BIOINFORMATICS

Vol. 20 no. 1 2004, pages 5–20
DOI: 10.1093/bioinformatics/btg364

Identifying periodically expressed transcripts in microarray time series data

Sofia Wichert[1], Konstantinos Fokianos[2] and Korbinian Strimmer[1,]*

[1]Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany and [2]Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus

Received on March 5, 2003; revised on June 17, 2003; accepted on July 8, 2003

ABSTRACT

Motivation: Microarray experiments are now routinely used to collect large-scale time series data, for example to monitor gene expression during the cell cycle. Statistical analysis of this data poses many challenges, one being that it is hard to identify correctly the subset of genes with a clear periodic signature. This has lead to a controversial argument with regard to the suitability of both available methods and current microarray data.

Methods: We introduce two simple but efficient statistical methods for signal detection and gene selection in gene expression time series data. First, we suggest the *average periodogram* as an exploratory device for graphical assessment of the presence of periodic transcripts in the data. Second, we describe an *exact statistical test to identify periodically expressed genes* that allows one to distinguish periodic from purely random processes. This identification method is based on the so-called g-statistic and uses the false discovery rate approach to multiple testing.

Results: Using simulated data it is shown that the suggested method is capable of identifying cell-cycle-activated genes in a gene expression data set even if the number of the cyclic genes is very small and regardless the presence of a dominant non-periodic component in the data. Subsequently, we re-examine 12 large microarray time series data sets (in part controversially discussed) from yeast, human fibroblast, human HeLa and bacterial cells. Based on the statistical analysis it is found that a majority of these data sets contained little or no statistical significant evidence for genes with periodic variation linked to cell cycle regulation. On the other hand, for the remaining data the method extends the catalog of previously known cell-cycle-specific transcripts by identifying additional periodic genes not found by other methods. The problem of distinguishing periodicity due to generic cell cycle activity and to artifacts from synchronization is also discussed.

Availability: The approach has been implemented in the R package `GeneTS` available from http://www.stat.uni-muenchen.de/~strimmer/software.html under the terms of the GNU General Public License.

Contact: strimmer@stat.uni-muenchen.de

INTRODUCTION

One of the earliest applications of gene expression experiments was the genome-wide monitoring of gene activity in a cell during cell division. Statistical analysis of this data poses many challenges due to the typically small number of measurements per gene (e.g. $N = 20$), high levels of non-normal random noise, and the large number of simultaneously assessed genes (usually $G > 1000$). Thus, gene expression time series data are quite different when compared with classical time series data.

One particular aim when analyzing microarray cell cycle data is to find statistical evidence of cyclicity or *periodicity*, and then to identify this subset of genes that is responsible during the cell cycle. However, usually only a small fraction of the genes under investigation exhibits some evidence of periodically varying expression during the cycle so that the overall signal in the data is dominated by non-periodic components. Consequently, the problem of identifying the subset of periodically expressed genes is quite involved and calls for new estimation methods.

The difficulties of the analysis of cell cycle data have echoed in a recent controversy about the statistical significance of results published for some widely used reference data sets. For instance, Shedden and Cooper (2002a,b) questioned the presence of generic cell-cycle-specific signal in two data sets concerning yeast and human cells that were previously analyzed by other authors, namely Spellman *et al*. (1998) and Cho *et al*. (2001), respectively. Shedden and Cooper suggested that the cyclicity observed by Cho *et al*. (2001) was due to chance fluctuation and criticized that cyclicity results for the yeast data lacked reproducibility across different synchronization methods.

In another seminal work, Whitfield *et al*. (2002) established a catalog of genes periodically expressed in the human cell cycle by conducting a series of large-scale microarray experiments. This paper also introduced a statistical test to identify

*To whom correspondence should be addressed.

genes relevant for the cell cycle. However, their approach is based on a criterion ('periodicity score') that is not only difficult to interpret but also does not take account of multiple testing.

These two examples show that a further development of statistical methodology for analyzing genetic time series data is necessary. The suggested methodology of this study is based on two simple and computationally inexpensive procedures for microarray time series analysis. First, we introduce a graphical exploratory device in order to elucidate the quality of cell cycle data prior to further analysis. Second, a statistical test is proposed to distinguish periodic signal from an otherwise random process. Individual genes are then screened for cell-cycle-specific activity using multiple testing under the criterion of false discovery rate (FDR). For data analysis purposes, both approaches are then combined to screen cell cycle microarray data for potentially relevant genes.

The rest of the paper is organized as follows. In the next section we explain in detail the mathematical and statistical background of our approach. Subsequently, we validate our approach using computer simulation. Next, we study molecular data from 12 gene experiments for yeast, bacterial and human cell lines (Spellman *et al*., 1998; Laub *et al*., 2000; Cho *et al*., 2001; Whitfield *et al*., 2002) where particular emphasis is given to the controversial issues raised by the Shedden–Cooper debate. Finally, we compare our approach with other related frameworks and we present some guidelines for the analysis of gene expression time series data.

METHODS

Outline

Our approach to the statistical analysis of gene expression time series data consists of two complementary procedures. First, we use a graphical device which will be called *average periodogram*, for exploring visually and detecting any periodic signatures in the data. The average periodogram is a simple extension of the standard periodogram—a tool that is widely used in time series analysis. Notice that the adjective 'average' refers to the averaging which takes place over all considered genes. Second, for inferring the subset of genes actually involved in the cell cycle and to quantize the results obtained by the average periodogram analysis we appeal to an exact test which is based on Fisher's g-statistic. This allows distinguishing periodic from purely random processes while the tests of significance are carried out for all genes simultaneously using the method of FDR for multiple testing of hypotheses.

Average periodogram

Consider the following simple model of periodic gene expression through time

$$Y_t = \beta \cos(\omega t + \phi) + \epsilon_t \qquad (1)$$

where β is a positive constant, $\omega \in (0, \pi)$, ϕ uniformly distributed in $(-\pi, \pi]$, and where $\{\epsilon_t\}$ is a sequence of uncorrelated random variables with mean 0 and variance σ^2, independent of ϕ. Then, the periodogram, which is denoted by $I(\omega)$, and defined as

$$I(\omega) = \frac{1}{N}\left|\sum_{t=1}^{N} Y_t \exp(-i\omega t)\right|^2, \quad \omega \in [0, \pi] \qquad (2)$$

can be used to detect periodic components in observed genetic time series data $Y_1, \ldots, Y_N$, where N is the sample size. A simple graphical device is to search for *significant peaks* in $I(\omega)$ by treating ω as a continuous variable or, much more often, as a discrete variable taking the values in $[0, \pi]$,

$$\frac{2\pi k}{N}, \quad k = 0, 1, \ldots, \left[\frac{N}{2}\right], \qquad (3)$$

known as Fourier frequencies. If a time series has a significant sinusoidal component with frequency $\omega_0 \in [0, \pi]$, then the periodogram exhibits a peak at that frequency with a high probability. Conversely, if the time series is a *purely random process*, that is to say $\beta = 0$ in Equation (1), then the plot of the periodogram against the Fourier frequencies reduces to a straight line (Priestley, 1981).

It turns out that the periodogram is a helpful exploratory device even when several short time series are observed over the same time span. This is the case for gene expression measurements where typically many thousands of genes are assessed in parallel. Suppose that Y_{it} denotes the i-th observed time series at time t where $i = 1, \ldots, G$ and $t = 1, \ldots, N$. The *average periodogram* can then be defined as

$$AI(\omega) = \frac{1}{G}\sum_{i=1}^{G} I_i(\omega), \qquad (4)$$

where $I_i(\omega)$ is the periodogram of the i-th time series. Then a plot of $AI(\omega_k)$ against $\omega_k = 2\pi k/N$ where $k = 0, 1, \ldots, [N/2]$ should reveal whether or not there are periodic components in the data. Note that in a related approach, Welch (1967) proposed the average periodogram for different blocks within a single time series whereas here we average for a fixed block size over multiple time series.

This simple graphical device is justified as follows: if the data follow a pure random process then the periodogram of all time series is uniform and therefore the average estimate should reduce to a straight line. However, if there are a few time series exhibiting strong periodicity, then their corresponding periodogram ordinates dominate Equation (4). Therefore, the shape of the otherwise flat average periodogram changes so that any visible peaks should indicate the presence of a periodic component.

Using computer simulations (see corresponding section in *Results*) we have validated these properties and shown

that the average periodogram is well suited to detect the presence of even minute amounts of periodically expressed transcripts. However, it will yield accurate results only when some of the time series exhibit the identical periodic behavior.

Fisher's test

Upon noticing that the periodogram (for a single time series) contains a peak, then a formal test should be carried out to determine whether or not this peak is significant or not. Most of the established time series theory deals with asymptotic results (Priestley, 1981) under the Gaussian assumption. However, there is an early result from Fisher (1929) that is valid also for finite samples, and this is what will be utilized next. Fisher derived an exact test of the maximum periodogram coordinate by introducing the g-statistic

$$g = \frac{\max_k I(\omega_k)}{\sum_{k=1}^{[N/2]} I(\omega_k)}. \quad (5)$$

Large values of g lead to the rejection of the null hypothesis of purely random process—that is $\beta = 0$ in Equation (1). To calculate the p-value of the test under the null hypothesis it is helpful to resort at the exact distribution of g which is given by

$$\mathrm{P}(g > x) = n(1-x)^{n-1} - \frac{n(n-1)}{2}(1-2x)^{n-1} + \cdots + (-1)^p \frac{n!}{p!(n-p)!}(1-px)^{n-1}, \quad (6)$$

where $n = [N/2]$ and p is the largest integer less than $1/x$. Hence, if $g^\star$ is the observed value of g, then Equation (6) yields a p-value $P(g > g^\star)$ that allows to test whether a gene behaves like a purely random process or whether it exhibits some periodic expression pattern (i.e. whether the maximum peak in the periodogram is significant).

Gene selection and multiple testing

Turning to the problem of identifying periodically expressed genes from multiple short time series data we also have to consider the problem of multiple testing. In the proposed methodology the g-statistic for each of the investigated genes is calculated. Hence a sequence of g-statistics is generated, say $g_1, g_2, \ldots, g_G$ with corresponding p-values $p_1, p_2, \ldots, p_G$ calculated by Equation (6). Subsequently, to answer the question which of these test statistics are significant we employ the method of FDR as multiple comparison procedure (Benjamini and Hochberg, 1995). FDR is a new approach to the multiple comparisons problem where one controls the *expected proportion* of false positives rather than the chance of any false positives (e.g. as in the standard Bonferroni correction). The FDR threshold is determined from the observed p-value distribution, and hence is adaptive to the actual data.

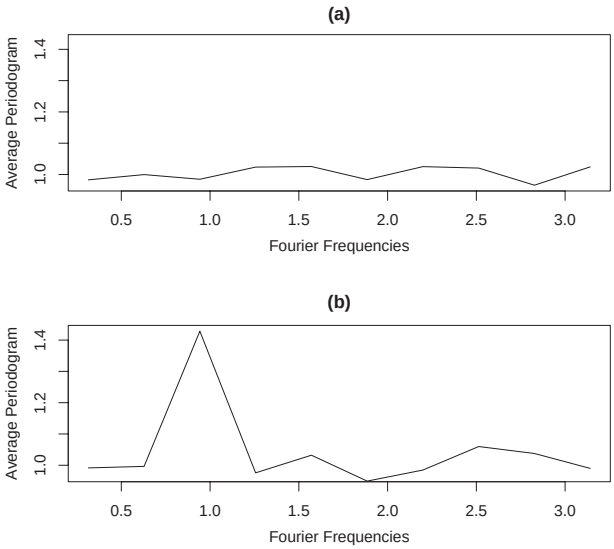

Fig. 1. Average periodogram for simulated data sets with 2000 time series (genes) of length 20. The upper part (**a**) corresponds to a white noise process for all 2000 genes. In the lower part (**b**) the batch of data includes 100 time series with frequency 1.

Accordingly, consider the set of ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(G)}$ with corresponding genes $g_{(1)}, g_{(2)}, \ldots, g_{(G)}$ and apply the following algorithm:

(1) Let i_q be the largest i for which $p_{(i)} \leq \frac{i}{G}q$,
(2) then reject the null hypothesis for all genes $g_{(1)}, g_{(2)}, \ldots, g_{(i_q)}$.

It can be shown that this procedure controls the FDR at level q (Benjamini and Hochberg, 1995).

Recipe for analysis

In summary, the time series methodology proposed here consists of the following simple steps:

(1) Using the average periodogram check graphically whether or not there are periodic components in the data.
(2) For each time series calculate Fisher's g-statistic.
(3) For each of the test statistic calculate the corresponding p-value.
(4) Identify the genes for which the null hypothesis is rejected under the desired FDR level (e.g. $q = 0.05$) as these are the genes that exhibit a statistically significant periodic component in the periodogram.

All algorithms are implemented in the R package `GeneTS`. It is freely available from the web page http://www.stat.uni-muenchen.de/~strimmer/software.html and is distributed under the terms of the GNU General Public License.

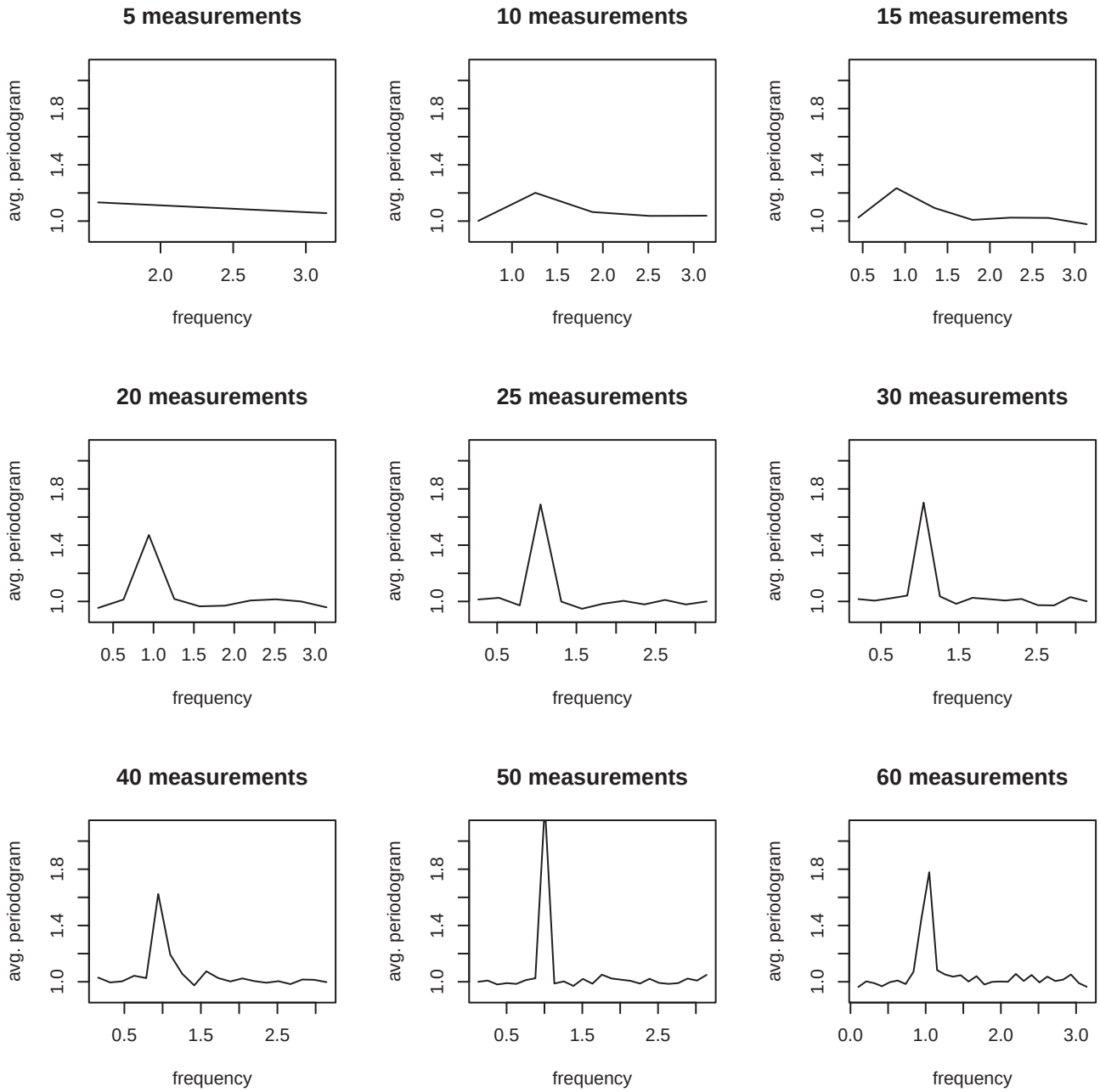

Fig. 2. Sensitivity of the average periodogram with regard to sample size ($N = 5, \ldots, 60$, with 100 cyclic and 1900 random genes).

RESULTS

Simulated data

First, we use simulated data to validate the proposed methodology. To investigate the accuracy and sensitivity of the average periodogram and the FDR gene selection strategy using the g-statistic we simulated time series data for the expression of $G = 2000$ genes as follows. Sequences of length $N = 10, 20, 50, 100$ and 200 were either drawn from the normal distribution with mean 0 and variance 1 or generated according to

$$Y_t = \cos(\omega t) + \sin(\omega t) + \epsilon_t \quad \text{for } t = 1, \ldots, N \qquad (7)$$

with normal noise ϵ_t and with one or more dominant frequencies (e.g. $\omega = 1, 2$). The number of genes realized with a periodic expression was varied between 0 and 200 (i.e. 0–10% of all simulated genes).

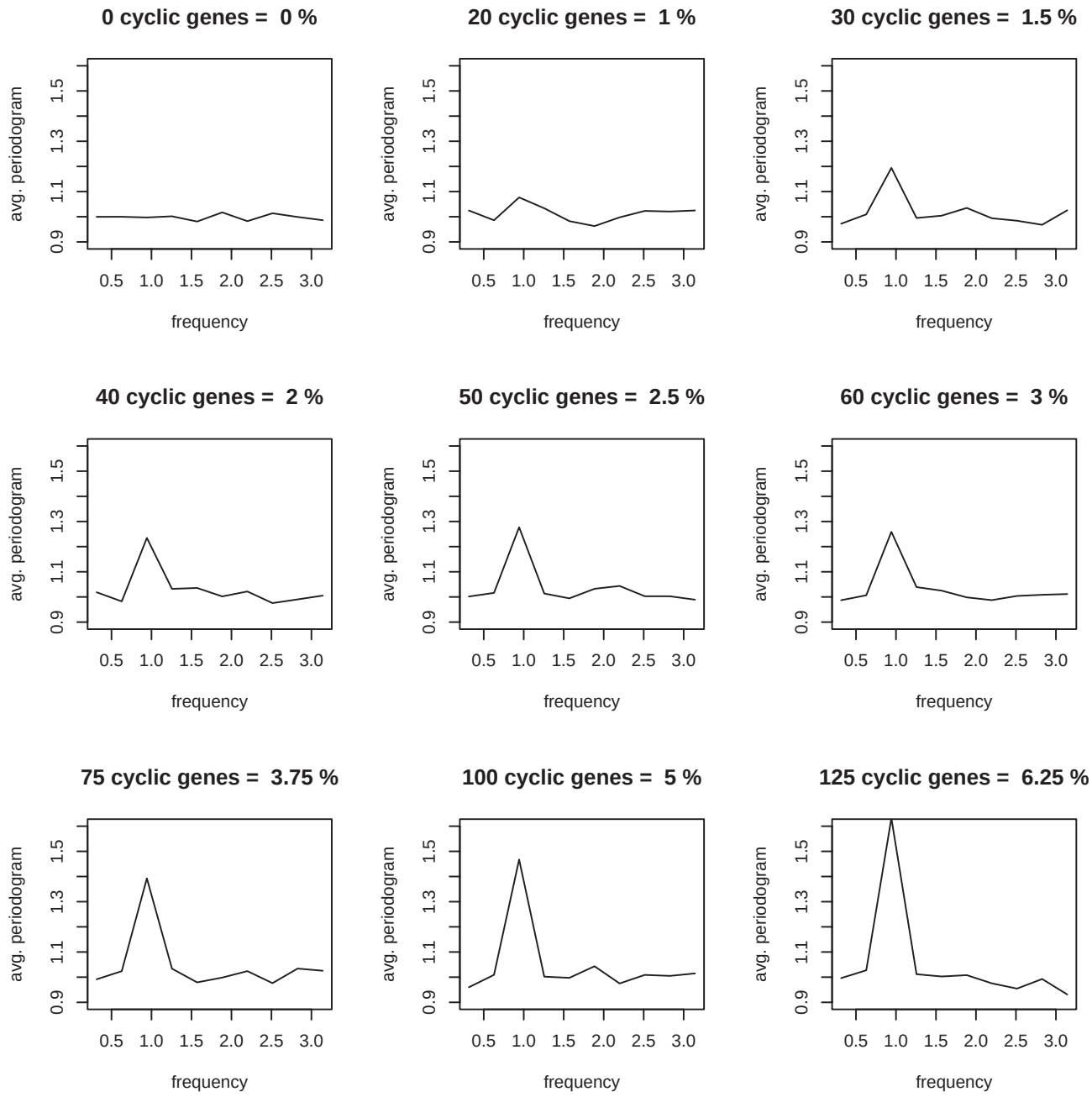

Fig. 3. Sensitivity of the average periodogram with respect to the number of cyclic genes ($N = 20$, with 0–125 cyclic genes and 2000–1875 random genes).

Average periodogram In Figure 1, the average periodogram for a completely random data set with $N = 20$ (a) is contrasted with a plot for a similar data set but containing 5% periodic genes of frequency $\omega = 1$ (b). The upper plot (a) shows a flat line whereas the lower plot (b) exhibits a peak at $\omega = 1$. This indicates that the average periodogram can detect the presence of a dominant frequency in a small subset of genes even if the length of each time series is not long. To investigate this further, we studied the sensitivity of the average periodogram with respect to the fraction of cyclic genes in the data set and the sequence length N. Figure 2 shows results from the analysis of data sets with N ranging from 5 time points to 60 longitudinal measurements per gene, and with 5% periodic genes. From N as low as 15 the peak at the dominant frequency is clearly visible. Figure 3 illustrates the impact of the size of the subset of cyclic genes in the data on

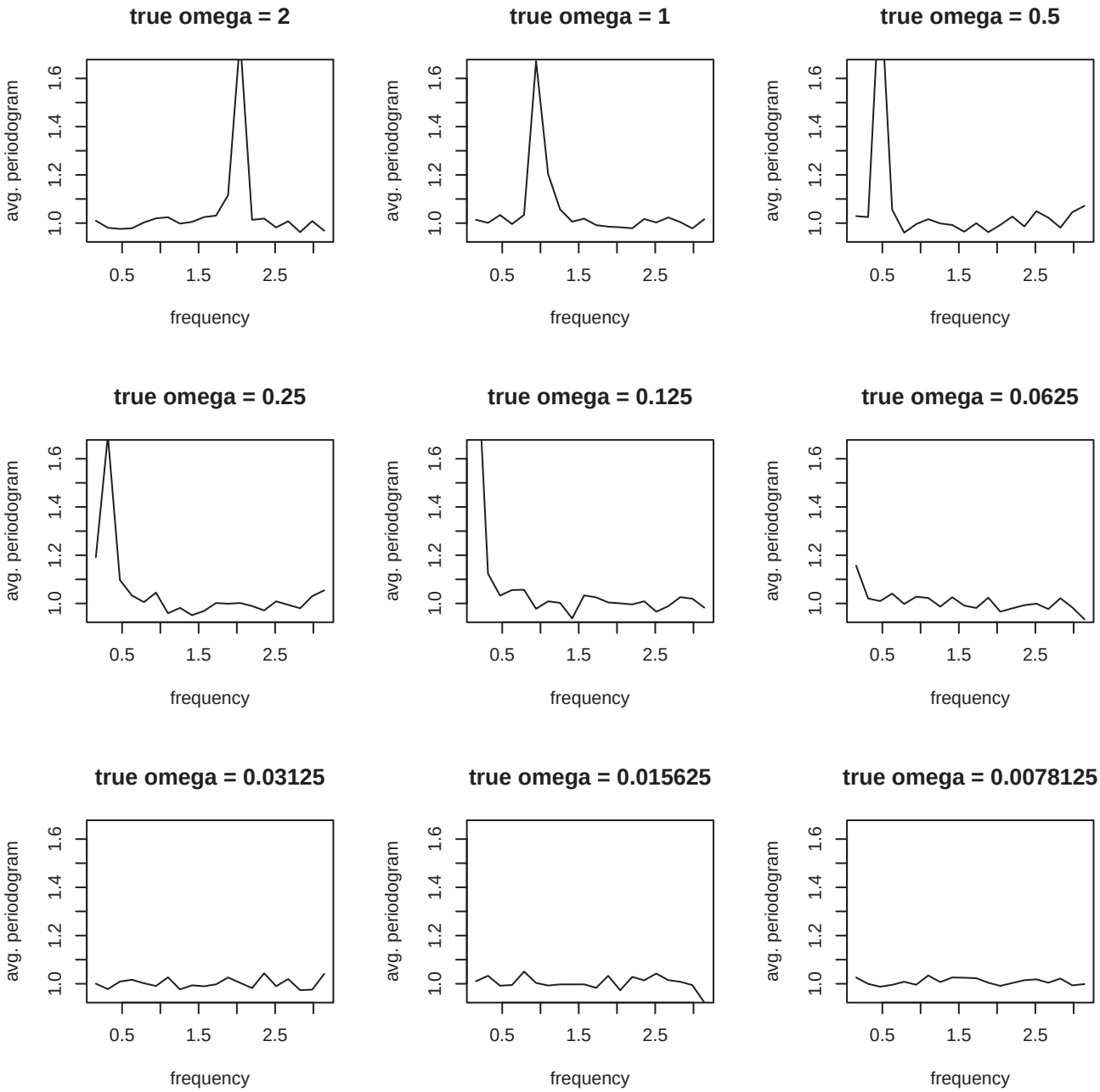

Fig. 4. Frequency resolution of the average periodogram. Plots are shown for simulated data for various true frequencies ($\omega = 2, 1, 1/2, 1/4, \ldots, 1/128$, $N = 40$, with 100 cyclic and 1900 random genes).

the average periodogram. At sample size $N = 20$ the presence of periodicity is detected already even when only 1.5% of all genes are cyclic. With larger sample size this threshold decreases further (data not shown).

Next, we investigated the frequency resolution of the periodogram and its impact on estimating the dominant frequency ω. In a series of simulations we generated data of different lengths N and assumed a variety of true frequencies ($\omega = 2, 1, 1/2, 1/4, \ldots, 1/128$). Figure 4 shows the results for $N = 40$ and 100 cyclic genes out of a total of 2000 genes. From Equation (3) the smallest positive estimate of ω for a time series of length N is $\omega_{\min} = 2\pi/N$. This can be seen clearly in Figure 4 where the average periodogram detects the true values of ω above and close to the threshold $2\pi/40 \approx 0.15$. Smaller values of ω cannot be inferred with the average periodogram (Fig. 4, bottom row).

In further simulations (data not shown) we additionally considered the occurrence of multiple peaks in the average

periodogram, i.e. the existence of two or more dominant frequencies. The resulting plots were similar as above, and the respective dominant frequencies were detected within the resolution limit. Hence, our simulations confirmed that the average periodogram is well suited as an exploratory device to detect periodic components even in moderately short multivariate time series data.

Gene selection Subsequently, we investigated the accuracy and power of the exact g-test, combined with FDR multiple testing, to distinguish generically periodic from random genes. Data were simulated as above with 100 periodic genes for different sequence lengths ($N = 10, 20, 40, 45, 50, 100, 200$). In each case we computed a p-value for each gene according to Equation (6) and applied the FDR rule to determine which genes were considered to exhibit cyclicity for a desired FDR level ($q = 0.15, 0.10, 0.05, 0.01, 0.001$).

The results are shown in Table 1. For sample sizes smaller than 40 measurements per gene, the power of the test to distinguish between a purely random and a periodic process is weak. For time series of this length the g-test underestimates the true number of periodically expressed genes. However, from 40 sample points onwards Fisher's g-test allows to identify most cyclic genes. Moreover, even when non-randomness cannot be rejected formally, the p-values can still be used to rank and thus to compare the genes relative to each other. As an example, the last line in Table 1 shows the number Z of true periodic genes among the first 100 genes ranked using the p-value. Z is surprisingly large even where the FDR test is formally failing.

We repeated this experiment several times using different fractions of cyclic genes (data not shown). Interestingly, the sample size threshold of approximately $N = 40$ was also valid for these data sets, and thus seems to be independent of the number of periodic genes in the data. Generally, however, the threshold will vary depending on the relative amplitude of the noise in the data [i.e. on the error term in Equation (7)].

Table 1. Number of inferred periodic genes using the FDR procedure

q	N						
	10	20	40	45	50	100	200
0.15	3	21	65	103	118	121	117
0.10	1	13	41	97	114	111	111
0.05	1	3	30	90	107	104	104
0.01	0	2	1	78	99	99	100
0.001	0	0	0	45	88	99	99
Z	10	52	64	93	98	100	100

The simulations were carried out as for Figure 1, with 1900 random genes and 100 periodic genes. N is the sequence length, q the desired FDR level expected type I error and Z the number of correctly identified periodic genes among the first 100 genes ranked according to their p-values.

Table 2. Data sets analyzed in this paper and results of FDR test

Cell type	Experiment	N	G	C	C/G (%)	Source
Yeast	cdc15	24	4289	766	17.9	(Spellman
Yeast	cdc28	17	1365	105	7.7	*et al.*, 1998)
Yeast	alpha	18	4415	468	10.6	
Yeast	elution	14	5695	193	3.4	
C.crescentus	bacteria	11	1444	44	3.0	(Laub *et al.*, 2000)
Human fibroblasts	N2	13	4574	0	0	(Cho *et al.*, 2001)
Human fibroblasts	N3	12	5079	0	0	
Human HeLa	score1	12	14 728	0	0	(Whitfield
Human HeLa	score2	26	15 472	134	0.9	*et al.*, 2002)
Human HeLa	score3	48	39 724	6043	15.2	
Human HeLa	score4	19	39 192	56	0.1	
Human HeLa	score5	9	34 890	0	0	

Notation: N is the sample size, G the total number of genes, C the number of periodic genes that are statistically significant for a FDR level of $q = 0.05$.

Molecular data

We now illustrate the theory by application to some real data which are available either on the web or in public databases. The number of experiments is 12, see Table 2 for an overview. The study includes gene expression data from experiments with small sequence length ($N \approx 10$), such as the *Caulobacter crescentus* bacterial cell cycle (Laub *et al.*, 2000) and the human fibroblasts data (Cho *et al.*, 2001), as well as those with a larger number of measured time points ($N \approx 20$–50), e.g. from yeast (Spellman *et al.*, 1998) and human cancer cell line experiments (Cho *et al.*, 2001). We adopt the normalization and prescreening procedures of the aforementioned papers. In particular, for the yeast data set `cdc15`, missing time points (columns) were imputed by interpolation for all genes. For all data considered, genes with missing values and constant expression levels for all sample points were removed from the study.

Yeast cell cycle The yeast *Saccharomyces cerevisiae* microarray experiments were performed by Spellman *et al.* (1998) and have been used as benchmark data set in previous studies. Spellman *et al.* (1998) produced four gene expression experiments data sets using three different cell cycle synchronization techniques, i.e. temperature arrest (`cdc15`, `cdc28`), alpha factor arrest (`alpha`) and elutriation synchronization (`elution`). The latter experimental technique is said to be the least stressful for the yeast cells, whereas the other two are supposed to perturb the internal state of the cells (Shedden and Cooper, 2002b).

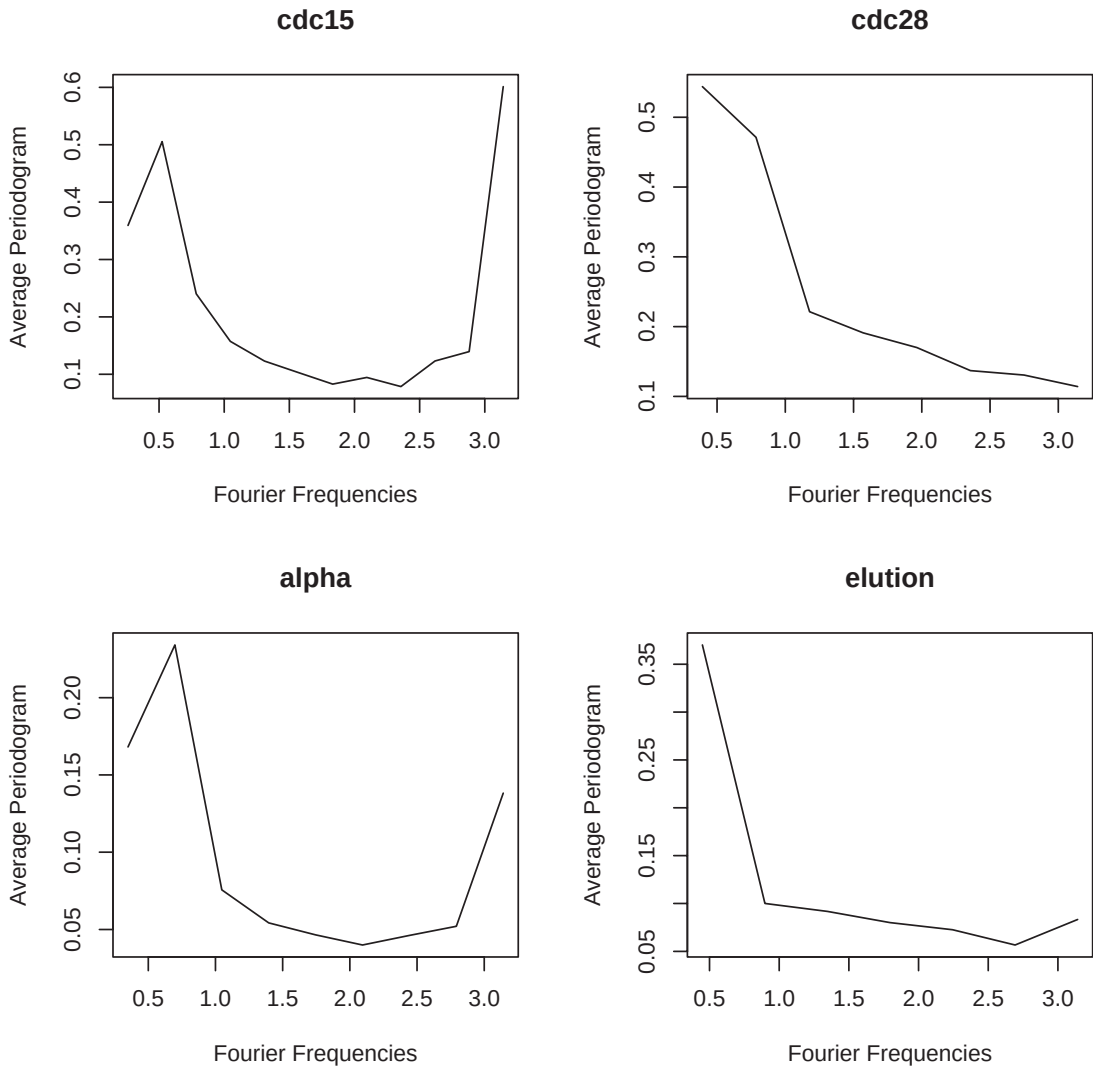

Fig. 5. Average periodogram for the four yeast data sets `cdc15`, `cdc28`, `alpha` and `elution` (see also Table 2).

The average periodograms computed for the data of the four experiments are shown in Figure 5. These plots indicate that for all synchronization techniques there is a clear signal of periodicity. The FDR procedure with an expected type I error (FDR level) of $q = 0.05$ yields the results of Table 2. For the `cdc15` experiment there are 766 genes with statistically significant periodic variation. This is in line with the approximately 800 periodically expressed genes identified in the original work (Spellman *et al*., 1998). However, for the other three techniques the number of periodically expressed genes was lower (Table 2). In particular, the `elution` data contained only 3.4% cyclic genes which supports the hypothesis of Shedden and Cooper (2002b) that the elution data provided only little statistically significant information with regard to cell cycle regulation.

As noted above, it is possible that the cells may be perturbed, and so it is difficult to distinguish cell-cycle-specific variation from an artifact of the method used to synchronize the cells. Shedden and Cooper (2002b) therefore maintained that the periodicity in the `cdc15`, `cdc28` and `alpha` data sets was only apparent and rather due to stress response than to cell cycle activity. In order to verify this claim, we compared the top 100 genes identified as periodic in the `cdc15`, `cdc28` and `alpha` experiments. While plots of the expression levels through time of these genes did confirm their periodicity, only four genes were identical across different synchronization methods. Therefore, if we subscribe to the assumption that the `elution` data set represents the least perturbed data set, this lends additional support to the arguments presented in Shedden and Cooper (2002b) against the presence of a large

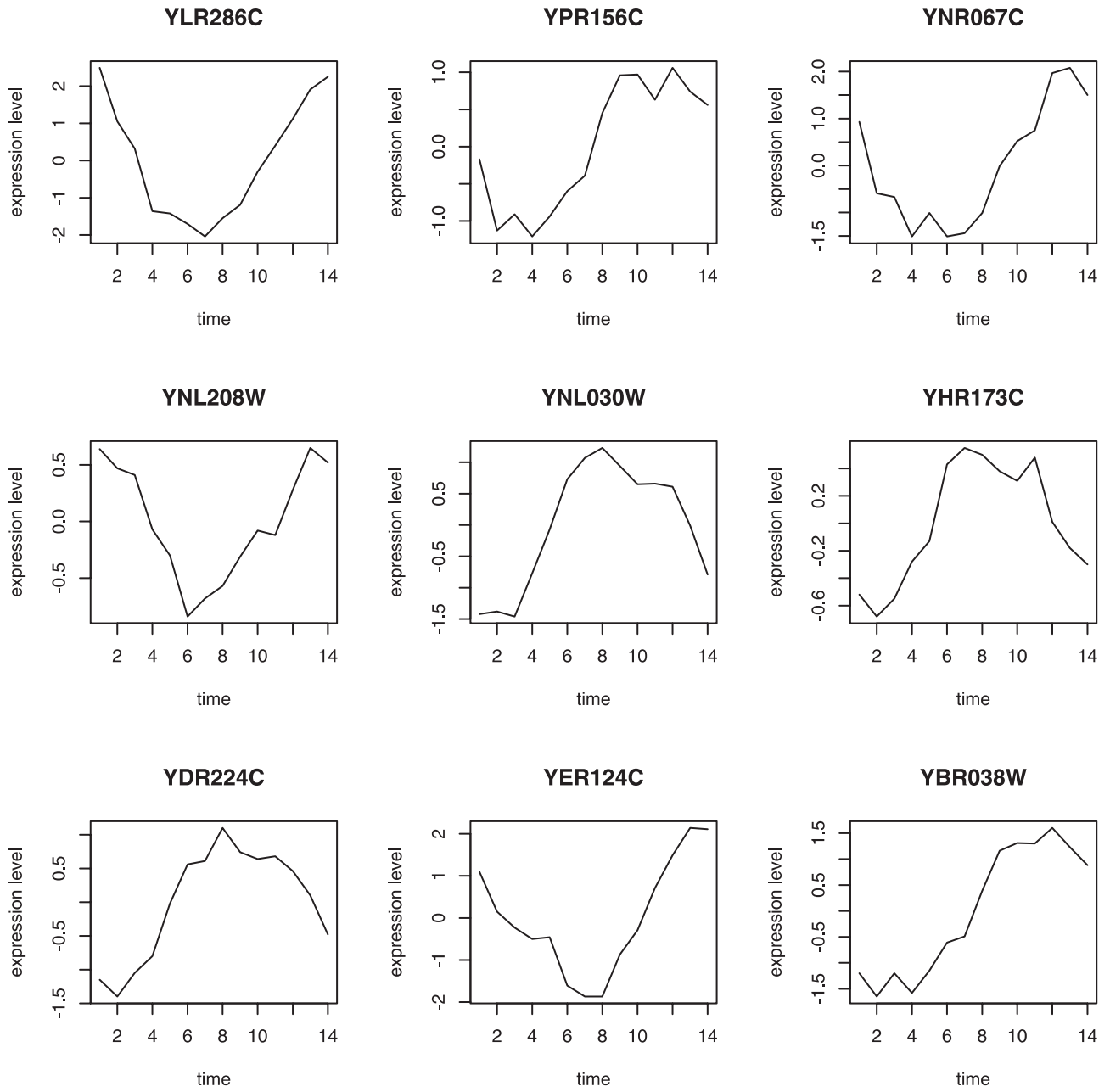

Fig. 6. The nine statistically most significant periodic genes in the `elution` data set according to p-values from Equation (6).

number of generic periodic cell-cycle-regulated genes in the data of Spellman *et al*. (1998). Nevertheless, our analysis does confirm 193 periodically expressed genes in `elution` and the temporal expression patterns of the nine statistically most significant periodic genes in `elution` can be inspected in Figure 6. However, the problem still exists as to whether these periodicities that we have identified in `elution` are due to artifacts or correspond to real cell-cycle-regulated expression pattern.

Bacterial cell cycle The next example is about gene expression data from synchronized cultures of the bacterium *C.crescentus* (Laub *et al*., 2000). The corresponding data matrix `bacteria` contains information on 1444 genes over 11 time points, and is therefore among the shortest time series considered here. The average periodogram computed for this data set, shown in Figure 7, is clearly indicative of the presence of cell-cycle-specific genes. It is worth noting that this contrasts positively with other microarray data of comparable

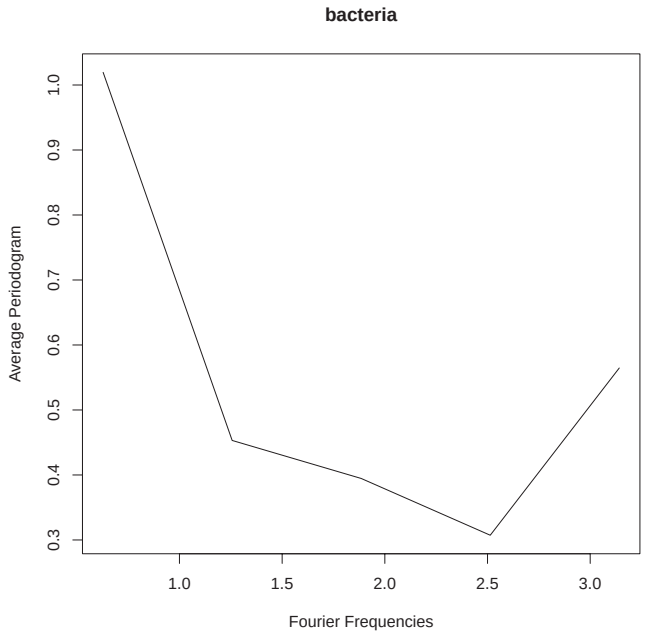

Fig. 7. Average periodogram for the *C.crescentus* `bacteria` data set (also Table 2).

sample size that are investigated elsewhere in this paper (`N2`, `N3`, `score5`) and that did not show any signs of periodicity in the average periodogram.

The suitability of this data set for statistical analysis was further confirmed in the application of the g-test to identify periodically expressed genes. Despite the availability of only 11 measurements per gene, 44 open reading frames were found where the null hypothesis of a random non-cyclic process could be ruled out on a FDR level of $q = 0.05$. For illustration, the expression levels through time of the nine genes with the smallest p-values are plotted in Figure 8. The complete list contains key regulatory genes such as crtA, indirectly regulated genes (e.g. for the flagella biogenesis) and repressors (e.g. lexA). Thus our results broadly confirm the findings in Laub *et al*. (2000). However, statistical support is provided only for the 44 top-ranked genes. We could not verify the claim of 553 cell-cycle-regulated genes in Laub *et al*. (2000). However, given the sparsity of available time points, this is not surprising.

Human fibroblasts Next, we consider data from cell cycle experiments using human fibroblasts cells. The design of the microarray experiments is described by Cho *et al*. (2001). The resulting data sets are two short time series `N2` and `N3` with 13 and 12 measurements per gene, respectively.

The results from an explorative inspection of these data using the method of average periodogram are shown in Figure 9. To complement the plots for `N2` and `N3`, we also show in Figure 9 the average periodogram for completely random data simulated with the same number of genes and sample points as in `N2` and `N3`. Interestingly, we do not find any evidence of periodic gene expression in the data that would be different from a purely stochastic process. A subsequent formal test using Fisher's g-statistic and FDR multiple testing also did not detect any periodicity.

This confirms Shedden and Cooper (2002a) who first raised doubts on biological grounds whether the cell cycle data from the human fibroblast experiments of Cho *et al*. (2001) was suited for statistical analysis. Note that the negative results cannot simply be due to the short sample size: the similarly short `elution` data from the yeast experiment showed clear evidence of periodicity both by graphical inspection and by formal testing. Another counter example is the even shorter `bacteria` data set discussed in the previous section. Our results therefore also indicate that there may have been a serious experimental problem, for instance failed forced cell synchronization (Cooper, 2003).

Human cancer cell line

In a pivotal paper, Whitfield *et al*. (2002) described a large-scale study on human cancer cells (HeLa S3) and established a catalog of human cell-cycle-regulated genes (http://genome-www.stanford.edu/Human-CellCycle/HeLa). Five experiments (`score1`, `score2`, `score3`, `score4`, `score5`) were conducted using microarray chips with 23 000–43 000 probes measuring the expression levels of up to approximately 30 000 genes. Three different cell cycle synchronization methods were used, a double thymidine block (`score1`, `score2`, `score3`), thymidine followed by arrest in mitosis with nocodazole (`score4`) and mitotic shake-off using an automated cell shake (`score5`). Measurements were taken for up to 48 time points (`score3`), which makes this study one of the most extensive microarray time series experiment so far.

The results for the data sets are summarized in Table 2 and in Figure 10. Visual exploration using the average periodogram gave clear evidence for periodicity in all data sets but `score5` which yields to a flat line (Fig. 10). In the subsequent gene selection test with an assumed FDR level of $q = 0.05$ the results were also similarly heterogenous. As expected `score5` did not contain any statistically significant periodic genes. The same was true for `score1` despite a promising average periodogram, whereas `score2` and `score4` contained a small fraction of detected periodic transcripts. For the longest time series `score3`, our FDR test detected a large number of statistically significant periodic elements.

These differences are likely to be mainly a consequence of sample size. For example, the mitotic shake-off synchronization used in `score5` is probably the least perturbing method but the short length of this time series (9 time points) prohibits detection of significant periodic genes. Similarly, as `score3` uses the same synchronization as `score1`, the lack

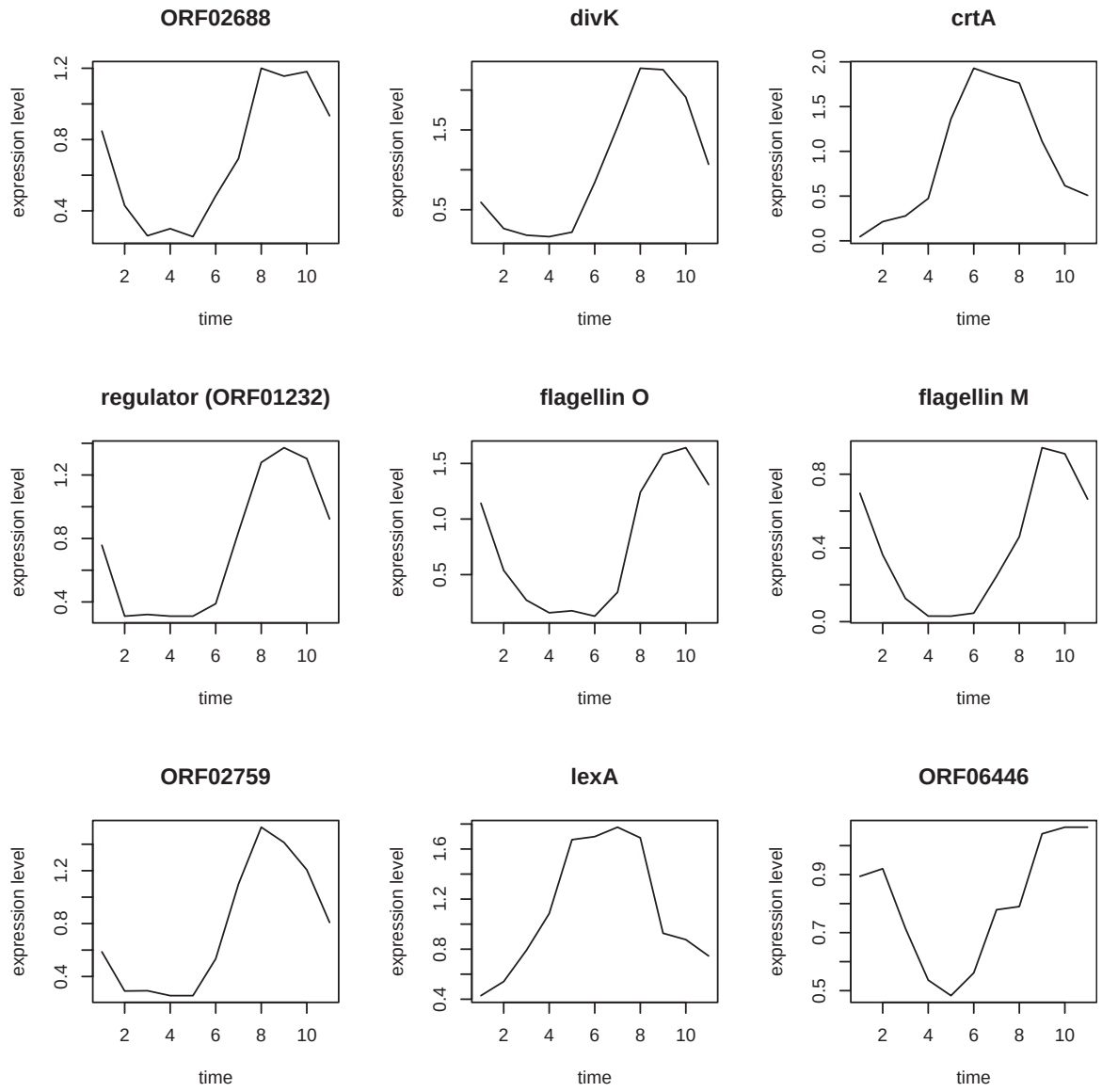

Fig. 8. The nine statistically most significant periodic genes in the `bacteria` data set.

of inferred periodic genes in `score1` must be a consequence of the small number (12) of measurements per gene.

Compared with the original analysis we were able to identify a substantial amount of additional periodic genes in `score3`. Figure 11 shows as an example the expression-through-time plots of the nine probe elements in `score3` with the smallest overall p-values found by our method. While these elements are clearly periodic they are not designated as such in the online database of Whitfield *et al*. (2002) at http://genome-www.stanford.edu/Human-CellCycle/HeLa. One reason for this discrepancy is that in contrast to the approach by Whitfield *et al*. (2002), our method is not based on the amplitude difference between the maximum and minimum expression level of each gene through time and, therefore, it also considers genes with small but still statistically significant amplitude changes.

Whitfield *et al*. (2002) tabulate their periodicity score and related test results for a number of selected inferred genes (their Table 1) and for cell-cycle-regulated genes known from the literature (their Table 2). First, we re-analyzed genes from their Table 1 using our approach, with the results shown in Table 3. Not all genes considered periodic in Whitfield *et al*. (2002) test positive in our setting. For instance genes PRIM1 and PSEN1 fail our test procedure. Interestingly, a simple plot of their expression levels against time reveals that the expression of these two genes does in fact exhibit a fairly random rather than a clear periodic

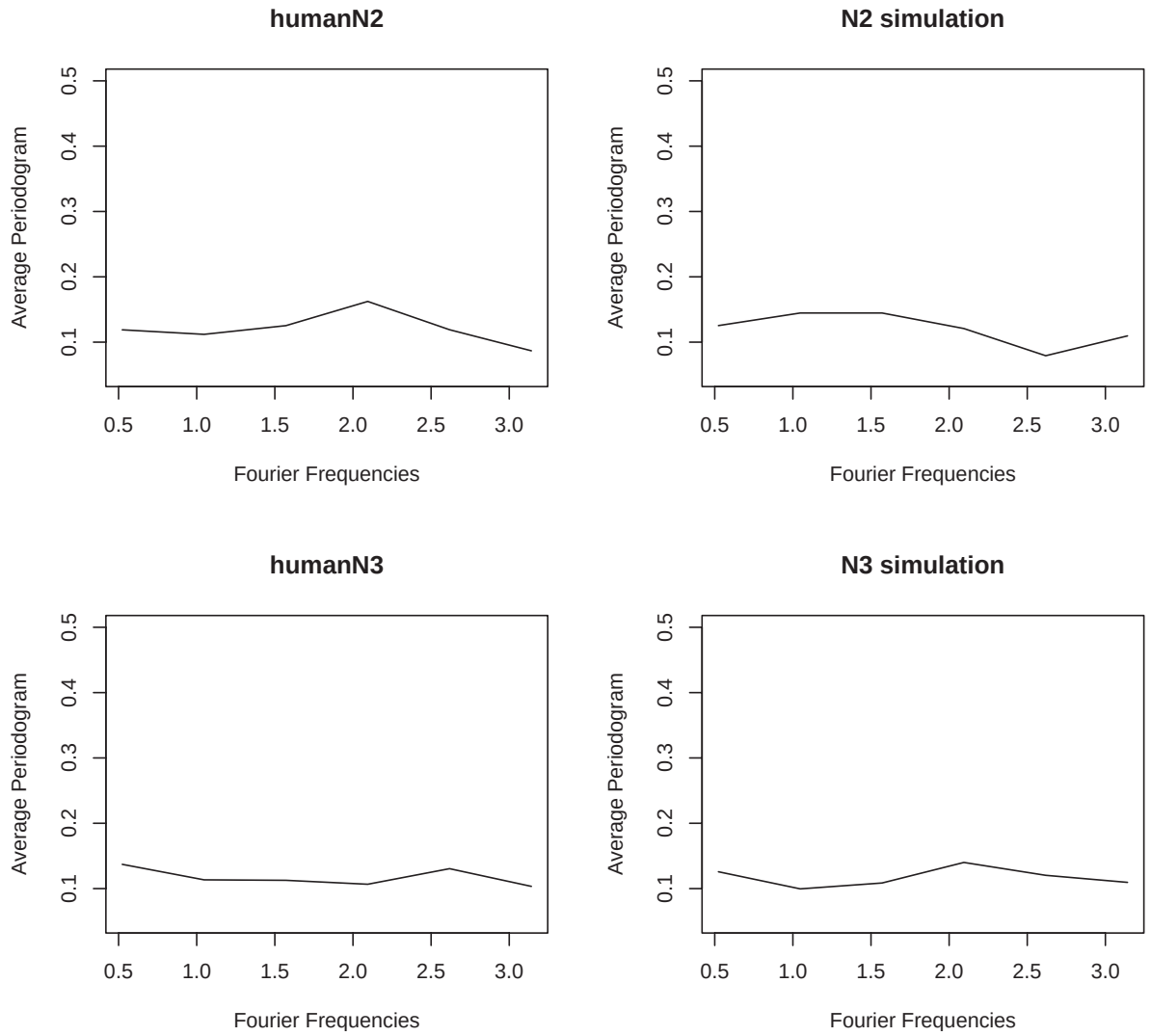

Fig. 9. Average periodogram for the two human fibroblast data sets `N2` and `N3`, compared with simulated random data (also Table 2).

pattern (Fig. 12). Second, we have also tested probes for all genes listed in Table 2 of Whitfield *et al*. (2002). Our results (data not shown) generally concur with those of Whitfield *et al*. (2002). However, in contrast to the approach used by these authors our method detect periodicity in the CDKN1A gene, in Histone H3 elements and rejects cyclicity for CDKN2D.

Finally, we would again like to add a word of caution. While periodic results may be obtained, as pointed out by Shedden and Cooper (2002a,b), such periodicities may be due to perturbations that have nothing to do with the cell cycle. What is most problematic about the data of Whitfield *et al*. (2002) is that the flow-cytometric data presented to support synchronization do not indicate any unperturbed cell synchronization. The initial cells in double-thymidine and the thymidine–nocodazole experiments have too much DNA per cell, and the subsequent samples do not indicate anything near appropriate synchronization. Hence, it is very likely that many of the genes identified by Whitfield *et al*. (2002), as well as the genes identified here, may be due to cell perturbations and not due to variations within the normal cell cycle.

DISCUSSION

We have presented two new promising tools for microarray time series analysis. First, the average periodogram was suggested as an exploratory device to assess whether a data set contains a signature indicative of the presence of periodically expressed transcripts. Second, we have developed a formal

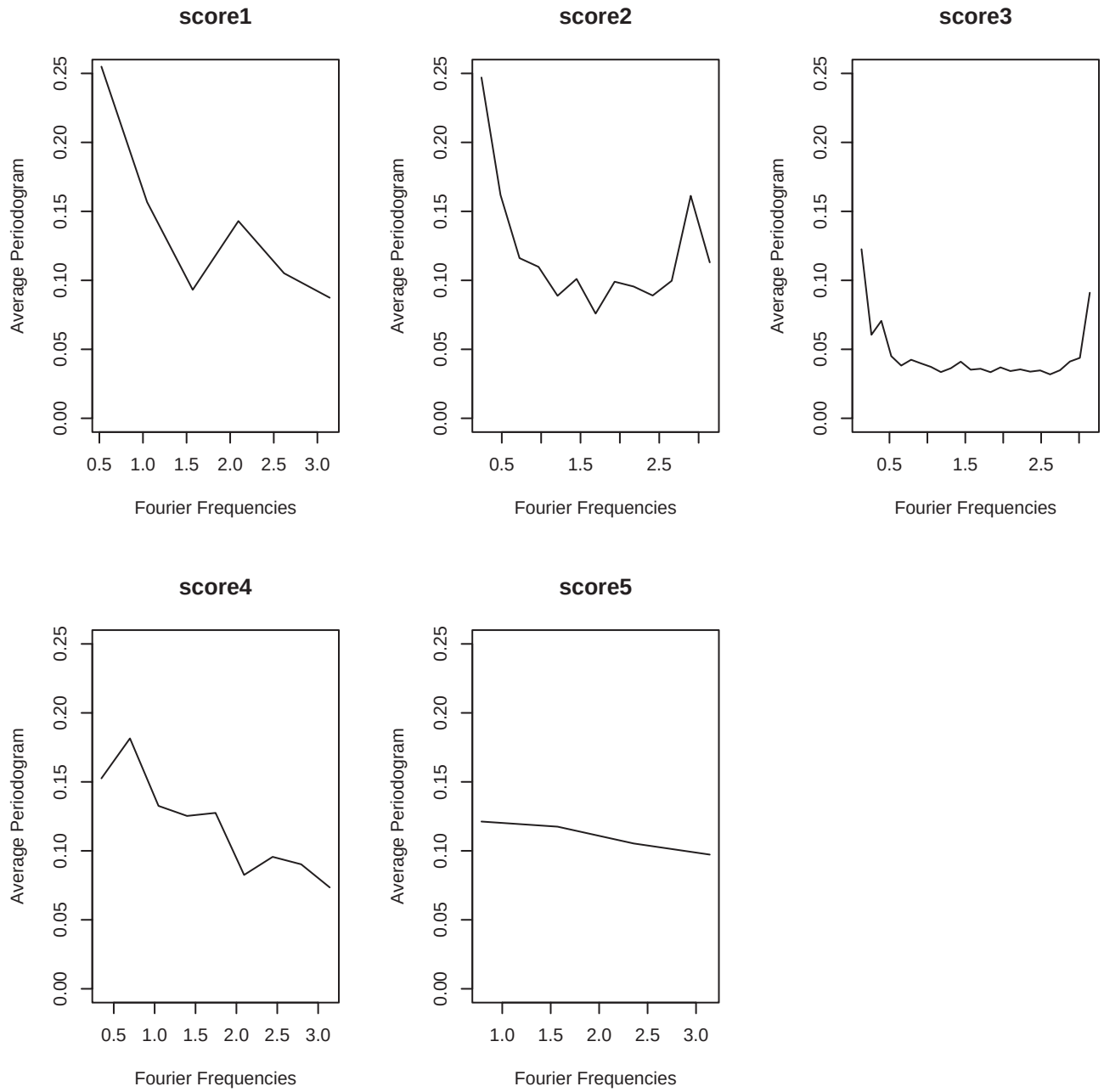

Fig. 10. Average periodogram for the five human HeLa data sets `score1`, `score2`, `score3`, `score4` and `score5` (also Table 2).

statistical test for gene selection based on Fisher's g-statistic and FDR multiple testing that allows to screen for individual periodic genes.

A novelty of our approach is that it combines visual inspection with a rigorous and exact testing procedure. Unlike most methods in classical time series analysis, our approach is also applicable to data sets with a comparatively small number of measurements per gene. The average periodogram as a non-parametric method of obtaining frequency estimators is suited to detect periodic signal in very short time series as it takes advantage of the parallel structure of the data. This distinguishes this method from other visualization strategies, such as geometric trajectories and correspondence analysis (Fellenberg *et al*., 2001)

The g-statistic requires works on the level of the individual gene and hence requires more sample points per gene than the average periodogram. However, unlike other related methods (Whitfield *et al*., 2002), it allows one to

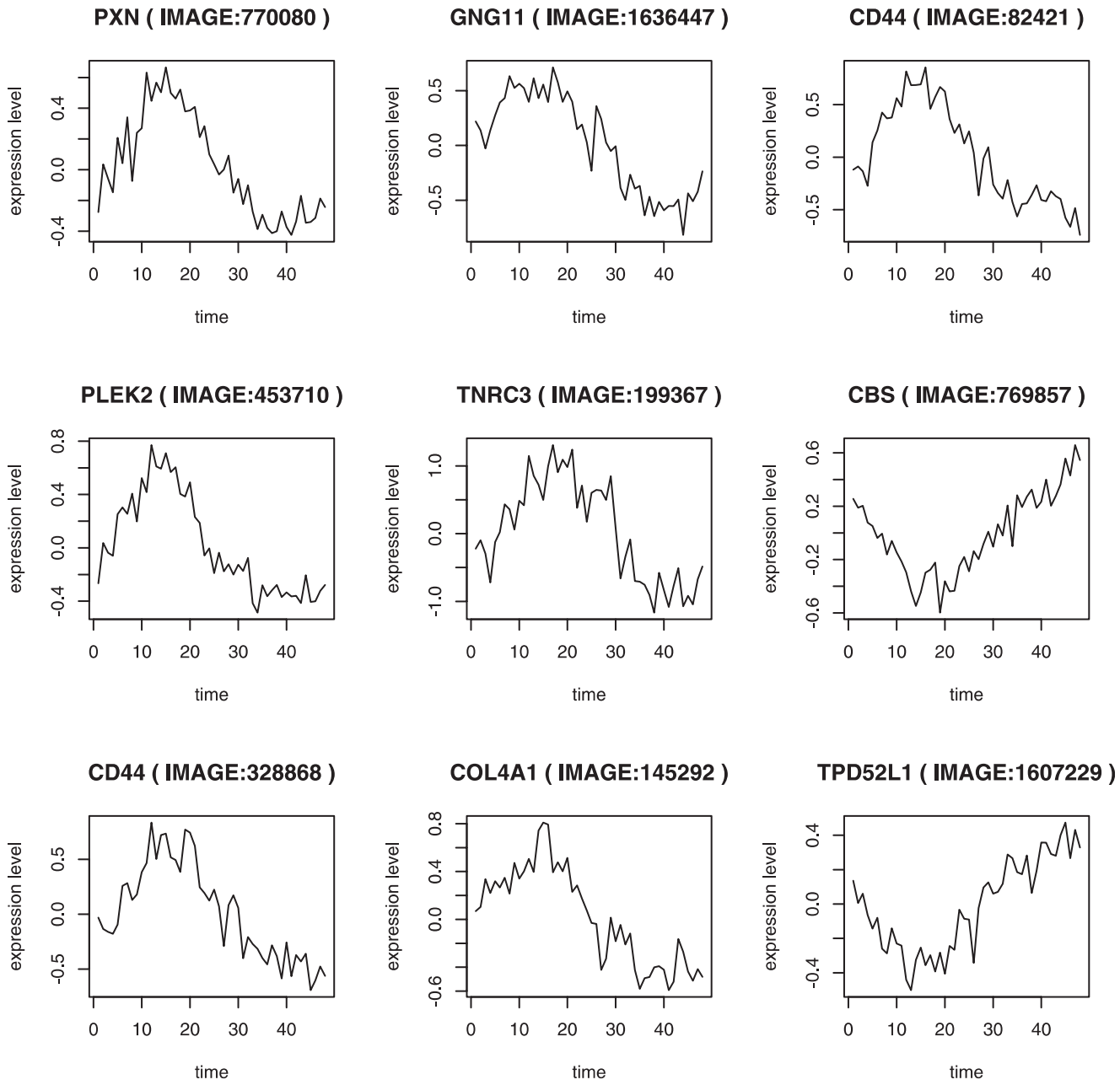

Fig. 11. The nine statistically most significant periodic genes in the `score3` data set.

detect statistically significant periodically expressed genes even with small amplitudes changes and is well defined for finite samples. Finally, the application for the method of FDR takes account of the multiple testing procedure.

While our approach performed well in identifying genes with periodic variation in the application to simulated and to 12 large microarray data sets, there are also some potential drawbacks. First, the g-statistic assumes as null-model a purely Gaussian process. It is unclear whether this hypothesis is a valid assumption for microarray data. However, high-level analysis such as gene selection will typically take place on preprocessed and transformed data [e.g. log-transformed to stabilize variance, see Strimmer (2003) for a list of references] so that the underlying stochastic processes may be well approximated by Gaussian assumptions. Second, the investigated genes are usually correlated and hence do not evolve independently. This may have an impact on our analysis. While it appears that the assumption of independence is not critical in the average periodogram or in FDR multiple testing, this would deserve further study.

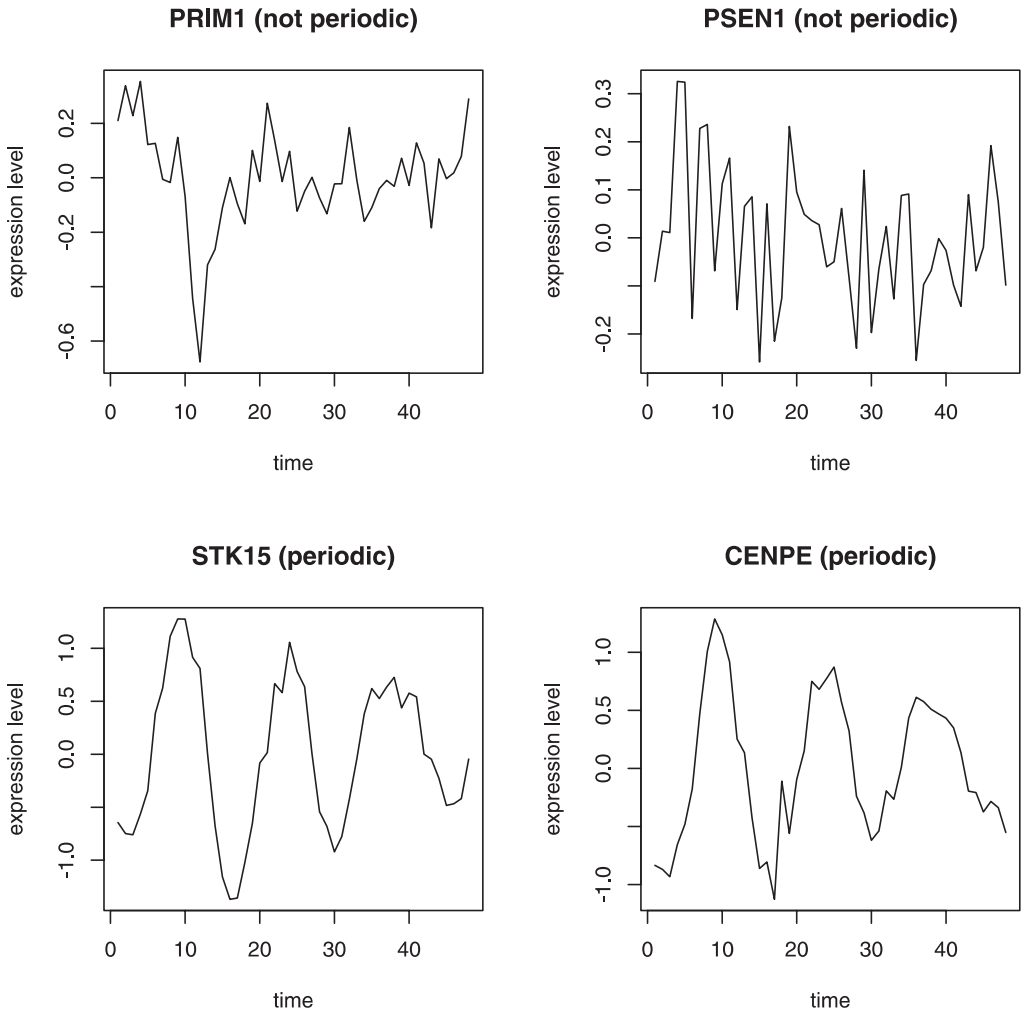

Fig. 12. Expression-through-time plot of `score3` genes PRIM1 and PSEN1 [top row, periodic in Whitfield *et al*. (2002), not considered periodic in this paper] and STK15 and CENPE (periodic, bottom row).

From the data analysis presented in this paper we conclude mainly two things. First, there seems to be a remarkable gradient of signal quality in the data sets publicly available. Looking at data sets with identical sample size we find some very well suited for statistical analysis (`bacteria`) while others (`N2`, `N3`) appear to contain mostly random noise. Hence, sample size is not everything but the experimental background, in particular the synchronization approach, seems to be at least equally important, see e.g. Cooper (2003) for a critique of forcing synchronization methods. Thus caution needs to be taken in the biological interpretation, as to whether the detected genes are linked to cell-cycle-specific tasks or to artifacts of the experiment. Second, having said this, we are bound to underestimate the number of cyclic genes in statistical test for small-length data sets. From our preliminary simulation results for the g-statistic, as well as from the analysis of the human HeLa cell line data, it would seem that for a reliable detection of cell-cycle-regulated genes at least 40 measurements per gene are desirable. We recommend that these considerations should be taken into account when planning future cell cycle studies.

ACKNOWLEDGEMENTS

Part of this work was carried out while K.F. was a visiting professor at the Department of Statistics of the University of Munich. We thank Anne-Laure Boulesteix and Juliane Schäfer for critical comments and discussion. We are also grateful to one anonymous referee whose extensive comments helped to improve greatly the presentation of our results. S.W. and K.S. were supported by an Emmy Noether research grant (STR 624/1-2) from the Deutsche Forschungsgemeinschaft (DFG).

Table 3. Test of periodicity of selected genes [in Table 1 in Whitfield *et al*. (2002)]

Gene	Periodic in		*p*-value
	Whitfield *et al*. (2002)	This paper	
STK	√	√	8.72e−09
PLK	√	√	1.78e−08
CENPE	√	√	4.89e−09
CKS2	√	√	5.29e−08
CCNA2	√	√	4.41e−08
CDC6	√	√	7.27e−11
CCNB1	√	√	4.79e−10
E2F1	√	√	2.17e−09
PCNA	√	√	8.20e−10
CIT	√	√	2.04e−03
GOLGIN-67	√	√	2.36e−05
ORC1	√	√	8.90e−10
PRIM1	√	—	9.92e−03
VCL	√	√	2.20e−06
BRCA1	√	√	6.84e−04
CCNE2	√	√	2.68e−03
PSEN1	√	—	1.16e−02
CDC42	√	√	9.67e−06
ZNF162	—	—	3.59e−02
YY1	—	—	1.79e−01

p-values are computed using the `score3` data set. For genes with multiple probes the result with the smallest *p*-value is reported. The FDR testing procedure gave a critical *p*-value of $p = 0.0076$ for a desired FDR level of $q = 0.05$. See also Figure 12 for a expression-through-time plot of of the non-periodic genes PRIM1 and PSEN1.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JR Statist. Soc. B*, **57**, 289–300.

Cho,R.J., Huang,M., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W., Lockhardt,D.J. and Campbell,M.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.

Cooper,S. (2003) Reappraisal of serum starvation, the restriction point, G0, and G1-phase arrest points. *FASEB J.*, **17**, 333–340.

Fellenberg,K., Hauser,N.C., Brors,B., Neutzner,A., Hoheisel,J.D. and Vingron,M. (2001) Correspondence analysis applied to microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 10781–10786.

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

Laub,M.T., McAdams,H.H., Feldblyum,T., Fraser,C.M. and Shapiro,L. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.

Priestley,M.B. (1981) *Spectral Analysis and Time Series*. Academic Press, London.

Rifkin,S.A. and Kim,J. (2002) Geometry of gene expression dynamics. *Bioinformatics*, **18**, 1176–1183.

Shedden,K. and Cooper,S. (2002a) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl Acad. Sci. USA*, **99**, 4379–4384.

Shedden,K. and Cooper,S. (2002b) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res.*, **30**, 2920–2929.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Strimmer,K. (2003) Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics*, **4**, 10.

Welch,P.D. (1967) The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, **AU-15**, 70–73.

Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murrary,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.