Gene expression

Bayesian detection of non-sinusoidal periodic patterns in circadian expression data

Darya Chudova1,*, Alexander Ihler1, Kevin K. Lin2, Bogi Andersen2 and Padhraic Smyth1
1Department of Computer Science and 2Departments of Medicine and Biological Chemistry,
University of California, Irvine, CA 92697

Received on June 30, 2009; revised on August 30, 2009; accepted on September 10, 2009
Advance Access publication September 22, 2009
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Cyclical biological processes such as cell division and circadian regulation produce coordinated periodic expression of thousands of genes. Identification of such genes and their expression patterns is a crucial step in discovering underlying regulatory mechanisms. Existing computational methods are biased toward discovering genes that follow sine-wave patterns.

Results: We present an analysis of variance (ANOVA) periodicity detector and its Bayesian extension that can be used to discover periodic transcripts of arbitrary shapes from replicated gene expression profiles. The models are applicable when the profiles are collected at comparable time points for at least two cycles. We provide an empirical Bayes procedure for estimating parameters of the prior distributions and derive closed-form expressions for the posterior probability of periodicity, enabling efficient computation.

The model is applied to two datasets profiling circadian regulation in murine liver and skeletal muscle, revealing a substantial number of previously undetected non-sinusoidal periodic transcripts in each. We also apply quantitative real-time PCR to several highly ranked non-sinusoidal transcripts in liver tissue found by the model, providing independent evidence of circadian regulation of these genes.

Availability: MATLAB software for estimating prior distributions and performing inference is available for download from http://www.datablab.uci.edu/resources/periodicity/.

Contact: dchudova@gmail.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Identifying periodic transcripts in large time course gene expression experiments is an important step in studying diverse biological systems, including the cell cycle, hair growth cycle, mammary cycle and circadian rhythms. The data from these studies are often characterized by a large number of genes with relatively coarse sampling in time (e.g. a few time points per cycle) and only a few measurements at each time point. The objective is to identify or rank which of these genes are most likely to be periodically regulated. In this article, we propose a simple probabilistic mixture model for identifying periodic expression in cyclic processes where cycle length is known a priori and expression levels can be profiled at comparable time points in multiple cycles. Such datasets are generated, for example, in experiments profiling circadian regulation in peripheral tissues (see Miller et al. (2007); Rudic et al. (2005); Storch et al. (2002) among others).

Existing techniques for detecting periodic expression patterns fall into two major categories: time domain and frequency domain analyses. Typical frequency domain methods compute the spectrum of the average expression profile for each probe, and test the significance of the dominant frequency against a suitable null hypothesis such as uncorrelated noise. However, frequency domain analysis is most effective on long time series and is not well suited for short time courses (Tai and Speed, 2007).

In time domain analysis, most methods rely on the identification of sinusoidal expression patterns (Andersson et al., 2006; Straume, 2004; Wijnen et al., 2006). These detectors are popular due to their simplicity and computational efficiency, but are not effective at finding periodic signals which violate the sinusoidal assumption. While this assumption can be appropriate for some data (such as the cell cycle), a significant number of profiles with non-sinusoidal shapes have been identified in the control of hair cycling (Lin et al., 2004) and in the circadian rhythms of Drosophila (Keegan et al., 2007). More general shapes could be modeled using, for example, B-spline representations (Luan and Li, 2004), but such approaches require a set of ‘guide genes’ to define the possible shapes of periodic patterns, which in practice may be unavailable or incomplete.

In this article, we propose a general statistical framework for detecting periodic profiles from time course microarray data by analyzing the similarity of observed profiles across the cycles. Using this framework, we identify a significant number of previously undetected circadianally regulated genes with non-sinusoidal profiles in peripheral mouse tissues. Figure 1 shows examples of expression profiles from the murine liver time course data set (Miller et al., 2007). Profiles shown in this figure were among those ranked most likely to be periodically expressed (in the top 25 profiles) by our proposed approach but were ranked much lower by a more traditional sine-wave detection algorithm (Miller et al., 2007). Notably, two of these probe sets (Nr1d1 and Arntl) correspond to well-established clock-control genes. In addition, circadian...
regulation of Cyp2a4 in liver has been established in Lavery et al. (1999), and Mfn2 has been identified as circadially regulated in liver in an independent microarray study by Oishi et al. (2003). Our quantitative PCR experiments validate circadian cycling for seven out of eight tested genes in this figure,2 demonstrating that these are likely true positives missed by previous analyses (see Section 3). Overall, we detect significant numbers of non-sinusoidal patterns that were missed by the original analyses using existing detection algorithms.

The rest of the article is organized as follows. In the next section, we describe our probabilistic model in detail and describe how it can be used to infer, for each probe set, the probability of its observed expression pattern being periodic. We also describe two simplified versions of the model, a (non-Bayesian) ANOVA test and a simplified Bayesian model which can be implemented using simplified versions of the model, a (non-Bayesian) ANOVA test

Consider a time course experiment that profiles expression of genes, the other corresponds to genes that remain constant throughout the experiment (background genes). To model periodic phenomena, we include an additional third component that encodes coordinated expression across multiple cycles (Fig. 2). Our task of identifying periodicity then reduces to a probabilistic inference problem: given the observed expression profiles, compute the posterior probability that a given probe set was generated by the periodic component.

2.1 A probabilistic model for periodicity

Consider a time course experiment that profiles expression of \( N \) probe sets over \( C \) cycles of known length. Each cycle is represented by the same grid of \( T \) time points, indexed from 1 to \( T \). Profiling is typically done using multiple observations or replicates at a given time point (e.g. 2 or 3) using a cross-sectional study design, i.e. all of the replicates at all of the time points originate from different biological subjects. We denote the number of replicate observations for probe set \( i \in [1, \ldots, N] \) at time point \( j \in [1, \ldots, T] \) of cycle \( c \in [1, \ldots, C] \) by \( n_{ij}^c \). Note that this number may be zero; for example, we may not make any observations at time \( t \) in some cycle \( c \), in which case \( n_{ij}^c \) will be zero for all \( i \). We use \( Y_{ij}^c \) to denote the expression intensity value for a particular probe set \( i \), time point \( j \) and replicate \( k \) for cycle \( c \), and let \( Y_{ij} \) be the entire set of observations
for probe set $i$. We assume that the intensity values $Y_{ik}$ have been estimated from raw data using a standard approach such as that of Wu et al. (2004), log-transformed and shifted to zero mean for each probe set's profile.

Our probabilistic model for expression, then, consists of three components: background ($b$), differentially expressed but aperiodic ($d$) and periodically expressed profiles ($p$). Let $Z_i \in \{b,d,p\}$ denote the component associated with probe set $i$. The forward or generative model is simple: to simulate an expression profile, one selects one of the three components according to their respective probabilities $[\pi_b, \pi_d, \pi_p]$, then samples a collection of observations according to the associated component model. Each of the three component models consists of a Normal/Inverse Gamma (NIG) prior distribution (Gelman et al., 1995) on the latent profile and additional Normal (i.e. Gaussian) noise on the observations. The components differ in the structure of latent profiles and in the parameters of their (NIG) model.

The NIG prior is a flexible and computationally convenient distribution commonly used as a prior model for latent expression levels and replicate variability (e.g. Smyth, 2004; Tai and Speed, 2006, 2007). In general, scalar variables $(\mu, \sigma)$ are distributed as NIG with parameters $(v, \nu, \eta, \alpha, \beta)$ if

$$P(\mu, \sigma) = \text{NIG}(\mu|v, \nu, \eta, \alpha, \beta)$$

where $\text{NIG}(v|\nu, \eta, \alpha, \beta)$ denotes a Gaussian distribution with mean $v$ and variance $\nu$ and $\text{Gamma}(\sigma|\alpha, \beta)$ denotes an inverse Gamma distribution with $\alpha$ degrees of freedom and scale parameter $\beta$, evaluated at $\sigma$.

Note that in what follows, we refer to three types of unknown quantities. The first are the prior parameters, denoted $\Theta$, which we determine via an empirical Bayesian procedure (details later) and are subsequently treated as known and fixed. The other two types are probe set-specific hidden variables: the latent profiles (consisting of a mean and variance) for each component, and the component identity $Z_i$, indicating from which component the data were generated.

### 2.2 Components of the mixture model

Our model is shown as a graphical model using plate notation in Figure 3 (Jordan, 2004). The plates, or rectangles, are used to group according to the associated component model. Each of the three components models consists of a Normal/Inverse Gamma (NIG) prior distribution (Gelman et al., 1995) on the latent profile and additional Normal (i.e. Gaussian) noise on the observations. The components differ in the structure of latent profiles and in the parameters of their (NIG) model.

The NIG prior is a flexible and computationally convenient distribution commonly used as a prior model for latent expression levels and replicate variability (e.g. Smyth, 2004; Tai and Speed, 2006, 2007). In general, scalar variables $(\mu, \sigma)$ are distributed as NIG with parameters $(v, \nu, \eta, \alpha, \beta)$ if

$$P(\mu, \sigma) = \text{NIG}(\mu|v, \nu, \eta, \alpha, \beta)$$

where $\text{NIG}(v|\nu, \eta, \alpha, \beta)$ denotes a Gaussian distribution with mean $v$ and variance $\nu$ and $\text{Gamma}(\sigma|\alpha, \beta)$ denotes an inverse Gamma distribution with $\alpha$ degrees of freedom and scale parameter $\beta$, evaluated at $\sigma$.

Note that in what follows, we refer to three types of unknown quantities. The first are the prior parameters, denoted $\Theta$, which we determine via an empirical Bayesian procedure (details later) and are subsequently treated as known and fixed. The other two types are probe set-specific hidden variables: the latent profiles (consisting of a mean and variance) for each component, and the component identity $Z_i$, indicating from which component the data were generated.

#### 2.2.1 The background component model

We model ‘background’ probe sets as having a constant expression over the experiment (denoted by $\mu_b$), with small fluctuations in the actual observations due to technical errors (variance $\sigma_b^2$). These variables are given a NIG prior shared by all background probe sets and parameterized by four scalars $\Theta_b = \{\mu_b, \sigma_b^2, \phi_b, \psi_b\}$.

---

Figure 3. A graphical model describing the observed profiles $Y$ and latent (unobserved) variables $Z$ (component identity) and $(\mu, \sigma)$ for each component using plate notation.

Since $\mu_b$ and $\sigma_b^2$ are shared across time, they are shown outside the cycle and time plates in Figure 3. The observations $Y_i$ are modeled as independent samples from a Gaussian distribution with mean and variance $(\mu_i^b, \sigma_i^b)$:

$$P(Y_i|\mu_i^b, \sigma_i^b, Z_i = b) = \prod_{c=1}^C \prod_{t=1}^T \text{N}(Y_{ict}|\mu_i^b, \sigma_i^b)$$

where the products are over the $C$ cycles, the $T$ time points within cycle $c$, and the observed replicate expression measurements for time point $i$ in cycle $c$, respectively.

#### 2.2.2 The differentially expressed component model

For differentially expressed genes, the true expression levels vary as a function of time. Accordingly, we let $\mu_d$ and $\sigma_d^2$ be $(C \times T)$-dimensional vectors characterizing the expression value and replicate variance at each of the time points. These variables are shown inside the cycle and time plates in Figure 3. We let the expression at each time point vary independently from the other time points, so that the prior distribution for this component is defined by four $(C \times T)$-dimensional parameters, $\Theta_d = \{\mu_d, \sigma_d^2, \phi_d, \psi_d\}$:

$$P(Y_i|\mu_i^d, \sigma_i^d, Z_i = d) = \prod_{c=1}^C \prod_{t=1}^T \text{N}(Y_{ict}|\mu_i^d, \sigma_i^d)$$

The independence assumption works well for relatively sparse sampling of the time axis, a common situation with expression data measurements in practise. Since the replicates are assumed to originate from different experimental units (cross-sectional design), we model observations as being independent given $(\mu_i^d, \sigma_i^d)$:

$$P(Y_i|\mu_i^d, \sigma_i^d, Z_i = d) = \prod_{c=1}^C \prod_{t=1}^T \text{N}(Y_{ict}|\mu_i^d, \sigma_i^d)$$

#### 2.2.3 The periodic component model

The periodic component assumes repeated expression of the same pattern across multiple cycles. The true, latent expression level at a single time point gives

---

1For more densely sampled data, one could extend this approach by adding dependency between the means, for example, by introducing covariance structure into the prior.
Bayesian detection of periodicity in circadian data

rise to the observed intensities in cycles 1 through C. We let \( \mu_p^i \) and \( \sigma_p^i \) be T-dimensional variables encoding expression levels and replicate variability in the ‘ideal’ cycle. These variables are shown inside the time plate but outside the cycle plate in Figure 3. Assuming sparsity of the time grid, we use independent NIG priors for each time point:

\[
P(\mu_p^i, \sigma_p^i | \theta_p) = \prod_{j=1}^T \text{NIG}(\mu_p^i, \sigma_p^i | \mu_p^i, \sigma_p^i, \nu_p^i, \theta_p^i)
\]

The periodic component is parameterized by four T-dimensional parameters \( \theta_p = \{\nu_p, \mu_p, \sigma_p, \theta_p^i\} \). Due to the cross-sectional study design, we again assume conditional independence of observations:

\[
P(Y_i | \mu_p^i, \sigma_p^i, Z_i = p) = \prod_{c=1}^C \prod_{j=1}^T \prod_{k=1}^{N_c} \text{N}(Y_{ij}^c | \mu_p^i, \sigma_p^i, \nu_p^i, \theta_p^i)
\]

The complete set of prior parameters \( \Theta \) includes the prior component probabilities \( \pi_z \) (corresponding to the relative frequencies of background, differentially expressed, and periodic probe sets), and prior parameters for each of the component models: \( \Theta = \{(\pi_z, \Theta_z), z \in \{b, d, p\}\} \).

2.3 Inference

Given the model, we can detect periodic expression by computing the posterior probability of the periodic component \( p[Z_i = p | Y_i, \Theta] \) conditioned on the prior parameters \( \Theta \) and the observed profile \( Y_i \):

\[
P(Z_i = p | Y_i, \Theta) = \frac{\pi_p P(Y_i | \theta_p, Z_i = p)}{\sum_{z \in \{b, d, p\}} \pi_z P(Y_i | \theta_z, Z_i = z)}
\]

Each of the three marginal likelihood terms in the denominator, for \( z \in \{b, d, p\} \), is computed by averaging over our uncertainty about the latent profiles \( \mu \) and replicate variances \( \sigma \). Since the priors for \( (\mu, \sigma) \) are conjugate to the Gaussian likelihood of \( Y_i \), the marginal likelihood can be computed in closed form as shown in Section A of the Supplementary Material.

2.4 An analysis of variance periodicity detector

Our Gaussian mixture model, and its resulting inferential test for periodicity, is quite close to a simplified, non-Bayesian test based on analysis of variance (ANOVA). We can construct a one-way ANOVA test for periodicity by dividing the data into groups, or factor levels, by their associated time point regardless of cycle number, so that all replicates \( Y_{ijk}^c \) for \( c = 1, \ldots, C \) and \( k = 1, \ldots, \nu_p^i \) fall into the same group. We then test whether the data support separation into these groups, i.e. whether the amount of variation between groups is significantly larger than the variation found within the groups. High values of the ratio of these quantities indicates that most of the variability in observations can be explained using a time-dependent, cycle-independent profile, i.e. that the profile appears periodic.

Like our Bayesian test, the ANOVA test has a number of desirable properties; for example, it considers both similarity among the raw replicate observations and the magnitude of overall changes (the average profile) over time. Both quantities are important—replicate variability is useful in assessing similarity among cycles relative to inherent biological variability, while the magnitude of change helps differentiate signals from random noise. The ANOVA test is also easy to implement using any standard statistical package.

However, there are also a number of disadvantages to the ANOVA test. For it to work as expected, we require a balanced experiment design in which the number of replicates is unchanged over time (\( |\nu_p^i| = n_i \)). It implicitly assumes that the data are Gaussian, with equal variance among the groups (i.e. over time). One can view our model as a Bayesian extension of the ANOVA test: both approaches discriminate based on the amount of variance in the data under models of different complexity, but the Bayesian model relaxes the assumption of equal variances over time and adds a prior term which regularizes the variance estimates when there are few data. Moreover, it can handle a variable number of replicates at each time—an important feature when the data may suffer from missing observations or insufficient replication at certain time points.

2.5 Estimating parameters of the prior distribution

Following Newton et al. (2004), Smyth (2004) and Tai and Speed (2006, 2007), we develop an empirical Bayes procedure to determine the prior parameters \( \Theta \) for our model. First we determine a tentative assignment of probe sets to each component, then use this assignment to find approximate maximum likelihood estimates of the location scale \( \eta \) and parameters of the inverse Gamma distribution \( (a, b) \); we set the location mean \( v \) to 0 in all three components.

To find a tentative initial assignment of probe sets for estimating prior parameters, we run one-way ANOVA detectors of differential expression and periodicity. Probe sets that vary significantly over time according to the first test \( (P < 0.01) \) are used to define parameters of the component for differential expression, while probe sets which fail this test \( (P > 0.1) \) are used to define the parameters of the background component. Similarly, we use the described ANOVA periodicity detector to identify probe sets for estimating the prior parameters of the periodic component. Choosing those probe sets with \( P < 0.001 \) results in a number of probe sets similar to that previously identified in the literature (Miller et al., 2007). The prior component probabilities \( \pi \) are set to the fraction of probe sets that were assigned to each component using this procedure.

The other parameters are then determined using a greedy maximum likelihood method. Briefly, the inverse Gamma parameters \( (a, b) \) are chosen to maximize the likelihood of the observed sums of squared deviations under an F-distribution. After the parameters \( a \) and \( b \) are fixed, \( \eta \) is chosen to maximize the likelihood of the observations \( Y \) under a NIG prior. While the resulting estimates do not necessarily maximize the joint likelihood with respect to \( a, b \) and \( \eta \), due to the two-step nature of the procedure, these estimates are fast to compute and we have found them to work well in practice. More details on this estimation process can be found in Appendices B1 and C1.

2.6 Implementation via Tai and Speed’s framework

We note in passing that a Bayesian model similar to our own can be implemented using the framework of Tai and Speed (2007) and the timecourse package in Bioconductor. Like the ANOVA test, we use only two hypotheses: periodic versus background, and again group together all replicates from the same relative time point regardless of cycle. We then apply the test from Tai and Speed (2007) for analysis of differential expression in one-sample cross-sectional experiments to the grouped data. Any aperiodic yet differentially expressed
signals should have high ‘in-group’ variation due to combining data across cycles, causing only periodic profiles to be ranked highly.

We believe this technique is less intellectually satisfying than our three-component Bayesian model, since it groups two sets of apparently different behaviors (background and periodic differential expression) under a single Gaussian model. However, in empirical comparisons, the methods often behaved similarly, and both models provide useful alternatives to traditional analyses that rely on identifying sinusoidal expression changes.

3 EXPERIMENTAL RESULTS

In this section, we demonstrate that our model can effectively identify both sinusoidal and non-sinusoidal periodic expression patterns in datasets profiling circadian expression in peripheral tissues, including the automatic discovery of genes which were not previously known to exhibit circadian patterns. It is widely believed that 5–10% of transcribed genes in these tissues may be under circadian regulation (Storch et al., 2004), log-transformed and normalized to zero-mean for each of the independent biological replicates were sampled every 4 h for two complete circadian cycles in wild-type mice, and every 4 h for a single circadian cycle (7 time points) in the Clock mutant. The raw intensity values were preprocessed using gcRMA software (Wu et al., 2004), log-transformed and normalized to zero-mean for each of the wild-type profiles.

Sine-wave detection: the original analysis of this data (Miller et al., 2007) used the sine-wave matching algorithm of Straume (2004). They identified 848 distinct rhythmic probe sets in liver and 383 such probe sets in skeletal muscle. The authors filtered out probe sets below a threshold value of intensity, resulting in a final ranked list of 714 probe sets in the liver and 252 probe sets in the skeletal muscle. A subsequent analysis of the skeletal muscle data using the same sine-wave matching algorithm but with a more stringent cut-off threshold resulted in 215 probe sets (McCarthy et al., 2007).

Model-based detection: using our model we ranked the probe sets by their posterior probability of belonging to the periodic component (Section A, Supplementary Material). The posterior probabilities inferred for each of the probe sets are available in the Supplementary Material. Among the top 25 probe sets there are nine that were not among the top 400 ranked by sine-wave matching. Many of their profiles (Fig. 1) peak or drop at a single time point, and are poorly matched to a sinusoidal shape. The fact that two of these are known core clock genes (Arntl and Nr1d1), suggests that such non-smooth measurements may be observed in true circadian genes due to the sparse sampling in time. The reverse list of probe sets, those ranked above 25 by the sine-wave method but below 400 by the model, contains just the single probe set Tns3. The profile conforms to the sine-wave pattern, but possesses a very small amplitude, and is assigned to the background component by the model. All of the other probe sets that were so highly ranked by the sine-wave method received posterior probabilities of periodicity >0.9 from our model.

PCR validation: we used quantitative real-time PCR to estimate fold changes over time of the nine probe sets with known gene identities from the combined difference sets. Eight of these genes correspond to probe sets ranked highly by our model but not by the sine-wave method, and the ninth (Tns3) was the gene ranked highly by the sine-wave method but not by our method. Details of the PCR experiment are described in Section D of the Supplementary Material.

Figure 4 shows estimated log-fold change at each of the 12 time points covering two complete circadian cycles. The ordering of genes in the panels is the same as in Figure 1, except that the gene Tubb2 (which was unidentified at the time we performed PCR) is replaced with the Tns3 gene. The PCR results for Tns3 indicate that the signal-to-noise ratio is smaller than 1: the variance of its mean profile over time (0.014) is smaller than the average replicate variability (0.0192) and thus quantitative PCR does not support circadian changes in this gene. This example illustrates how
an explicit background model can use replicate variability to filter out noisy profiles that may appear periodic to methods that do not weigh the magnitude of changes in the averaged profiles against the variability of the replicates.

In contrast, all of the genes identified by the model except for Zfp929 show profiles consistent with circadian regulation. They change significantly over time and the changes are consistent across the cycles. P-values (from an ANOVA periodicity detector) for these seven profiles are \( <2.13 \times 10^{-6} \); the largest value corresponds to Rnas4. The profile of Zfp929 shows substantially smaller variation over time than the other seven genes, and little similarity across the cycles (\( P = 0.082 \)). In the microarray experiment, this gene peaks at a single time point within each cycle (Fig. 1) and may be an example of a false positive arising from the random background process.

FDR analysis: we estimate the false discovery rate (FDR) to characterize the number of false positive probe sets that exceed a particular threshold on the posterior probability of periodicity. Assuming a correct model, the FDR for a given threshold can be estimated directly as the average of the posterior probability of non-periodicity, taken over probe sets above the threshold (Newton et al., 2004). A threshold of 0.9 selects 468 probe sets in the liver and 97 in the skeletal muscle corresponding to an estimated FDR of 2.28% and 2.23%, respectively.

However, this estimate of the FDR is likely to be optimistic, since it assumes a correct model. As an alternative, we can estimate the FDR using a permutation test. We simulate data from the background distribution by permuting the time labels of our original data within each cycle. This permutation removes correlations in time, but preserves the overall magnitude and observed replicate variability. Both the original and permuted data are then scored under the model. The FDR estimate is defined as the ratio of the number of permuted probe sets that exceed the posterior threshold to the number of the original probe sets that exceed the same threshold (Keegan et al., 2007; see Section E in the Supplementary Material). As expected, the FDR estimates based on time-permuted data are higher than those computed directly using the posterior probabilities, and suggests that for our threshold of 0.9 we can expect an FDR of approximately 14% (liver) and 17% (skeletal muscle). These rates are consistent with the PCR findings in the previous section, where there is evidence that one out of eight detections from our model is a false positive and the other seven are likely to be true positives. That these rates are not closer to zero may be due to the short and sparse nature of the time course datasets.

Comparison with mutant time series data: to further validate the rankings from our model, we evaluate the influence of the Clock mutation on the top ranking profiles. Our hypothesis is that a large fraction of genes under circadian regulation will change their expression patterns in response to the mutation. We perform a two-sample comparison between wild-type and mutant time series, using the Bioconductor timecourse package (Tai and Speed, 2007) and comparing mutant time points 1–7 to their corresponding circadian times in the wild-type (time points 2–8). In this analysis, we do not normalize the individual time courses to zero mean, so that a shift in absolute intensity can be detected as well.

Each plot in Figure 5 shows how many probe sets ranked in the top k by periodicity are also in the top 5% ranked according to differential expression in mutants, in the liver (left) and skeletal muscle (right). Our model consistently selects more probe sets with altered expression patterns between the wild type and the mutant than the sine-wave method. Since temporal profiles of non-rhythmic genes are also affected by the mutation (McCarthy et al., 2007; Miller et al., 2007), this evaluation should be interpreted with caution. Nonetheless, in the absence of ground truth, these results provide additional (albeit indirect) evidence to indicate that the Bayesian model is able to consistently extract more relevant information from the data than a sine-wave approach.

4 DISCUSSION

Our Bayesian model for detecting periodic expression has a number of inherent simplifying assumptions which ensure a fast estimation process. Primarily, these assumptions are:

- All probe sets are independent.
- Latent expression profiles (\( \mu, \sigma \)) are Gaussian and independent across time, except as constrained by the component type.
- Replicate measurements are Gaussian given the latent profile.

However, there are a number of possible extensions to the model which could lead to more robust detection, at the expense of increased computational costs.

Distributional assumptions: conjugate prior distributions such as the NIG form assumed here ensure closed-form computation. However, some authors have suggested that non-Gaussian forms such as Gamma–Gamma models are more appropriate for expression data (Lewin et al., 2007; Newton et al., 2004). Our model can be easily recast with alternative priors, but may require numerical approximations in the computation of posterior probabilities.

Dependence across time: The assumption of a sinusoidal shape regularizes (or smoothens) the estimated true profiles of periodic patterns. In contrast, the periodic component in our model does not have any such regularization (it treats sequential time points independently). Adding a non-diagonal covariance structure (correlation in time) such as that in Tai and Speed (2007) might increase the specificity of the model in detecting periodic probe sets with lower magnitude changes. For very sparsely sampled time points such as those in our datasets, however, this seems to be unnecessary.

Shared expression patterns: computation is greatly simplified by assuming that all genes are independent, but many genes share
similar patterns of expression (Do et al., 2005). Including a higher level mixture model which groups similar periodic profiles together could help identify weak patterns that appear in many expression profiles by sharing information across genes. Although this change is easy in principle, it greatly complicates the inference process. In this case, the estimates of periodicity for each gene become coupled and must be computed jointly rather than individually, and requires more complex methods such as Markov chain Monte Carlo.

5 CONCLUSIONS

In this article, we present an alternative to sinusoid or frequency-based testing for identifying periodic patterns in gene expression time series data. We argue that in typical experiments with only a small number of samples per cycle, we should test for arbitrary patterns which are repeated between cycles, rather than parametric shapes. To this end, we propose a Bayesian mixture model for identifying patterns of unconstrained shape, which stand out as shapes. To this end, we propose a Bayesian mixture model for identifying periodically expressed genes based on time course microarray gene expression profile data by using replicate variance. Proc. Natl Acad. Sci. USA, 101, 13955–13960.


