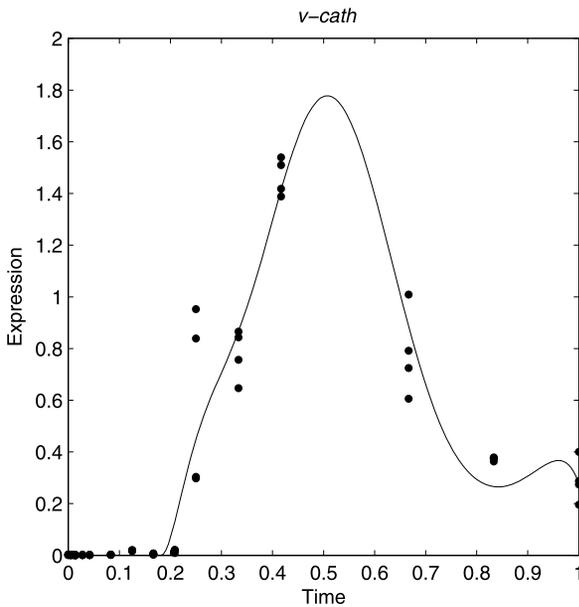| Estimand		Mode	Mean	Stdv	Support
Ton	Posterior	0.1819	0.1836	0.0072	(0.1197, 0.2079)
	Prior		0.1329	0.0510	(0.0023, 0.2498)
Tmax	Posterior	0.5093	0.4990	0.0082	(0.4444, 0.5231)
	Prior		0.7902	0.2278	(0.2083, 1.0000)
Max	Posterior	1.7779	1.6797	0.0877	(1.2713, 1.9397)
	Prior		2.1139	0.5189	(0.3668, 3.0793)
Tslope	Posterior	0.2176	0.2358	0.0420	(0.1944, 0.4074)
	Prior		0.4236	0.3499	(0.0509, 1.0000)
Slope	Posterior	8.1613	8.7394	1.0778	(5.6079, 13.5625)
	Prior		15.0351	9.1052	(1.7144, 59.4057)
L_1-norm	Posterior	0.6206	0.6005	0.0298	(0.5004, 0.7357)
	Prior		1.0878	0.3419	(0.1173, 2.2368)
Tend	Posterior	0.8380	0.8400	0.0720	(0.7500, 1.0000)
	Prior		0.9366	0.1217	(0.3611, 1.0000)

FIG. 2. *The data and the posterior mode of the time course expression of the gene v-cath.*

maximum are reported in Table 3b; similarly, the last two rows give respectively the ratio of the posterior probability to the prior probability and the Bayes factor; Table 3b strongly suggests that the expression profile increases before its global maximum.
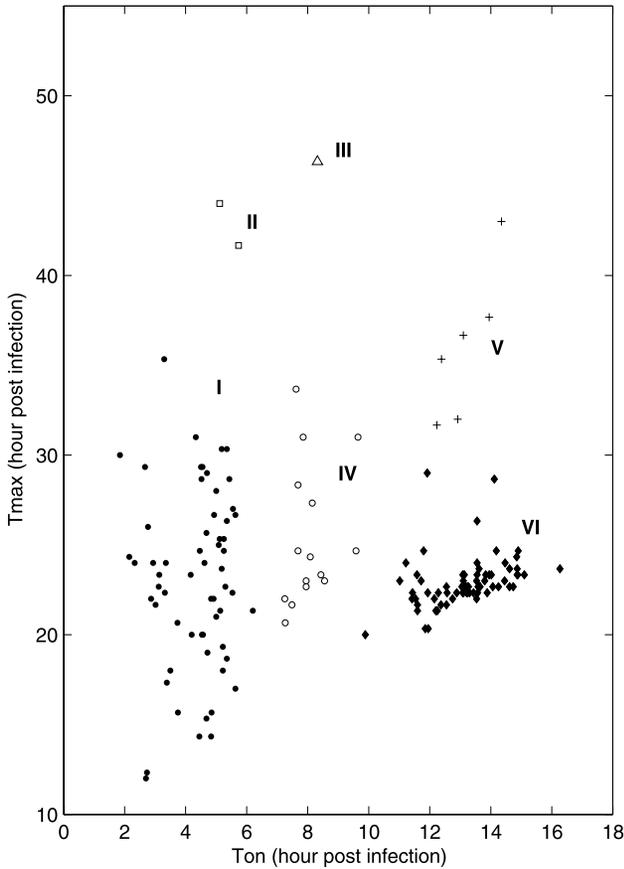
Let τ_0 (Tend) denote the largest time point t such that the time course expression profile is unimodal on $[0, t]$. Let L_1-norm denote the area under the time course expression profile on $[0, \tau_0]$. Table 3c reports Ton, Tmax, Max, Tslope, Slope, L_1-norm and Tend of the mode of the posterior density ν in (2.1) and the sample mean, sample standard deviation (Stdv) and support of these features on the sample respectively from the posterior and prior distributions. Comparing the Stdv and the support from the posterior and the prior, we know that the data have substantial contribution in the inference on these features.

It is customary in microarray literature to cluster genes according to their expression profiles for biologists to use. Using the Ton and Tmax of the mode of the posterior distribution, we apply the cluster analysis algorithm proposed by Hall and Heckman (2002) to cluster the 153 genes into six groups, which are I (early onset and early to maximum), IV (mid-course onset and early to maximum), V (late onset and mid-course to maximum), VI (late onset and early to maximum), and II and III (mid-course onset and late to maximum). The scatterplot in Figure 3 reports the cluster analysis result; genes with known functions are listed according to the clusters to which they belong.

While Figure 3 helps to shed light on the gene groups, it would be interesting to see if genes in the same group have a more similar overall expression profile. Using the rank correlation of two time course expression profiles as the distance between two genes, Table 4 shows that the means of the rank correlation for two genes randomly chosen from the same one of the clusters are smaller than that from the set of all 153 genes. We note that the rank correlation is a measure of similarity between functions studied by Heckman and Zamar (2000). This seems to suggest that genes in the same group have a more similar expression profile.

Based on the time course expression profile of the 153 genes obtained by the posterior mode, we use the K-means algorithm along with the sample rank correlation matrix to cluster them; as in Jiang et al. (2006), we also consider five clusters. The five gene clusters are contained in Figure 4.

We note that clustering is an important step toward gaining insights from high-throughput expression data and there is usually some arbitrariness in forming clusters. Since clustering in Figure 3 is based only on onset times and times to maximum, it is easier to cluster and to interpret, but Figure 4 is more informative in general. For example, Cluster 5 in Figure 4 consists of three genes; one of the most obvious features of these three genes seems to be their large expression levels; thus, it is interesting to note that they are also in such close proximity to each other in Figure 3 and they form exactly the Groups II and III in Figure 3.

FIG. 3. *A classification methodology for the* 153 *genes based on Ton and Tmax. Selected known genes in each classified group are listed at the bottom.*

Group I (early onset and early to maximum) *35K/p35, egt, me53, 39K/pp31, pcna, 94K, ie-2, lef1, pnk/pnl, he65, ie-01, ie-1, lef6, pk-2, DNA-pol, gp64, pe38, lef3, p48, lef7, p26, ctx, helicase, lef11, lef2, p15, tlp, orf-603*

Groups II and III (mid-course onset and late to maximum) *orf-1629, p10, p74*

Group IV (mid-course onset and early to maximum) *gta, p40, ptp, iap1, p43, alk-exo, cg30, odv-e18, PE/pp34*

Group V (late onset and mid-course to maximum) *pk-1, v-cath*

Group VI (late onset and early to maximum) *gp41, p47, p6.9, vlf-1, chitinase, ie-0, pkip, sod, lef9, odv-ec27, lef5, env-prot, lef4, lef8, p95, vp39, gp16, 38K, bro, fgf, fp, HisP, iap2, odv-e56, v-ubi, 49K, odv-e25, vp80, gp37, lef10, p24, odv-e66*

TABLE 4
*Mean and standard deviation of the rank correlation of the time
course expression of two genes chosen from specific groups*

		Rank correlation	
Group	Number of genes	Mean	Stdv
I	60	0.8070	0.1578
II	2	0.9793	0.0000
III	1	NA[*]	NA
IV	15	0.8852	0.0807
V	6	0.9108	0.0594
VI	69	0.8981	0.0867
All	153	0.7717	0.2023

*NA means not applicable.

3.3. *Total expression amount and structure genes.* It is of great interest to study the widely discussed conjecture that the virus has a great demand of structural proteins. While we cannot provide a definitive answer to this question, we think the method of this paper can shed some light on it. One of the salient features of the expression profile obtained by our method is the area under the time course expression profile (L_1-norm); roughly speaking, the L_1-norm of a gene is the sum of the lives of all the mRNA molecules transcribed during the time interval ended at Tend; the life of an mRNA molecule is the time length from its transcription to its degradation or its Tend. Although the relation between the L_1-norm and the total number of the proteins translated is complex, we expect they are positively correlated. We indicate in the following that structure genes seem to have larger L_1-norms. There are 74 baculovirus genes with known names, in which 15 of them are structure genes and the rest are not. We find that, in terms of the L_1-norm, four of the five largest genes are structure genes, giving an odds ratio of 21.1; among the ten largest genes, five of them are structure genes, giving an odds ratio of 5.4; among the 20 largest genes, 7 of them are structure genes, giving an odds ratio of 3.1. We also study by the Wilcoxon statistic the null hypothesis that there is no difference in the L_1-norm between structural genes and nonstructural genes. We find the statistic has value 1.73 and using the one-sided Wilcoxon test, it has p-value 0.0418. This seems to reinforce the conjecture that structural genes tend to have larger L_1-norms. We note it seems hard to estimate the L_1-norms and to study this conjecture by the method of Jiang et al. (2006).

3.4. *Motif and onset time.* Biologists tend to think that genes participating in the same biological process may be transcriptionally coregulated. One preliminary step in studying this phenomenon might be to examine whether upstream sequence motifs of a gene have something to do with its transcription time. In the baculovirus
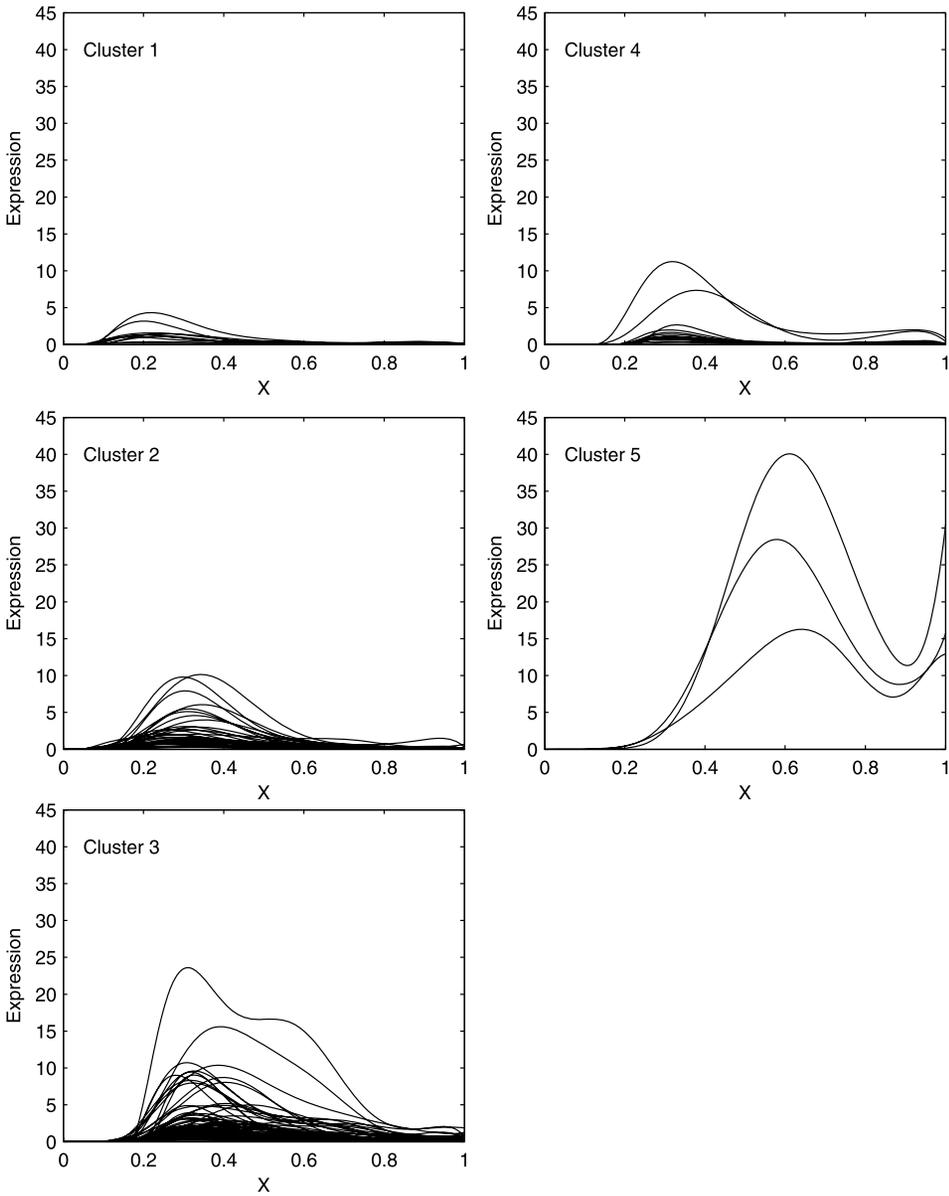
FIG. 4.   *Cluster analysis for the* 153 *viral gene expression profiles.*

literature [Ayres et al. (1994) and Friesen and Miller (2001), for example], motifs A(A/T)CGT(G/T) and CGTGC are called the early motif; motif TAAG is called the late motif; genes having motif CATG are usually believed to express early. Jiang et al. (2006) studies this by reporting the proportions of these motifs in the 5 gene clusters obtained from clustering the time course expression crude data.

TABLE 5
Motifs have to do with onset time. Comparing the onset times of genes having specific motifs with those without by the Wilcoxon statistic, which is asymptotically standard normal. Minus (plus) values indicate the former (latter) is smaller (larger)

Motif	With	Without	Wilcoxon statistic	*p*-value
Early*	64	66	−2.65	0.00402
TAAG	70	60	4.04	0.00003
CATG	69	61	−2.66	0.00391
Early/CATG	110	20	−2.54	0.00554

*The early motif (Early) consists of motifs A(A/T)CGT(G/T) and CGTGC.

While we can conduct a similar study by means of the clusters obtained from our Bayesian method, we propose to ignore the clusters and take a more direct and relevant approach to address this issue.

Based on the onset times of this paper, we study the hypotheses that, with a given motif, there is no difference between the onset times of the genes with this motif and those without this motif. We study them by the Wilcoxon statistic. Table 5 summarizes the numbers of genes having or not having these motifs and reports the Wilcoxon statistics and their *p*-values for testing the corresponding one-sided null hypothesis. For example, the second row shows that 70 genes have TAAG and 60 genes do not have it, its Wilcoxon statistic is 4.04 and the *p*-value is smaller than 0.0001, which seem to suggest that the genes having TAAG tend to have later onset times. It seems Table 5 supports the idea that motifs have something to do with onset times.

3.5. *Colocalization.* Because functionally correlated or coregulated genes in an operon of a bacterial genome may be located in nearby loci of the physical genome [Lagreid et al. (2003)], Jiang et al. (2006) investigated whether a similar gene organization exists in the Ac*M*NPV genome. Based on the time course expression normalized data, Jiang et al. (2006) clustered genes into five clusters and noted six colocalized clusters. A colocalized cluster is defined as a genome region that contains at least five consecutive genes from the same gene cluster where no more than one interruption occurs by a gene from other gene clusters in either the plus or minus strand. Using the same definition of a colocalized cluster, we find there are nine colocalized clusters, based on the five clusters exhibited in Figure 4. These nine colocalized clusters are shown in Figure 5. This seems to suggest that expression profiles from our sophisticated method reveals more signals than the naive method.

The phenomenon that genes with similar expression profile tend to be located near each other is referred to as colocalization in Jiang et al. (2006). Since the above definition of a colocalized cluster is somewhat arbitrary, we present a more
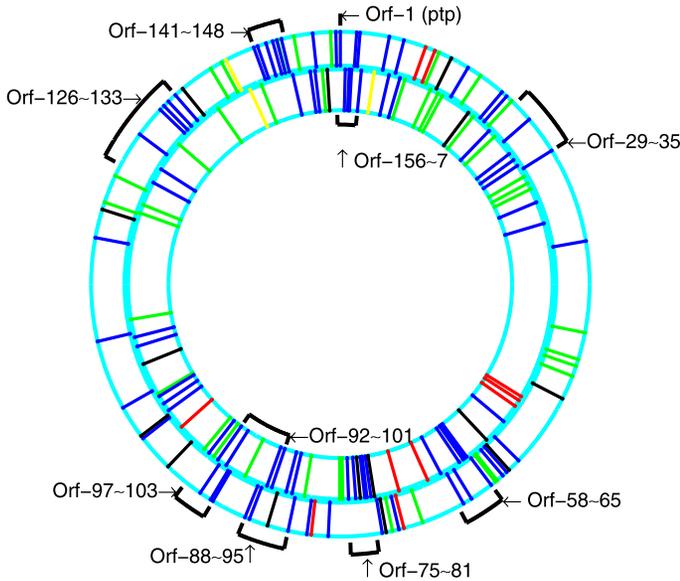
FIG. 5. *Genome map view of the five gene clusters color tagged in the baculovirus genome. Red, green, blue, black and yellow represent respectively genes in the cluster* 1, 2, 3, 4, 5 *in Figure* 4.

systematic study on this in Table 6. Column two and column three of Table 6 give respectively the probability of two (three, four, five) randomly chosen genes that belong simultaneously to the same one of the five clusters and the probability of two (three, four, five) randomly chosen neighboring genes that belong simultaneously to the same one of the five clusters. Because the numbers in column 2 are smaller than those in column 3, it seems that colocalization does exist.

From the viewpoint of evolution, it might also be appealing to see if genes close to each other on the genome have a similar expression pattern. One relevant null hypothesis would be that there is no difference in the rank correlation of expression profiles from nearby genes and that from far away genes. For integer $0 \leq Z \leq 76 = (153 - 1)/2$, let $\text{Nei}(g, Z)$ denote the set of genes whose distance from gene g is

TABLE 6
The probability that N randomly chosen (*neighboring*)
genes belong to the same cluster in Figure 4

N	Randomly chosen	Neighboring
2	0.3835	0.4837
3	0.1820	0.2680
4	0.0926	0.1373
5	0.0484	0.0719

TABLE 7

Comparing the rank correlation of the time course expression profiles from nearby genes and that from far away genes. Rn(Z) is the set of rank correlations for genes having no more than Z genes lying between them; RCn(Z) is that for genes having at least Z genes lying between them

Rn(Z)	RCn(Z)	Wilcoxon statistic	*p*-value
Rn(2)	RCn(12)	8.81	0.0000
Rn(4)	RCn(14)	5.82	0.0000
Rn(6)	RCn(16)	5.09	0.0000
Rn(8)	RCn(18)	2.72	0.0033
Rn(10)	RCn(20)	2.01	0.0221
Rn(12)	RCn(22)	0.84	0.2002
Rn(14)	RCn(24)	1.25	0.1051
Rn(16)	RCn(26)	2.31	0.0105
Rn(18)	RCn(28)	1.59	0.0558
Rn(20)	RCn(30)	0.23	0.4078
Rn(22)	RCn(32)	0.55	0.2913
Rn(24)	RCn(34)	1.45	0.0737
Rn(26)	RCn(36)	0.25	0.4014

no larger than Z; here the distance between two genes is the number of genes lying strictly between them. Let $Rn(Z)$ denote the set of rank correlations of the time course expression profile of a gene g and that of a gene in $Nei(g, Z)$. Let $RCn(Z)$ denote the set of rank correlations of the time course expression profile of a gene g and that of a gene not in $Nei(g, Z)$. In terms of this notation, the null hypothesis becomes that there is no statistical difference between $Rn(Z_1)$ and $RCn(Z_2)$. We studied the hypothesis by the Wilcoxon statistic for many choices of Z_1 and Z_2. Table 7 reports the Wilcoxon statistics and their p-values for testing the corresponding one-sided null hypothesis for several choices of Z_1 and Z_2. It suggests that nearby genes do have a higher chance to have a similar expression pattern.

4. Discussion. We have illustrated a hierarchical Bayesian shape restricted regression method for the inference on the genome-wide time course expression of virus genes and, based on the profiles provided by this method, we have examined salient features on the time course expression curves, studied some hypotheses on and thus brought insights into baculovirus biology. It is to be noted that our method helps to formulate biological questions quantitatively so as to make modern statistics methods applicable. Although we looked at colocalization, the relation between upstream motifs and onset times, and that between area under curve and gene function, these are, nevertheless, preliminary investigations. Further studies are needed to give a more complete account of these aspects of the baculovirus.

In view of the facts that genome-wide expression studies of virus genes are gaining popularity, all the previous works in this area use at most crude statistics

for biological interpretation, and the existing discrepancies between the studies need to be resolved, we think our method is useful not only in one single expression study of virus genes but also in comparing these studies, which would enhance our understanding of the gene regulation network. We note that our method can be used to provide comprehensive comparison of the time course transcription profiles from different experiments when even their time points are not identical, as long as there are enough of them to capture their respective main features.

As for future methodological development, we think the Bernstein–Dirichlet prior of Petrone (1999) and the related samplers are also useful in this context; studies in this line and comparison with the approach in this paper deserve our attention.

APPENDIX: METROPOLIS-WITHIN-GIBBS ALGORITHM FOR THE POSTERIOR

Let $B_n = \{b_n \in \mathbb{R}^{n-1} : F_{c,b_n} \in I_n$ for some $c \in [0, 1)\}$. Denote $(b_{2,n,g}, \ldots, b_{n,n,g}) = b_{n,g}$ by $(a_{2,g}, \ldots, a_{n,g}) = a_g$. Let $\mathbf{c} = (c_1, \ldots, c_{154})$; $\mathbf{a} = (a_1, \ldots, a_{154})$; $\mathbf{u} = (\mu_1, \ldots, \mu_{154})$.

Let $B = \{\phi, \mathbf{c}, \mathbf{a}, \mathbf{u} \mid \phi = (\phi_1, \phi_2, \phi_3, \phi_4) \in [2.2771, 3.2771] \times [1.9481, 2.9481] \times [0.5, 1.5] \times [0.5, 1.5], c_g \in [0, \hat{X}_g], a_g \in B_n, \mu_g \in [0, 2\overline{Y}_{0g}]\}$. Our computational strategy consists of the following five MCMC algorithms to update ϕ, $\mathbf{c}$, $\mathbf{a}$ and $\mathbf{u}$ consecutively. Let $x^{(t)} = (\phi^{(t)}, \mathbf{c}^{(t)}, \mathbf{a}^{(t)}, \mathbf{u}^{(t)}) \in B$ denote the current state of the MCMC chain for sampling the posterior distribution.

(i) *Update ϕ_1 and ϕ_2*

1. Let $\tilde{\phi}_1$ and $\tilde{\phi}_2$ be two random samples from $Uniform(\phi_{11}, \phi_{12})$ and $Uniform(\phi_{21}, \phi_{22})$ respectively;
2. let $y = (\tilde{\phi}_1, \tilde{\phi}_2, \phi_3^{(t)}, \phi_4^{(t)}, \mathbf{c}^{(t)}, \mathbf{a}^{(t)}, \mathbf{u}^{(t)})$;
3. set

$$x^{(t+1)} = \begin{cases} y, & \text{with prob. } \rho = \min\left\{1, \dfrac{v(y)}{v(x^{(t)})}\right\}, \\ x^{(t)}, & \text{otherwise.} \end{cases}$$

(ii) *Update ϕ_3 and ϕ_4*

1. Let $\tilde{\phi}_3$ and $\tilde{\phi}_4$ be two random samples from $Uniform(\phi_{31}, \phi_{32})$ and $Uniform(\phi_{41}, \phi_{42})$ respectively;
2. let $y = (\phi_1^{(t)}, \phi_2^{(t)}, \tilde{\phi}_3, \tilde{\phi}_4, \mathbf{c}^{(t)}, \mathbf{a}^{(t)}, \mathbf{u}^{(t)})$;
3. set

$$x^{(t+1)} = \begin{cases} y, & \text{with prob. } \rho = \min\left\{1, \dfrac{v(y)}{v(x^{(t)})}\right\}, \\ x^{(t)}, & \text{otherwise.} \end{cases}$$

(iii) *Update* **c**

There are 154 components $(c_1, \ldots, c_{154})$ in **c**; we update them one at a time in the order of the coordinates. Suppose $c_1^{(t)}, \ldots, c_{g-1}^{(t)}$ have been just updated and we now want to update $c_g^{(t)}$.

1. Let U be a random sample from $Beta(\phi_1^{(t)}, \phi_2^{(t)})$;
2. let $\tilde{c}_g = \hat{X}_g \times U$; let $\pi_{1,g}(\tilde{c}_g | \phi_1^{(t)}, \phi_2^{(t)})$ denote the prior density $\pi_{1,g}$ of $\tilde{c}_g$ given $\phi_1^{(t)}$ and $\phi_2^{(t)}$;
3. let $y = (\phi^{(t)}, c_1^{(t)}, \ldots, c_{g-1}^{(t)}, \tilde{c}_g, c_{g+1}^{(t)}, \ldots, c_{154}^{(t)}, \mathbf{a}^{(t)}, \mathbf{u}^{(t)})$;
4. set

$$
x^{(t+1)} = \begin{cases} y, & \text{with prob. } \rho = \min\left\{1, \dfrac{\nu(y)\pi_{1,g}(c_g^{(t)}|\phi_1^{(t)}, \phi_2^{(t)})}{\nu(x^{(t)})\pi_{1,g}(\tilde{c}_g|\phi_1^{(t)}, \phi_2^{(t)})}\right\}, \\ x^{(t)}, & \text{otherwise.} \end{cases}
$$

(iv) *Update* **a**

We update one coordinate of **a** each time in the order of the coordinates. Suppose we have updated $a_{2,g}^{(t)}, \ldots, a_{i-1,g}^{(t)}$ and we now want to update $a_{i,g}^{(t)}$.

1. Let V be a random sample from $Beta(\phi_3^{(t)}, \phi_4^{(t)})$;
2. let $\tilde{a}_{i,g} = 2Y_{(4[15])g} \times V$; let $\pi_{3,g}(\tilde{a}_{i,g}|\phi_3^{(t)}, \phi_4^{(t)})$ denote the prior density $\pi_{3,g}(\cdot|n)$ of the coefficient $\tilde{a}_{i,g}$ given $\phi_3^{(t)}$ and $\phi_4^{(t)}$;
3. let y be the same vector as $x^{(t)}$ except replacing $a_{i,g}^{(t)}$ by $\tilde{a}_{i,g}$;
4. set

$$
x^{(t+1)} = \begin{cases} y, & \text{with prob. } \rho = \min\left\{1, \dfrac{\nu(y)\pi_{3,g}(a_{i,g}^{(t)}|\phi_3^{(t)}, \phi_4^{(t)})}{\nu(x^{(t)})\pi_{3,g}(\tilde{a}_{i,g}|\phi_3^{(t)}, \phi_4^{(t)})}\right\}, \\ x^{(t)}, & \text{otherwise.} \end{cases}
$$

(v) *Update* **u**

There are 154 components $(\mu_1, \ldots, \mu_{154})$ in **u**; we update them one at a time in the order. Suppose we have updated $\mu_1^{(t)}, \ldots, \mu_{g-1}^{(t)}$ and we now want to update $\mu_g^{(t)}$.

1. Let $\tilde{\mu}_g$ be a random sample from $Uniform(0, 2\overline{Y}_{0g})$;
2. let $y = (\phi^{(t)}, \mathbf{c}^{(t)}, \mathbf{a}^{(t)}, \mu_1^{(t)}, \ldots, \mu_{g-1}^{(t)}, \tilde{\mu}_g, \mu_{g+1}^{(t)}, \ldots, \mu_{154}^{(t)})$;
3. set

$$
x^{(t+1)} = \begin{cases} y, & \text{with prob. } \rho = \min\left\{1, \dfrac{\nu(y)}{\nu(x^{(t)})}\right\}, \\ x^{(t)}, & \text{otherwise.} \end{cases}
$$

Acknowledgments. We are grateful to Professor Xiao-Li Meng for his comments on an earlier version of this paper, which led to improvements of the paper in several ways. We are also grateful to two anonymous referees for their valuable comments that led to a more focused and balanced treatment of the subjects.

SUPPLEMENTARY MATERIAL

Profiling time course expression of a single virus gene (DOI: 10.1214/09-AOAS258SUPP; .pdf). This nonhierarchical Bayesian method, using also Bernstein polynomials, allows nontrivial prior probability on the order of the Bernstein polynomial and is amenable to simulation studies, which indicate its excellent numerical performance.

REFERENCES

AYRES, M. D., HOWARD, S. C., KUZIO, J., LOPEZ-FERBER, M. and POSSEE, R. D. (1994). The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus. *Virology* **202** 586–605.

BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616. MR0073894

CHANG, I. S., CHIEN, L. C., HSIUNG, C. A., WEN, C. C. and WU, Y. J. (2007). Shape restricted regression with random Bernstein polynomials. In *Complex Datasets and Inverse Problems* (R. Liu, W. Strawderman and C. H. Zhang, eds.). *IMS Lecture Notes—Monograph Series* **54** 187–202. Inst. Math. Statist., Beachwood, OH. MR2459189

CHANG, I. S., CHIEN, L. C., GUPTA, P. K., WEN, C. C., WU, Y. J. and HSIUNG, C. A. (2008). Supplement to "Profiling time course expression of virus genes—an illustration of Bayesian inference under shape restrictions."

CHANG, I. S., HSIUNG, C. A., WU, Y. J. and YANG, C. C. (2005). Bayesian survival analysis using Bernstein polynomials. *Scand. J. Statist.* **32** 447–466. MR2204629

DETTE, H., NEUMEYER, N. and PILZ, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli* **12** 469–490. MR2232727

DUNSON, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *J. Amer. Statist. Assoc.* **100** 618–627. MR2160564

DUPLESSIS, M., RUSSELL, W. M., ROMERO, D. A. and MOINEAU, S. (2005). Global gene expression analysis of two *Streptococcus thermophilus* bacteriophages using DNA microarray. *Virology* **340** 192–208.

FRIESEN, P. D. and MILLER, L. K. (2001). Insect viruses. In *Fields' Virology*, 4th ed. (D. M. Knipe, P. M. Howley, D. E. Griffin, M. A. Martin, R. A. Lamb, B. Roizman and S. E. Straus, eds.) 608–609. Lippincott Williams & Wilkins, Philadelphia.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton. MR2027492

GELMAN, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Statist. Rev.* **71** 369–382.

GELMAN, A., MENG, X. L. and STERN, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

GIJBELS, I. (2003). Monotone regression. Discussion Paper 0334, Institute de Statistique, Université Catholique de Louvain. Available at http://www.stat.ucl.ac.be.

HALL, P. and HECKMAN, N. E. (2002). Estimating and depicting the structure of a distribution of random funcions. *Biometrika* **89** 145–158. MR1888371

HECKMAN, N. E. and ZAMAR, R. H. (2000). Comparing the shapes of regression functions. *Biometrika* **87** 135–144. MR1766834

HILDRETH, C. (1954). Point estimate of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. MR0065093

HILL, J. E., KUZIO, J. and FAULKNER, P. (1995). Identification and characterization of the v-cath gene of the baculovirus, CfMNPV. *Biochimica et Biophysica Acta* **1264** 275–278.

IWANAGA, M., TAKAYA, K., KATSUMA, S., OTE, M., TANAKA, S., KAMITA, S. G., KANG, W. K., SHIMADA, T. and KOBAYASHI, M. (2004). Expression profiling of baculovirus genes in permissive and nonpermissive cell lines. *Biochemical and Biophysical Research Communications* **323** 599–614.

JIANG, S. S., CHANG, I. S., HUANG, L. W., CHEN, P. C., WEN, C. C., LIU, S. C., CHIEN, L. C., LIN, C. Y., HSIUNG, C. A. and JUANG, J. L. (2006). Temporal transcription program of recombinant *Autographa californica* multiple nucleopolyhedrosis virus. *Journal of Virology* **80** 8989–8999.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KNEBEL-MORSDORF, D., FLIPSEN, J. T., RONCARATI, R., JAHNEL, F., KLEEFSMAN, A. W. and VLAK, J. M. (1996). Baculovirus infection of *Spodoptera exigua* larvae: *lacZ* expression driven by promoters of early genes pe38 and me53 in larval tissue. *Journal of General Virology* **77** 815–824.

LAGREID, A., HVIDSTEN, T. R., MIDELFART, H., KOMOROWSKI, J. and SANDVIK, A. K. (2003). Predicting gene ontology biological process from temporal gene expression patterns. *Genome Research* **13** 965–979.

LAVINE, M. and MOCKUS, A. (1995). A nonparametric Bayes method for isotonic regression. *J. Statist. Plann. Inference* **46** 235–248.

LAVINE, M. and SCHERVISH, M. J. (1999). Bayes factors: What they are and what they are not. *Amer. Statist.* **53** 119–122. MR1707756

MAJTAN, T., HALGASOVA, N., BUKOVSKA, G. and TIMKO, J. (2007). Transcriptional profiling of bacteriophage BFK20: Coexpression interrogated by "guilt-by-association" algorithm. *Virology* **359** 55–65.

MILKS, M. L., WASHBURN, J. O., WILLIS, L. G., VOLKMAN, L. E. and THEILMANN, D. A. (2003). Deletion of *pe38* attenuates AcMNPV genome replication, budded virus production, and virulence in *Heliothis virescens*. *Virology* **310** 224–234.

MISHRA, G., CHADHA, P. and DAS, R. H. (2008). Serine/threonine kinase (pk-1) is a component of *Autographa californica* multiple nucleopolyhedrovirus (*Ac*MNPV) very late gene transcription complex and it phosphorylates a 102 kDa polypeptide of the complex. *Virus Research* **137** 147–149.

PETRONE, S. (1999). Random Bernstein polynomials. *Scand. J. Statist.* **26** 373–393. MR1712051

REILLY, L. M. and GUARINO, L. A. (1994). The *pk-1* gene of *Autographa californica* multinucleocapsid nuclear polyhedrosis virus encodes a protein kinase. *Journal of General Virology* **75** 2999–3006.

SMITH, I. (2007). Misleading messengers? Interpreting baculovirus transcriptional array profiles. *Journal of Virology* **81** 7819–7821.

VAN MUNSTER, M., WILLIS, L. G., ELIAS, M., ERLANDSON, M. A., BROUSSEAU, R., THEILMANN, D. A. and MASSON, L. (2006). Analysis of the temporal expression of *Trichoplusia ni* single nucleopolyhedrovirus genes following transfection of BT1-Tn-5B1-4 cells. *Virology* **354** 154–166.

YANG, W. C., DEVI-RAO, G. V., GHAZAL, P., WAGNER, E. K. and TRIEZENBERG, S. J. (2002). General and specific alterations in programming of global viral gene expression during infec-

tion by VP16 activation-deficient mutants of herpes simplex virus type 1. *Journal of Virology* **76** 12758–12774.

L.-C. CHIEN
P. K. GUPTA
C. A. HSIUNG
DIVISION OF BIOSTATISTICS
   AND BIOINFORMATICS
NATIONAL HEALTH RESEARCH INSTITUTES
35 KEYAN ROAD, ZHUNAN TOWN,
   MIAOLI COUNTY 350
TAIWAN
E-MAIL: hsiung@nhri.org.tw

S. S. JIANG
NATIONAL INSTITUTE OF CANCER RESEARCH
NATIONAL HEALTH RESEARCH INSTITUTES
35 KEYAN ROAD, ZHUNAN TOWN,
   MIAOLI COUNTY 350
TAIWAN

I-S. CHANG
DIVISION OF BIOSTATISTICS
   AND BIOINFORMATICS
AND
NATIONAL INSTITUTE OF CANCER RESEARCH
NATIONAL HEALTH RESEARCH INSTITUTES
35 KEYAN ROAD, ZHUNAN TOWN,
   MIAOLI COUNTY 350
TAIWAN

C.-C. WEN
DEPARTMENT OF MATHEMATICS
TAMKANG UNIVERSITY
151 YINGCHUAN ROAD, TAMSUI TOWN,
   TAIPEI COUNTY 251
TAIWAN

Y.-J. WU
DEPARTMENT OF APPLIED MATHEMATICS
CHUNG YUAN CHRISTIAN UNIVERSITY
200 CHUNG PEI ROAD, CHUNGLI CITY 320
TAIWAN