How to cluster gene expression dynamics in response to environmental signals

Yaqun Wang, Meng Xu, Zhong Wang, Ming Tao, Junjia Zhu, Li Wang, Runze Li, Scott A. Berceli and Rongling Wu

Submitted: 25th February 2011; Received (in revised form): 16th May 2011

Abstract
Organisms usually cope with change in the environment by altering the dynamic trajectory of gene expression to adjust the complement of active proteins. The identification of particular sets of genes whose expression is adaptive in response to environmental changes helps to understand the mechanistic base of gene–environment interactions essential for organismic development. We describe a computational framework for clustering the dynamics of gene expression in distinct environments through Gaussian mixture fitting to the expression data measured at a set of discrete time points. We outline a number of quantitative testable hypotheses about the patterns of dynamic gene expression in changing environments and gene–environment interactions causing developmental differentiation. The future directions of gene clustering in terms of incorporations of the latest biological discoveries and statistical variation of complex traits and processes.
innovations are discussed. We provide a set of computational tools that are applicable to modeling and analysis of dynamic gene expression data measured in multiple environments.

**Keywords:** dynamic gene expression; functional clustering; gene–environment interaction; mixture model

**INTRODUCTION**

The development of high-throughput technologies, such as DNA microarrays and proteomics platforms, has made it possible to ask and address many fundamental but difficult questions in developmental biology and biomedicine. For example, variation in the pattern of gene expression may reflect unique physiological or pathological properties of individual cells, organs or organisms that cannot be observed readily or directly [1, 2]. By screening approximately 1000 proteins in individual cancer cells, Cohen et al. [3] detected a subset of proteins whose expression displays different dynamic patterns between seemingly identical cancer cells that actually have different fates. To identify distinct patterns of gene expression dynamics from a flood of microarray data, powerful computational tools for clustering genes or proteins based on their dynamic profiles have become essential. In the past decade, enormous efforts have been made to develop computational methods for cataloguing gene expression dynamics and use these distinct patterns to assess developmental functions and mechanisms of biological phenomena [4–15]. There is also a pressing need for computational approaches to cluster analysis of dynamic gene expression that interacts with the environment, because the activation and expression of many genes is environment-contingent [16]. However, to our best knowledge, such approaches have not been systematically described in the literature, obstructing geneticists to analyze time series data in a way that can detect complex effects on expression patterns.

In biology, there is a widespread phenomenon that organisms cope with biotic and abiotic environments by controlling gene expression to harness the complement of active proteins [17–19]. When the environment alternates between discrete states, organisms will stimulate their regulatory system through adjusting gene transcription rates to best adapt to the environment. For this reason, by detecting the difference in the pattern of gene expression trajectories between discrete environments, we will be in a better position to study interactions between genes and environments and dictate a comprehensive map of gene–environment relationships. Traditional approaches for studying gene–environment interactions are based on quantitative trait locus (QTL) mapping usually with experimental crosses. Significant gene–environment interactions are identified if specific QTLs are detected, through statistical tests, to display different effects between environments [20, 21]. More recently, gene transcript abundance has been used to study gene–environment interactions in many organisms such as yeast [16, 22] and worms [23], but all these studies are limited to gene expression data measured at single time points of development.

The purpose of this article is to describe a general framework for identifying environment-specific clusters of gene expression. The use of gene expression dynamics to understand gene–environment interactions is highly informative because of its capacity to identify development-related genes. However, this is a challenging work in terms of gene clustering across two-dimensional spaces, time and environmental state. The framework described in this article integrates developmental and environment-dependent programs of gene expression. Mathematical aspects of gene expression dynamics are implemented into a mixture model setting by considering the impact of environment on gene expression. The patterns of gene expression related to specific physiological functions can be parsimoniously modeled using a set of mathematical parameters [14, 24]. Thus, by estimating the parameters that determine mathematical functions, the pattern of how genes change their level of expression over time and environment can be estimated and tested. The results from these models, therefore, can better be interpreted in a biologically sensible way. In addition, the framework considers the intrinsic structure of time-dependent correlations based on an optimal statistical process, which increases the power of detecting significantly differentiated patterns. To demonstrate its usefulness and utilization in practice, we use this framework to analyze a real data set from a rabbit hemodynamic study in which gene expression is observed in two distinct blood flow environments [25, 26]. We evaluated the advantages of this tool by
performing simulation studies. The simulation results illustrated that the tool has favorable statistical properties and can be used in any environment-dependent gene expression data.

**GENE-CLUSTERING FRAMEWORK**

**Statistical model**

Suppose there are $n$ genes each measured at $T$ time points in $L$ environments. Let $y_i = (y_i(t_1^l), \ldots, y_i(t_j^l))$ denote the gene expression data for gene $i$ in environment $l$. Combining all the environments, we have $y_i = (y_i(t_1^1), \ldots, y_i(t_j^l); \ldots; y_i(t_1^L), \ldots, y_i(t_j^L))$. If these genes are grouped into $J$ clusters, it means that any one of genes $(i)$ is assumed to arise from one (and only one) of the $J$ possible clusters. Thus, the phenotypic value of gene $i$ expressed at time $t_i^l$ in environment $l$ is written as

$$y_i(t_i^l) = \sum_{j=1}^J \xi_{ij} \mu_j(t_i^l) + \sum_{c=1}^C \beta_c u_c + e_i(t_i^l) \quad (1)$$

where $\xi_{ij}$ is an indicator for gene $i$, defined as 1 if this gene belongs to cluster $j$ and 0 otherwise, $\mu_j(t_i^l)$ is the mean of all genes belonging to cluster $j$ at time $t_i^l$ in environment $l$, $u_c$ is the value of covariate $c$ ($c = 1, \ldots, C$) for gene $i$, $\beta_c$ is the effect of covariate $c$, and $e_i(t_i^l)$ is the residual assumed to follow a Gaussian distribution with mean zero and variance $\sigma^2(t_i^l)$. For longitudinal data, residual errors at different time points may be correlated with covariance $\sigma(t_1^l, t_2^l)$ ($l \leq 1, \ldots, L$). At least one of the inequalities, $l_1 \neq l_2, \tau_1 \neq \tau_2$, holds. The residual variances and covariance comprise a $(TL \times TL)$ covariance matrix $\Sigma$.

The distribution of gene expression data is expressed as the $J$-component mixture probability density function, i.e.

$$y_i \sim f(y; \omega, \mu, \Sigma) = \sum_{j=1}^J \omega_j f_j(y; \mu_j, \Sigma), \quad (2)$$

where $\omega = (\omega_1, \ldots, \omega_J)$ is a vector of mixture proportions which are nonnegative and sum to unity; $\mu = (\mu_1, \ldots, \mu_J)$ contains the mean vector of cluster $j$; and $\Sigma$ contains residual variances and covariances among $T$ time points over $L$ environments which are common for all clusters. The probability density function of cluster $j$, $f_j(y; \mu_j, \Sigma)$, is assumed to be multivariate normally distributed with $TL$-dimensional mean vector

$$\mu_j = \left(\mu_j(t_1^l) + \sum_{c=1}^C \beta_{c,j} u_c; \ldots; \mu_j(t_{TL}^l) + \sum_{c=1}^C \beta_{c,j} u_c; \ldots\right) \quad (3)$$

and covariance matrix $\Sigma$. Notice that $\mu_j$ contains gene-specific covariate effects.

The likelihood based on a mixture model containing $J$ clusters can be written as

$$L(\Theta|y) = \prod_{i=1}^n \sum_{j=1}^J \omega_j f_j(y; \mu_j, \Sigma). \quad (4)$$

where $\Theta$ is a vector of unknown parameters including the mixture proportions, cluster-specific mean vectors and covariance.

Different from traditional treatments, we will incorporate mathematical and statistical models to fit the mean-covariance structures. Instead of estimating all elements in the vectors and covariance, we estimate the mathematical and statistical parameters that model the mean-covariance structures. Thus, the question of clustering gene expression dynamics becomes how to find a set of parameters (arrayed in $\Theta_{\mu_j} = (\Theta_{\mu_j}^{1,1}, \ldots, \Theta_{\mu_j}^{J,1})$) that models cluster-specific expression profiles in a biologically and statistically meaningful way and to find a set of parameters (arrayed in $\Theta_{\Sigma}$) that models the covariance structure both parsimoniously and flexibly.

**Structural modeling of mean vectors**

Since the transcript levels of DNA microarrays generally vary in a time course, we may use mathematical and statistical models to approach their dynamic changes. Below is a list of approaches for modeling time-dependent gene expression profiles:

**Parametric modeling**

For clock gene cases, the amount of mRNAs within the cell division cycle changes periodically, coincident with the cell cycle, which helps to maintain proper order during cell division or to conserve limited resources. The oscillation of cell cycle-regulated genes can be mathematically described by periodic Fourier functions or other functions [27–32].
The Fourier function can be approximated by its first $K$ terms, expressed as

$$F_k(t) = a_0 + \sum_{k=1}^{K} \left( a_k \cos \left( \frac{2\pi kt}{\lambda} \right) + b_k \sin \left( \frac{2\pi kt}{\lambda} \right) \right).$$

The coefficients $a_k$ and $b_k$ determine the times at which the expression level achieves maximums and minimums, $a_0$ is the average expression level of the gene and $\lambda$ specifies the periodicity of the regulation. From Equation (3), the mean expression value of gene cluster $j$ at time $t_i$ in environment $l$ is expressed as

$$\mu_j^l(t_i) = F_k(t_i; \Theta^l_j)$$

where $\Theta^l_j = (a_{0j}, a_{1j}, \ldots, a_{Kj}, b_{0j}, \ldots, b_{Kj}, \theta_j)$ denotes the vector of Fourier parameters of the first $K$ orders. Thus, by estimating the parameters that define the periodic curves for individual clusters, we can determine the differences in the temporal pattern of gene expression [14].

There are many other biologically well-justified curves that can be used to model gene expression dynamics. These include sigmoid equations for gene expression related to biological growth [33–37], triple-logistic equations for gene expression related to human body growth [38], bi-exponential equations for gene expression related to HIV dynamics [39], sigmoid Emax models for gene expression related to pharmacodynamic response [40], biological thermal dynamics [41], aerodynamic power curves for gene expression related to bird flight [42, 43], hyperbolic curves for gene expression related to photosynthetic reaction [44], etc.

**Nonparametric modeling**

If time-varying expression of genes does not obey an explicit mathematical function, nonparametric approaches, such as kernel estimators or B-splines, can be used [9, 45, 46]. Kernel estimators are based on local polynomial regression, whereas smoothing splines use a piece-wise polynomial function. As shown by Silverman [47], kernel smoothing and smoothing-spline smoothing are asymptotically equivalent for independent data and splines are higher-order kernels. More recently, Legendre orthogonal polynomials (LOP) have been used to model dynamic changes of complex traits that cannot be explicitly described by a mathematical equation [48–51]. Since the LOP are orthogonal to each other and integrate to 0 in the interval $[-1, 1]$, nonparametric estimates derived from this approach display favorable asymptotic properties [52, 53]. The LOP have been successfully used to model time-varying phenotypic or genetic changes for many complex traits, such as milk production [54] and plant height growth [48, 49]. As will be seen from an example below, it should be equally useful for modeling the dynamic pattern of gene expression profiles in a time course.

**Semiparametric modeling**

If gene expression spans multiple distinct stages (see [13]), at some of which the expression values follow a parametric form but at others of which they do not, we can implement a semiparametric model that combines the parsimony and biological relevance of parametric approaches with the flexibility of nonparametric approaches. As explained by Cui et al. [49], such a semiparametric approach was used to model the growth process and death process of tiller number in a lifetime of rice. A similar semiparametric approach can also be implemented in the clustering framework of multistage gene expression dynamics. This will enable us to study dynamic changes of gene expression by relating its temporal profiles from different developmental stages.

**Structural modeling of covariance**

Unstructured estimates of a longitudinal covariance matrix may be highly unstable for large matrices. This, in conjunction with the fact that the covariance among repeated measures over time has an inherent structure [55], implies that structuring a covariance matrix with few parameters may be crucial for parsimonious and efficient parameter estimates in dynamic gene clustering. An extreme of covariance structuring is compound symmetry and autoregression of order one, but this may be far from the true covariance, leading to severe bias. The best covariance estimator should be at the balance between its variance and bias. Below, we list several commonly used approaches for covariance structure.

**Autoregressive moving-average process($p,q$)**

The autoregressive moving-average process, ARMA($p, q$) [56], is flexible to provide a robust estimate of gene expression covariance structure [38]. The zero-mean residual error $e_i(t_i)$ in environment $l$ (1) is generated according to the following process

$$e_i(t_i) = \eta_i^l + \sum_{l=1}^{p} \varphi_i^l e_i(t_i - t_i) + \sum_{l=1}^{q} \theta_i^l \eta_{l-i}^l$$

where $\varphi_1^l, \ldots, \varphi_p^l$ and $\theta_1^l, \ldots, \theta_q^l$ are unknown parameters and $\{\eta_i^l\}$ is a sequence of independent
and identically distributed normal random variables with zero mean and variance $\sigma^2_i$. The ARMA($p,q$) model parameters are arrayed in $\Theta = \{ \phi_0^1, \ldots, \phi_p^s, \theta_0^1, \ldots, \theta_q^s, \sigma^2_1, \ldots, \sigma^2_n \}$. The merit of the ARMA model includes the existence of closed forms for the estimates of the inverse and determinant of the structured covariance matrix [57, 58], which enhances computational efficiency.

By various constraints, the ARMA model can be reduced to a simple autoregressive (AR) model and structured antedependence (SAD) model [59]. Both the first-order AR and first-order SAD models use only two parameters, but the latter is more flexible than the former since the latter allows the variance and correlation to change over time. The SAD model has been successfully incorporated in functional mapping of dynamic complex traits, Yap et al. [60] has successfully incorporated Cressie and Huang’s nonseparable model to estimate the covariance of photosynthetic rate over temperature and irradiance. A similar idea can be developed to use this nonseparable model to fit the spatiotemporal covariance of gene expression.

### Estimation and tests

A hybrid EM-simplex algorithm was implemented to estimate the parameters, $\Theta$, contained in the likelihood (4). The EM algorithm provides a platform for estimating the proportions of different gene clusters, within which the simplex algorithm is embedded to estimate base vectors for each cluster and the covariance-structuring parameters. This can be described as follows:

In the E step, we define and estimate the posterior probabilities of gene $i$, with which it belongs to a particular expression pattern $j$, by

$$
\Omega_{ij} = \frac{\omega_j f_j(y_i; \mu_j, \Sigma)}{\sum_{j=1}^{T} \omega_j f_j(y_i; \mu_j, \Sigma)}.
$$

In the M step, the proportion of expression pattern $j$ is calculated by

$$
\omega_j = \frac{\sum_{i=1}^{n} \Omega_{ij}}{n}.
$$

Mean-covariance structuring parameters in $\Theta_{\mu_j}$ and $\Theta_{\Sigma}$ are estimated in this step, but no closed forms can be derived for their estimators. The EM algorithm

---

**Kernel smoothing**

The kernel smoothing method has been used to estimate longitudinal covariances [61]. The advantage of this method lies in its flexibility to specify any form of covariances and asymptotic properties. Under the homogeneous assumption, the covariance of gene expression between any two time points $t_{l_1}$ and $t_{l_2}$ in environment $l$ is written as a function of time interval, i.e. $\text{Cov}(t_{l_1}, t_{l_2}) = P(t_{l_1} - t_{l_2})$. Kernel smoothing describes this covariance by

$$
f(\frac{t_{l_1} - t_{l_2}}{h}) = \frac{1}{n} \sum_{i=1}^{T} \sum_{l=1}^{T} K \left\{ \frac{y_i(t_{l_1}) - y_i(t_{l_2})}{h} \right\}
$$

where $n$ is the total number of genes, $T$ is the total number of time points, $K(\cdot)$ is a kernel function, $h$ is a bandwidth and $y_i^l$ is the mean of gene $i$, given by $y_i^l = \frac{1}{T} \sum_{t=1}^{T} y_i(t)$. One of the mostly used kernel functions is Gaussian kernel, i.e. $K(d) = \exp(-d^2)$. A variety of statistical methods have been developed to choose an optimal kernel and appropriate bandwidth (see [61]).

**Modeling covariance over time and environment**

In this article, we consider gene expression dynamics in multiple environments. The approaches described above are used to model time-dependent covariances of gene expression separately for each environment. These approaches can be similarly used to model covariances over different environments at each time point. The covariance structure over time and environment is then modeled by taking the product of purely temporal and environmental covariances. This so-called separable approach is simple but has many undesirable properties since it does not allow environment–time interactions. We can implement a nonseparable stationary model (see [62–64]) to structure time–environment covariance of gene expression. A nonseparable covariance is not expressed as a Kronecker product of two matrices like separable structures. An advantage of covariance modeling in this context is in providing a better characterization of the random process to obtain optimal kriging or prediction of unobserved portions of it. In functional mapping of dynamic complex traits, Yap et al. [65] has successfully incorporated Cressie and Huang’s nonseparable model to estimate the covariance of photosynthetic rate over temperature and irradiance. A similar idea can be developed to use this nonseparable model to fit the spatiotemporal covariance of gene expression.
was derived to estimate these parameters when gene expression dynamics is modeled by Fourier series approximation and the covariance modeled by the ARMA process [14, 38]. In general, the simplex algorithm, which does not depend on explicit equations, can be implemented to estimate these structuring parameters. The framework for clustering gene expression dynamics over multiple environments allows geneticists to address many biologically meaningful questions. First, an optimal number of gene clusters in terms of their different expression dynamics over all environments can be determined using AIC or BIC approaches (see the example shown below). Second, we need to determine the optimal number of gene clusters in a specific environment. For two given patterns $j$ and $j'$, they may be identical in an environment $l$, although different for all the $L$ environments. This can be tested by

$$H_0: u_j^l(t) = u_{j'}^l(t)$$ versus

$$H_1: u_j^l(t) \neq u_{j'}^l(t)$$

over $\tau = 1, \ldots, T$ for $j < j' = 1, \ldots, J$ \hspace{1cm} (9)

The $H_0$ states that gene expression over an entire time course is identical for the two gene clusters $j$ and $j'$. If the $H_1$ is rejected, it means that the optimal number of clusters in environment $l$ is equal to, or less than, $J - 1$. By performing this pairwise test for all gene clusters, this approach allows the identification of the optimal number of expression patterns for environment $l$.

Third, we can test the significance of gene–environment interactions. This can be done by testing

$$H_0: u_j^l(t) = u_j^{l'}(t)$$ versus

$$H_1: u_j^l(t) \neq u_j^{l'}(t)$$, over $\tau = 1, \ldots, T$ for $l < l' = 1, \ldots, L$ \hspace{1cm} (10)

The $H_0$ states that the gene expression of a cluster over an entire time course is identical between the two environments $l$ and $l'$. If the $H_0$ is rejected, it means that the expression of gene cluster $j$ displays significant gene–environment interactions. This test provides a quantitative way to study the relationship between genes and environments.

Fourth, we are interested in testing how genes interact with time in a specific environment. For two given clusters $j$ and $j'$, this test is expressed as

$$H_0: u_j^l(t) = c \cdot u_j^l(t)$$ versus

$$H_1: u_j^l(t) \neq c \cdot u_j^l(t)$$, over $\tau = 1, \ldots, T$ for $j < j' = 1, \ldots, L$ \hspace{1cm} (11)

where $c$ is a constant. The $H_0$ states that time-dependent gene expression of cluster $j'$ can be expressed as a linear function of that of cluster $j$ in environment $l$. If the $H_0$ is rejected, it means that this pair of gene clusters has a significant gene–time interaction under this particular environment.

**WORKED EXAMPLE**

**Data analysis**

The new tool is demonstrated by analyzing a data set of microarray genes associated with response to vein bypass grafting designed to treat arterial occlusive disease. The data were obtained using a rabbit bilateral vein graft construct, as previously described by Jiang et al. [25]. New Zealand White rabbits (weighing 3.0–3.5 kg) were treated by bilateral jugular vein interposition grafting and unilateral distal carotid artery branch ligation to create two distinct flows, i.e. two different environments. Through ligation of the internal carotid and three of the four primary branches of the external carotid artery, an immediate 6-fold difference in blood flow between the right and left vein grafts was obtained. A segment of the vein was retained at the time of implantation for baseline morphometric measurements. Vein grafts were harvested at 1, 3, 7, 14, 28, 90 and 180 days after implantation. Expression of 14 958 genes was recorded for each of these time points under both treatments, high flow and low flow. Other parameters related to hemodynamic behavior, such as graft flow rate, intraluminal pressure, mean circumferential wall stress and shear stress, were also measured or estimated.

An initial step is the selection of an appropriate model that fits the dynamic change of gene expression over time. Figure 1 shows the plotting of 10 randomly selected genes expressed over time in the two treatments, from which we found that there is a great variability in gene expression trajectories. Some are curvaceous, while others are quite flat. Thus, we used a flexible nonparametric approach based on LOP to model gene expression dynamics. Let $P_i(t^*) = [P_i(0), P_i(t^*), \ldots, P_i(t^*)_r]$ denote a family of LOP with a particular order $r$ derived from a special differential equation, where $t^*$ is a scaled time with a range $[-1,1]$. Let $u^l_{j'} = [u^l_{j0}, u^l_{j1}, \ldots, u^l_{j_r}]$ denote a vector of base values for cluster $j$ in environment $l$. Then, time-varying mean values for cluster $j$ in environment $l$ in Equation (2) can be expressed as a
linear combination of \( u_{jr} \) weighted by the family of LOP, i.e.

\[ \mu_{jr}(t^*) = P_j(t^*) u_{jr}. \]

Our task now is to estimate the base vector \( u_{jr} \) from the given data. The variance of expression among genes seems to be broadly consistent over time points, suggesting that the first-order AR model may fit the data. To combine the expression data from the two flows, we used a separable model to structure the covariance over time and environment.

In Figure 2, a plot of BIC values is illustrated against varying numbers of gene clusters under different LOP orders, from which we chose eight clusters and three orders that provide a best combination for curve fitting. Implementing this combination, we estimated the expression trajectories of each of the eight gene clusters for both high and low flows.

We need to detect if any of these eight clusters, labeled from A to H (see Supplementary Figure S1), overlap in a flow. Pairwise tests using hypothesis test (9) indicate that no pair of clusters is identical for expression trajectories in each flow (\( P < 0.0001 \)). This result suggests that the optimal number of clusters should be eight for both flows. Table 1 gives the estimated proportions and their standard errors of each cluster from these genes.

By calculating the posterior probabilities of each gene that belongs to different clusters using Equation (7), we can determine the most likely cluster of this gene. Thus, we can draw gene expression trajectories for all genes that belong to a particular cluster A–H, separately for the two flows and the mean trajectory of the cluster for each flow using the estimates of curve parameters (Figure 3). The 95% confidence intervals of each estimated trajectory are generally within the variation of temporal gene expression profiles among individual genes, suggesting that our estimates are reasonably accurate. In general, most individual gene expression trajectories display similar time-varying trends between the two flows, but marked discrepancies in expression trajectories were detected for some particular clusters.

Hypothesis test (10) allows the detection of gene–environment interactions in expression profiles over high and low flows. Table 1 gives the results of significance tests for gene–environment interactions. Except for clusters B, C and E, all other clusters are expressed differently between high and low flows. To better show treatment-dependent expression differences, we draw the mean trajectories of each cluster from high and low flows in the same plot (see Supplementary Figure S2). The expression of cluster A has the highest level right after...
implantation, decreases drastically within 25–35 days and then increases gradually. This cluster displays a more pronounced change of expression over time in high flow than low flow. Cluster B has a similar trend of gene expression profiles although its time-varying change is milder compared to cluster A. It appears that clusters C and D have a minimum level of expression throughout experimental time. Clusters E, F and H are expressed at a low level in the beginning of treatment, reach a peak at Day 50 after implantation. Cluster H changes its expression level over time most abruptly, followed by clusters F and E. The expression of cluster G increases over time monotonously until Days 100–120, after which its expression decreases although it is more striking in low flow than high flow.

Simulation

We performed simulation studies to examine the statistical properties of the clustering model. The expression data were simulated by mimicking the structure of the real data analyzed above. A total of 4800 genes were assumed to include eight different clusters in two environments specified by mean trajectories as shown in Figure 3. These genes each have an expression trajectory over four time points as the sum of the mean trajectory of the underlying cluster and residual errors whose covariance structure follows the first-order AR model, but assuming the value of variance that triples the estimated variance. The proportions of eight clusters in Table 1 were used to simulate gene expression profiles.

Simulated data were analyzed by the model. Based on BIC values, the model can detect the correct number of clusters and the correct order of LOP. The model estimates the proportions of clusters A–H precisely. Also, expression trajectories of each gene cluster can be reasonably well estimated (Figure 4), despite a tripled variance used. This shows that the results from the real data set analyzed by the model are convincing from a statistical point of view.

An additional simulation was conducted to test the power of the model to detect gene–environment interactions and its false positive rates (FPR). Consider clusters A, D, F, G and H obtained from rabbit gene expression data which display a certain level of gene–environment interactions. By repeating the simulation and estimation procedure 100 times, we detected the numbers of cases in which hypothesis tests for gene–environment interactions using test (10) are significant, are 85–97 for clusters A, D, F, G and H, respectively. To test the FPR, we used the same parameters for clusters B, C and E to simulate gene expression data for high and low flows. Of 100 simulation replicates, less than 5 were detected to be significant. This suggests that the FPR of our tool is acceptably low.

DISCUSSION

There is a pressing need for computational tools that can unravel the developmental machinery of time-dependent gene expression profiles, despite a vast body of literature presenting these tools [4–6, 8, 10–12]. One significant lack is the unavailability of models for analyzing gene–environment interactions for gene expression dynamics. A few pioneering studies have demonstrated the capacity of gene transcript abundance to comprehend the genetic architecture of gene–environment interactions [16, 22, 23].

In this article, we describe a computational tool that can test gene–environment interactions on a genomic level using dynamic gene expression data. The developmental dynamics of cells, organs or organisms are related with many fundamental phenomena in biology, such as growth and phenotypic plasticity. Our understanding of how developmental

<table>
<thead>
<tr>
<th>Cluster</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>0.0116</td>
<td>0.0123</td>
<td>0.3354</td>
<td>0.3831</td>
<td>0.1134</td>
<td>0.0359</td>
<td>0.0100</td>
<td>0.0083</td>
</tr>
<tr>
<td>Gene–environment interaction test</td>
<td>$&lt;0.0001$</td>
<td>$&gt;0.100$</td>
<td>$&gt;0.250$</td>
<td>$&lt;0.0001$</td>
<td>$&gt;0.400$</td>
<td>$&lt;0.0001$</td>
<td>$&lt;0.001$</td>
<td>$&lt;0.0001$</td>
</tr>
</tbody>
</table>
Figure 3: Mean expression trajectories (thick purple curve) of each of the eight gene clusters, labeled (A)–(H), over time under high flow and low flow. The two broken curves in each plot are the 95% confidence intervals of each estimated mean trajectory. Expression trajectories of all those genes that belong to a particular cluster, determined by the posterior probabilities, are shown in light blue.
Figure 4: Comparison of estimated (solid) and true (dot) mean expression trajectories for each of the eight clusters, labeled (A)–(H), under two hypothesized treatments, high flow and low flow. The simulated data mimicked the structure of the rabbit data, but assuming an error variance that triples the estimated variance. The 95% confidence intervals (broken curves) of each estimated mean trajectory are given. Expression trajectories of all those genes that belong to a particular cluster are shown in light blue.
Gene expression dynamics are regulated through a balance of gene and environment. This helps to reveal the mechanistic origins of these phenomena. The new tool may provide an important means for analyzing temporal expression data and construct a map of gene–environment interactions. As an example, we used Legendre-based nonparametric fitting to model dynamic changes of gene expression within a mixture-model framework. One advantage of the Legendre approach lies in its flexibility for curve fitting [52], computational efficiency and avoidance of knot choice essential for B-splines. For gene expression data with explicit curves, such as periodic transcriptional profiles, robust mathematical equations with better biological relevance and better parsimony can be used.

Gene–environment interactions are important to understand biology and can now be tested in a quantitative way by the tool presented. By applying it to a real data set of microarray genes associated with response to vein bypass grafting between high and low flows in rabbits, this tool identified eight distinct patterns of expression trajectories in a time course. Each of these patterns may be related with a particular biological function operational in hemodynamic processes. Although many patterns display a similar trend of time-varying expression between high and low blood flows, many of them are essentially different based on hypothesis tests by our tool. A further molecular study may help identify specific biochemical pathways related to each of these different gene expression patterns. In a recent study, Cohen et al. [3] detected the discrepancy of dynamic trajectories for a few proteins, which corresponds to cell death or survival, between seemingly identical cells. The tool presented should gain more quantitative insights into the distinction of seemingly trivial differences in genomic and proteomic dynamic studies.

While modeling gene expression dynamics jointly over time and environment, we assume no interactions between time and environment. Although it facilitates our modeling and computing, this assumption may not be realistic in some cases. A general model should consider the dependence of gene expression profiles in different times and environments by nonseparable covariance structuring approaches (see [62–65] for examples). Different approaches for covariance structure based on nonparametric or semiparametric models should be incorporated and compared using penalized likelihood criteria. In our software package for environment-dependent functional clustering, we implement many of these approaches for users to choose the most parsimonious one given their particular data sets. The computer code for the tool developed is available at Penn State Center for Statistical Genetics web site, http://statgen.psu.edu.

SUPPLEMENTARY DATA
Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points
- An organism adapts to the environment where it grows through the alteration of gene expression dynamics.
- The identification of dynamic gene expression patterns in response to environmental change can facilitate our understanding of the mechanistic base of gene–environment interactions essential for organismic development.
- This article presents a comprehensive computational framework for clustering gene expression dynamics over a range of environments. The framework model is validated through simulation and real data analysis.
- Our goal is to provide interested researchers with the computational tools that are applicable to modeling and analysis of dynamic gene expression data derived from multiple environments.

FUNDING
NIH/NHLBI (grant R01 HL095508); NIDA/NIH (grants R21-DA024260 and P50-DA10075); National Natural Science Foundation of China (grant 11028103). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NHLBI, NIDA, or the NIH.

References


