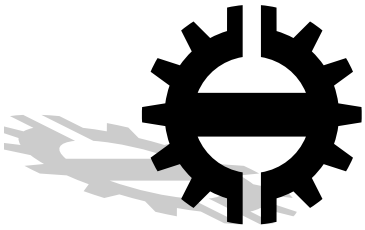
TAMPEREEN TEKNILLINEN KORKEAKOULU
Tietotekniikan osasto
Signaalinkäsittelyn laitos

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology
Institute of Signal Processing



Laitosraportti
Technical Report
2-2002

Harri Lähdesmäki, Tommi Aho, Heikki Huttunen,
Marja-Leena Linne, Jari Niemi, Juha Kesseli,
Ron Pearson, Olli Yli-Harja

Using Signal Processing Tools to Improve the Quality of Microarray Time-Series Measurements

Tampere, Finland 2002

ISSN 1455-142X
ISBN 952-15-0810-8

Laitosraportti 2-2002
Technical Report 2-2002

Using Signal Processing Tools to Improve the Quality of Microarray Time-Series Measurements

Harri Lähdesmäki, Tommi Aho, Heikki Huttunen,
Marja-Leena Linne, Jari Niemi, Juha Kesseli,
Ron Pearson, Olli Yli-Harja

Tampere University of Technology
Institute of Signal Processing
P.O. Box 553
FIN-33101 TAMPERE
FINLAND

ISSN 1455-142X
ISBN 952-15-0810-8

Tampere University of Technology
Institute of Signal Processing
P.O. Box 553
FIN-33101 TAMPERE
FINLAND

Tampere 2002

Using Signal Processing Tools to Improve the Quality of Microarray Time-Series Measurements

Harri Lähdesmäki [a] Tommi Aho [a] Heikki Huttunen [a]
Marja-Leena Linne [a] Jari Niemi [a] Juha Kesseli [a] Ron Pearson [a]
Olli Yli-Harja [a]

[a] *Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, FIN-33101 Tampere, Finland*

Abstract

We introduce several approaches to improve the quality of microarray data obtained from time-series measurements by applying signal processing tools. Performance of the proposed methods are examined using both simulated and real yeast gene expression data. In detail, we concentrate especially on a smoothing effect caused by the phase distribution of the cell population and introduce several methods of inverting this phenomenon.

The proposed methods rely on the partition of the genes, as well as the corresponding expression profiles, into the cell cycle regulated and non-cell cycle regulated genes. For that purpose, we first study the cell cycle regulated genes and introduce a method that can be used to estimate the period length of those genes. We also estimate the spreading of the underlying distribution of the cell population based solely on the observed gene expression data. After the preliminary experiments we introduce some methods for estimating the underlying true distribution of the cell population instead of the spreading rate of the distribution. These methods assume certain additional measurements, such as flow cytometry (e.g. fluorescent-activated cell sorter (FACS)) or bud counting measurements, to be available. We also apply the standard blind deconvolution method for estimating the true distribution of the cell population.

The found estimates of the spread of the cell distribution and the distributions of the cell population themself are used to invert the smoothing effect. To that end, we discuss some inversion approaches applicable to the problem in hand.

Key words: microarray, time-series, cell cycle regulated genes, distribution of cell population, density estimation, inverse filtering, regression problem

1 INTRODUCTION

Modelling of genetic networks based on molecular biological data is becoming a very important issue in trying to understand the functional aspects of the genomics. We discuss the use of signal processing tools to improve the quality of the *gene expression* data produced by *DNA microarray technique*. Microarray technique is presently the most advanced technique for monitoring gene expression for thousands of genes in parallel. It is gaining more and more interest in molecular biology and biomedical sciences after its first introduction in 1995 (see e.g. [12] and [13]). In this paper, a special emphasis is paid on considering the problems related to the microarray *time-series measurements* and suitability of data for modeling purposes.

The biological term gene expression describes the phenomena at which the necessary parts of the DNA (genetic material) located in the chromosomes of the cell nucleus are transcribed (copied) into messenger RNA to be used in the translation (production) of proteins. Gene expression is a fundamental cellular process for all living creatures, because the production of new proteins is not only necessary for the development but also for the whole life span of an organism. *Functional genomics*, in turn, is a discipline which is expanding the molecular biological investigation from studies of single genes or proteins to studies of functions and interactions of all genes or proteins in a high-throughput, systematic manner (such as the microarray technique). It will turn the biological genomics studies towards *system biology* studies which is significantly different from investigating one or a few genes at a time.

Microarray measurements make it possible to study the expression levels of thousands of genes simultaneously. Presently, however, most of the microarray studies are used for profiling or classifying genes, e.g. to identify genes involved / expressed in specific types of cancer. In such studies comparisons are done between two different types of tissue (e.g. normal and unhealthy), without paying attention on time dimension of examined phenomenon. However, when trying to model genetic networks we are forced to follow the gene expression profiles of specific genes for a longer time and sample data with short enough time intervals. For that purpose the time-series microarray measurements provide a way of obtaining necessary data.

In time-series microarray measurements, the cell population (such as the yeast cell population) is set in *synchrony* by using specific chemical or other treatments. In a biological context, the term synchrony is used for describing the phenomenon in which all cells in the chosen cell population are exactly in the same phase of the *cell cycle*. Furthermore, a cell cycle can be defined as a series of events which make it possible for a cell to grow and

divide into two daughter cells containing the same genetic material. The cell division cycle consists of four phases [1]: G1, S, G2 and M. The S phase involves DNA replication, i.e. the double-stranded DNA string is doubled to form two identical DNA strings called sister chromatids. The M phase involves the mitosis, i.e. the DNA strings are segregated into the dividing "mother" and "daughter" cells. The G1 and G2 are "temporal gap phases" between M and S, and S and M, respectively. The cell cycle involves several checkpoints at which the decision about continuing towards the execution of the next phase is done. These checkpoints are regulated by complex *biochemical machinery*.

At this moment, microarray experiments are not done using a single cell, because it does not contain enough extractable mRNA to measure. Instead, the procedure requires a sample that typically contains millions of cells. In principle, every cell has an effect on the gene expression measurement result. This is the reason why the whole cell population should be in exactly the same phase of the cell cycle. For biological reasons, however, the rate of cell growth is not constant. Some cells are passing the cell cycle phases faster while some others are passing them slower. This is mostly dependent on the developmental status of the cell (e.g. developing vs. mature cell). Therefore, the cells gradually distribute themselves into different cell cycle phases and the cell population looses its synchrony with time. Because of this widening of cell cycle phase distribution, the measured gene expressions become mean values of neighboring cell cycle phases. With the help of signal processing tools we aim at improving the quality of the microarray time-series measurements by taking into account the effect caused by the biology, the widening of the cell cycle phase distribution during the microarray time-series measurements.

The rest of the paper is organized as follows. The general problem statement is given in Section 2. The results of measuring the period length of the cell cycle regulated genes and estimating the spreading rate of the underlying distribution of the cell population are also shown in Section 2. The proposed methods for estimating the underlying distribution of the cell population are introduced in Section 3. Some approaches for inverting the smoothing effects are given in Section 4. Finally, some conclusions are given and directions of the future research are discussed in Section 5.

2 PROBLEM STATEMENT

In this section we define the biological problem in mathematical terms. Let us assume that gene expression time-series data consists of m parallel measurements, i.e., the measurements have been taken at m discrete time points

t_i, $i = 1, \ldots, m$, and each of them have measurements for n genes. Let $x_j(t)$, $t \in [t_1, t_m]$, denote the (continuous) underlying true expression profile of the jth gene. We omit index j in most cases since we focus only on a single gene at a time, and we also write $x(i) = x(t_i)$, $i = 1, \ldots, m$ for brevity. Distribution of the cell population at the ith measurement time is denoted by $p_i(t)$, where t denotes "place", or using the word from the biological context, "phase", in the cell cycle. We also assume each $p_i(t)$ to be a discrete distribution since measurements are known to be taken from a finite cell population. Now, we assume that each cell has an equal contribution to the resulting measurement. This assumption forces the measurement to be an average over the distribution of the cell population and therefore, we can express the measurements as

$$y(i) = \sum_k p_i(t_k) x(t_i + t_k) + v(i), \tag{1}$$

for all $i = 1, \ldots, m$, k runs over all cells in the cell population (distribution) and v is an additive noise term. The same equation applies to all the genes although the index j is omitted here. Essentially, Equation (1) defines a discrete inner product between the cell distribution and the true expression profile. We could also use an integral for modelling purposes since each measurement is known to be taken from a huge number of cells. In the strict sense that would not be quite correct since the sample population used in the experiments is finite. (See also Section 5 for further discussion of the measurement process.)

If the cells were in perfect synchrony, $p_i(t)$ would be the impulse function and observed measurements would be $y(i) = x(i) = x(t_i)$ for all $i = 1, \ldots, m$. However, that is not the case in reality since each cell has its own pace of living (like each watch has its own speed of operation) resulting in a spread of the cell population along the time. So, whenever $p_i(t)$ is not the impulse function, measurements are "smoothed" by the distribution of the cell population $p_i(t)$ as shown in Equation (1). This smoothing phenomenon catches most attention in this paper.

Because we have only coarse-scale discrete measurements available, we find it convenient to form a discrete approximation of the sum shown in Equation (1). Another possible approach could be based on the approximation of continuous functions and the use of numerical integration methods. However, this would lead to the interpolation methods where all points between the measured ones are "artificial". Further operations with such data may result to unpredictable results. The accuracy of the approximation depends on the number of terms in the sum. Assume for now that we know the distributions of the cell population $p_i(t)$ at different time instants $i = 1, \ldots, m$ and let $h_i(j)$ denote their discrete approximations. Then, Equa-

tion (1) is approximated as

$$y(i) \approx \sum_j h_i(j)x(i+j) + v(i), \tag{2}$$

where the sum is computed over those j that satisfy $h_i(j) \neq 0$.

A simple way of computing the inner product coefficients h_i is as follows. The jth element of h_i is found simply by summing $p_i(t)$ over the interval $[t_j - (t_j - t_{j-1})/2, t_j + (t_{j+1} - t_j)/2]$, i.e.,

$$h_i(j) = \sum_{t \in \left[t_j - \frac{(t_j - t_{j-1})}{2}, t_j + \frac{(t_{j+1} - t_j)}{2}\right] : p_i(t) > 0} p_i(t). \tag{3}$$

Basically, one could use any discretization method for the same purpose. The above procedure defines, however, a reasonable way of computing h_i since Equation (3) guarantees that $\sum_j h_i(j) = 1$ for all i, assuming that each p_i is really a distribution.

Equations from (1) to (3) form the basis for further analysis in this paper. The rest of this section is devoted to estimating the period length of the cell cycle and the spreading of the distribution of the cell population.

2.1 Cell Cycle Regulated Genes

One of the most important definitions in this context is the concept of cell cycle regulation (CCR). This property of a single gene has been extensively studied e.g. in [15] and in the references therein. A gene is said to be cell cycle regulated if its expression profile is "regulated" by the cell cycle. From the practical point of view, cell cycle regulated genes are usually found by inspecting their expression profiles and searching for some regular behavior, e.g., peaks at a certain point (phase) of the cell cycle (for instance, in S or M phase [15]). In mathematical terms, cell cycle regulation means that the expression profile of a particular gene is periodic (cyclic) and the period length is equal to length of the cell cycle. So, the CCR property can be stated as follows: the jth gene is CCR if and only if $x_j(t) = x_j(t + L)$ for all t, and the period length, L, is equal to the length of the cell cycle. We also assume that L remains the same between consecutive cell cycles. In our discrete time approximation setting the same definition can be written as (ignoring the index j) $x(i) = x(i + L)$, for all $i = 1, \ldots, m - L$. Note that this discrete version makes two implicit assumptions on the measurements, i.e., they are sampled regularly, and the measurements from different cell cycles are taken precisely from the same places (phases) of the cell cycle. Fortunately, we are able to control both of these. In practise, we may need to consider

the previous definition with $\approx$ -sign instead of pure equality since most of the biological processes are presently assumed to be stochastic in nature [8].

Spellman *et al.* [15] found approximately 800 out of about 6000 genes in the yeast to be cell cycle regulated. For simplicity of notation, we let $I_{all} = \{1, \ldots, n\}$ denote the set of indices for all genes and $I_{all} = I_{CCR} \cup I_{non-CCR}$ its disjoint partition into cell cycle regulated and non-cell cycle regulated genes, according to the work by Spellman *et al.* [15]. We utilize this partition of the genes in the analysis of the gene expression time-series.

Since the following sections rely heavily on the existence and validity of cell cycle regulated genes we give an extra attention for this issue. That is, we also estimate the length of the cell cycle directly from the observed data in this section.

Spellman *et al.* use the discrete Fourier transform (DFT) together with two other statistics for finding the genes that are cell cycle -regulated [15]. The difference is that they only want to see if a gene is cyclic or not. The DFT can be used to find the length of the cell cycle, as well. Indeed, a well known fact is that the DFT is the minimum variance unbiased estimator for the frequency content of a discrete signal (under some assumptions, see e.g. [7]).

In order to get more accurate estimates of the cell cycle length, the measurements for each gene are first appended with zeros, resulting in a signal whose length is a power of two. In the following experiments we have used the fast Fourier transform (FFT) of length 512.

After finding the FFT for each time series, the largest (absolute value of the) FFT coefficient is used as an estimate of the cycle length for each gene separately. More precisely, the largest coefficient among the $N/2$ first was found and its index tells the most significant frequency component (denote this by $k \in \{0, 1, 2, \ldots, 255\}$). The cycle length is related to the index k by

$$\text{cycle length} = \frac{512}{k}.$$

Figure 1 shows the cycle lengths in the CDC15 experiment, where the measurements are (with a few exceptions) 20 minutes apart.

A better view of the cycle lengths are found from the histogram of the cycle lengths. Spellman et al. [15] refer to four sets of measurements, which we have also been using. Figure 2 shows the cycle length histograms for all of them. The right most peak in each (sub)figure contains the cumulative sum of all the longer cell cycle lengths.

The first three experiments give promising results; the estimated cycle length

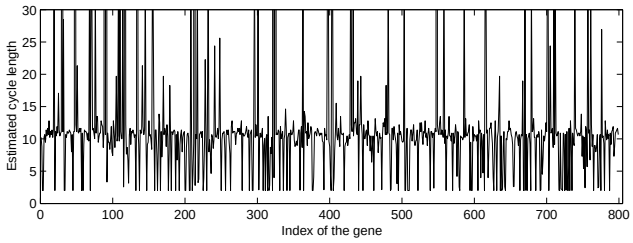
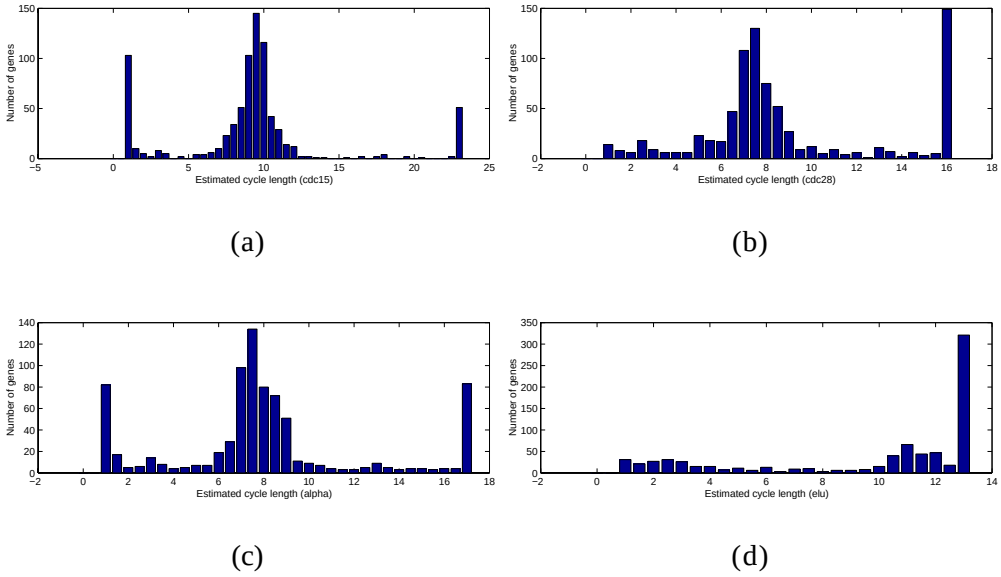


Figure 1. The estimated cell cycle lengths for 801 genes.



(a) (b)

(c) (d)

Figure 2. The histograms of the estimated cell cycle lengths for (a) experiment “CDC15”, (b) experiment “CDC28”, (c) experiment “Alpha” and (d) experiment “Elutriation”.

seem to be rather easy to find from the mean of the Gaussian-like distribution. The peaks at very low and high cycle lengths indicate that for these genes the cell cycle cannot be found by the Fourier transform. The tests were run for those 801 genes Spellman found to have a cyclic behavior. We also tested the same algorithm on all the genes (roughly 6000), but found out that the results were a lot worse. It was impossible to tell from the resulting images where the peak actually was.

2.2 Spreading of the Distribution of the Cell Population

At this point we assume to have no additional measurements, only the gene expression data. It may be fairly difficult, if not impossible, to estimate the underlying distributions of the cell population from this basis. However, there exists applicable tools if we are considering only the spread of the distribution of the cells. This approach is also important since most of the

current gene expression measurements are not associated with any additional data.

An intuitive assumption of the behavior of the cell distribution could be the use of 0-mean normal distributions with increasing variance, although we do not qualify this as the working assumption. Normal distribution has several interesting properties, especially, the convolution of two Gaussians is another Gaussian, apart of normalization. That is, the distribution of the cells at some time instant t would be a convolution between a properly normalized Gaussian kernel and the cell distribution at some previous time instant $s < t$.

Another interpretation for the widening effect of the cell population could be as follows. Consider a single cell in a distribution of the cell population at a certain time instant. Assume further that the amount of disturbance for that single cell from its normal operation pace remains constant. In other words, each cell appearing in a certain position of the distribution of the cell population, let say in $p_i(t)$ for some fixed t, is exposed to a stationary random process (disturbance) that defines the position of the cell at the next time instant. This setting corresponds to the addition of two random variables. The probability distribution of such a sum is known to be a convolution of the distributions of the random variables to be added. So, it is interesting to study whether or not we are able to model the spread of the cell distribution by a convolution. In fact, this turns out to be the case to some extent (see simulation results below).

In detail, we seek for a convolution kernel h that maps the cell distribution from any time point (place) t at the cell cycle, to the corresponding time (place) $t+L$ at the next cycle, i.e., $p_{t+L} = h * p_t$, where $*$ stands for the convolution. It may be that nature does not obey such a naïve model. Especially, a critical question is whether or not we can use a constant (stationary) impulse response h. On the other hand, our experiments with real data seem to support stationarity assumption, and therefore, we use a constant h for simplicity. Due to the practical reasons and/or the discrete approximation stated in Equation (2), we can use only discrete versions of p_t. Therefore, the above convolution is replaced by

$$h_{i+L} = h * h_i \tag{4}$$

for all $i = 1, \ldots, m-L$, assuming stationary h. Under the above assumption we can write the sequence of the discretized distributions of the cell population as $h_1, \ldots, h_L, h * h_1, \ldots, h * h_L, h * h * h_1, \ldots, h * h * h_L, \ldots$ and so on. It is worth noting that we may need to define separate scaling factor for each time point.

Convolution kernel h must satisfy certain conditions in order to be able to

map a distribution h_i into another distribution h_{i+L}. If we assume the cell population to be in the perfect synchrony in the beginning, $h_1 = \delta(n)$, the mapped distribution h_{L+1} is directly the kernel h that spreads the distribution. So, the kernel h should sum up to unity, i.e., $\sum_k h(k) = 1$. Since distributions must be positive we have an extra requirement that $h(j) \geq 0$ for all j. The kernel h should also be symmetric and centered around the origin so that the same holds for the output distribution. The symmetry assumption, however, is not necessary in this application. Proposed estimation methods for h are presented next.

The approximated measurement equation was introduced in Equation (2). Essentially, that equation defines an inner product between the sequences h_i and x. In order to make the following notations more consistent, we formalize the approximated measurement equation as a convolution. This can be done simply by time-reversing the inner product kernel which is denoted as $\overleftarrow{h_i}$. Equation (2) can then be written as $y(i) = (\overleftarrow{h_i} * x)(i)$. The measurement convolution kernel for the cell cycle numbers larger than 1 (measurement time instants $i' > L$) can be expressed as $\overleftarrow{h_{i'}} = \overleftarrow{h * \ldots * h * h_{i''}} = \overleftarrow{h} * \ldots * \overleftarrow{h} * \overleftarrow{h_{i''}}$, with proper number of repetitions of $h*$ and $L \geq i'' \geq 1$. This can further be written as $h * \ldots * h * \overleftarrow{h_{i''}}$ if h is symmetric. We will use the last notation although we do not necessarily assume h to be symmetric. In the case of non-symmetric h we can simply time-reverse the estimate of h. So, since we model the spread of the cell population by a convolution, we can rewrite the measurement equations for the kth cell cycle as

$$y(i) = (\underbrace{h * \ldots * h}_{k-1 \text{ times}} * \overleftarrow{h_{i-(k-1)L}} * x)(i) + v(i), \tag{5}$$

where $i = (k-1)L + 1, \ldots, kL$. We assume $v(i)$ to be white noise sequence with variance σ^2 and independent of the true expression profile x. For $k = 1$, the $k-1$ times repeated convolution is considered as the identity operator. We are particularly interested in the measurements of the first and the second cell cycle since most of the current gene expression time-series measurements are taken from that time-frame. These measurements are $y(i) = (\overleftarrow{h_i} * x)(i) + v(i)$, $i = 1, \ldots, L$, and $y(i) = (h * \overleftarrow{h_{i-L}} * x)(i) + v(i)$, $i = L+1, \ldots, 2L$, respectively.

Let us consider the cell cycle regulated genes only, previously denoted by x_j, $j \in I_{CCR}$ (we omit the index j in the following). The situation is now quite familiar from the context of adaptive signal processing (see e.g. [4]). That is, we are given two signals $y(i)$, $i = 1, \ldots, L$, and $y(i)$, $i = L+1, \ldots, 2L$, taken from consecutive cell cycles. Because we are considering only the CCR genes, the previous two expression profiles should be the same in ideal case, i.e., in the case of no smoothing. Then the goal is to find a linear time-invariant mapping $\hat{h}$ from the first cell cycle to the second cell cycle. This

mapping is directly an estimate of the kernel h that spreads the distribution of the cell population. So, using the above notation we want to find a mapping from $(\overleftarrow{h_i} * x)(i)$, $i = 1, \ldots, L$ to $(h * \overleftarrow{h_{i-L}} * x)(i)$, $i = L+1, \ldots, 2L$. However, we can not do that directly since our measurements are deteriorated by the additive noise term v. We simplify our problem slightly and find an optimal mapping from $(\overleftarrow{h_i} * x)(i) + v(i)$, $i = 1, \ldots, L$ to $(h * \overleftarrow{h_{i-L}} * x)(i) + v(i)$, $i = L+1, \ldots, 2L$. We give the relation between these two cases after we have shown the standard solution for the latter one.

To follow the standard notation [4], we set the measurements from the first cell cycle to be the input signal,

$$u(i) = (\overleftarrow{h_i} * x)(i) + v(i) =_{\text{def.}} z(i) + v(i),$$

for all $i = 1, \ldots, L$, and the measurements from the second cell cycle to be our reference signal,

$$d(i) = (h * \overleftarrow{h_i} * x)(i + L) + v(i + L) = (h * z)(i + L) + v(i + L),$$

for all $i = 1, \ldots, L$. The computed output is then defined as

$$o(i) = (\hat{h} * u)(i) = \hat{h} * (\overleftarrow{h_i} * x + v)(i) = (\hat{h} * (z + v))(i),$$

for all $i = 1, \ldots, L$. Note that h denotes the true spread of the distribution of the cell population, and $\hat{h}$ is an estimate of h. The error signal $e(i)$ is then the difference between $d(i)$ and $o(i)$, $i = 1, \ldots, L$. Let us formalize this in vector form for notational simplicity and consistency. Let the computed output be

$$o(i) = \hat{\mathbf{h}}^T \mathbf{u}(i) = \hat{\mathbf{h}}^T (\mathbf{z}(i) + \mathbf{v}(i)),$$

where $\hat{\mathbf{h}} = (\hat{h}(-k), \ldots, \hat{h}(-1), \hat{h}(0), \hat{h}(1), \ldots, \hat{h}(-k))^T$, and $\mathbf{u}(i) = (u(i + k), \ldots, u(i + 1), u(i), u(i - 1), \ldots, u(i - k))^T$ ($\mathbf{z}(i)$ and $\mathbf{v}(i)$ are defined in the same way), and their length are $2k + 1$. The reference signal can be expressed, using similar notation, as

$$d(i) = \mathbf{h}^T \mathbf{z}(i) + v(i).$$

The optimal solution minimizing the expectation of the squared error, $E(e^2(i))$, i.e., the minimum mean-square error solution is the well known Wiener filter solution [4]. Optimal FIR-filter coefficients are usually given in matrix form $\hat{\mathbf{h}} = \mathbf{R}_u^{-1} \mathbf{p}_{du}$, where $\hat{\mathbf{h}}$ contains the optimal filter coefficients, $\mathbf{R}_u = E(\mathbf{u}(n)\mathbf{u}(n)^T)$ is the autocorrelation matrix of the input signal u that is assumed to be non-singular, and $\mathbf{p}_{du} = E(d(n)\mathbf{u}(n))$ is the cross-correlation vector between the input signal u and the reference signal d.

Because $\mathbf{R}_u$ and $\mathbf{p}_{du}$ are unknown, they must be estimated from the data. Clearly, we can not estimate $\mathbf{R}_u$ and $\mathbf{p}_{du}$ for each gene separately if we as-

sume h to be non-stationary. Due to the shortness of the available time-series, we can not get reliable estimates for $\mathbf{R}_u$ and $\mathbf{p}_{du}$ even in the stationary case if $\mathbf{R}_u$ and $\mathbf{p}_{du}$ are estimated for each gene separately. Therefore, and since the distribution should have similar effect on all the genes we simply average them over all the cell cycle regulated genes, that is,

$$\hat{\mathbf{R}}_u = \frac{1}{C} \sum_{i \in I_{CCR}} \mathbf{u}_i(n)\mathbf{u}_i(n)^T \tag{6}$$

$$\hat{\mathbf{p}}_{du} = \frac{1}{C} \sum_{i \in I_{CCR}} d(n)\mathbf{u}_i(n)^T, \tag{7}$$

where $C = |I_{CCR}|$ is a proper scaling factor, and index i distinguishes separate CCR genes. Note that, $\hat{\mathbf{R}}_u$ and $\hat{\mathbf{p}}_{du}$ are actually time-varying estimates. In the case of stationary h, $\hat{\mathbf{R}}_u$ and $\hat{\mathbf{p}}_{du}$ are averaged over time, too. That is, an extra sum over n is added to Equations (6) and (7).

The found 3-length impulse responses $\hat{h}$ for time indices $i = 2, \ldots, 10$ are shown in Figure 3, and ordered from left to right and from top to bottom. In detail, the following impulse responses $\hat{h}$ map the cell distributions from time instances $i = 2, \ldots, 10$ to time instances $i = 11, \ldots, 19$. We used data from the CDC15 experiments. The kernels $\hat{h}$ are estimated based of those cell cycle regulated genes only whose DFT give the largest values in the 9th or 10th position (see Section 2.1 for further details).

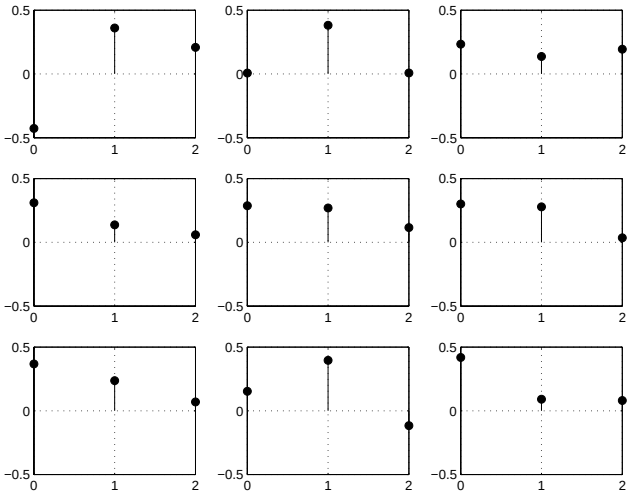


Figure 3. The found 3-length impulse responses h for time indices $i = 2, \ldots, 10$.

The same investigations were made for the 5-length impulse responses and the results are shown in Figure 4. That is, the following impulse responses $\hat{h}$ map the cell distributions from time instances $i = 3, \ldots, 11$ to time instances $i = 12, \ldots, 20$. The found impulse responses seem to support the assumption of a stationary kernel $\hat{h}$.

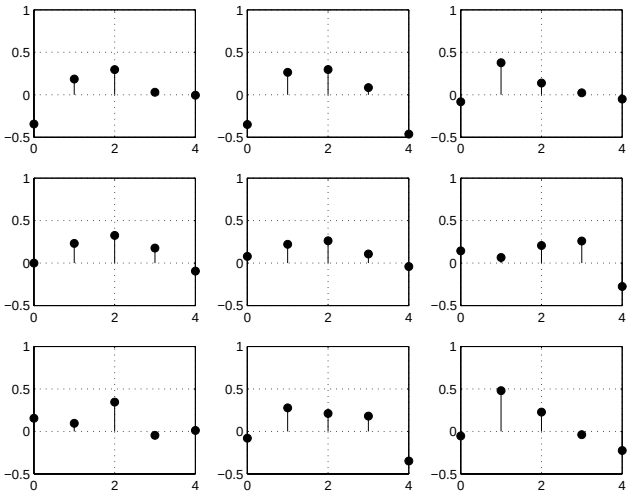
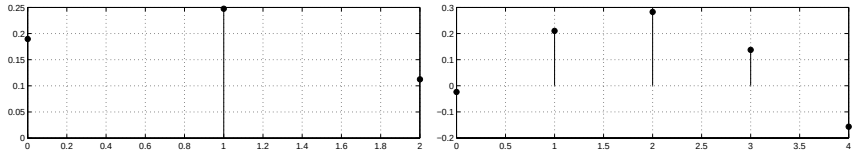


Figure 4. The found 5-length impulse responses h for time indices $i = 3, \ldots, 11$.



(a) (b)

Figure 5. Stationary estimates for the kernels h having length (a) 3 and (b) 5.

One is able to get more reliable estimates for $\hat{\mathbf{R}}_u$ and $\hat{\mathbf{p}}_{du}$, and further for $\hat{h}$, by assuming stationary kernel $\hat{h}$. The stationary estimates for the kernels having length 3 and 5 are shown in Figure 5.

It was mentioned above that we made a minor simplification in our approach to find $\hat{h}$. We now examine the effects of this simplification. It is fairly easy to see that

$$\begin{aligned} E(e^2(i)) &= E(d^2(i)) - 2E(d(i)o(i)) + E(o^2(i)) \\ &= \underbrace{\mathbf{h}^T\mathbf{R}_z\mathbf{h} + \sigma^2}_{E(d^2(i))} - 2\underbrace{\mathbf{h}^T\mathbf{R}_z\hat{\mathbf{h}}}_{E(d(i)o(i))} + \underbrace{\hat{\mathbf{h}}^T\mathbf{R}_u\hat{\mathbf{h}}}_{E(o^2(i))} \end{aligned}$$

and further that

$$\frac{\partial}{\partial\hat{\mathbf{h}}}E(e^2(i)) = -2\mathbf{h}^T\mathbf{R}_z + 2\hat{\mathbf{h}}^T\mathbf{R}_u = -2(\mathbf{p}_{du}^T - \hat{\mathbf{h}}^T\mathbf{R}_u).$$

It also follows that $\mathbf{R}_u = \mathbf{R}_z + \mathbf{R}_v = \mathbf{R}_z + \sigma^2\mathbf{I}$ since z and v, or actually x and v are statistically independent (by assumption). Thus, when solving

the optimal filter coefficients we get

$$\hat{\mathbf{h}} = \mathbf{R}_u^{-1}\mathbf{p}_{du} = (\mathbf{R}_z + \sigma^2\mathbf{I})^{-1}\mathbf{R}_z\mathbf{h}.$$

If we were able to get a reliable estimate $\hat{\sigma}^2$ for σ^2, we could find an estimate of $\mathbf{R}_z$ as $\hat{\mathbf{R}}_z = \hat{\mathbf{R}}_u - \hat{\sigma}^2\mathbf{I}$. Then we could solve for $\mathbf{h}$ as

$$\mathbf{h} = \hat{\mathbf{R}}_z^{-1}\hat{\mathbf{R}}_u\hat{\mathbf{h}}$$

assuming all inverse matrices exist.

It is fairly easy to design a proper separate experiment for estimating the noise variance σ^2, assuming there exists only additive noise. Basically, one only needs to perform a simple repetition experiment for gene expression measurements at a single time instant. By repetition we mean that one should measure the expression value of a gene (or certain genes) multiple times. Since all the measurements should have the same expression value, the fluctuation in the measured values are then caused by the noise. Further discussion of this topic is left out of the scope of this paper.

3 PROPOSED SOLUTIONS FOR FINDING THE DISTRIBUTION OF CELL POPULATION

Inversion of the smoothing effect, as introduced by Equations (1) and (2), can be made essentially easier by collecting certain additional information during the standard microarray time-series measurements [12][13]. Such additional measurements can be e.g. flow cytometry, such as fluorescent-activated cell sorter (FACS) [8][10], or bud counting data [8]. The use of such additional data sources are introduced in this Section. We also introduce a way of applying the blind deconvolution methods to estimate the distribution of the cell population in Section 3.5.

3.1 Additional Information for Finding the Distribution of the Cells

In general, flow cytometry is an experimental technique for measuring certain physical and chemical characteristics of cells as they travel in cell suspension past a sensing point. We are able to measure cellular parameters related to cell size, shape and internal complexity, as well as any cell component or function which can be defected by a fluorescent compound. One of the measures of the flow cytometry can be the variation of the amount of DNA in cells during the cell cycle. Basically, the amount of DNA grows from one unit up to two units, corresponding to the beginning of the cell

cycle and the time before cell divides [8]. That is, the amount of DNA in a single cell defines its place in the cell cycle, assuming we know exactly the growth of the amount of DNA during the cell cycle. Flow cytometry can be used to measure the distribution of the DNA amount of the cell population during the microarray time-series measurements. Similar argument applies to the bud counting as well. The main difference is that bud counting measures the distribution of the size of the buds.

As the first possibility, we have measurements of a rapidly changing parameter. For instance, the amount of DNA changes rapidly from two units to one unit per cell during the cytokinesis (cell division) [8]. Similarly, budding occurs quickly, although at a different place of the cell cycle. So, assume we have measured parameters changing rapidly in ideal case. Observed smooth changes in those parameters are then caused by the loss of synchrony. A method that utilizes such data is introduced in Section 3.2.

In a bit more complicated problem setting, we know only the distributions of a binary valued feature. That is, we can only measure the number of cells with one or two copies of the DNA strand. Similar arguments apply to other methods as well, such as bud counting data, if we can only measure the number of cells with or without a bud. A method that utilizes such binary-valued data is introduced in Section 3.3.

Alternatively, we may know both the whole distribution (histogram) of a certain cell cycle dependent parameter and its growth during the cell cycle, such as the amount of DNA in a cell. Then, we can map the measured histogram directly into a distribution (histogram) of the cell population. This method is shown in Section 3.4.

3.2 Estimation Method Based on a Rapidly Changing Parameter

The distribution of the cells in time can be estimated if there are parameters, whose values have sharp discontinuities, or, the value of the parameter changes rapidly from one level to another. This parameter could be, for instance, the amount of DNA, which drops from two units to a single unit as the cell divides. To see whether this method actually works, let us consider the simulated data shown in Figure 6.

Near the sharp discontinuity, the above graph can be approximated by a function well known in signal processing, the step function $u(n)$,

$$u(n) = \begin{cases} 1, & \text{if } n \geq 0, \\ 0, & \text{if } n < 0. \end{cases}$$

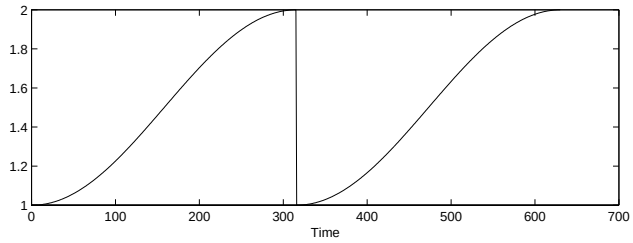


Figure 6. Simulated ideal data of a parameter changing rapidly from one level to another.

Since $u(n)$ increases from zero to one at $n = 0$, we need to consider its time-reversed version and add unity to the result. Thus, near the edge, we can approximate the parameter by $1 + u(-n)$.

Due to the cell asynchrony, the actual measurements no longer result in a sharp discontinuity. Instead, the edge is spread over a longer period of time, as shown in Figure 7. In addition to the spreading, we have added white Gaussian noise with variance 0.001.

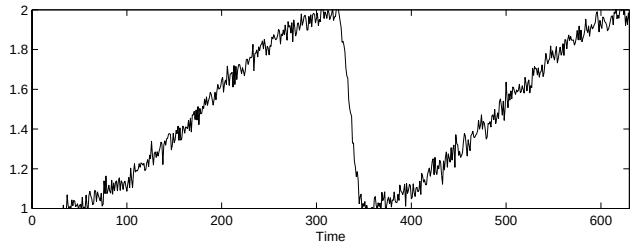


Figure 7. Simulated noisy and "smoothed" data of a parameter changing rapidly from one level to another.

If we denote the distorted data of Figure 7 without noise by $y(n)$, the following equality holds,

$$y(n) = F(1 + u(-n)),$$

where $F(\cdot)$ is the unknown system behind the distortion. If we assume that $F(\cdot)$ is linear and time-invariant (LTI), as can be assumed according to Equations (1) and (2), we can find its impulse response easily. The connection between an impulse and the time-reversed step edge is given by $\delta(n) = u(-n-1) - u(-n)$. Using the linearity and time-invariance properties, we can get the impulse response from the measured step response $y(n)$ as

$$\begin{aligned} h(n) &= F(\delta(n)) = F(u(-n-1) - u(-n)) \\ &= F(1 + u(-n-1) - (1 + u(-n))) \\ &= F(1 + u(-n-1)) - F(1 + u(-n)) \\ &= y(n-1) - y(n). \end{aligned}$$

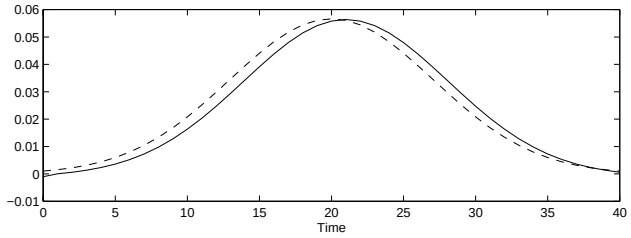


Figure 8. Simulated cell distribution (dashed line) and the distribution approximated from simulated data without noise.

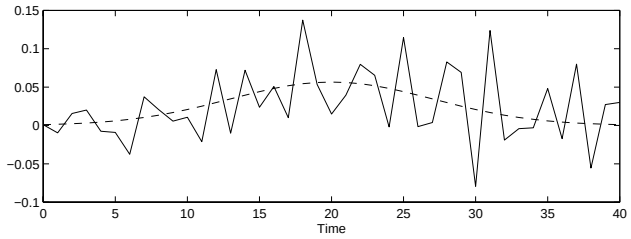


Figure 9. Simulated cell distribution (dashed line) and the distribution approximated from simulated data with Gaussian noise ($\sigma^2 = 0.001$).

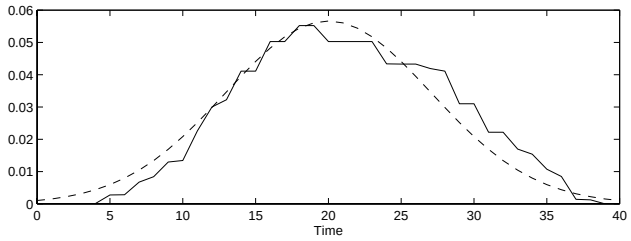


Figure 10. Simulated cell distribution (dashed line) and the distribution approximated from simulated data with Gaussian noise ($\sigma = 0.001$). The data has been prefiltered with a median filter with $N = 15$ and the resulting impulse response has also been smoothed by a median filter with $N = 15$.

Figure 8 confirms that the method works efficiently in case the measurements are noise-free. The estimated distribution differs from the theoretical one because the ideal input was not exactly the time-reversed step function.

If the measurements are noisy, then we have $y(n)+\nu(n)$ instead of $y(n)$ and therefore $h(n) = y(n-1) + \nu(n-1) - y(n) - \nu(n)$, i.e., the noise $\nu(n)$ in the measurements directly affects the estimated $h(n)$. If the input measurements are corrupted by Gaussian noise with variance 0.001, the impulse response is far away from the theoretical optimum. Figure 9 shows that the estimation is not very robust in the presence of noise. There are several alternatives for improving the estimate of the impulse response. For instance, in the case of Gaussian additive noise $\nu(n)$, one could use e.g. mean filters. However, we do not want to further smooth the edges, i.e., we need to apply some edge-preserving filters. Thus, some edge-preserving nonlinear filter may be a good choice in our case (see e.g. [2]). However, without

knowing the characteristics of the true measured data, it is difficult to make any assumptions on the filter type that should be used in this application. Using a simple median filter for filtering the measurements and for filtering the estimated impulse response results in a significantly more reliable impulse response, see Figure 10.

The method introduced above assumes that the underlying distribution remains constant during the time when the cell population passes the "checkpoint". Previously, we considered fully time-varying widening phenomenon for the cell population. Once more information about the underlying biological processes are obtained we can consider more tailored methods for estimating the distribution of the cell population. Therefore, we leave the modification of this method for time-invariant purposes as part of the future work.

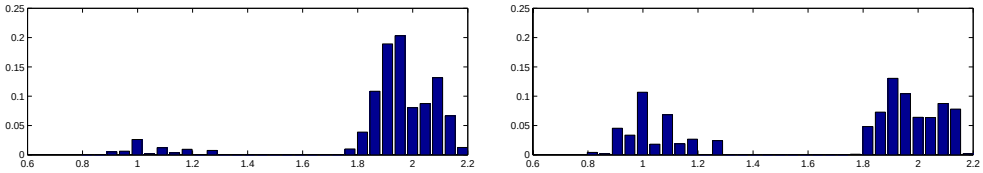
3.3 Estimation Method Based on a Binary-Valued Parameter

FACS equipment produces distributions (histograms) of the amount of DNA in the cell population. Sampling the population in time with FACS device therefore produces a time-series of histograms. It appears that this method can provide rather robust estimates of the distribution even in noisy conditions.

Similarly to Section 3.2, the preferable situation is such that we measure the FACS data during such a time interval where the measured parameter experiences a sudden change from one level to another, as shown in Figure 6. In ideal conditions the measurement would result in a single peak, i.e., the device would give exactly the correct value of the parameter in a histogram form.

However, due to the spread of the population in time and the noise in the measurements, the result is not the ideal one. Instead, near the edge the histogram is bimodal since some of the cells are already on the lower level while others are still on the upper level. Furthermore, the spreading of the population and the noise in the measurements together causes the values to spread around the two peaks. An example of resulting simulated histogram is shown in Figure 11. The figure on the left shows the histogram before the majority of the cells have moved to the right hand side of the edge. The figure on the right has a larger mass of the histogram near the unity indicating that a larger amount of cells have gone past the edge. The distribution of the cells is the one shown in Figure 8 and the measurements were distorted by Gaussian noise with variance $\sigma^2 = 0.01$.

In order to analyze the bimodal histogram, we need to find the point that



(a) (b)

Figure 11. An example of a FACS histogram. (a) The smaller peak is caused by cells that are already moved to level 1 and the larger peak is caused by cells still on the level 2. (b) The peak on the left has grown and the one on the right has shrunk indicating that the population is moving to the unity level.

most naturally divides the histogram in two. This can be done either manually or by finding the index k minimizing

$$\min_k \left(\mathrm{var}(h(1), h(2), \ldots, h(k)) + \mathrm{var}(h(k+1), h(k+2), \ldots, h(N))\right),$$

where $h(1), h(2), \ldots, h(N)$ is the histogram with N bins and $\mathrm{var}()$ denotes the sample variance. Other binarization methods can also be used for the same purpose.

The spread of the population can easily be determined from the FACS histograms sampled frequently enough. For example, in Figure 11 (a), the peak near unity contains 7.2% of the total mass of the histogram, whereas in Figure 11 (b) the mass is 34.8%. This means that between these two sampling instants, 27.6% of the population has moved across the edge. Thus, we can approximate the density of the distribution at this point by 0.276. Going through all the sample histograms, we can build an estimate of the true population distribution. Naturally, the accuracy of the estimate depends on the sampling period. In Figure 12, we have assumed that the total spread of the population spans 81 sampling instants. Figure 12 shows the estimated spread for different noise variances.

The above method assumes that the distribution remains constant during the time when cells travel through the checkpoint. We again leave the modification of this method for time-invariant purposes as part of the future work.

3.4 Direct Conversion of the Distribution of a Cell Cycle Regulated Parameter

This method provides a way of estimating the true distribution of sample cell population directly, assuming we know the growth of the corresponding parameter during the cell cycle. We assume to know both the whole

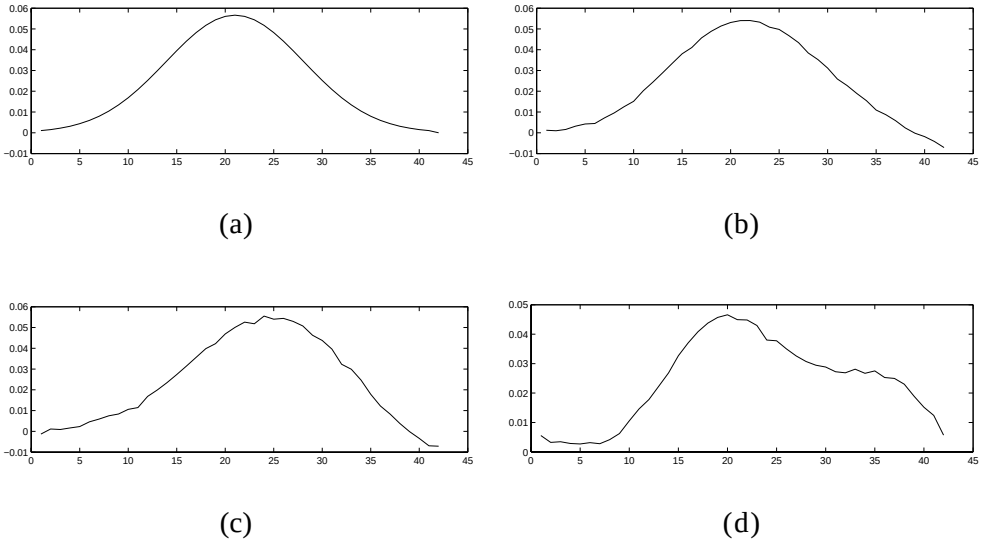


Figure 12. Estimated cell population spreads. The simulated data has been corrupted by zero-mean Gaussian noise with variances a) $\sigma^2 = 0.01$, b) $\sigma^2 = 0.1$, c) $\sigma^2 = 0.3$ and d) $\sigma^2 = 0.5$.

distributions of a certain cell cycle dependent parameter, or an estimate of that, and its growth during the cell cycle, like in the case of the amount of DNA in a cell introduced in Section (3.1). Although we concentrate only on FACS measurements for a while the same methods applies to other measurements as well, assuming the growth curve of the measured parameter for a single cell is known during the cell cycle.

Let us denote the output of FACS device at time instant i as $f_i(t)$, where t denotes the position (phase) at the cell cycle, and the growth of the amount of the DNA as $g(t)$. Let us further denote the cell cycle by an interval $[0, t_{max}]$, where 0 and t_{max} denote the beginning and the end of the cell cycle. One hypothetical graph representing the growth of the DNA amount during the cell cycle is shown in Figure 6 with t_{max} approximately equal to 310 (units are unimportant here). The function g is also assumed to be strictly increasing, i.e., $(0 \leq t_a < t_b \leq t_{max}) \Rightarrow (g(t_a) < g(t_b))$. The assumption of a strictly monotonic g may have some biological arguments but it is given mostly for mathematical convenience. That is, under these assumptions the output of the FACS device can be converted into a distribution of the cell population by using a combined mapping as

$$p_i(t) = (f_i \circ g)(t) = f_i(g(t)) \tag{8}$$

for all $t \in [0, t_{max}]$.

We previously discussed the sharp discontinuities in the graph of the growth of some cell cycle dependent parameters (see e.g. Section 3.2). This contra-

dicts with the assumption on the strictly increasing g unless the discontinuity is located at the end of the cell cycle, as is the case for the growth of the DNA amount. However, one is able to handle one discontinuity in the middle of the cell cycle, as well. The distribution p_i can be formed by using Equation (8) as long as the function g is injective.

In practise we will not be working with continuous distributions. Basically, this follows from the fact that e.g. FACS device outputs only a discrete approximation of the distribution of the amount of DNA in the cell population. In discrete case one only needs to consider a discrete version of Equation (8).

Equation (8) will also produce a noisy estimate of the underlying distribution. Unfortunately, there seems to be no information about the characteristics of the noise present in f_i. Those noise characteristics may also be rather difficult to estimate. So, it is fairly hard to give arguments for the use of any particular noise removal method on f_i, not to speak of the optimal parameter values for that method. However, if estimates for f_i turn out to be corrupted by significant amount of noise, one possible noise removal method could utilize f_i, or p_i, for all $i = 1, \ldots, m$ altogether. It could be possible to tailor a method, based on *a priori* knowledge of the underlying biological process, where estimates of p_i are modified so that the resulting distributions would most probably be originated from the corresponding process.

3.5 Blind Deconvolution Methods

During the last decade, blind channel identification methods have been succesfully used in digital communications. The communication channel distorts the signal, and this distortion can be modeled by convolution. In order to recover the original data, the impulse response needs to be found. Traditional methods use training data for this purpose, but this causes inefficient use of the communication channel. A significant amount of the bandwidth is used for transmitting training data, i.e., a fixed sequence known by both the transmitter and the receiver. The bandwidth is used more efficiently if blind methods for channel identification can be used. Such methods estimate the channel impulse response usually from statistical properties of the output only. Therefore, no training data needs to be transmitted.

The channel identification problem in digital communications is similar to the problem of finding the distribution of the cell population. Now, the measured data is distorted by the spread of the cells in time. Blind methods are useful here, since the identification using training sequences is difficult. In

order to use training we would need a way of generating known input sequences, which is often impossible.

Many recent blind channed estimation techniques use so called subspace methods. The idea behind the subspace methods is that the distortion caused by the channel lies in a subspace of the observation statistics. Moreover, this subspace is orthogonal to the noise subspace. In order to find the channel behind the distortion in the gene expression profiles, we use the method of Moulines *et al.* [9]. Moulines' method assumes a single data source of which several distorted versions are acquired at the receiving end. In our case, each cell cycle represents one measurement of the underlying true events of one cycle.

The identification of a communication channel relies on large amounts of available data. In our case, the time span of the measurements is small, but this can be compensated by the large amount of measurements made on each time instant. Because each one of yeast's roughly 6000 genes is measured on every time instant, there are 6000 examples of the distortion on all cell cycles. This can be exploited by rearranging the data into a vector. For instance, suppose, that there are measurements from two cell cycles,

$$\mathbf{X}_1 = \begin{pmatrix} x_1(1) & x_1(2) & \cdots & x_1(L) \\ x_2(1) & x_2(2) & \cdots & x_2(L) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(1) & x_n(2) & \cdots & x_n(L) \end{pmatrix}$$

and

$$\mathbf{X}_2 = \begin{pmatrix} x_1(L+1) & x_1(L+2) & \cdots & x_1(2L) \\ x_2(L+1) & x_2(L+2) & \cdots & x_2(2L) \\ \vdots & \vdots & \ddots & \vdots \\ x_n(L+1) & x_n(L+2) & \cdots & x_n(2L) \end{pmatrix},$$

where $\mathbf{X}_1$ contains measurements of a total of N genes during the first cell cycle (L first time instants) and $\mathbf{X}_2$ contains the measurements during the second. Then, two one-dimensional signals are constructed as follows,

$$\begin{aligned} \mathbf{x}_1 = (&x_1(1), x_1(2), \ldots, x_1(L), x_2(1), x_2(2), \ldots, x_2, \ldots, \\ &x_n(1), x_n(2), \ldots, x_n(L))^T, \\ \mathbf{x}_2 = (&x_1(L+1), x_1(L+2), \ldots, x_1(2L), x_2(L+1), x_2(L+2), \ldots, \\ &x_2(2L), \ldots, x_n(L+1), x_n(L+2), \ldots, x_n(2L))^T, \end{aligned}$$

i.e., the rows of both $\mathbf{X}_1$ and $\mathbf{X}_2$ are read one by one. This way we have two

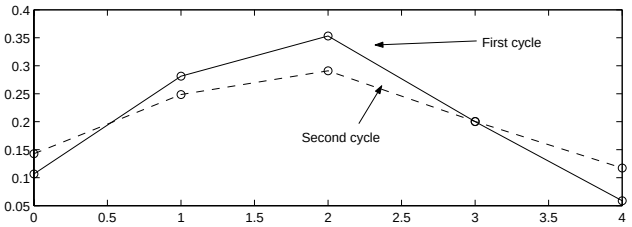


Figure 13. The channel estimates obtained by blind techniques.

measurements of the same signal corrupted by different channels. Now we can apply the subspace based identification of the channel as explained by Moulines. Since the original signal has to be same for both channels, we'll have to restrict the data to those genes that are cell-cycle dependent and have the same cycle length. As shown in Section 2.1, the length of the cycle can be estimated using the Fourier transform. Since the most frequent cell cycle length in CDC15 test seems to be ten, we use exactly those 171 genes. The results for the blind channel estimation using this data with window length 5 are shown in Figure 13.

4 INVERSION OF THE SMOOTHING CAUSED BY CELL POPULATION ASYNCHRONY

Ultimately, we would like to invert the smoothing effect caused by the distribution of the cells as shown in Equations (1) and (2). So, basically we aim to invert a discrete inner product with time-varying kernel, or a discrete convolution with time-reversed and time-varying kernel, where the kernel corresponds to the discrete approximation of the distribution of the cell population.

Once we have found estimates for either the underlying distribution of the cells themself or the spreading rate of the distribution of the cells we need to consider proper approaches for inverting the "smoothing" effects as shown in Equations (1) and (2). We can use several different methods for that purpose. The standard inverse filtering method and a regression-type approach are introduced in Sections 4.1 and 4.2, respectively.

4.1 Inverse filtering

The standard approach for inverting the effects of convolution is the so-called inverse or Wiener filter approach (see e.g. [5]). In order to be able to implement the optimal inverse filter one should know the spectral density function of the true signal x and the noise v, often denoted as S_{uu} and S_{vv},

respectively. One also needs to take into account that the convolution kernel h_i is time-varying, resulting in a time-varying inverse filter. Although the inverse filter is time-dependent we only need to compute the outputs for each time instant with different inverse filters. So, the Fourier transform of the inverse filter for the time instant i is then

$$G_i(\omega) = \frac{H_i^*(\omega)S_{xx}(\omega)}{|H_i(\omega)|^2 S_{xx}(\omega) + S_{vv}(\omega)} = \frac{H_i^*(\omega)}{|H_i(\omega)|^2 + S_{vv}(\omega)/S_{xx}(\omega)}, \quad (9)$$

where $H_i(\omega)$ denotes the frequency response of the convolution kernel h_i and the spectral density functions S_{uu} and S_{vv} are assumed to be stationary.

Most often S_{uu} and S_{vv} are unknown and therefore, they have to be estimated by some means. On the other hand, if we assume v to be white noise, as was done in Section 2.2, $S_{vv}(\omega)$ will be constant and equal to $1/2\pi$, where 2π refers to the normalized sampling frequency. The right hand side of Equation (9) in turn provides a simple way of estimating the spectrum. That is, one only needs to replace $S_{vv}(\omega)/S_{xx}(\omega)$ by a proper estimate. In principle, a model for the $S_{xx}(\omega)$ could be "invented" in the auto-correlation domain r_{xx}, and then converted it into S_{xx} by the Fourier transform.

4.2 Regression-type Approach

Even though the inverse filter provides an optimal method of inverting the effects of convolution, its implementation may be troublesome. The problems arise since the spectral density functions S_{uu} and S_{vv} are usually unknown in practise. We also need to apply the inverse filter in frequency domain. If we applied the inverse filter in spatial domain we would lost a major part of our signal since the measured expression profiles are short. Therefore, we also introduce another method based on solving a set of linear equation. This method is applicable due to the same reason as the spatial-domain inverse filter is inappropriate, i.e., the gene expression profiles are of little length. It is also easy to form the set of linear equations such that we are able to utilize the cyclic behavior of the cell cycle regulated genes, assuming the measurements are taken from appropriate time points.

Let us first summarize our assumptions on the available measurements. We assume to have the gene expression measurements $y(i)$ (for n genes) and estimates for the distribution of the cell population p_i, or more precisely, corresponding discrete approximations h_i for all time instants $i = 1, \ldots, m$. Discrete kernels h_i may be formed e.g. by using Equation (3). We also assumed the measurements be corrupted by an additive noise term v. In order to make this even more clear for the reader, we rewrite our approximated

measurement equation

$$y(i) = \sum_j h_i(j)x(i+j) + v(i), \tag{10}$$

that holds for all $i = 1, \ldots, m$.

One can formalize the above set of equations in matrix form by simply expanding Equation (10) for all i, resulting in the following equation

$$\begin{pmatrix} y(1) \\ \vdots \\ y(m) \end{pmatrix} = \begin{pmatrix} \cdots & h_1(-1) & h_1(0) & h_1(1) & \cdots \\ & \ddots & \ddots & \ddots & \ddots \\ & \cdots & h_m(-1) & h_m(0) & h_m(1) \ \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ x(0) \\ x(1) \\ \vdots \\ x(m) \\ x(m+1) \\ \vdots \end{pmatrix} + \begin{pmatrix} v(1) \\ \vdots \\ v(m) \end{pmatrix}. \tag{11}$$

The same can be written in more compact form as $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{v}$, where definitions of the variables $\mathbf{y}$, $\mathbf{H}$, $\mathbf{x}$ and $\mathbf{v}$ are evident. The noise v was previously assumed to be white. Under this assumption, we can write $E(\mathbf{u}) = \mathbf{0}$ and $V(\mathbf{u}) = \sigma^2\mathbf{I}$, where $V(\mathbf{u})$ denotes the covariance of $\mathbf{u}$.

At first, we assume to have reliable estimates for all h_i, $i = 1, \ldots, m$, such that we can consider the kernel matrix $\mathbf{H}$ to be known. Further, we concentrate first on the cell cycle regulated genes. Let us also assume that measurements from consecutive cell cycles are taken precisely from the same phases of the cell cycle. Fortunately, we are able to control this requirement in practise. Since the genes under consideration are periodic, we have that $x(i) = x(i + L)$ for all $i = 1, \ldots, m - L$. Combining that information into Equation (11) one can rewrite it as

$$\begin{pmatrix} y(1) \\ \vdots \\ y(m) \end{pmatrix} = \begin{pmatrix} h_1(0) & h_1(1) & h_1(2) & \cdots & h_1(-1) \\ h_2(-1) & h_2(0) & h_2(1) & \cdots & h_2(-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_L(1) & h_L(2) & h_L(3) & \cdots & h_L(0) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_m(k) & h_m(k+1) & h_m(k+2) & \cdots & h_m(k-2) \end{pmatrix} \begin{pmatrix} x(1) \\ \vdots \\ x(L) \end{pmatrix} + \mathbf{v}, \tag{12}$$

where $k = (m \bmod L) + 1$, $k \pm 1$ and $k \pm 2$ are computed in modulo L, and the kernel matrix $\mathbf{H}$ belongs to $\mathbf{R}^{m \times L}$. In essence, equation (12) sets up the standard regression problem. Because we assume to have measurements from consecutive cell cycles, we have that $m > L$, and the rank of the kernel matrix $\mathbf{H}$ in Equation (12) is most probably full, i.e., $\operatorname{rank}(\mathbf{H}) = L$. According to the well-known corollary of the Gauss-Markov Theorem [6], the best

linear unbiased estimate of the true expression profile $\mathbf{x}$ is

$$\hat{\mathbf{x}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}. \tag{13}$$

If we are unable to control the measurements from consecutive cell cycles to be taken precisely from the same positions, the solution becomes slightly more complicated. However, this can be formalized in a matrix form very similar as the one shown above. Let the cell cycle be denoted by a time interval $[0, t_{max}]$ as previously and $t'_i = t_i \bmod t_{max}$,$i = 1, \ldots, m$, be measurement times in modulo t_{max}. Let us also sort the wrapped measurement times t'_i and denote them as $t'_{(i)}$. The coefficients of the kernels h_i are now formed as in Equation (3) except that integrals are computed over the intervals $[t'_{(j)} - (t'_{(j)} - t'_{(j-1)})/2, t'_{(j)} + (t'_{(j+1)} - t'_{(j)})/2]$. The discrete measurement time indices $i = 1, \ldots, m$ are permutated in the same order as $t'_{(i)}$s and they denoted as $i' = 1', \ldots, m'$. Equation (10) can again be written in matrix form as in Equation (12) with an exception that the kernel matrix $\mathbf{H}$ belongs now to $\mathbf{R}^{m\times m}$. In detail, the matrix formulation can be given as

$$\begin{pmatrix} y(1') \\ \vdots \\ y(m') \end{pmatrix} = \begin{pmatrix} h_{1'}(0) & h_{1'}(0) & h_{1'}(2) & \cdots & h_{1'}(-1) \\ h_{2'}(-1) & h_{2'}(0) & h_{2'}(1) & \cdots & h_{2'}(-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{m'}(1) & h_{m'}(2) & h_{m'}(3) & \cdots & h_{m'}(0) \end{pmatrix} \begin{pmatrix} x(1') \\ \vdots \\ x(m') \end{pmatrix} + \begin{pmatrix} v(1') \\ \vdots \\ v(m') \end{pmatrix}, \tag{14}$$

If $\mathrm{rank}(\mathbf{H}) = m$ the optimal solution is given as in Equation (13).

Equation (11) can not be put in more compact form for the genes that are not cell cycle regulated. As a consequence, the rank of the kernel matrix $\mathbf{H}$ is not full, i.e., $\mathrm{rank}(\mathbf{H}) < m$. Equation (11) shows, however, that the noise-free smoothed expression profile (vector $\mathbf{Hx}$) must lie in the range of the $\mathbf{H}$, $R(\mathbf{H})$, i.e., in the space spanned by the columns of $\mathbf{H}$. The standard Gauss-Markov Theorem [6] can be applied to equation $\mathbf{y} = \mathbf{Hx} + \mathbf{v}$ in order to get the best linear unbiased estimate for $\mathbf{z} =_{\mathrm{def.}} \mathbf{Hx}$. This is achieved by projecting the measured vector $\mathbf{y}$ orthogonally into the space $R(\mathbf{H})$. The orthogonal projection can be computed using a matrix multiplication $\hat{\mathbf{z}} = P(R(\mathbf{H}))\mathbf{y}$, where $\hat{\mathbf{z}}$ is the optimal linear estimate of $\mathbf{Hx}$ and $P(R(\mathbf{H}))$ is the orthogonal projector into the $R(\mathbf{H})$. The estimate of $\mathbf{x}$ itself is not unique in this case since $\mathrm{rank}(\mathbf{H}) < m$. In other words, proper solutions for $\mathbf{x}$ constitute a space $S_{\mathbf{x}} = \{\mathbf{x} | \mathbf{x} = \hat{\mathbf{x}} + \mathbf{x}_0,\ \mathbf{x}_0 \in \mathcal{N}(\mathbf{H}))\} \subset \mathbf{R}^m$, where $\hat{\mathbf{x}}$ is an optimal solution for $\mathbf{x}$ and $\mathcal{N}(\mathbf{H}))$ is the null space of the $\mathbf{H}$. In order to limit the space $S_{\mathbf{x}}$, or even squeeze it into a single point, we should add some extra constraints on the solution, e.g., constraints on the smoothness.

We made some experiments with hypothetical cell distributions and gene expression data to test the regression-type approach. For now, we concentrate only on cell cycle regulated genes. For experimental purposes, we gen-

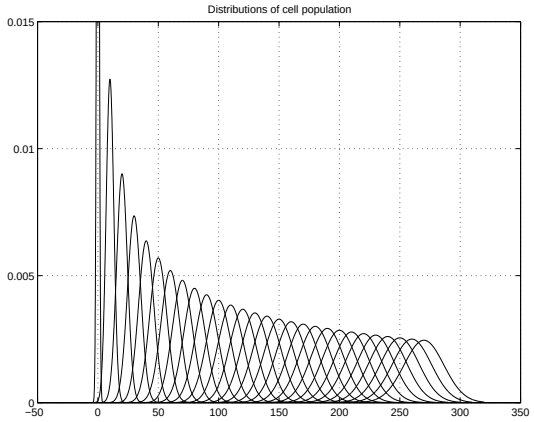


Figure 14. Cell distributions used in experiments.

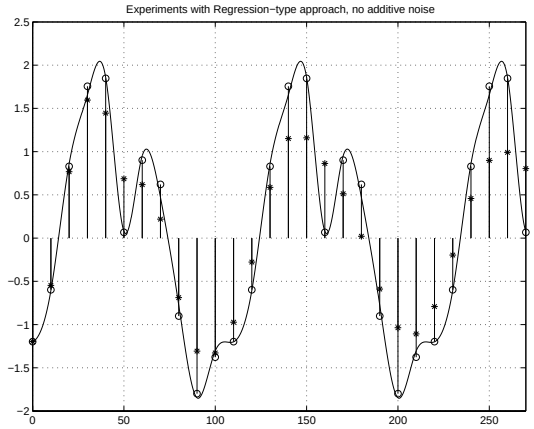


Figure 15. Optimal solution of the gene expression profile for the case with no additive noise.

erated a (fairly random) gene expression profile x which is supposed to represent the underlying signal for the first cell cycle. The signal was then forced to be periodic by repeating it certain number of times, corresponding to the number of observed cell cycles. The number of the observed cell cycles was set to be approximately two and half. Cubic spline interpolation was used to obtain a continuous version of the discrete signal. The measurements were generated according to Equation (1), first by ignoring the additive noise term v and later by adding white noise such that the resulting signal-to-noise ratio is approximately 20. The kernels h_i were computed from the true distributions as shown in Equation (3). The optimal solution was obtained by using Equation (13). The cell distributions used in our experiments are shown in Figure 14. We used truncated Gaussian distributions with increasing variance for that purpose. The optimal solution for the case with no additive noise is shown in Figure 15. The solid curve shows the continuous true gene expression profile, stars denote the observed measurements and circles show the optimal linear unbiased solution, i.e., the corrected measurements. The optimal solution for the case

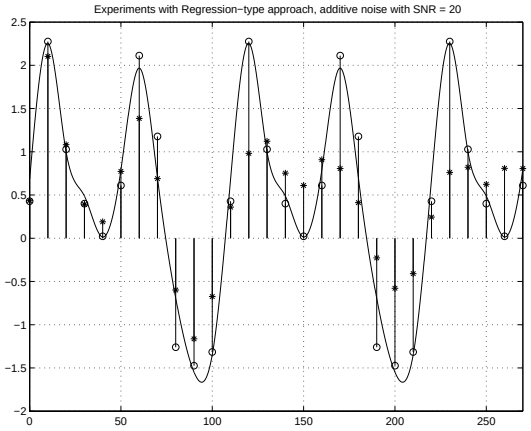


Figure 16. Optimal solution of the gene expression profile for the case with additive white noise.

with additive white noise is shown in Figure 16. The graphs are encoded as above.

So far we have considered the regression-type approach only by assuming that the kernel matrix **H** is known (deterministic). This is not quite true since its elements are measured / estimated. So, matrix **H** is actually stochastic and has a certain probability distribution. It is evident that elements in **H** are not independent random variables. For instance, elements in each row are dependent because they must sum up to unit (rows in **H** represent distributions). Moreover, nor different rows of **H** can be independent since they are defined by the underlying biological process. However, stochastic extensions as illustrated above are left out of the scope of this paper.

5 DISCUSSION

It is worth noting that Equation (1) may represent an oversimplified model of the measurement process, even though it contains an additive noise term. A promising model was recently introduced in [11], where the authors suggest the measurement to contain both additive and multiplicative noise terms as well as certain offset terms and scaling factors between samples from different time points. However, they ignore the averaging effect completely, although its existence is apparent due to the loss of synchrony and the averaging type measurement arrangements, i.e., the mRNA is extracted from a cell population instead of a single cell. On the other hand, we may consider the averaging to take place even before a “standard” measurement process, e.g. the one in [11], so that the two measurement processes can be considered to be nested. In order to keep our discussion simple enough, so that we are able to work with the above smoothing effect, we ignore any other error sources and / or models. Due to the nested-type interpretation

of the errors we are allowed to assume that all multiplicative errors and scalings are compensated well enough in advance. To that end, we simply refer to recent papers that consider the normalization and inversion of such phenomena, see e.g. [3], [11] and [14]. Once more complicated measurement processes become validated we can then consider more advanced approaches to this problem, or even merge our current averaging model, Equation (1), into a general measurement model.

It can be assumed that the quality of the microarray measurements will be improved and the amount of measurements will keep on increasing in the future. The synchronization problem addressed in this paper is, however, independent of the measurement equipment. In other words, even though measurement equipments will get more accurate the synchronization problem will remain due to the biological reasons. This emphasizes the importance of this approach.

There is still plenty of work to be done along the lines of preprocessing and (algorithmically) improving the quality of the microarray measurements. Some possible directions of the future work were already discussed above. For instance, the problem of estimating the noise variance was mentioned in Section 3. Another possibilities of the future work, as were discussed in Sections 3.2 and 3.3, consists of designing truly time-varying methods for estimating the distribution of the cell population, if that turns out to give significantly better results. Some of the methods introduced in this paper are already applied to real gene expression time-series. However, we will apply all these techniques to real gene expression data in the future.

Acknowledgement

The authors wish to thank Dr. Meelis Kolmer and Dr. Christophe Roos from MediCel Ltd. for fruitful discussions.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed., Garland Publishing, New York, 1994.

[2] J. Astola and P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering,* CRC Press, 1997.

[3] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young, “Maximum likelihood estimation of optimal scaling factors for expression array normalization,” *In Microarrays: Optical Technologies and Informatics, Proc. SPIE,* volume 4266, 2001.

[4] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice Hall, 1996.

[5] A. K. Jain, *Fundamentals of Digital Signal Processing*, Prentice Hall, 1989.

[6] R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, 4th ed., 1998.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[8] H. Lodish, A. Berk, L. S. Zipursky, P. Matsudaira, D. Baltimore and J. E. L. Darnell, *Molecular Cell Biology*, 4th ed., New York: W H Freeman & Co, 1999.

[9] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace Methods for the Blind Identification of Multichannel FIR Filters," *IEEE Trans. on Signal Proc.*, pp. 516–525, vol. 43, no. 2, February 1995.

[10] M. Rieseberg, C. Kasper, K. F. Reardon, and T. Scheper "Flow Cytometry in Biotechnology," *Appl Microbiol Biotechnol*, 56(3-4):350-60, 2001.

[11] D. M. Rocke and B. Durbin, "A Model for Measurement Error for Gene Expression Arrays," *Journal of Computational Biology*, 8:557-69, 2001.

[12] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* 270(5235): 467-70, 1995.

[13] D. Shalon, S. J. Smith, and P. O. Brown, "A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization," *Genome Res.*, 6(7): 639-45, 1996.

[14] I. Shmulevich and W. Zhang, "Binary Analysis and Optimization-Based Normalization of Gene Expression Data," (in press), *Bioinformatics*.

[15] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell*, Vol. 9, p. 3273–3297, December 1998.