# Are we overestimating the number of cell-cycling genes? The impact of background models for time series data

Matthias E. Futschik and Hanspeter Herzel

Institute for Theoretical Biology, Charité, Humboldt-Universität,
Invalidenstrasse 43
10115 Berlin, Germany
m.futschik@staff.hu-berlin.de

**Abstract:** 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 a large number of genes has been detected as periodically expressed. However, the small overlap of identified genes between different studies has shaded considerable doubts about the reliability of the detected periodic expression. In this study, we show that a major reason for the lacking agreement is the use of an inadequate background model 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.

## 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 defined.

Microarray technologies 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 indicated that periodic expression may not be restricted to a small number of genes, but that a substantial part of transcriptome underlies periodic activation during the cell cycle [Ch98,Sp98]. 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 [FC04]. 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 [AM02]. 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 [Ch01]. The authors detected several known and many apparently novel cell-cycle regulated genes. However, Shedden and Cooper could convincingly demonstrate in a follow-up analysis that most of these detected genes do not show a reproducible periodic pattern [SC02b]

Thus, stringent statistical methods are essential to assure the reliability of detected periodic expression. Several approaches for detection have been proposed based on time series analysis and statistical modeling [Jo03,Sp98,Wi04,Zh01]. (For a recent comparison of their performance, please refer to the study by de Lichtenberg and collegues [Li05].) 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 detected usually 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 [Li05,SC02a,Zh01]. As our study indicates one major reason for the observed lack in agreement is likely the overestimation of the number of periodically expressed genes in these datasets due to the use of inadequate background models.

## 2 Materials and Methods

### 2.1 Expression studies of the yeast cell cycle

As a case study we re-analyze two yeast cell cycle microarray experiments. The first study included the expression of over 6000 genes derived by employing Affymetrix chips [Ch98]. Synchronization was achieved using temperature sensitive yeast cells (CDC28). At the non-permissive temperatures of 37°C, cells are arrested in the late $G_1$ phase. Shifting the temperature back to the permissive range of 25°C, the cells enter the cell cycle again. Samples of cells were taken every 10 minutes for 160 minutes. 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 with less than 75% of the measurements. 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 [Tr01]. Data were standardized to have mean values equal to zero and standard deviation equal to one for subsequent time series analysis. Optionally, additional scaling

by quantile normalization was performed [Bo03]. The later step proved to have considerable impact on the detection of periodically expressed genes (see table 1). The density distributions for the datasets before and after scaling can be found in the Supplementary Materials.

As second dataset, we use the microarray experiments of the yeast cell cycle by Spellman and colleagues [Sp98]. 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 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*. The data was downloaded from the authors' webpage [Sp98].

**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, a number of different approaches have been put forward ranging from simple visual expection [Ch98] to elaborated statistical models [Lu04]. An extensive comparison surprisingly showed that a relatively simple permutation-based method using Fourier analysis performs superior to other approaches. This method was also chosen for the detection of periodically expressed genes in this study. It is based on the Fourier score defined as

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

where **g** is the standardized expression vector (mean(**g**)=0; sd(**g**) =1) for the gene, T is the period and $g_i$ is the measured expression at time point $t_i$ . The score F is larger the closer a gene's expression follows a (possibly shifted) cosine curve. 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 [Sp98].

**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. This means that the expression data follow a normal distribution

$$f(g_i) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[ -\left( \frac{g_i}{2 \cdot \sigma} \right)^2 \right] \qquad [2]$$

where $g_i$ is the expression of the gene at time point *i*. In case 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 are 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 $Z$. For a discrete time-series, autoregressive processes of the order p have generally the form

$$X_t = \alpha_1 \cdot X_{t-1} + \alpha_2 \cdot X_{t-2} + \ldots + \alpha_p \cdot X_t + Z_t$$

where the parameters $\alpha_i$ determining how strongly $X_t$ depends on past values and $Z_t$ is an independent random with a mean value of zero and variance $\sigma_z^2$. Of special interest here are autoregressive processes of order (AR(1)):

$$X_t = \alpha_1 \cdot X_{t-1} + Z_t \qquad [3]$$

for which $\alpha_1$ 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. The variance of $Z_t$ is given by $\sigma_z^2 = \sigma_x^2 (1- \alpha_1^2)$. Having determined $\alpha_1$ and $\sigma_z$, 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 the 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 for the original one. Therefore, AR(1) processes allow us to construct a background distribution that capture 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.

An important (and in this context crucial) characteristic for 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:

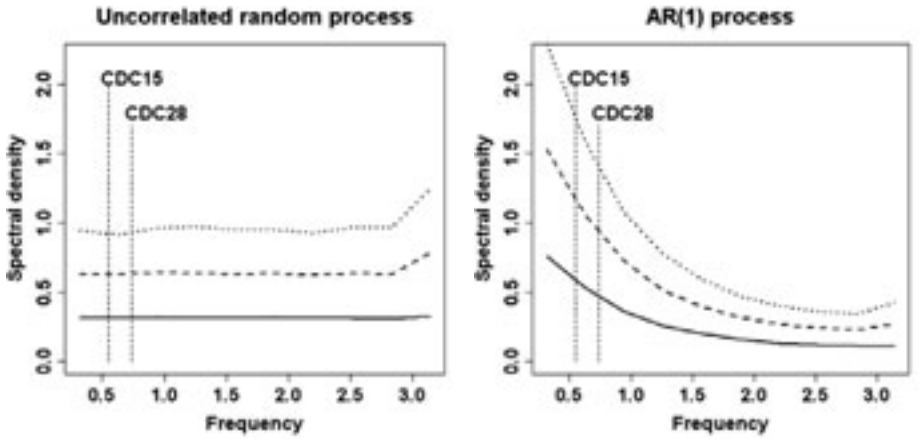


**Figure 1: Spectral density distributions for uncorrelated random and AR(1) process. The distributions were calculated based on 10000 independent simulations of a time series with length 20. The frequency was scaled so that the maximum detectable frequency (i.e. Nyquist frequency) is equal to π. Solid lines represent the mean spectral density; dashed lines represent the mean plus the standard deviation; and dotted lines indicate the upper 90% level of the distributions. For the AR(1) process, an autocorrelation coefficient of 0.5 was chosen.**

$$I[f_p] = \left[ \left( \sum_i \cos(2 \cdot \pi \cdot f_p) \cdot g_i \right)^2 + \left( \sum_i \sin(2 \cdot \pi \cdot f_p) \cdot g_i \right)^2 \right] / N\pi \quad [4]$$

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 random 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 depend on the underlying autocorrelation coefficient with negative autocorrelation yielding to larger power at higher frequencies.

### 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 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 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 *t* for the Fourier score:

$$FDR(t) = \frac{\sum \delta_i (F_b \geq t) / n}{\sum \delta_j (F_o \geq t)} \quad [4]$$

where $F_o$ and $F_b$ are the Fourier scores derived for the original and background distribution respectively, n is the 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 could be obtained by comparison with the generated background distribution. For example, a FDR-value of 0.01 would indicate that a Fourier score larger than or equal to the measured one was observed in one out of hundred generated background time courses.

## 3. Results

To study the influence of background models on the detection of periodic patterns, we re-analyzed two microarray experiments of the yeast cell cycle. 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 generated derived from sampling from 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 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. 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 cluster showing periodic patterns, many other expression profiles occur (see Fig. 2). They frequently display prominent non-periodic trends which might have been evoked by the release of cell cultures after synchronization. Such processes are 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 such trends may arise by

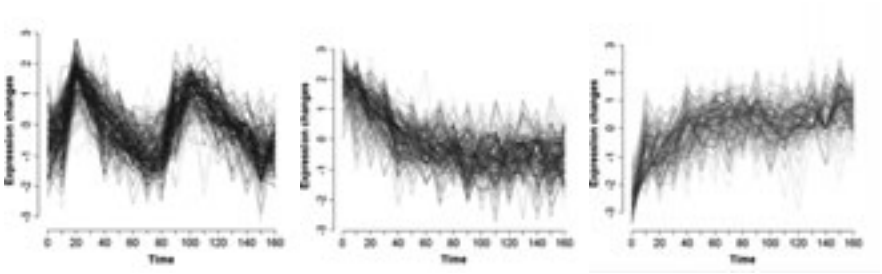

**Figure 2: Examples of periodic (left) and aperiodic (middle, right) expression patterns in the CDC28 dataset. The clusters were detected by a soft clustering approach which allows differentiation of cluster membership. Darker shades of gray correspond to larger membership. Details of the clustering can be found in Futschik and Charlisle [FC05].**

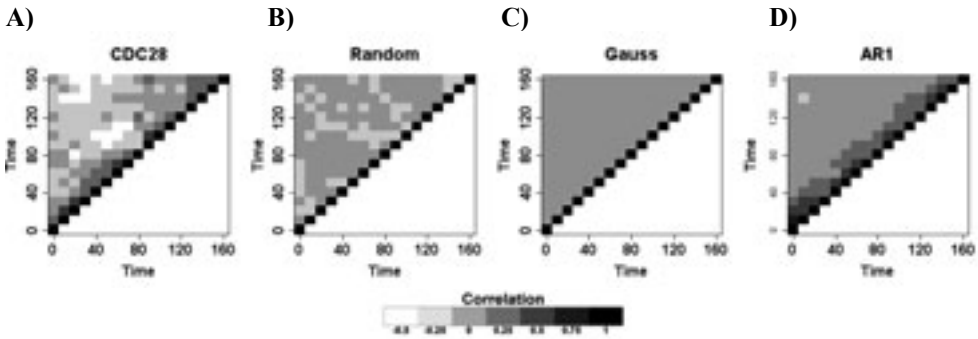


**Figure 3: Autocorrelation in original dataset and background distributions: The upper triangle of the correlation matrix is displayed with respect to the temporal ordering of arrays. The original dataset CDC28 was standardized and scaled.**

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 (see figure 3A). Directly successive measurements generally showed a clear correlation (e.g. Pearson correlation of 0.29 ± 0.17 for CDC28). 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 random and Gaussian background distributions (figure 3 B and C). For these generated datasets, the autorcorrelation generally was neglectable. For example, a Pearson correlation between directly successive arrays of 0.05 ± 0.01 and of 0.004 ± 0.02 was calculated for random and Gaussian background distributions respectively. For AR(1)-based background distributions, however, we obtained clear correlation patterns (figure 2D). Similar to the original data, directly successive measurements were significantly correlated (0.39 ± 0.07). Note that the detected 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

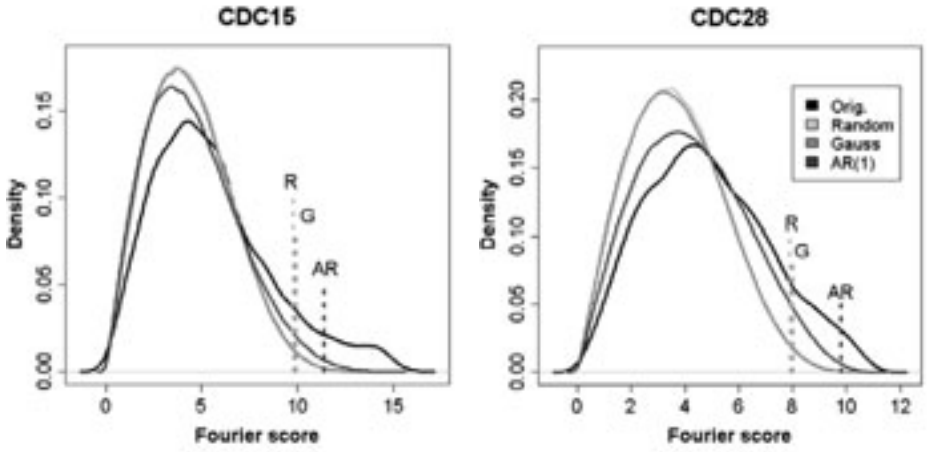


**Figure 4: The distribution of Fourier scores for the original datasets and the different background datasets are shown. Dashed lines indicate the thresholds for FDR=0.1 as determined in section 2.4.**

method. Nevertheless, the comparison shows that the AR(1)-based background reflects the important feature of autocorrelation as observed in the original data sets. Thus, it 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 and each gene in the two original data sets. 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 can be found in the Supplementary Materials.

Figure 4 compares the distribution of Fourier scores obtained for the original datasets and the corresponding background distributions. The following general patterns emerge: Random and Gaussian background led to very similar distributions of Fourier scores. Notably, the proportion of expression vectors with large scores (signifying strong periodicity) is considerable 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 achieves a similar distribution for the high scoring range as the original CDC28 dataset.

What is the underlying cause for such difference 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 processes. However, this behavior depends on the value of the autocorrelation coefficient. In fact, only positive autocorrelation up to a cycle period dependent value (CDC28: 0.75, CDC15: 0.85) yield higher spectral density. But this is also the range where we observe an enrichment of autocorrelation coefficients for the datasets (see supplementary figure

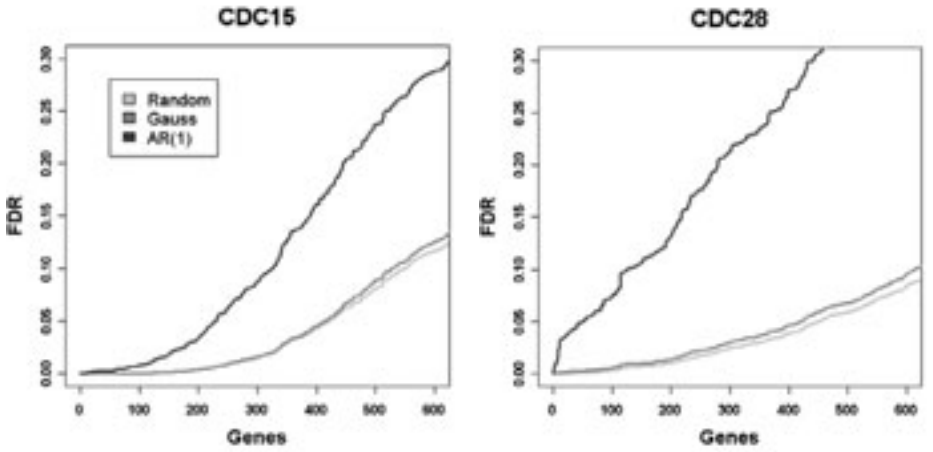


**Figure 5 FDR for periodic expression. The dependency between number of significant periodically expressed genes and the significance level is shown. Lowering the threshold for the Fourier score leads to an increase of the number of significant genes but also to larger FDRs.**

4). Therefore, we can conclude that the autocorrelation in the analyzed datasets can cause spurious periodicities.

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 applying 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 random and Gaussian background result in very similar number 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 random or Gaussian background distribution. For FDR=0.01, this is reduced to 150-250 genes. Choosing the AR(1) background, we obtain considerable lower numbers. For the CDC15 dataset, the number of significant genes was reduced up to about 50%. Even more drastic was the reduction for the CDC28 dataset. For a FDR=0.01, only 3 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.

Besides the strong influence of the choice of background model, we also noted the importance of data preprocessing for the significance testing. The scaling to the same distribution generally results in an increase of detected periodically expressed genes. 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.

| Background Model | CDC15 | | CDC28 | | FDR |
|---|---|---|---|---|---|
| | Standardized | Standardized & Scaled | Standardized | Standardized & Scaled | |
| **Random** | 258 | 302 | 192 | 201 | 0.01 |
| **Gauss** | 257 | 307 | 152 | 215 | |
| **AR(1)** | 119 | 129 | 3 | 14 | |
| **Random** | 420 | 497 | 448 | 454 | 0.05 |
| **Gauss** | 413 | 488 | 419 | 445 | |
| **AR(1)** | 257 | 280 | 52 | 106 | |
| **Random** | 551 | 672 | 649 | 685 | 0.10 |
| **Gauss** | 527 | 649 | 614 | 671 | |
| **AR(1)** | 326 | 383 | 126 | 200 | |

**Table 1: Number of genes detected as significantly periodically expressed. 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.**

### 3.3 Assessment of detected significance

Our comparison indicated so far that the AR(1)-based background represents more adequately the data structure in the 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 [Li05]: 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 MIPS as "cell cycle and DNA processing" after the exclusion of genes included in the two other benchmark sets. The quality of identification ofperiodically expressed genes can be assessed by variety of measures such as specificity (Sp=TN/(TN+FP)) and sensitivity (Se=TP/(TP+ FN)) where TP are the true positives, FP are the false positives and FN are the false negatives detected. Here, we are using the positive predictive value (PPV), since this measure tends to be more informative when the prior probability of finding a positive is low [JG04]. Expressed as functions of specificity and sensitivity, the PPV takes the form:

$$\mathrm{PPV} = \frac{P \cdot Se}{P \cdot Se + N \cdot (1 - Sp)} \left[ = \frac{TP}{TP + FP} \right]$$

where $P$ and $N$ are the numbers of real positives and negatives respectively.

The PPVs were calculated for genes significant at several FDRs and shown in table 2.

| Benchmark Set | CDC15 | | CDC28 | | FDR |
|---|---|---|---|---|---|
| | Random | AR(1) | Random | AR(1) | |
| **Small scale experiments** | 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 experiments** | 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 experiments** | 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 | |

**Table 2: Positive predictive value (PPV) derived for the use of different background models for significance testing. PPVs based on Gaussian backgrounds were similar to the ones based on random 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.**

For the first benchmark sets, a clear improvement was achieved for both the CDC15 and CDC28 datasets when using AR(1)-based background. 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 as genes of the other benchmark sets are excluded [Li05]. 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 random or Gaussian background models.

## 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 how the data 'looks like' if no true periodic processes underlie the observed expression patterns. Frequently, random or Gaussian background models are used assuming that non-periodic genes display no autocorrelation. However, if 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. Note that the study of the yeast cell cycle is not only of purely academic interest since it has become clear that the underlying fundamental processes of DNA replication and cell division show a high similarity among eukaryotes. The analysis of gene regulation in yeasts can offer valuable insights into the genetic origin of human diseases [St02].

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 random 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 adequate representation of correlations between measurements.

We also demonstrated that the choice of background model has drastic effect on the number of genes detected as significantly periodically expressed. Random and Gaussian background models led to around 600-700 genes being 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 random 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 compared to the selection of the detection methods, our results demonstrate that it is of major importance. Random and Gaussian background models may give overoptimistic number of significant periodically expressed genes. In contrast, the use of the more adequate 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 detection of cell-cycling genes.

### Supplementary Materials

Supplementary materials can be found in *https://itb.biologie.hu-berlin.de/Members/futschik/GCB2007/.*


### Acknowledgements

The work presented was supported by the *Deutsche Forschungsgemeinschaft (DFG)* by the SFB 618 grant.


## Literature


[AM02] Ambroise, C. ; McLachlan G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 99, 2002; S. 6562-6566

[Bo03] Bolstad, B.M. et al.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 2003; S. 185-193.

[Ch01] Cho, R.J. et al.: Transcriptional regulation and function during the human cell cycle. Nat Genet 27, 2001; S. 48-54.

[Ch98] Cho, R.J. et al.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2, 1998; S. 65-73.

[FC04] Futschik, M.; Crompton T.: Model selection and efficiency testing for normalization of cDNA microarray data. Genome Biol 5, 2004; S. R60.

[FC05] Futschik, M.E.; Carlisle B.:Noise-robust soft clustering of gene expression time-course data. J Bioinform Comput Biol 3, 2005; S. 965-988.

[JG04] Jansen, R.; M. Gerstein: Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr Opin Microbiol 7, 2004; S. 535-545.

[Jo03] Johansson, D. et al.: A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics 19, 2003; S. 467-473.

[Li05] de Lichtenberg, U. et al.: Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics 21, 2005; S. 1164-1171.

[Lu04] Lu, X. et al.: Statistical resynchronization and Bayesian detection of periodically expressed genes. Nucleic Acids Res 32, 2004; S. 447-455.

[SC02a] Shedden, K.; Cooper S.: Analysis of cell-cycle gene expression in Saccharomyces cerevisiae using microarrays and multiple synchronization methods. Nucleic Acids Res 30, 2002; S. 2920-2929.

[SC02b] Shedden, K.; Cooper S.: Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. Proc Natl Acad Sci U S A 99, 2002; S. 4379-4384.

[Sp98] Spellman, P.T. et al.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9, 1998; S. 3273-3297.

[St02] Steinmetz, L.M. et al.:Systematic screen for human disease genes in yeast. Nat Genet 31; 2002; S. 400-404.

[Tr01] Troyanskaya, O. et al.: Missing value estimation methods for DNA microarrays. Bioinformatics 17, 2001; S. 520-525.

[Wi04] Wichert, S. et al.: Identifying periodically expressed transcripts in microarray time series data. Bioinformatics 20, 2004; S. 5-20.

[Zh01] Zhao, L.P. et al.: Statistical modeling of large microarray data sets to identify stimulus-response profiles. Proc Natl Acad Sci USA 98, 2001; S. 5631-5636.