BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 8 2008, pages 1063–1069
doi:10.1093/bioinformatics/btn072

Gene expression

Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis

Matthias E. Futschik* and Hanspeter Herzel

Institute for Theoretical Biology, Charité, Humboldt-Universität, Invalidenstrasse 43, 10115 Berlin, Germany

Received on October 30, 2007; revised and accepted on February 20, 2008

Advance Access publication February 29, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Periodic processes play fundamental roles in organisms. Prominent examples are the cell cycle and the circadian clock. Microarray array technology has enabled us to screen complete sets of transcripts for possible association with such fundamental periodic processes on a system-wide level. Frequently, quite large numbers of genes have been detected as periodically expressed. However, the small overlap between genes identified in different studies has cast some doubts on the reliability of the periodic expression detected.

Results: In this study, comparative analysis suggests that the lacking agreement between different cell-cycle studies might be due to inadequate background models for the determination of significance. We demonstrate that the choice of background model has considerable impact on the statistical significance of periodic expression. For illustration, we reanalyzed two microarray studies of the yeast cell cycle. Our evaluation strongly indicates that the results of previous analyses might have been overoptimistic and that the use of more suitable background model promises to give more realistic results.

Availability: R scripts are available on request from the corresponding author.

Contact: matthias.futschik@charite.de

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

1 INTRODUCTION

Periodicity is an important phenomenon in molecular biology and physiology. Many fundamental processes follow periodic patterns of activation. One intensely studied periodic process is the cell cycle. In all organisms, it underlies growth and reproduction, the distinct features of life. On the microscopic level, this comprises the replication of DNA and the division of cells into daughter cells equipped with the structure necessary for correct functioning. Although the core machinery of the cell cycle is well studied, the effects on the whole system have been less well defined.

Microarray technologies have enabled us to measure genome-wide changes in expression, thus, permitting a system-wide assessment of periodic patterns. Microarray studies of the cell cycle in different organisms have indicated that periodic expression may not be restricted to a small number of genes, but that a substantial part of the transcriptome undergoes periodic activation during the cell cycle (Cho *et al.*, 1998; Spellman *et al.*, 1998). However, it should be noted that microarrays have their limitations: The produced data are frequently compromised by a high inherent level of noise as well as by various experimental biases (Futschik and Crompton, 2004). Furthermore, special caution in the interpretation of microarray data has to be taken, since the large amount of generated data leads to the emergence of many kinds of patterns merely due to chance (Ambroise and McLachlan, 2002). This increases the risk of detecting patterns that satisfy the assumptions of researchers but which may have arisen at random. A prominent example of this 'self-fulfilling prophecy' might be the study of the human cell cycle by Cho and co-workers (Cho *et al.*, 2001). The authors detected several known and many apparently novel cell-cycle-regulated genes. However, Shedden and Cooper (2002a) could convincingly demonstrate in a follow-up analysis that most of these detected genes do not show a reproducible periodic pattern.

Thus, stringent statistical methods are essential to assure the reliability of the periodic expression detected. Several approaches for detection have been proposed based on time-series analysis and statistical modeling (Johansson *et al.*, 2003; Spellman *et al.*, 1998; Wichert *et al.*, 2004; Zhao *et al.*, 2001). [For a recent comparison of their performance, please refer to the study by de Lichtenberg and collegues (2005).] To assess the significance of the identified periodic expression, most of the proposed methods rely on data normality or the extensive use of permutation tests. However, this neglects the fact that time-series data exhibit generally a considerable autocorrelation i.e. correlation between successive measurements. Therefore, neither the assumptions of data normality nor for randomizations may hold.

We show in this study that this failure can substantially interfere with the significance testing, and that neglecting autocorrelation can potentially lead to a considerable overestimation of the number of periodically expressed genes. For illustration, we re-examined two microarray studies of the yeast cell cycle which have been intensively analyzed by various methods. While these methods usually detected a large number of periodically expressed genes (ranging from about 300 to 800), there was remarkably little agreement in the set of genes identified in different experiments (de Lichtenberg *et al.*, 2005;

*To whom correspondence should be addressed.

Shedden and Cooper, 2002a; Zhao *et al.*, 2001). Our study suggests that one reason for the observed lack in agreement could be an overestimation of the number of periodically expressed genes due to the use of inadequate background models.

2 METHODS

2.1 Expression studies of the yeast cell cycle

As case studies we re-analyze two yeast cell-cycle microarray experiments. The first study included the expression of over 6000 genes derived by employing Affymetrix chips (Cho *et al.*, 1998). Synchronization was achieved using temperature sensitive yeast cells (CDC28). Samples of cells were taken every 10 min for 160 min. This period of time included two cell cycles. By visual inspection of expression patterns, Cho *et al.* found over 400 genes showing periodicity.

We excluded genes if more than 25% of the expression measurements during the time course were missing. Affymetrix signals were converted into ratios by dividing the expression of genes by the average value. After $\log_2$-transformation, missing values were replaced by estimates derived by the knn-method (Troyanskaya *et al.*, 2001). Data were standardized to have mean values equal to zero and SD equal to one for subsequent time-series analysis. Optionally, additional scaling by quantile normalization was performed (Bolstad *et al.*, 2003). The distributions of expressions values for the datasets before and after scaling are shown in Figure S6.

As second dataset, we use the microarray experiments of the yeast cell cycle by Spellman and colleagues (1998). Synchronization of the cell cultures was similarly achieved as in the experiment by Cho *et al.*, but using the mutant CDC15 strain. Sampling was performed over almost three cell cycles (290 min). Transcript levels were measured using two-color cDNA arrays including over 6000 genes. For reference RNA, cells were grown without synchronization. Using Fourier analysis and additional experiments, Spellman and colleagues (1998) found 800 cell-cycle-regulated yeast genes. Except for the conversion in ratios, we performed the same pre-processing as for the dataset by Cho *et al.*

2.2 Detection of periodic signals in microarray data

The described microarray experiments deliver time-series data i.e. gene-expression values in a well-defined order. To detect periodic signals within the large datasets, several different approaches have been put forward ranging from simple visual inspection (Cho *et al.*, 1998) to elaborated statistical models (Lu *et al.*, 2004). Recently, an extensive comparison showed that a relatively simple permutation-based method using Fourier analysis performs better than other approaches. It is based on the Fourier score defined as

$$F[\mathbf{g}] = \sqrt{\left(\frac{\sum_i \cos(2\pi t_i)}{T}\cdot g_i\right)^2 + \left(\frac{\sum_i \sin(2\pi t_i)}{T}\cdot g_i\right)^2} \qquad (1)$$

where $\boldsymbol{g}$ is the vector of standardized expressions g_i ($\mathrm{mean}(\boldsymbol{g})=0$; $\mathrm{sd}(\boldsymbol{g})=1$), T is the period of the cell cycle and g_i is the expression measured at time t_i. The closer a gene's expression follows a (possibly shifted) cosine curve of period T, the larger is the score F. To identify periodicity, Fourier scores were calculated for the temporal expression of each gene. For the cell-cycle period, the values were taken from the original publications, i.e. $T=85$ min for CDC28 and $T=115$ min for CDC15 (Spellman *et al.*, 1998).

2.3 Background models for time-series data

Microarray data comprise the measurements of transcript levels for many thousands of genes. Due to the large number of genes, it can be expected that some genes show periodicity simply by chance. To assess therefore the significance of periodic signals, it is necessary first to define what distribution of signals can be expected if the studied process exhibits no true periodicity. In statistical terms this is equivalent with the definition of a null hypothesis of non-periodic expression. The most simple model for non-periodic expression is based on randomization of the observed times series. A background distribution can then be constructed by (repeated) random permutation of the sequentially ordered measurements in the experiment. Alternatively, non-periodic expression can be derived using a statistical model. A conventional approach is based on the assumption of data normality. In case that the time-series data has been standardized ($\sigma=1$), a background distribution can be readily generated.

In time-series analysis, an important class of stochastic processes is the autoregressive processes for which the value of the time-dependent variable X_t depends on past values of X up to a normally distributed random variable Z. Of special interest here are autoregressive processes of order (AR(1)):

$$X_t = \alpha_1 \cdot X_{t-1} + Z_t \qquad (2)$$

for which α_1 is equal to the correlation coefficient of X_t and X_{t-1} (i.e. the autocorrelation of X_t with a time lag of one) and Z_t is an independent random with a mean value of zero and variance σ_z^2. In our case, X_t denotes the expression of a gene at time t. The value of α_1 and σ_z^2 can be estimated for each gene separately using maximum likelihood estimation (see Supplementary Materials). Thus, we can approximate the observed time series X_t as AR(1) process. It is important to note in this context, that AR(1) processes cannot capture periodic patterns except for alternations with period two. Since Z_t is a random variable, we can readily generate a collection of time series with the same autocorrelation as in the original dataset. Therefore, although AR(1) processes constitute random processes, they allow us to construct a background distribution that captures the autocorrelation structure of original gene-expression time series without fitting the potentially included periodic patterns. An illustration of the different background models can be found in the Supplementary Materials (Fig. S1–S3).

An important (and in this context crucial) characteristic of time series is their power spectrum. The power spectrum (or spectral density distribution) I represents the strength of periodic components in a signal with respect to their frequency. It can be calculated for a time series of length N using Fourier analysis:

$$I[f_p] = \frac{\left[\left(\sum_i \cos(2\pi f_p t_i) g_i\right)^2 + \left(\sum_i \sin(2\pi f_p t_i) g_i\right)^2\right]}{N\pi} \qquad (3)$$

The frequencies are $f_p = p/N$ with integer p ranging from 1 to $N/2$. Note that the Fourier score defined in Equation (1) is equal the square root of the spectral density at the cell-cycle frequency (up to a normalization constant).

Figure 1 shows the power spectra for an uncorrelated random and an AR(1) process. The spectrum for an uncorrelated random process (which is assumed for the randomized and Gaussian background model) is constant over the frequency range. This is in remarkable contrast to the spectrum obtained for an AR(1) process with autocorrelation of 0.5 which shows larger power at lower frequencies (Fig. 1B). It should be noted, however, that the spectrum of AR(1) processes depends on the underlying autocorrelation coefficient with negative autocorrelation yielding to larger power at higher frequencies.

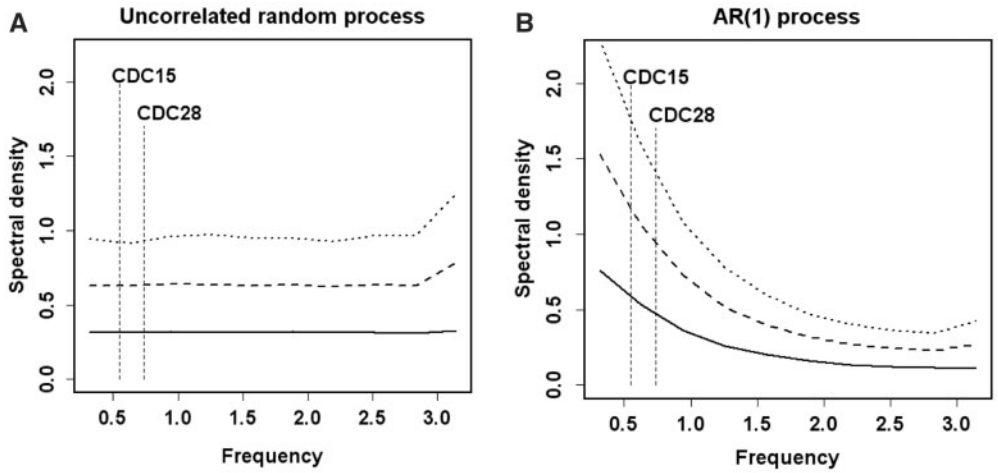

Fig. 1. Spectral density distributions for uncorrelated random and AR(1) processes. The distributions were calculated based on 10 000 independent simulations of time series with length 20. They closely follow the analytical expressions for power spectra that can be derived for processes of infinite length: $I(f) = \sigma_x^2/\pi$ for Gaussian and $I(f) = \sigma_x^2(1-\alpha^2)/(\pi\,(1-2\alpha\cdot\cos\omega+\alpha^2))$ for AR(1) processes. (Chatfield, 1995). The frequency was scaled so that the maximum detectable frequency (i.e. Nyquist frequency) is equal to π. Solid lines represent the mean spectral density; dashed lines represent the mean plus the SD; and dotted lines indicate the upper 90% level of the distributions. Vertical dashed lines indicate the cell-cycle frequency of the two analyzed yeast strains. For the AR(1) process, an autocorrelation coefficient of 0.5 was chosen.

2.4 Significance of periodic signals

To assess the significance of the Fourier score obtained for the original gene-expression time series, the probability has to be calculated of how often such a score would be observed by chance based on the chosen background distribution. Since multiple testing is involved, we used the false discovery rate (FDR) to represent the statistical significance. It is defined here as the expected proportion of false positives among all genes detected as periodically expressed. Thus, we can calculate the empirical false discovery rate for a chosen threshold c for the Fourier score:

$$\mathrm{FDR}(c) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{N}\delta(F_{ij}^{b} \geq c)/n}{\sum_{i=1}^{N}\delta(F_{i}^{o} \geq c)} \quad (4)$$

where F_i^o is the Fourier score derived for the ith gene for the original observation, F_{ij}^b is the Fourier scores for the ith gene in the jth independent generation of background series, N is the total number of genes, n is the total number of generated background series for each gene and $\delta(x) = 1$ for $x \geq 0$, respectively, $\delta(x) = 0$ for $x < 0$. Thus, the significance of the measured periodicities can be obtained by comparison with the generated background distribution.

3 RESULTS

To study the influence of background models on the detection of periodic patterns, we re-analyzed two yeast cell-cycle microarray experiments. After preprocessing of the two datasets (CDC15 and CDC28) we generated background distributions on following procedures: (i) *Randomized* background distributions were produced by repeated random permutation of the observed time series for every gene; (ii) *Gaussian* background distributions were derived from sampling of the normal distribution and (iii) *AR(1)*-based background distributions were constructed by fitting the original data to AR(1) processes and subsequent generation of random time-series based on the obtained fitting parameters.

3.1 Autocorrelation in cell-cycle datasets

Significance of periodicity in microarray data is often assessed by comparison of the observed data with background distributions. Most approaches so far use randomized data or assume data normality to construct background distributions (Spellman *et al.*, 1998; Wichert *et al.*, 2004). Their usage implies that no correlation occurs between successive measurements within the time series for non-periodic genes. However, many time series in nature exhibit autocorrelation. A first indication that this is also true for the yeast cell-cycle datasets is given by cluster analysis. Besides clusters showing periodic patterns, many other expression profiles occur (Fig. 2). The displayed prominent non-periodic trends might have been caused by the applied synchronization procedure inducing initial stress responses and slowly decaying perturbations. Such trends are also biologically meaningful as transcript levels within a cell at a certain time are at least partially determined by their levels in the past. However, as these trends may arise by chance, a more stringent assessment of the data structure is needed. Therefore, we calculated the gene-wise correlation matrix between all measurements (i.e. arrays). For both datasets, considerable autocorrelation was detected (Fig. 3A). Directly successive measurements generally showed a clear correlation (e.g. Pearson correlation of 0.29 ± 0.17 for CDC28). Note, that this the pattern remains prominent even after exclusion of highly periodic genes (Fig. S7). Temporally more distant measurements seemed to be anti-autocorrelated supporting the existence of long-term trends as indicated by cluster analysis. In summary, both time series exhibit clear autocorrelation.

This was contrasted by the correlation matrix that we calculated for randomized and Gaussian background distributions (Fig. 3B and C). For these generated datasets, the autocorrelation generally was negligible. For example, a Pearson correlation between directly successive arrays of

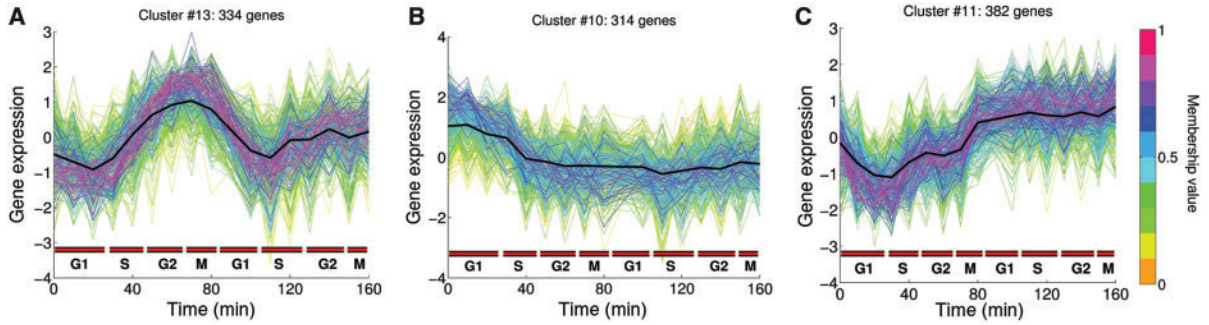

Fig. 2. Examples of periodic (left) and aperiodic (middle, right) expression patterns in the CDC28 dataset. The clusters were detected by a soft-clustering approach which allows differentiation of cluster membership. The membership values are color-encoded according to the color-bar on the far right site. Details of the applied clustering method can be found in Futschik and Carlisle (2005).

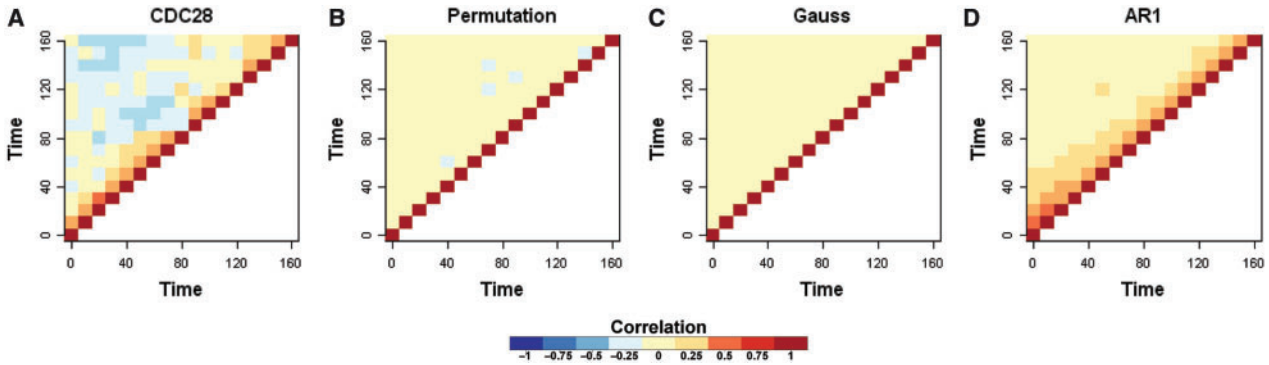

Fig. 3. Autocorrelation in original dataset and background distributions: The upper triangle of the correlation matrix is displayed with respect to the temporal ordering of arrays. The color-bar shows the color-coding of the observed correlation. The original dataset CDC28 was standardized and scaled.

-0.06 ± 0.02 and of 0.001 ± 0.02 was calculated for randomized (permutated) and Gaussian background distributions, respectively. For AR(1)-based background distributions, however, we obtained clear correlation patterns (Fig. 3D). Similarly to the original data, directly successive measurements were significantly correlated (0.39 ± 0.07). Note that the anti-correlation detected in the original dataset between distant measurements is not reflected. This is not surprising as we have restricted the order of the autoregressive process to one to avoid interference with the detection of periodicity. Nevertheless, the comparison shows that the AR(1)-based background reflects the important feature of autocorrelation as observed in the original datasets. This is also supported by a comparison of the distributions of autocorrelation coefficients α for the different background models (Fig. S5). Thus, the AR(1) model can provide a more accurate background model for significance testing.

3.2 Impact of background models on significance testing

To examine the impact of background models on significance testing, we generated 100 independent distributions for each type of background model for the two original datasets. These independently generated distributions were subsequently merged for each background model and used for the calculation of the Fourier score. Examples of time series generated by different background models are shown in Fig. S1–3.

Figure 4 displays the distributions of Fourier scores obtained for the original datasets and the corresponding background models. Randomized and Gaussian background led to very similar distributions of Fourier scores. Notably, the proportion of expression vectors with large scores (signifying strong periodicity) is considerably smaller than for the original datasets. In contrast, the AR(1)-based background model yielded a larger number of high-scoring expression vectors. Remarkably, it leads to a similar distribution for the high-scoring range as the original CDC28 dataset.

What is the underlying cause for such differences between the background models? As Figure 1 shows, AR(1) processes can lead to a higher spectral density, and thus larger Fourier scores, for the observed cell-cycle frequencies compared to random uncorrelated processes. However, this increase depends strongly on the value of the autocorrelation coefficient α. Calculations of the power spectrum for different α show that only positive autocorrelation below a certain threshold can cause larger higher spectral densities. More specifically, higher spectral densities are achieved by values for α either between 0 and 0.75 for the CDC28 experiment or between 0 and 0.85 for CDC15, respectively (Fig. S4A). Remarkably, this is also the range where we observe a prominent enrichment of autocorrelation coefficients for the datasets (Fig. S4B). Therefore, we can conclude that the autocorrelation in the analyzed datasets can lead to spurious periodicities.

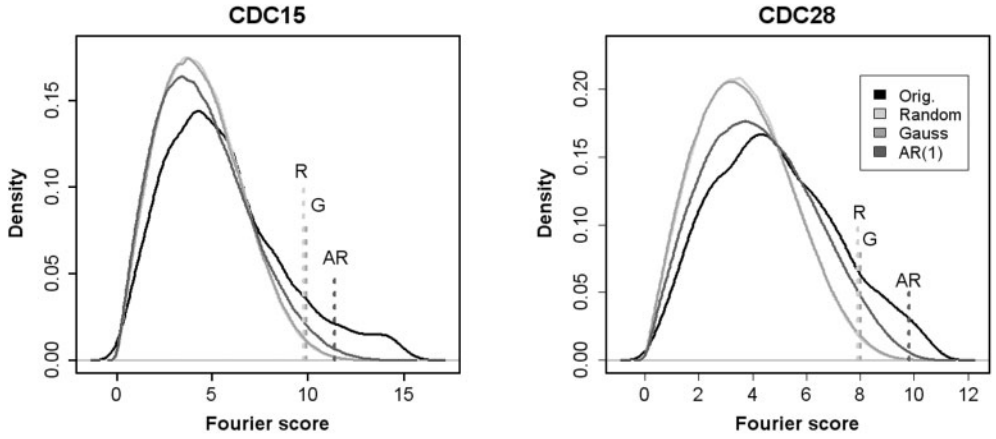

Fig. 4. The distribution of Fourier scores for the original datasets and the different background datasets are shown. Dashed lines indicate the threshold for FDR = 0.1 as determined in section 2.4.

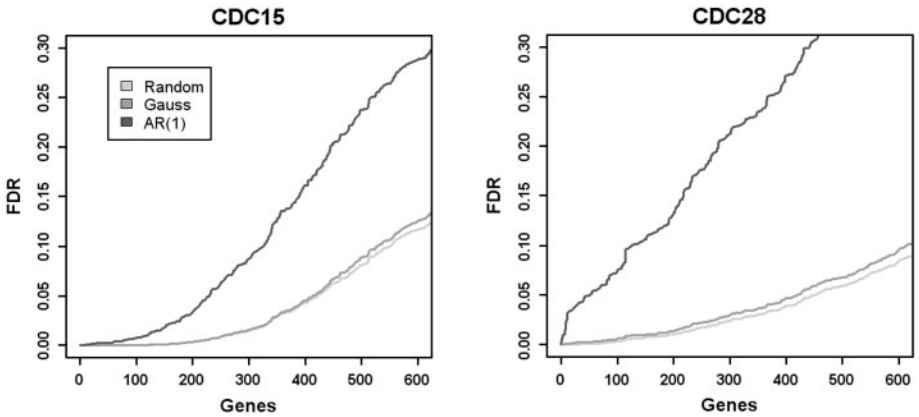

Fig. 5. FDR for periodic expression. The dependency between number of significantly periodically expressed genes and the significance level is shown for different background models. Lowering the threshold for the Fourier score leads to an increase of the number of significant genes but also to larger FDRs.

To evaluate quantitatively the obtained Fourier scores, we assessed the significance based on the empirical FDR. By shifting a threshold for the Fourier score and using Equation (4), the number of significant genes for different FDR can be obtained. The dependency is visualized in Figure 5. The influence of the choice of background model is striking: Whereas the randomized and Gaussian backgrounds result in very similar numbers of significant genes, using the AR(1) background leads to a considerable reduction of the number of significant genes independent of the chosen FDR. Note that this is especially the case for the CDC28 dataset. The exact numbers of significant genes can be found in Table 1. For a less stringent FDR of 0.1, we obtain for both datasets about 500–600 significant genes in the case of randomized or Gaussian background distribution. For FDR = 0.01, 150–250 genes remained significant. Choosing the AR(1) background, we obtain considerably lower numbers. For the CDC15 dataset, the number of significant genes was reduced by up to 50%. Even more drastic was the reduction for the CDC28 dataset. For a FDR = 0.01, only three genes were identified as significantly periodically expressed. Choosing FDR = 0.1 leads to 126 significant genes. The difference between the two datasets might arise from the fact that the CDC15 spans three cell cycles and thus periodic expression may be easier to detect in contrast to the CDC28 with only two cell cycles monitored. Notably, adjusting the threshold also increases the overlap of significant genes between the two datasets based on the AR(1) background model.

Besides the strong influence of the choice of background model, we also noted the importance of data preprocessing for the significance of periodicity. Scaling to the same distribution generally results in an increase of periodically expressed genes detected. For CDC15, an increase of up to 20% was observed, whereas for CDC28 this effect strongly depended on the significance level and the chosen background model.

3.3 Assessment of detected significance

Our comparison indicated so far that the AR(1)-based background most adequately represents the data structure in the

Table 1. Number of genes detected as significantly periodically expressed

Background model	CDC15		CDC28		Overlap	FDR
	Standardized	Standardized & scaled	Standardized	Standardized & scaled		
Randomized	258	302	192	201	0.34	0.01
Gauss	257	307	152	215	0.33	
AR(1)	119	129	3	14	0.21	
Randomized	420	497	448	454	0.36	0.05
Gauss	413	488	419	445	0.36	
AR(1)	257	280	52	106	0.39	
Randomized	551	672	649	685	0.35	0.10
Gauss	527	649	614	671	0.35	
AR(1)	326	383	126	200	0.40	

'Standardized' refers to standardization of gene expression values (mean = 0, SD = 1). 'Scaled' refers to the scaling of the dataset to the same distribution. The significance is shown as empirical FDR as described in Methods and Materials. The overlap is defined here as the percentage of significant genes found in CDC28 that are also determined as significant for CDC15.

yeast cell-cycle experiments. But do we improve the quality of the detection of periodicity? To asses this issue, we compared the sets of significant genes found using different background models with three previously compiled benchmark datasets of cell-cycle genes (de Lichtenberg *et al*., 2005): (i) The first benchmark set comprises a total of 113 genes identified as periodically expressed in small scale experiments; (ii) the second set consists of 352 genes which underlie the control of known cell-cycle transcription factors and (iii) the third set comprises 518 genes annotated in the Munich Information Center for Protein Sequences (MIPS) database as 'cell cycle and DNA processing' after the exclusion of genes included in the two other benchmark sets. The quality of identification of periodically expressed genes was assessed using the positive predictive value (PPV), since this measure tends to be more informative when the prior probability of finding a positive is low (Jansen and Gerstein, 2004). It can be defined as $PPV = TP/(TP + FP)$, where TP is the number of true positives and FP is the number of false positives. The PPVs were calculated for several FDRs and shown in Table 2.

For the first benchmark set, a clear improvement was achieved for both the CDC15 and CDC28 datasets when using the AR(1)-based background model. For the second set, the PPV increased strongly for the CDC28 dataset and only slightly for the CDC15 dataset. The comparison is less conclusive for the MIPS benchmark set. It should be noted, however, that the MIPS dataset is expected to include a lower proportion of periodically expressed genes, since cell-cycle genes of the other benchmark sets were excluded from the MIPS dataset (de Lichtenberg *et al*. 2005). In summary, the use of AR(1)-background models improved the PPV in most cases. It also indicates that we might overestimate the number of periodically expressed genes using randomized or Gaussian background models.

Table 2. Positive predictive value (PPV) derived for the use of different background models for significance testing

Benchmark set	CDC15		CDC28		FDR
	Randomized	AR(1)	Randomized	AR(1)	
Small scale exp.	0.21	0.31	0.19	0.66	0.01
Chromatin IP	0.17	0.18	0.18	0.33	
MIPS	0.10	0.06	0.20	–	
Small scale exp.	0.15	0.23	0.11	0.32	0.05
Chromatin IP	0.16	0.17	0.12	0.19	
MIPS	0.11	0.10	0.18	0.19	
Small scale exp.	0.13	0.18	0.09	0.21	0.10
Chromatin IP	0.14	0.18	0.10	0.19	
MIPS	0.10	0.11	0.16	0.25	

PPVs based on Gaussian backgrounds were similar to the ones based on randomized background (data not shown). Details regarding the benchmark sets are given in the text. No true positives were detected using the AR(1) background model for the MIPS benchmark set at a FDR of 0.01.

4 DISCUSSION AND CONCLUSIONS

In this study, we examined the impact of the choice of background model on the detection of periodically expressed genes in microarray data. These background models manifest what we would expect the data to 'look like' if no true periodic processes underlie the observed expression patterns. Frequently, randomized or Gaussian background models are used, assuming that non-periodic genes display no autocorrelation. However, whether such an assumption holds has not been examined so far in the literature. Thus, we scrutinized different background models and their implications using two yeast cell-cycle microarray datasets as case studies. We assessed the data structure of the cell-cycle experiments by means of autocorrelation which is an important tool to describe the evolution of a process through time. Our analysis shows that randomized and Gaussian background models neglect the dependency structure within the observed data. In contrast, the use of AR(1)-based background models gave a more accurate representation of correlations between measurements.

The observed residual (non-cell-cycle dependent) correlation could have several sources. One likely source is the applied synchronization method which is based on shifting the cell cultures from a permissive to a non-permissive temperature range. The induced stress responses can evoke the up-regulation of a variety of genes such as heat shock proteins. These initial conditions might then lead to slowly decaying perturbations—which are manifested as autocorrelative patterns in the measured gene expression.

We also demonstrated that the choice of background model has drastic effects on the number of genes detected as significantly periodically expressed. Randomized and Gaussian background models led to around 600–700 genes being determined to be significantly periodically expressed (FDR = 0.1). Using an AR(1)-based background, however, we detected around 400 genes for the CDC15 and around 200 for CDC28 as significant. A subsequent assessment using benchmark datasets indicated that the use of randomized or Gaussian background models can lead to overestimating the number of periodic genes.

Although the choice of the background model has generally been given less consideration than the selection of the detection methods, our results demonstrate that it is of major importance. Randomized and Gaussian background models may overestimate number of significant periodically expressed genes. In contrast, the use of the more accurate AR(1)-background led to a considerable reduction of the number. That does not mean that only a small number of genes is periodically expressed but rather it reflects the inherent noise in microarray data and may give a more realistic picture of current capacities for the detection of cell-cycling genes.

Finally, we would like to note that this presented framework is not restricted to the study of the cell cycle, but should apply generally to the detection of periodic signals in high-throughput data. A further prominent example is the detection of circadian expression using microarray technology (Bozek *et al.*, 2007; Storch *et al.*, 2002). As—similar to the presented cell-cycle studies—a strikingly poor overlap between different microarray experiments was observed (Bozek *et al.*, 2007), we are currently investigating the impact of background models on the detection of genes controlled by the circadian clock.

ACKNOWLEDGEMENTS

The work presented was supported by the *Deutsche Forschungsgemeinschaft (DFG)* by the SFB 618 grant. We would like to thank Bronwyn Carlisle (University of Otago, NZ) for careful and critical reading of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bozek,K. *et al.* (2007) Promoter analysis of mammelian clock genes. *Genome Inform.*, **18**, 65–74.

Chatfield,C. (1995) *The Analysis of Time Series*. Chapman & Hall, London.

Cho,R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Cho,R.J. *et al.* (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.

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

Futschik,M. and Crompton,T. (2004) Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol.*, **5**, R60.

Futschik,M.E. and Carlisle,B. (2005) Noise-robust soft clustering of gene expression time-course data. *J. Bioinform. Comput. Biol.*, **3**, 965–988.

Jansen,R. and Gerstein,M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.*, **7**, 535–545.

Johansson,D. *et al.* (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.

Lu,X. *et al.* (2004) Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Res.*, **32**, 447–455.

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

Shedden,K. and Cooper,S. (2002b) 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.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Storch,K.F. *et al.* (2002) Extensive and divergent circadian gene expression in liver and heart. *Nature*, **417**, 78–83.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Wichert,S. *et al.* (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5–20.

Zhao,L.P. *et al.* (2001) Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl Acad. Sci. USA*, **98**, 5631–5636.