BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 182–188
doi:10.1093/bioinformatics/bts672

Gene expression

Advance Access publication November 19, 2012

An empirical Bayes approach for analysis of diverse periodic trends in time-course gene expression data

Mehmet Kocak[1,2,*], E. Olusegun George[3], Saumyadipta Pyne[4,5] and Stanley Pounds[2]

[1]Department of Preventive Medicine, University of Tennessee Health Sciences Center, Memphis, TN 38105, USA, [2]Department of Biostatistics, St Jude Children's Research Hospital, Memphis, TN 38105, USA, [3]Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152, USA, [4]C.R. Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad, Andhra Pradesh 500046, India and [5]Broad Institute of MIT and Harvard University, Cambridge, MA 02142, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: There is a substantial body of works in the biology literature that seeks to characterize the cyclic behavior of genes during cell division. Gene expression microarrays made it possible to measure the expression profiles of thousands of genes simultaneously in time-course experiments to assess changes in the expression levels of genes over time. In this context, the commonly used procedures for testing include the permutation test by de Lichtenberg *et al.* and the Fisher's *G*-test, both of which are designed to evaluate periodicity against noise. However, it is possible that a gene of interest may have expression that is neither cyclic nor just noise. Thus, there is a need for a new test for periodicity that can identify cyclic patterns against not only noise but also other non-cyclic patterns such as linear, quadratic or higher order polynomial patterns.

Results: To address this weakness, we have introduced an empirical Bayes approach to test for periodicity and compare its performance in terms of sensitivity and specificity with that of the permutation test and Fisher's *G*-test through extensive simulations and by application to a set of time-course experiments on the *Schizosaccharomyces pombe* cell-cycle gene expression. We use 'conserved' and 'cycling' genes by Lu *et al.* to assess the sensitivity and CESR genes by Chen *et al.* to assess the specificity of our new empirical Bayes method.

Availability and implementation: The SAS Macro for our empirical Bayes test for periodicity is included in the supplementary materials along with a sample run of the MACRO program.

Contact: mkocak1@uthsc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 20, 2012; revised on October 31, 2012; accepted on November 12, 2012

1 INTRODUCTION

There is a substantial body of works in the biology literature that seeks to characterize the cyclic behavior of genes during cell division. In cancer cells, genomic instability can arise when cell division becomes irregular and identifying genes that have cyclic behavior during cell division may add to the understanding of the biological process of these cells and open doors to targeted therapies. Gene expression microarray technology made it possible to measure the expression levels of thousands of genes simultaneously and open ways to conducting time-course experiments which aim at describing the change in the expression levels of genes during successive phases of cell division. Of the several methods proposed for testing periodicity in time-course gene expression profiles, the two most commonly used methods are Fisher's *G*-test by Fisher (1929) and the permutation test described by de Lichtenberg *et al.* (2005a, b), both of which we briefly describe in Sections 2.1 and 2.2.

Futschik and Herzel (2008) provide comparative analysis of methods for assessing periodicity on time-course gene expression data on yeast cells. They argue that analyses based on commonly used methods might have resulted in an overestimation of the number of cyclic genes as quite a large number of genes were identified as periodic by these methods and the overlap among the resulting list of genes is poor. This finding motivates our investigation to construct a new periodicity test that can more accurately distinguish periodicity from not only noise but also from other 'non-periodic' patterns.

Toward this, we present a new test for periodicity that uses a polynomial mode as the null hypothesis against a pattern of expression modeled by a flexible family of periodic functions. We use an empirical Bayes approach to estimate the parameters of the periodic mode. We then compare the performance of our approach with that of the Fisher's *G*-test and the permutation test through extensive simulations. We apply our method to 10 publicly available gene expression time-course experiments (Supplementary Table S1) conducted by Rustici, Olivia and Peng on the fission yeast, *Schizosaccharomyces pombe* (Oliva *et al.* 2005; Peng *et al.* 2005; Rustici *et al.* 2004).

In Section 2, we describe the Fisher's *G*- and the permutation tests followed in Section 3 by an introduction to our empirical Bayes procedure for testing for periodicity and its algorithm. In Section 4, we describe simulation setups to compare the three methods and present the results of these simulations in Section 5. In Section 6, we apply our method to the *S.pombe* time-course gene expression. We end in Section 1 with a conclusion and some discussions in Section 7.

*To whom correspondence should be addressed.

2 METHODS

2.1 The Fisher's *G*-test

A detailed description of Fisher's *G*-test proposed by Fisher in 1929 was given by Wichert *et al.* in 2004. The method is based on the periodogram spectral estimator defined as

$$I(w_k) = \frac{1}{N}\left|\sum_{n=1}^{N} y_n e^{-iw_k n}\right|^2 = \frac{1}{N}\left|\sum_{n=1}^{N} y_n[\cos(w_k n) + i\sin(w_k n)]\right|^2$$

where N is the time series length, y_n is the gene expression level at time n, $n = 1, \ldots, N$, $w_k = 2\pi k/N, k = 1, \ldots, a$ and α is the integer part of $(N-1)/2$. Based on the periodogram evaluated at w_k's, the Fisher's *G*-statistic for periodicity is given by

$$G = \frac{\max_{1\le k\le a} I(w_k)}{\sum_{k=1}^{a} I(w_k)}$$

for which Fisher also provided an expression for calculating the exact distribution as follows:

$$P(G>x) = a(1-x)^{a-1} - \frac{a(1-a)}{2}(1-2x)^{a-1} + \cdots + (-1)^b \frac{a!}{b!(a-b)!}(1-bx)^{a-1}$$

where b is the largest integer less than $1/x$ and x is the observed value of the *G*-statistic. Small *P*-values indicate evidence for a periodic pattern. The '*fisher.g.test*' function in *GeneCycle or GeneTS* libraries developed by Ahdesmaki, Fokianos and Strimmer (2009) in R is a very efficient tool for computing the exact *P*-values for Fisher's *G*-test provided in the Supplementary Materials. We have also written a SAS macro to perform the Fisher's *G*-test.

Despite its popularity and computational efficiency, Fisher's *G*-test has the following deficiencies:

- It tests for periodic pattern against noise.
- It ignores the actual experimentation time points and uses the rank order of the times which reduces its sensitivity when experiment times are not equally spaced. This is shown through simulations in Section 5.
- It ignores information on interdivision times and is evaluated at Fourier frequencies; thus, if the time-course data are not close to complete cycles, the sensitivity of Fisher's *G*-test to periodicity decreases as we show through simulations in this article.

2.2 The permutation test

de Lichtenberg *et al.* (2005a) gives details for the following permutation test. Specifically, let

$$F_g = \sqrt{\left(\sum_{k=1}^{n_g} \sin\left(2\pi\frac{t_k}{T}\right) y_g(t_k)\right)^2 + \left(\sum_{k=1}^{n_g} \cos\left(2\pi\frac{t_k}{T}\right) y_g(t_k)\right)^2}$$

denote the Fourier score at time t_k for a gene g, where t_k is the time at which the gene expression $y_g(t_k)$ is measured, $k = 1,2,\ldots,n_g$, and T is the estimated interdivision time. To test whether or not gene g has a periodic pattern of expression over time, the Fourier score F_g calculated from the actual time-course data for gene g is compared with Fourier scores calculated based on N artificial profiles, say $N = 100\,000$, generated by randomly shuffling the time points while keeping the expression values intact or vice versa. As can be expected, de Lichtenberg's permutation test is not as computationally efficient as the Fisher's *G*-test. In this article, we shall refer to de Lichtenberg's permutation test as the permutation test.

Similar to the Fisher's *G*-test, the permutation test is designed to test for periodicity against noise. Thus, when a time-course data results in a significantly small *P*-value from the Fisher's *G*-test or from the permutation test, it suggests that there may exist in the data expression profiles white noise. However, periodicity is just one of the possible patterns of gene expression and it should be distinguished from other expression profiles. This observation motivates us to seek a novel method for testing for periodicity using an empirical Bayes approach.

3 AN EMPIRICAL BAYES APPROACH FOR TESTING FOR PERIODICITY

A real-valued function $f(t)$ is a periodic function with period T if $f(t) = f(t+kT)$ for all real t and for any integer k.

In the microarray literature, the most commonly referred periodic functions are trigonometric functions, specifically sine or cosine functions, as these functions are defined on a unit circle. The most commonly used tests, the Fisher's *G*- and the permutation tests, are constructed based on these trigonometric functions. In this article, we build a more flexible class of periodic functions for modeling periodic expressions by using these trigonometric functions as building blocks. Any continuous function $f(\cdot)$ defined on the interval $[0,T]$, with period T, can be used as a periodic function by extending its range beyond $[0,T]$ for any x by requiring $f(x) = f(x \ mod \ T)$. In our approach, we construct periodic functions with the *sine* function as the building block.

Let $y_{a,1}, y_{a,2}, \ldots, y_{a,n}$ denote the observed gene expression values for a given gene at experimentation time points $t_{a,1}, t_{a,2}, \ldots, t_{a,n}$, respectively. Let T denote the interdivision time. In the interest of notational and computational ease, we transform the experimentation times to 'cycle times' by $t_j = t_{a,j}/T, j = 1, 2, \ldots, n$ and the corresponding gene expression values to y_j, where

$$y_j = \frac{y_{a,j} - \min(y_{a,1}, y_{a,2}, \ldots, y_{a,n})}{\text{range}(y_{a,1}, y_{a,2}, \ldots, y_{a,n})}, j = 1, 2, \ldots, n$$

With $t_j^* = t_j \bmod T, j = 1, 2, \ldots, n$, this process converts data values to triplets (y_j, t_j, t_j^*).

We start by assuming a periodic model based on the sin(+) function:

$$y_j = \alpha_0 + 0.5\sin(2\pi(t_j + \alpha_1)) + \varepsilon_j, j = 1, 2, \ldots, n, \qquad (1)$$

where $t_j, j = 1, 2, \ldots, n$ are cycle times, defined above, and $\{\varepsilon_j\}_{j=1}^n$ are independent and identically normally distributed with mean = 0 and variance σ_ε^2. As the range of sin() is [−1,1], with the length of 2 units, we multiply the sin(+) term in (1) by 0.5 to reduce the range to 1 unit as y_j's are defined on the interval [0,1] with a length of 1 unit. The parameter α_1 in (1) represents the zero of the sin(+) function in (1). The intercept parameter is self-explanatory.

To expand the family of periodic patterns formed by the model (1), we introduce a modification inside the sin(+) function as

$$y_j = \alpha_0(1 + t_j^*) + 0.5\sin(2\pi(t_j + \alpha_1 t_j^*)) + \varepsilon_j, j = 1, 2, \ldots, n, \quad (2)$$

where $t^* = t \ mod \ 1$. The addition of $\alpha_1 t_j^*$ in (2) changes the underlying period 'within' each cycle, creating multiple peaks by reducing the period or flattening the underlying shape of the periodic function through increasing the period. When the effect of $\alpha_0(1 + t_j^*)$ is added to the model, the periodic structure within

each cycle is broken again as $\alpha_0(1+t_j^*)$ changes the amplitude of each sub-cycle differently at a given value of t_j^*, thus preserving the underlying period as shown in Supplementary Figures S1 and S2.

Further gain in horizontal and vertical flexibility of the regression model in (2) can be obtained by introducing another multiplicative term to the model as

$$y_j = [\alpha_0(1+t_j^*) + 0.5\sin(2\pi(t_j+\alpha_1 t_j^*))](1+\alpha_2 t_j^*) + \varepsilon_j, \quad j = 1, 2, \ldots, n, \tag{3}$$

where the multiplicative term $(1+\alpha_2 t_j^*)$ stretches the extrema of the periodic function vertically as necessary within each cycle. Supplementary Figures S1 and S2 illustrate the increase flexibility introduced by the two modifications proposed above.

For estimation of the parameters in the models introduced above, we use an empirical Bayes procedure. This facilitates development of tests for periodicity against an alternative that does not represent just noise, as has been customarily done in existing publications based on the Fisher's *G-test* and the permutation test for periodicity.

To generate a family of models under which the distribution of gene expression will be assumed to be non-periodic, we limit all models under consideration to those that can be embedded into a power series regression model of the form

$$y_j = \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + \beta_3 t_j^3 + \beta_4 t_j^4 + \cdots + \varepsilon_j. \tag{4}$$

Considering the data limitations, the number of parameters that can be fitted with a polynomial regression model is limited. However, polynomial functions contain white noise as a trivial case and belong to a much wider class of non-periodic functions for comparison against the periodic model specified by (3). In utilizing such models, higher order polynomials can be selected as more and more cycles of data become available.

We have two competing models; the periodic model described in Equation (3), which is flexible enough to capture a wide range of periodic patterns, and the polynomial regression model shown in (4) which includes constant (i.e. white noise), linear, quadratic, cubic and higher order patterns that are not periodic will be generated.

For constructing a test procedure, one possible approach is to use a likelihood ratio test. This is possible with a family of parameterized non-periodic and periodic functions. Non-parametric alternatives would involve the use of splines. However, to increase the utility of the periodic model, we do not want to treat the interdivision time as fixed since it is just an estimate with some expected variability. Instead, we reconstruct the above periodic model in an empirical Bayes setting where we can define a prior distribution for the interdivision time using its estimate. We then assess the goodness-of-fit of the periodic model based on posterior realizations of the model parameters and assess the goodness-of-fit of the polynomial regression model based on the least-squares estimates of its parameters. The algorithm for this procedure is outlined below.

3.1 Algorithm for the empirical Bayes test of periodicity

Step 1 (Data transformation): Without loss of generality, transform the original data as described above to obtain the data triplets $(y_j, t_j, t_j^*), j = 1, 2, \ldots, n$, where y_j is the transformed gene expression values onto [0,1] interval, t_j is the cycle time based on the estimated interdivision time and t_j^* is the remainder of t_j/T.

Step 2 (Prior specification and construction of the Bayesian model): Assume that $y_j \sim N(\mu(t_j), \sigma_\varepsilon^2)$, where

$$\mu(t_j) = (\alpha_0(1+t_j^*) + 0.5\sin(2\pi(t_j+\alpha_1 t_j^*)/\alpha_{\mathrm{IDT}}))(1+\alpha_2 t_j^*)$$

For each of the parameters in this periodic model, we introduce the following prior distributions:

- $\alpha_k \sim N(\tilde{\alpha}_k, \tilde{\sigma}_k^2), k = 0, 1, 2,$ where $(\tilde{\alpha}_k, \tilde{\sigma}_k^2)$ is the non-linear least squares estimate of $\alpha_k, k = 0, 1, 2,$ and its variance multiplied by 100, using the nonlinear periodic regression model by fixing the interdivision time in the model as 1.0 as $y_j = (\alpha_0(1+t_j^*) + 0.5\sin(2\pi(t_j+\alpha_1 t_j^*)/1))(1+\alpha_2 t_j^*) + \varepsilon_j$,
- $\alpha_{IDT} \sim \mathrm{Gamma}(1/0.0001 + 1, 0.0001)$,
- $\sigma_\varepsilon^2 \sim \mathrm{InverseGamma}(100/\tilde{\sigma}_y^2 + 1, 100)$, where $\tilde{\sigma}_y^2$ is the mean squared error obtained from the non-linear fit of the periodic model, $y_j = (\alpha_0(1+t_j^*) + 0.5\sin(2\pi(t_j+\alpha_1 t_j^*)/1))(1+\alpha_2 t_j^*) + \varepsilon_j$, where $j = 1,2,\ldots,n$, using least-squares approach for the gene under investigation.

We note that we have used a highly informative prior for α_{IDT} because the experimentation times have been converted into 'cycle times' using the estimated interdivision time. For this purpose, we assume a small variance for the interdivision time and express this by assigning a small variance to the prior α_{IDT}. Specifically, we assign a gamma prior with the mode 1.0. We show through simulations as illustrated in Supplementary Figure S3, where we compared the sensitivity and specificity under randomly generated time-course samples from five periodic and five non-periodic patterns which will be discussed in detail in Section 5, and using a given choice of the scale parameters of 0.000001, 0.0001, 0.01 and 1.0, that the above choice of 0.0001 for the scale parameters for the gamma distribution for α_{IDT} has excellent operating characteristics. However, if prior information regarding the variability of the estimated interdivision time is available, then the researcher is encouraged to define the prior distribution of α_{IDT} based on this prior information accordingly.

For the prior distribution for σ_ε^2, an inverse gamma distribution with parameters estimated by the least square fit of the periodic model was used. Specifically, it was found that a prior with mode equal to the variance of the residuals obtained from the non-linear fit of the periodic model using least-squares approach which is inflated by 100 has excellent convergence characteristics as shown in Supplementary Figure S4, where we compare the sensitivity and specificity under randomly generated time-course samples from five periodic and five non-periodic patterns, which will be discussed in details in Section 5 and using a set of scale parameters of 1, 10, 100 and 1000.

Using the likelihood function

$$L(y|\theta) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(\frac{-1}{2\sigma_\varepsilon^2}(y_j - \mu(t_j))^2\right)$$
$$= (2\pi\sigma_\varepsilon^2)^{-n/2} \exp\left(\frac{-1}{2\sigma_\varepsilon^2}\sum_{j=1}^{n}(y_j - \mu(t_j))^2\right)$$

where $\theta = \{\alpha_0, \alpha_1, \alpha_2, \alpha_{IDT}, \sigma_\varepsilon^2\}$ and $\mu(t_j) = (\alpha_0(1 + t_j^*) + 0.5 \sin(2\pi(t_j + \alpha_1 t_j^*)/\alpha_{IDT}))(1 + \alpha_2 t_j^*)$, and the prior distributions assigned above and under the independence assumption, the joint posterior distribution can be expressed as

$$\Pi(\theta|y) \propto (\sigma_\varepsilon^2)^{-n/2} \exp\left[\frac{-1}{2\sigma_\varepsilon^2}\sum_{j=1}^{n}(y_j - \mu(t_j))^2\right] \times \prod_{k=0}^{2}\exp\left(\frac{-(\alpha_k - \tilde{\alpha}_k)^2}{2\tilde{\sigma}_k^2}\right) \times \frac{(\alpha_{IDT})^{10\,000}}{\Gamma(10001) * 0.0001^{10\,001}}\exp(-\alpha_{IDT}/0.0001) \times \frac{100^{100/\tilde{\sigma}_y^2+1}(\sigma_\varepsilon^2)^{-(100/\tilde{\sigma}_y^2+2)}}{\Gamma\left(100/\tilde{\sigma}_y^2+1\right)}\exp(-100/\sigma_\varepsilon^2)$$

Closed form fully conditional posterior distributions of the parameters were obtained when possible and presented in the Appendix in the Supplementary Materials. Since two of the parameters, α_1 and α_{IDT}, in the Bayesian periodic model do not have fully conditional conjugate posteriors, we used an MCMC procedure in SAS® Version 9.2 which utilizes an adaptive blocked random-walk Metropolis algorithm with target acceptance rate equal to 0.20 and acceptance tolerance 0.075. A t-distribution (degrees of freedom = 3) was used as a proposal distribution. This approach allows for computational convenience.

Step 3 (Obtaining samples from the posterior distribution): Generate 10 000 posterior realizations of the set of parameters in the Bayesian periodic model following 100 000 burn-in runs and using a thinning parameter of 10 to reduce the correlation between successive samples. This posterior sample can be represented by the matrix $\{\alpha_{0,m}, \alpha_{1,m}, \alpha_{2,m}, \alpha_{IDT,m}, \sigma_{\varepsilon,m}^2\}_{m=1}^{10\,000}$, where $\sigma_{\varepsilon,m}^2$ is a posterior realization of the error variance which will be used to compare the performance of the Bayesian periodic model with that of the polynomial model as discussed in Step-5 below.

Step 4 (Constructing the null space): Using the data pairs (y_j,t_j), $j = 1,2,\ldots,n$, obtain the least-squares estimates of the parameters of the polynomial model of the form $y_j = \beta_0 + \beta_1 t_j + \beta_2 t_j^2 + \cdots + \beta_K t_j^K + \varepsilon_j$, where $\varepsilon_j \sim N(0, \sigma_{\beta\varepsilon}^2)$. The degree of the polynomial, K, is recommended as follows: Starting with the constant (a zero degree polynomial, i.e. the white noise case) for one cycle of data, an additional term is added to the model for each half cycle of data; that is, if there is only one cycle of data, only the intercept term, β_0, is used any given pattern including a linear pattern may be repeated in later cycles in the same fashion and thus cannot be eliminated as 'non-periodic'. As the amount of data in cycles increase, the null space can be expended to include any linear, quadratic, cubic and other higher order polynomial patterns by adding an additional term to the polynomial model for each half-cycle of data. So, the number of terms used in the polynomial model representing the null space can be considered as a function of the number of cycles of data and can be summarized as follows:

- 1.0 cycle → constant, which is β_0
- 1.5 cycles → linear polynomial model with β_0 as intercept
- 2.0 cycles → quadratic polynomial model with β_0 as intercept
- 2.5 cycles → cubic polynomial model with β_0 as intercept
- 3.0 cycles → polynomial model of degree 4 with β_0 as intercept

Obtain the least squares estimates of the parameters of the polynomial model and compute the mean squared errors (MSE_0).

Step 5 (Summarizing the evidence for periodicity): Estimate the empirical probability that posterior realization of the error variance, σ_ε^2, from the Bayesian periodic model is larger than the mean squared errors (MSE_0) from the polynomial model representing the null space, by $P_{Bayes} = \sum_{m=1}^{10\,000} I(MSE_0 \le \sigma_{\varepsilon,m}^2)/10\,000$. Small P_{Bayes} represents empirical evidence against the non-periodic model and the null hypothesis is rejected accordingly.

4 IMPLEMENTATION OF THE ALGORITHM AND SIMULATION DESIGN

For the Bayesian computations, we have written a SAS macro which utilizes the MCMC procedure in SAS® Version 9.2. The MCMC procedure in SAS is dedicated to Bayesian computations and has several built-in diagnostic tools, including Geweke diagnostic and Heidelberger-Welch diagnostics, examples of which are presented in the Supplementary Materials. The macro program also optionally performs the Fisher's *G*-test and the permutation test.

In Supplementary Figures S5 and S6, we present a sample model fit and diagnostic graph for one of the parameters of interest, respectively, to show that the algorithm has desirable operating characteristics. Diagnostic graphs for the other parameters in the Bayesian model have similar attributes. In addition to the diagnostic plots, the Bayesian periodicity test procedure has desirable diagnostic characteristics based on the Geweke diagnostic test by Geweke (1992) and Heidelberger-Welch diagnostics tests by Heidelberger and Welch (1981, 1983), which include the Stationarity and Half-width tests.

The performance of the empirical Bayes test for periodicity was compared with that of the Fisher's *G*- and the permutation tests using randomly generated time-course data from five periodic (sine, spike, double spike, beta and double beta) and five non-periodic (noise, linear, low-plato, high-plato and random spikes) patterns for which examples are shown in Supplementary Figure S7. These patterns were identified by reviewing the time-course profiles of hundreds of genes in the *S.pombe* experiments.

Using each of these 10 patterns, we generated 500 random samples (a total of 5000 time-course samples under each sample size), where the sample sizes ranged from 1.5 cycles to 3.0 complete cycles with 10 observation per cycle. We then applied the three methods of periodicity testing on each sample and compared the empirical cumulative distribution functions (CDFs) of *P*-values from the three methods in pairs of patterns with one periodic and one non-periodic patterns. The CDF of *P*-values under the periodic pattern was plotted agains the CDF of *P*-values under the non-periodic pattern (CDF-CDF graph) and the area under the curve formed by the CDF-CDF graph.

We also compared the performances of three methods under missing data where 20% of the observation were randomly

removed in each sample and the test results were obtained based on the remaining observations. All simulation data were generated and all computations were performed in SAS® (Version 9.2) environment.

5 RESULTS

In Figure 1, we compare the permutation, Fisher's G and empirical Bayes tests using a plot of the empirical CDF of P-values under a given periodic pattern versus the empirical CDF of P-values under a given non-periodic pattern. We call this plot a 'CDF-CDF curve'.

The empirical CDF is computed by $\mathrm{CDF}(\alpha) = \sum_{n=1}^{N_p} (P_n \leq \alpha)/N_p$ for a sample of P-values $\{P_n\}_{n=1}^{N_p}$. As in the receiver operating characteristics (ROC) curves with sensitivity and specificity, the closer the curve to the upper left corner, the better operating characteristics it has in terms of correctly distinguishing periodic patterns from non-periodic ones and vice versa.

Clearly, none of the methods is uniformly best under various pairing of a periodic and a non-periodic setting. However, under the 'sinusoidal' pattern, the permutation test seems to be better against 'noise' and 'random spikes', where the Fisher's G-test fails against 'linear' and 'high plateau' patterns. This is because any linear or high plateau patterns can be fit easily to a small part of a sine function and the Fisher's G-test is sensitive to such patterns and falsely accepts them as periodic without considering the period. Under the 'spike' and 'double spike' patterns, the Bayesian periodic model seems to be the best against any of the five non-periodic patterns. Under the 'beta' pattern, all three methods had high performance. However, the Fisher's G-test did not differentiate a periodic pattern from a 'high plateau' pattern as expected. Under the 'bi-modal' pattern, the Bayesian approach and the Fisher's G-test had similar performance while the permutation test did not perform well against any of the five non-periodic patterns. For each combination of a periodic and non-periodic pattern as depicted in Supplementary Figure S7, we present the area under curve in Supplementary Table S2.

To compare the performance of the three methods under the case of irregularly spaced experimentation times, as is commonly encountered in real experiments, we randomly removed 20% of the data points. Under this 'missingness', the performance of the Bayesian approach is superior to that of the Fisher's G-test and the permutation test as illustrated by Figure 2. For each combination of a periodic and non-periodic pattern, we present the AUC of any given CDF-CDF curve in Supplementary Table S3.

The results when the data were drawn from 1.5, 2.5 and 3 cycles were very similar to the above results both with full data and when some data were randomly missing. (Simulation results for 1.5 cycles of data are presented in Supplementary Fig. S8.)

6 APPLICATION TO *S.POMBE* CELL-CYCLE GENE EXPRESSION EXPERIMENTS

We used 10 microarray experiments (Supplementary Table S1) measuring genome-wide time-course gene expression during the cell division cycles of the fission yeast *S.pombe* based on two synchronization protocols—elutriation (*Elut* on the figures) and Cdc25 block-release (Cdc25 on the figures) (Oliva *et al.* 2005; Peng *et al.* 2005; Rustici *et al.* 2004). The data were normalized and median centered by the original experimenters.

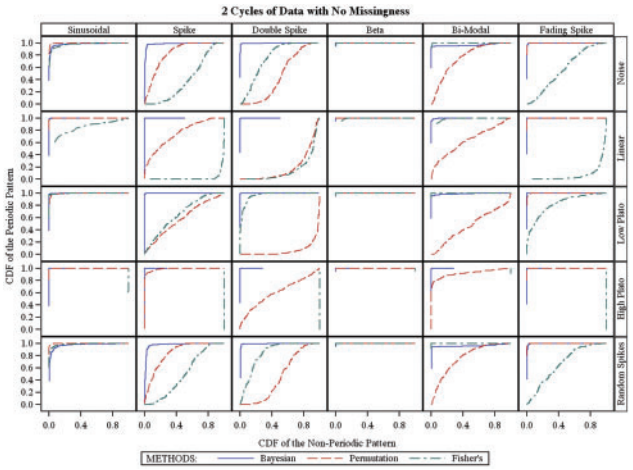

Fig. 1. CDF-CDF graph of P-values from the three tests for periodicity in a given pair of a periodic and a non-periodic pattern with no missingness

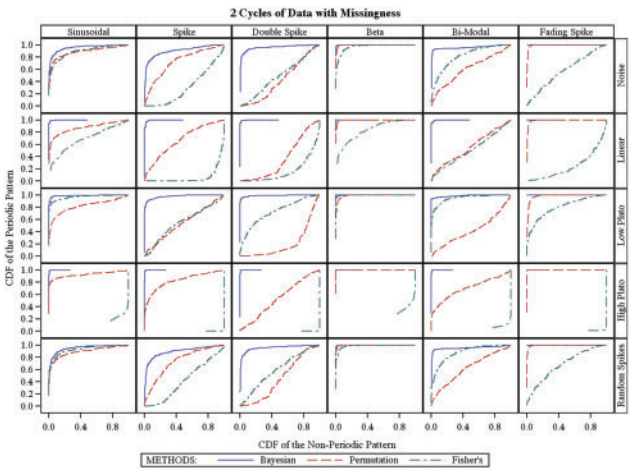

Fig. 2. CDF-CDF plot of the three tests for periodicity using the six periodic patterns against the five non-periodic patterns where the time increments between the experimentation times were not the same

We have estimated the interdivision time (IDT) using the 35 genes which are known to be periodic (Rustici *et al.* 2004) and obtained the mean estimate for IDT along with its variance as presented in Supplementary Table S3. For the Rustici experiments, 29–33 genes of those 35 genes were used in the estimation process. For the other five experiments, we had the time-course data for only 11 of those 35 genes. Overall, inter-division time estimate provided by the experimenters are close to what we estimated based on the Fourier score approach with varying degree of variation.

We then applied the empirical Bayes test to the time-course gene expression data from the 10 experiments, using the variance estimates of IDT in defining the prior distribution of α_{IDT}, to test the null hypothesis that a given gene is not cell-cycle regulated (i.e. not periodically expressed). Similarly, P-values for each gene

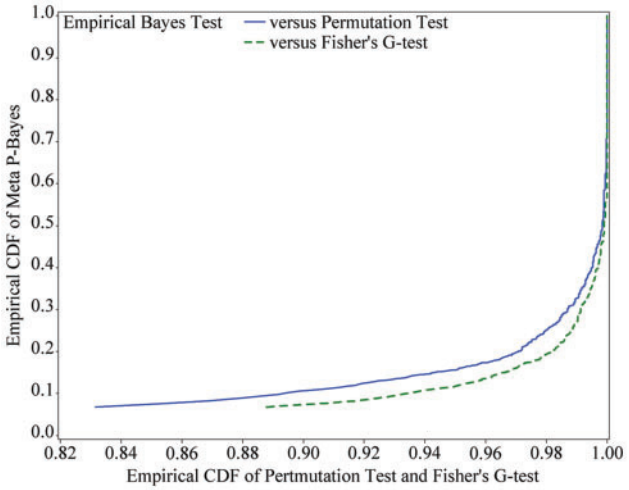

Fig. 3. CDF-CDF graph comparing the Meta *P*-values from the empirical Bayes test for periodicity versus the permutation test and Fisher's *G*-test

in each experiment were obtained using the Fisher's *G*-test and the permutation test. We then applied the logit method by George and Mudholkar (1983) to pool the *P*-values across the 10 experiments. In Figure 3, we present the CDF-CDF graph comparing the Meta *P*-values from the empirical Bayes test for periodicity versus the permutation test and Fisher's *G*-test.

Clearly, the empirical CDF of Meta P_{Bayes} is much lower compared with the empirical CDF of the Meta *P*-values of the other two methods at a given threshold, which results in overwhelmingly more 'periodic' gene calls in the permutation test and Fisher's *G*-test than in the empirical Bayes test for periodicity. For example, at significance threshold of 0.05, the Bayesian approach calls 797 (16%) genes as 'periodic' while the permutation test and Fisher's *G*-test calls 4700 (95%) and 4789 (97%) genes as 'periodic' (Supplementary Fig. S9).

We used a benchmark set of 40 periodic genes reported by Marguerat *et al*. (2006) to assess the sensitivity and specificity of the three approaches. Twenty-five, 27 and 27 of these 40 genes were ranked within the top 100 genes by the empirical Bayes test for periodicity, the permutation test and Fisher's *G*-test, respectively. Similarly, 28, 31 and 30 of these 40 genes were ranked within the top 200 of the genes by the empirical Bayes test for periodicity, the permutation test and Fisher's *G*-test, respectively. We present the expression profiles of the lowest ranked gene among the 40 benchmark genes by the empirical Bayes test for periodicity in Supplementary Figure S10.

We have also compared the three methods using a set of 52 'conserved' genes and a set of 235 'cycling' genes reported by Lu *et al*. (2007). Among the top 100 genes, 12, 14 and 14 conserved genes were identified by the Bayesian approach, the permutation test and Fisher's *G*-test, respectively. Similarly, 36, 39 and 38 'cycling' genes were identified by the Bayesian approach, the permutation test and Fisher's *G*-test, respectively, among the top 100 genes. We present the gene expression profiles for the lowest ranked 'conserved' gene by the Bayesian approach in Supplementary Figure S11, where we observe that the periodicity of these genes is not consistently supported by the independent experiments. Similarly, we present the gene expression profiles for the lowest ranked 'cycling' gene by the Bayesian approach in Supplementary Figure S12.

Table 1. Ranks for the CESR genes by the three periodicity tests

Tests	Ranks for the CESR genes ($N = 126$)				
	Min.	Q1	Median	Q3	Max.
Empirical Bayes	82	1837	2795	4255	4920
Permutation	79	876	1819	2988	4682
Fisher's *G*	46	710	1512	2668	4829

Based on the 'benchmark', 'conserved' and 'cycling' gene lists described above, we have looked into the Gene Ontology (GO) terms for biological processes in the genes that were ranked within the top 1000 by the permutation test and Fisher's *G*-test while ranked at 2000 and higher by the empirical Bayes test. Based on the reported GO terms in biological processes, we argue that most of these genes, although ranked within the top 1000 by the permutation test and Fisher's *G*-test, do not have convincing evidence of being part of cell-cycle as shown in Supplementary Tables S4–S6.

We have also compared the ranks of Core Environmental Stress Response (CESR) genes (Chen *et al*. 2003) which are expected to be expressed or repressed as a response to environmental stress and are not expected to be periodic. We had time-course data for 126 CESR genes and the median rank by the Bayesian approach was 2795 while the median ranks were 1819 and 1512 for the permutation test and Fisher's *G*-test, respectively, as shown in Table 1.

Along the same lines, only 20 CESR genes were ranked within the top 1000 by the empirical Bayes test while 36 and 41 genes were ranked within the top 1000 by the permutation test and Fisher's *G*-test, respectively. The counts for the top 2000 genes were 37, 70 and 78, respectively. These results show the more desirable sensitivity and specificity of the empirical Bayes periodicity test compared with its counterparts.

We have also performed gene enrichment analysis on biological processes for the top 1000 genes by each method (Supplementary Table S7). Based on the hyper-geometric test, we see that for most biological processes, the top 1000 genes by the Bayesian approach has much more genes than the other two methods. For example, for DNA replication, while the Bayesian method has 10 genes in the top 1000, permutation test and Fisher's *G*-test have only three genes.

7 DISCUSSION

We have demonstrated through simulations that a goodness-of-fit comparison of the Bayesian periodic model and the polynomial regression model has better sensitivity to distinguish periodic patterns from non-periodic patterns when compared with the permutation test and Fisher's *G*-test. This property extends to the case where there are data missing at random. The property is also observed when there are deviations of the interdivision time from the true interdivision time. The Bayesian approach seems to be able to distinguish periodic patterns from pure noise as well as from any linear, quadratic, cubic and

higher order polynomial patterns. As a result of its superior specificity, the Bayesian approach identified only 797 (16%) of the 4940 genes as 'periodic' at the significance threshold of 0.05, which we believe is much closer to the percentage of genes that are periodic in fission yeast, while the permutation test and the Fisher's exact test identified >95% of the genes as 'periodic', which is not realistic and shows that these two tests are not sufficiently specific as they erroneously show evidence of periodicity if the time-course data show something different than noise. This point was also eloquently raised by Futschik and Herzel (2008) as discussed in Section 1.

Our GO review of the genes that were ranked ≥2000 by the Bayesian approach while ranked within the top 1000 by the permutation test and Fisher's *G*-test shows that most of these genes are not reported being involved in the cell division process. Therefore, it is highly critical to have a testing procedure that has high sensitivity in recognizing any periodic pattern as well as high specificity in eliminating patterns (or no patterns) that are not periodic, including noise, from the final list of genes. Using extensive simulation studies and application to 10 cell-cycle gene expression experiments as well as evaluating the ranks of CESR genes by each periodicity test, we have shown that the Bayesian test of periodicity has such desirable sensitivity and specificity characteristics.

However, we feel that computational efficiency of the empirical Bayes approach can be improved. In its current form, a high level of computations is required to obtain the posterior realizations of the model parameters and perform further calculations. For example, obtaining the Bayesian probabilities for each gene in each cell-cycle experiment by Rustici may take about 4 hours in SAS® Version 9.2 on a 3.00 GHz PC with Intel® Core™2 Duo CPU and 3.25 GM of RAM PC; however, first, considering the limited resources both in terms of financial and human resources, a more sensitive and specific gene list for further small-scale studies should not be sacrificed for obtaining such a list rapidly. Second, we believe that the implementation of the Bayesian test of periodicity on a different platform such as OpenBugs can definitely help reduce the computation time which makes such an issue less and less relevant.

Although the main application area for the empirical Bayes Periodicity test is the time-course gene expression data, it can also be used for the inventory management of consumer goods. Bensoussan *et al*. (2011) describes the need of using cyclical demand signals to maximize the profit. Thus, our testing approach for periodicity can be applied to time series demand data on a given product over a time period and identify the products among tens of thousands of products which have periodic consumer demand. The empirical Bayes periodicity test can also be used to identify 'cyclical' stocks which may potentially provide valuable information for investors. As an example, Akar and Baskaya (2011) use univariate spectral analysis to identify cyclic behavior of the Turkish stock market. Our method can also be used to test for cyclic behavior of stock markets at given inter-cycle times.

Conflict of Interest: none declared.

REFERENCES

Akar,C. and Baskaya,Z. (2011) Detecting the long term cyclical behaviour of the Turkish stock market by means of spectral analysis. *Int. Res. J Finance Econ*., **67**, 160–167.

Bensoussan,A. *et al*. (2011) Achieving a long-term service target with periodic demand signals: a newsvendor framework. *Manuf. Serv. Oper. Manag*., **13**, 73–88.

Chen,D. *et al*. (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell*, **14**, 214–229.

de Lichtenberg,U. *et al*. (2005a) Comparison of computational methods for the identification of cell-cycle-regulated genes. *Bioinformatics*, **21**, 1164–1171.

de Lichtenberg,U. *et al*. (2005b) New weakly expressed cell-cycle regulated genes in yeast. *Yeast*, **22**, 1191–1201.

Fisher,R.A. (1929) Tests of significance in harmonic analysis. *Proc. R. Soc. Lond*., **125**, 54–59.

Futschik,M.E. and Herzel,H. (2008) Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis. *Bioinformatics*, **24**, 1063–1069.

George,E.O. and Mudholkar,G.S. (1983) On the convolution of logistic random variables. *Metrika*, **30**, 1–14.

Geweke,J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4*. Clarendon Press, Oxford, UK, pp. 169–193.

Heidelberger,P. and Welch,P.D. (1981) A spectral method for confidence interval generation and run length control in simulations. *Commun. ACM*, **24**, 233–245.

Heidelberger,P. and Welch,P.D. (1983) Simulation run length control in the presence of an initial transient. *Oper. Res*., **31**, 1109–1144.

Lu,Y. *et al*. (2007) Combined analysis reveals a core set of cycling genes. *Genome Biol*., **8**, R146.

Marguerat,S. *et al*. (2006) The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast*, **23**, 261–277.

Oliva,A. *et al*. (2005) The cell-cycle regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol*., **3**, e225.

Peng,X. *et al*. (2005) Identification of cell-cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell-cycle. *Nat. Genet*., **36**, 809–817.