December 2005

Bayesian Analysis of Cell-Cycle Gene Expression Data

Chuan Zhou
Vanderbilt University, chuan.zhou@vanderbilt.edu

Jon Wakefield
University of Washington, jonno@u.washington.edu

Linda Breeden
Fred Hutchinson Cancer Research Center, Seattle, lbreeden@fhcrc.org

8

Bayesian Analysis of Cell-Cycle Gene Expression Data

Chuan Zhou,
Department of Biostatistics, Vanderbilt University

Jon C. Wakefield,
Department of Statistics and Biostatistics, University of Washington

Linda L. Breeden,
Fred Hutchinson Cancer Research Center

Abstract

The study of the cell-cycle is important in order to aid in our understanding of the basic mechanisms of life, yet progress has been slow due to the complexity of the process and our lack of ability to study it at high resolution. Recent advances in microarray technology have enabled scientists to study the gene expression at the genome-scale with a manageable cost, and there has been an increasing effort to identify cell-cycle regulated genes. In this chapter, we discuss the analysis of cell-cycle gene expression data, focusing on a model-based Bayesian approaches. The majority of the models we describe can be fitted using freely available software.

8.1 Introduction

Cells reproduce by duplicating their contents and then dividing into two. The repetition of this process is called the cell cycle, and is the fundamental means by which all living creatures propagate. On the other hand, abnormal cell divisions are responsible for many diseases, most notably cancer. Therefore studying cell cycle control mechanisms and the factors essential for the process is important in order to aid in our understanding of cell replication, malignancy, and reproductive diseases that are associated with genomic instability and abnormal cell divisions.

For decades, biologists have been studying the cell cycle, using the model organism budding yeast *Saccharomyces cerevisiae*. This focus on budding yeast is due to the fact that it exists as a free living, single cell, which has the same general architecture and control pathways as the cells of its highly complex, multi-cellular relatives (e.g. humans). Moreover, a number of conditions have been identified that enable researchers to arrest yeast cells at a specific point in the cell cycle and then release them from that

state in order to follow a population of cells that are progressing through the cell cycle in synchrony. Until technologies are available to follow the molecular events in individual cells, synchronizing populations of cells is our only means to follow and characterize the key events in the cell cycle.

Duplication of a complex structure like a living cell requires the organization and coordinated activity of thousands of components. These components are built from plans coded in the genes of the cell (DNA). This code is accessed and duplicated or transcribed into RNA and then read and translated to generate the components, which are called proteins. As with any assembly process, each component is required in different amounts and at different times. One universal strategy that has evolved to simplify this process is the regulation of transcription, which means that a gene is not transcribed (and translated) until the component is needed. It is believed that up to 20% of the genes of organisms as diverse as bacteria and humans may be transcriptionally regulated during the cell cycle and many of the components encoded by these genes participate in or control specific events in the cell cycle. For reviews of cell-cycle regulation, see for example Kelly and Brown (2000) [10] and Morgan (1997) [15].

Recent breakthroughs in microarray technology have enabled biologists to measure the number of transcripts made from every gene in an organism's DNA. This microarray technology allows an unprecedented look at the state of a cell at a particular time within the cell cycle. Due to the importance of understanding the cell duplication process, studies of transcriptional regulation during the cell cycle of yeast were among the first experiments to be carried out using microarray technology. These pioneering efforts provided far more information than had been gleaned from the previous 20 years of research in the area. They also highlighted the need for computational methods for analyzing microarray data and for identifying statistically significant patterns in time series gene expression.

8.2 Previous Studies

As one of the first genome-wide gene expression studies, Cho et al. (1998) [4] used Affymetrix microarrays and visual inspection to identify 416 out of 6,000 yeast genes as cell-cycle regulated. Spellman et al. (1998) [19] conducted a set of experiments using cDNA arrays and three different synchronization methods to obtain three more data sets. By fitting these profiles to sinusoidal functions and correlating those profiles with the profiles of transcripts already known to be cell cycle regulated, these authors identified 800 genes as cell-cycle regulated. These data have further served as a testing

ground for dozens of new computational methods; the earliest among these were a number of clustering algorithms (Eisen et al., 1998 [6]; Quackenbush, 2002 [16]; Tamayo et al., 1999 [22]).

Recently, there has been increasing interest in developing model-based approaches for analyzing gene expression data. The clustering algorithms are useful exploratory tools, but they lack the ability to model the variability at various levels of the microarray experiments, the structure to take into account covariates and external information, a distributional framework for formal statistical inference, and also have difficulties with missing data. As a contrast, many of the problems associated with these *ad hoc* clustering algorithms can be overcome by assuming specific functional forms on the expression pattern or distributional assumptions on model parameters, leading to more informative analysis and principled inference. Zhao, Prentice and Breeden (2001) [27] employed a single pulse model along with generalized estimating equation techniques to re-examine the three data sets by Spellman et al. (1998). Johansson, Lindgren and Berglund (2003) [9] used a partial least squares regression approach on the three data sets individually and in combination. Lu et al. (2004) [12] used a two-component mixture-Beta model with an empirical Bayesian method to detect periodic genes. Wakefield, Zhou and Self (2003) [24] proposed a fully Bayesian hierarchical models for the analysis of cell cycle expression data, and their approach was subsequently extended by Zhou and Wakefield (2005) [28]. Other approaches using mixed-effect models and smoothing techniques have also been applied to these data, see for example, Luan and Li (2004) [13]. However, the agreement between these methods is remarkably poor. As reported in a comparison study by Lichtenberg et al. (2005) [11], in total nearly 1,800 different genes have been proposed to be periodic – which is almost one third of the *S. cerevisae* genome. These results suggest that more powerful statistical methods, more accurate data, or the incorporation of biological information are required to resolve these problems.

When applying model-based approaches to the time course gene expression data, it is important to specify the model in such a way that it captures the systematic behavior of the regulation process as much as possible, otherwise important information might be missed. The incorporation of additional information is important due to the noise inherent in these time series data sets with no replicates, and also from the difficulties in comparing and combining the results from different data sets. The four experiments reported in Spellman et al. (1998) were carried out with different synchronization methods, in the hope that analysis of the combined data would minimize the effect of artifacts due to any one synchronization method.

However, it is not clear how many periodic transcript profiles would be obscured by synchrony artifacts in any one data set, nor is it clear what other complexities would arise in combining them. In addition to the cycle lengths being different across experiments, the cycles themselves are slightly out of phase, because the points of arrest differ. Moreover, the synchrony at release is not perfect and it decays with time. An additional problem is that the arrested cells continue to grow and accumulate key cell components even during the arrest, so the first cycle after release may be shorter than the second one.

We emphasize that these experimental artifacts should be carefully considered in the analysis, as they are often systematically reflected in the expression levels throughout experiments course. Failure to recognize them may lead to unreliable results and erroneous conclusion. We have three major goals for this work: first, to extend and apply the model framework proposed in Wakefield et al. (2003) to cell-cycle time course gene expression data with the characteristics described above; second, to provide a streamlined analysis of such data including evaluation of measurement error, filtering and partitioning; third to demonstrate that with carefully specified models, we can extract important biological information from such analysis.

8.3 Data

The working data is provided by Tata Pramila and Linda Breeden at the Fred Hutchinson Cancer Research Center. It was collected from cDNA microarrays and was normalized using GenePix software (Axon Instruments, Inc.) [1]. It has the advantages of refined microarray technology compared to that obtained six years earlier and a shorter sampling interval. Microarray experiments were also performed to directly assess measurement error. The three cell cycle data sets we used monitor all yeast transcripts and each involves the same α-factor method of synchronization; α-factor was used because it is a physiological arrest of wild type cells from which cells recover rapidly. Since α-factor is a natural inhibitor of the cell cycle, we can assume that all cellular processes that might interfere with the viability or recovery of these cells from the arrest are stopped. The quality of the synchronous release can be inferred from the fact that periodic transcripts can be followed for up to four cell cycles after release from the arrest (Breeden, 1997 [3]). The timing of release is also highly reproducible, thus enabling multiple experiments to be compared.

The data was collected with the following design: cells were first synchronized by α-factor arrest; then the cells were released to progress through

the cell cycle. Gene expression levels through the cell cycle relative to asynchronized cell samples were measured at 5 minutes time intervals from $t = 0$ to $t = 120$ minutes. This length covers approximately two full yeast cell cycles. The 5 minute intervals offer finer resolution in time compared to those of Spellman et al. (1998) and Cho et al. (1998). Two microarrays were performed with the RNA collected from this experiment. In the first case (referred to as 38wt), the cell cycle transcripts were labeled with red dye, and the reference transcripts from asynchronous cells were labeled with green dye. A second microarray (30wt) was then performed with the dyes swapped. This dye-swapped data set is treated as a replicated experiment. The duplicated experiment provided valuable additional information regarding the variability and magnitude of the expression patterns.

Another important data set consists of six arrays with expression measures of all transcripts relative to themselves to give a so-called self-self hybridization. Deviations from a ratio of 1 in these measurements indicate measurement error. Using the fully Bayesian model-based approach, we were able to incorporate the additional information gathered from these data into our main analysis, using informative prior distributions.

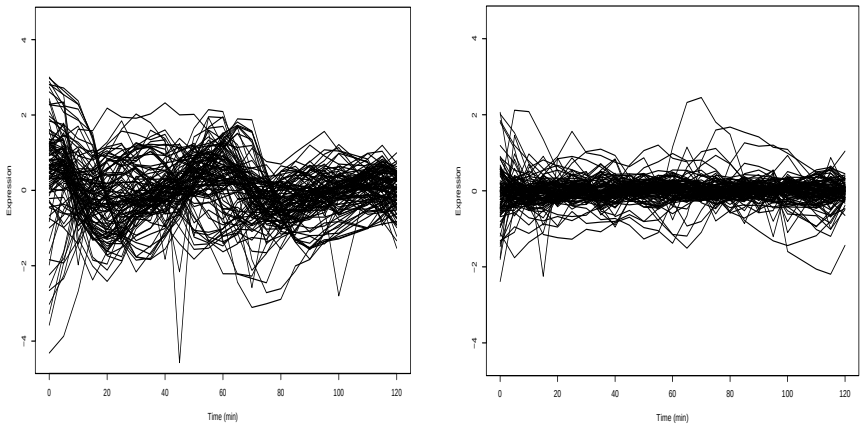
All three data sets use the same 6216 yeast transcripts, which cover the complete yeast genome. An initial exploratory analysis, which was confirmed by closer examination, revealed that the mRNA sample at 105 minutes was contaminated, therefore the data generated from that array were dropped from subsequent analyses.

The left panel in Figure 8.1 shows expression of 100 genes which are known to be cell cycle regulated (CCR) from previous studies. It appears that they do demonstrate strong cyclic signals in our data set. As a contrast, a large portion of the genes do not show strong signals as we see from the random sample of 100 genes shown in the right panel of Figure 8.1.

8.4 Bayesian Analysis of Cell Cycle Data

8.4.1 Measurement Error

There are various sources of variation involved in microarray experiments, and their identification and evaluation have proven to be crucial for making accurate inference. Other than variations which we can attribute to certain systematic sources, the remaining variability is often referred to as measurement error. To estimate measurement error, we use data from six microarrays with mRNAs collected at 0, 25, 35, 45, 60 and 100 minutes. These mRNAs were copied into cDNA, split and then coupled to either Cy5 or Cy3 dyes. The two samples were mixed and hybridized to cDNA arrays.



(a) 100 CCR genes (b) 100 random genes

Fig. 8.1. Expression of 100 CCR genes and 100 randomly selected genes.

Fluorescence measured from each dye is expressed as a ratio and its deviation from unity provides an estimate of measurement error. This is often referred to as a same versus same measurement.

We now summarize the analysis of this reference data, based on which the prior distribution on the measurement error was specified. Figure 8.2 shows the boxplots of the data from these 6 chips; we can see that the average gene expressions of these asynchronized samples are close to zero. There were genes which exhibited large variations across time, but they did not appear to be cyclic under closer inspection. The samples appear to be more spread out at later times, suggesting that measurement error may increase with time. This observation supports our speculation that using only early data could under-estimate the measurement error. Therefore we proceeded to carry out a Bayesian analysis using the pooled data from all six chips.

Let $\boldsymbol{y} = \{y_1, \cdots, y_N\}$ denote the *pooled* reference data, y_i denote the ith observation. We assume a simple normal model for the data

$$y_i \,|\, \mu, \sigma^2 \sim_{i.i.d.} \mathrm{N}(y_i \,|\, \mu, \sigma^2). \tag{8.1}$$

We assume a "non-informative" prior on (μ, σ^2) with $p(\mu, \sigma^2) \propto 1/\sigma^2$, which leads to the following posterior distribution

$$p(\sigma^{-2} \,|\, \boldsymbol{y}) = \mathrm{Ga}(\sigma^{-2} \,|\, a, b), \tag{8.2}$$

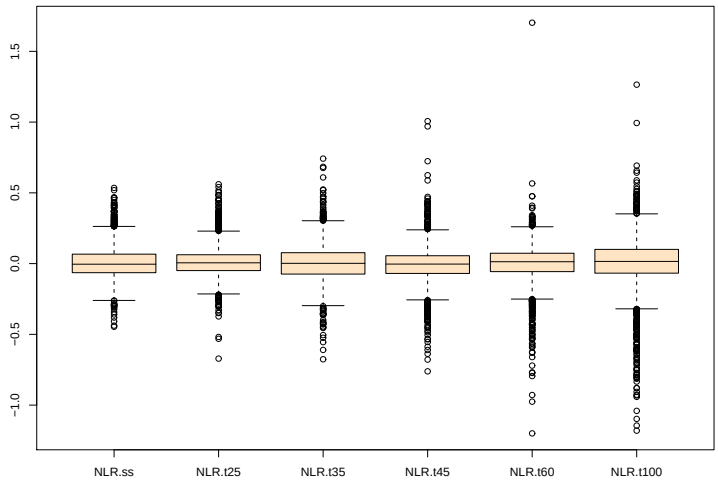


Fig. 8.2. Boxplots for the data from each of the six chips.

where $a = \frac{1}{2}(N-1)$, $b = \frac{1}{2}ns^2$ with $ns^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

We could use the parameter values from this posterior analysis as a way of obtaining a prior specification for later analyses, but the large sample size from pooling the six chips leads to a highly concentrated posterior distribution on the standard deviation σ. The sampling posterior median of σ is 0.151, with 95% sampling interval $(0.150, 0.153)$. To avoid being too restrictive, we calibrated a and b to allow larger variation. We set the modal value for σ to be 0.15, and an upper bound 0.5 so that $\Pr(0 < \sigma < 0.5) = 0.95$. Solving the resultant equations gave $a = 1.52$ and $b = 0.05$, under which the 95% sampling interval is $(0.10, 0.68)$. These values were then used as priors in subsequent filtering and partitioning analysis.

8.4.2 Filtering

In cell cycle analysis, our main interest lies in identifying and characterizing genes that are cell-cycle regulated. For those genes which show differential expression but do not coincide with cell cycle events, we do not consider them as cell cycle regulated, and consequently exclude them from later analysis. In this section, we apply a filtering procedure to cell cycle data. The aim is to first identify candidate periodic genes, then perform more reliable analysis

on these candidates, using a more sophisticated model tuned to the cell-cycle nature of the data.

Let y_{ij} denote gene expression at time t_j for gene i, $i = 1, \cdots, n$, $j = 1, \cdots, T$. We assume a first order Fourier model for the data,

$$y_{ij} = R_i \cos 2\pi(f_0 t_j + \phi_i) + \epsilon_{ij}, \tag{8.3}$$

where $\epsilon_{ij} \sim_{i.i.d.} \mathrm{N}(0, \sigma_e^2)$ are the measurement errors, and (R_i, ϕ_i) are gene specific parameters, R_i is the amplitude, i.e., the magnitude of the cyclic signal, and ϕ_i is the phase, governing where the signal peak. The cell cycle frequency is denoted by f_0, fixed at $1/58$ minutes^{-1}, and assumed to be common to all genes. The cell cycle span is estimated to be 58 minutes using the known CCR genes [27].

For the purpose of filtering, we want to test the following hypothesis independently for each gene i,

$$M_0 : R_i = 0 \ v.s. \ M_1 : R_i \neq 0.$$

To carry out the filtering procedure, we need to specify the prior distributions. For measurement error, we assume $\sigma^{-2} \sim \mathrm{Ga}(a, b)$, where a and b are determined from the reference data analysis described in Section 8.4.1.

We assume models M_0 and M_1 are equally probable *a priori*. Under M_0, the parameter ϕ_i is redundant. Under M_1, we assume R_i and ϕ_i are independent with the following prior distributions,

$$R_i \sim_{i.i.d.} \mathrm{Exp}(\lambda) \tag{8.4}$$

$$\phi_i \sim_{i.i.d.} \mathrm{Unif}(-0.5, 0.5) \tag{8.5}$$

Because the trigonometric functions in the Fourier model are periodic, ϕ_i is restricted to $(-0.5, 0.5)$ for identifiability, so the uniform prior on ϕ_i is "non-informative". We chose an exponential prior on the amplitude R_i because it has a simple form and reasonably reflects prior belief based on data. The parameter λ was based upon an exploratory analysis of the 100 known CCR genes. We have found that the 100 known cell cycles genes showed consistently strong signals in both the main and the dye-swapping experiments, and believed their expression levels were representative of genes with strong signals. So we extracted data for the 100 known cell cycle regulated genes from the dye-swapping experiment, transformed them into the same format as the 38wt data set by changing the signs of the log ratios. Model (8.3) can be re-parameterized as

$$y_{ij} = A_i \cos(2\pi f_0 t_j) + B_i \sin(2\pi f_0 t_j) + \epsilon_{ij}, \tag{8.6}$$

with $A_i = R_i \cos 2\pi\phi_i$, and $B_i = R_i \sin 2\pi\phi$. Given f_0 and t_j, it is just

a simple linear model, for which we can obtain least squares estimates of (A_i, B_i) and transform them back to (R_i, ϕ_i). We chose λ to be 1.43 so that the mean amplitude is 0.7 with variance 0.5 on the basis of the least square estimates. We believe that amplitudes of these known CCR genes are within the upper range of the signals, we would expect many CCR genes to have smaller amplitude than these genes. Figure 8.3 shows the expression of the 100 CCR genes, with fitted curves based on the least squares estimates. The distributional and independence assumptions were checked by inspecting the histograms and scatter plots of the parameter estimates.

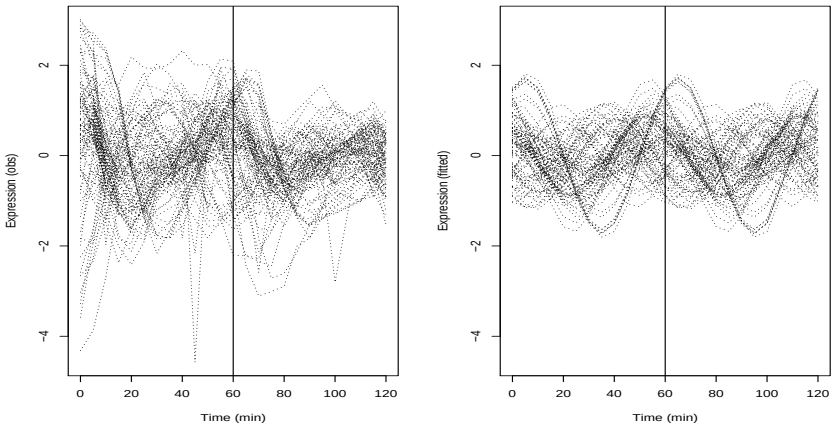


Fig. 8.3. Observed gene expression of 100 known CCR genes, and their fitted values based on least square estimates using model (8.3).

Figure 8.4 shows 100 simulated gene expression time series from the above priors including measurement error. It suggests our prior choices are reasonable, as we see patterns in the simulated data match quite closely to what we see in the main data (as seen in Figure 8.1).

We sampled parameter values from the prior distributions and used importance sampling technique to estimate the posterior probabilities $p_i = \Pr(M_1 \mid \boldsymbol{y}_i)$, then ordered genes based on these probabilities. Figure 8.5 displays the 100 highest ranked genes and the 100 lowest ranked genes. It appears that the filter was able to pick out genes with large variations. Because the model (8.3) allows cyclic oscillation in the data, genes showing cyclic patterns tend to be ranked higher than genes that are not cyclic even though they may show differential expression. So the higher a gene is ranked by this filtering procedure, the more likely it is cyclic and thus a candidate for cell cycle regulation.

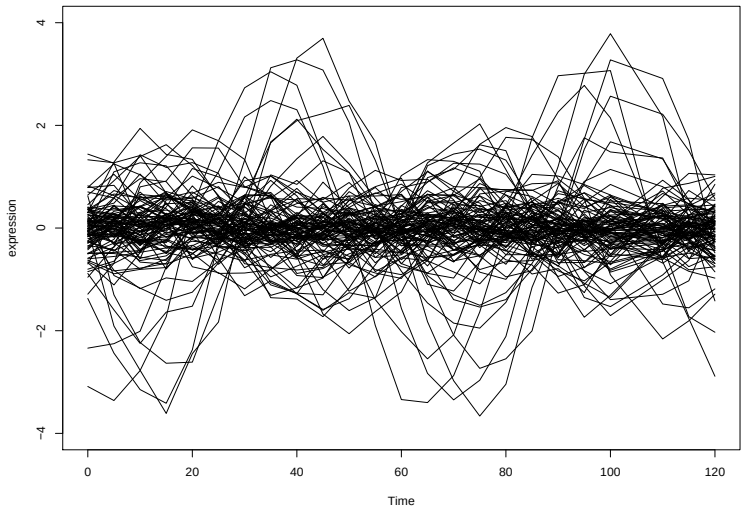


Fig. 8.4. $N = 100$ simulated gene expression time series based on the following priors: $R_i \sim \text{Exp}(1.43), \phi_i \sim \text{Unif}(-0.5, 0.5), \sigma_e^2 = 0.2^2$.

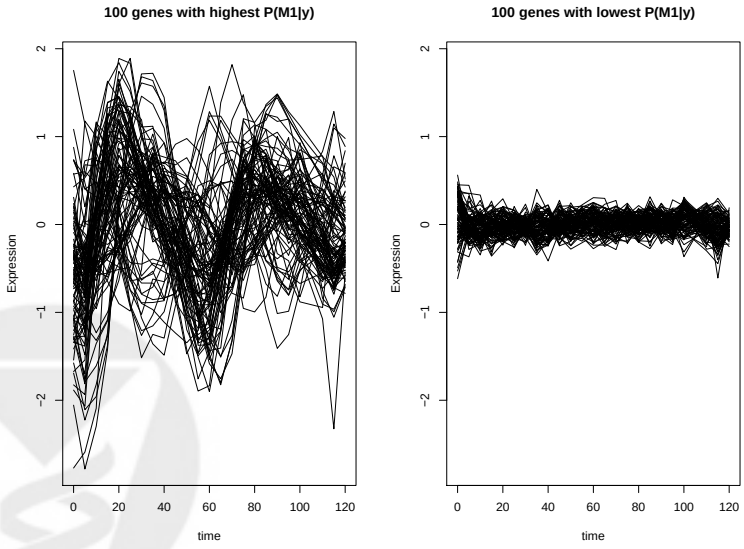


Fig. 8.5. Expression of the 100 highest ranked genes (left panel) and lowest ranked genes (right panel).

At this point, we can either pick a cutoff point subjectively, and proceed with genes above the threshold, or we can choose the cutoff point based

on some more formal criteria, such as controlling the false discovery rate (FDR) and false negative rate (FNR). The concepts of FDR and FNR, and Bayesian procedures for controlling them have been discussed in Storey (2002) [21]. Note FDR and FNR are two competing concepts, optimal results for minimizing both error rates cannot be achieved at the same time. We would miss nothing by rejecting all hypotheses and concluding all genes are cell-cycle regulated, so FNR=0, but clearly FDR would be high in this case, and vice versa. Therefore some compromise has to be made, depending on the scientific question and our subsequent preference for making the two types of errors. In our analysis, we feel we are in a "discovery" mode, and therefore a certain amount of false discovery is tolerable as long as we do not miss too many cell-cycle genes.

Figure 8.6 illustrates various thresholds from minimizing the loss function $cFDR+FNR$, where c is a positive number chosen to reflect our preference in controlling FDR and FNR. For example, if we are twice as concerned with FDR as with FNR, we could set $c = 2$ and consider the top 1340 genes (bottom left panel). Of course, choosing an appropriate value c is not a trivial task.

As an alternative Figure 8.7 shows the optimal number of rejections for minimizing Bayesian FNR while controlling Bayesian FDR at the 0.05 level. This is similar to the frequentist practice of maximizing the power while controlling the significance level. Based upon this result, we decided to identify the top 1680 genes as candidates for cell cycle regulation, and the cutoff for marginal posterior probability $\Pr(M_1 \mid \boldsymbol{y}_i)$ was set to be 0.78†. More sophisticated Bayesian methods for differential gene expression have been proposed, see for example Do, Müller and Tang (2005) [5].

8.4.3 Model-based Partitioning

This first-order Fourier model requires model refinement since it does not account for the attenuation in the cell-cycle data. This synchronization causes an intrinsic difficulty in a cell-cycle study. To effectively observe the cell-cycles, yeast cells have to be initially synchronized. In addition, our ability to observe the true cell-cycle span is impeded because the cell-cycle can be altered by the synchronization. This fact has long been recognized by biologists, and has been addressed in gene expression analyses as well (Lu et al., 2004 [12]). α-factor synchronization is considered as a better choice compared to other synchronization methods because of its relative ease, sensitivity and gentleness to cells. α-factor is a mating pheromone

† The discrepancy of 5 is due to the rounding error in 0.78.

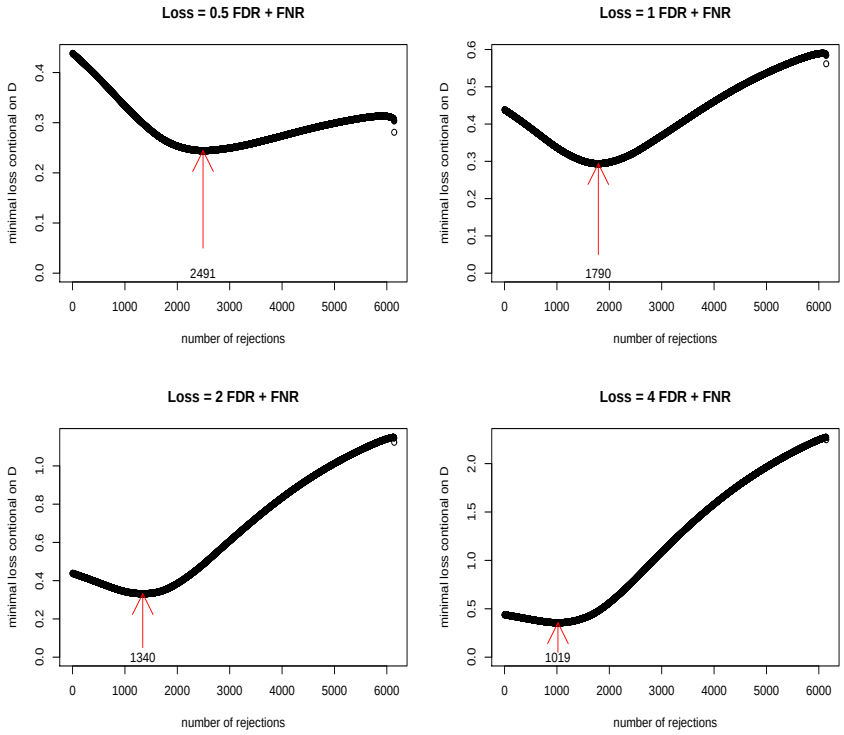


Fig. 8.6. Optimal solutions to different loss functions in the form of $cFDR+FNR$.

that is secreted by haploid *S. cerevisiae* cells of the α mating type. It blocks cell division in G1 and induces mating-specific gene expression. Even when transcriptions are held at START‡, during this time cell mass increases and cell wall growth continues, resulting in enlarged and frequently distorted cells. After the release the large size of cells leads to near elimination of the G1 phase and hence an abbreviated cell cycle. This is consistent with our observation that there tends to be shortened cell-cycle span early on after release, but the difference decreases over time. Breeden (1997) recommends that with α-factor arrest, the first cycle after release should be considered a recovery cycle, which may differ from the normal mitotic cycle in specific ways. Any oscillating activity that persists through the second and third cycles after recovery is most likely to be a property of the normal mitotic cell cycle.

There are drug-induced cell cycle arrests, which are unnatural and poten-

‡ An important checkpoint in the eukaryotic cell cycle. Passage through START commits the cells to enter S phase.

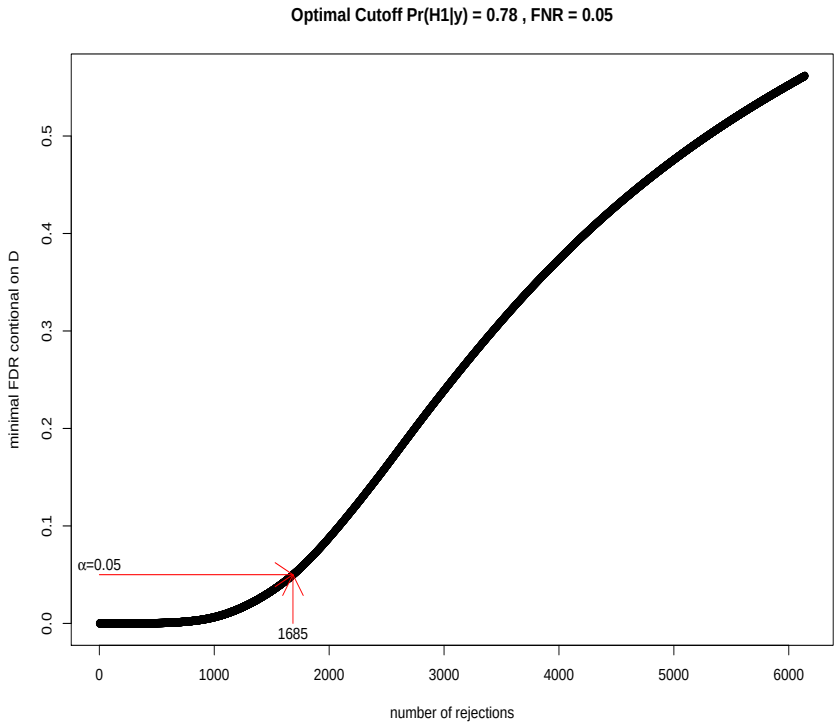


Fig. 8.7. Optimal solutions to minimizing FNR, subject to $FDR \leq 0.05$.

tially toxic and non-specific. Genetically induced arrests using *cdc* mutants are more specific and two such arrests (*cdc28, cdc15*) have been used by Spellman et al. (1998). However, the arrests evoked by these mutations are abnormal in the sense that they are caused by the loss of a critical gene product. The cells arrest in an apparently uniform state, but it cannot be assumed that all cell-cycle specific progresses are halted, or that recovery from the arrest occurs under balanced growth conditions. Even with the elutriation synchronization, which collects G1 cells based on size and introduces minimal perturbation, cells need some time before they resume normal mitotic cell cycles. With these synchronization methods, the first cell cycle should also be considered a recovery cycle, as with α-factor synchronization.

So if the first cycle cannot be trusted, why not run the experiments longer and only look at later cell cycles? This brings up a second point: the number of observable cell cycles is limited. Most of the time the cyclic signals dissipate after three or four cycles. There are several factors that could contribute to this phenomenon. One is how well the cells are synchronized. But

even with a perfect synchrony, after two doublings only one of four of the cells experienced the initial conditions. This, in addition to random fluctuation in the transcription of each gene, means that soon the cells become asynchronous and we are unable to observe the cyclic patterns any more.

To make the matter even more complicated, certain signals we observe could be artifacts of the synchronization. Even with a perfect release from the arrest, this budding yeast divides asymmetrically yielding a new daughter cell that is smaller than the mother cell. This daughter cell must grow during the next G1 before it can enter S phase. The mother cell has no growth requirement and as a result has a shorter G1 interval. This asymmetry precludes perfect synchronization. For example, in the case of α-factor synchrony, because α-factor is a mating pheromone, it will induce mating-specific gene expression. As a consequence, many mating-related genes will either be induced or repressed, leading to increased or decreased transcript levels. In some extreme cases, the changes in expression level are so dramatic that the cyclic signals are totally obscured.

In the following we extend the first-order Fourier model to allow variable frequency and time-dependent amplitude. Let y_{ij} denote the expression level of gene i measured at time t_j, and let $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{iT_i})$ denote the expression profile for gene i measured across T_i time points, so that genes are allowed to be measured at different sets of time points or have missing values under our model.

- **Stage 1**: We assume each observed gene expression profile follow a multivariate normal distribution,

$$\boldsymbol{y}_i \,|\, \boldsymbol{\theta}_i, \boldsymbol{S}_i \sim \mathrm{N}_{T_i}(\boldsymbol{\theta}_i, \boldsymbol{S}_i), \tag{8.7}$$

  where $\boldsymbol{\theta}_i$ is the $T_i \times 1$ mean vector, and $\boldsymbol{S}_i$ is the $T_i \times T_i$ covariance matrix, for $i = 1, \cdots, n$.
- **Stage 2**: We introduce partition label z_i, which indicates the partition that gene i belongs to. We assume the mean vector is a context specific function of covariates $\boldsymbol{X}_i$ and partition specific parameter vector $\boldsymbol{\mu}_k$, with $\boldsymbol{\theta}_i = h(\boldsymbol{X}_i, \boldsymbol{\mu}_k)$ if $z_i = k$. For the cell cycle data, the covariate is time, and the mean structure has the form

$$h(t_j, \boldsymbol{\mu}_k) = e^{-\gamma_k t_j} \left\{ A_k \cos[2\pi f_{t_j}(\phi_k) t_j] + B_k \sin[2\pi f_{t_j}(\phi_k) t_j] \right\}, \tag{8.8}$$

  where $f_{t_j}(\phi_k) = f_0(\frac{t_j}{t_{max}})^{\phi_k}$, with $\boldsymbol{\mu}_k = (A_k, B_k, \gamma_k, \phi_k)$ characterizing the mean trajectory, parameters A_k and B_k account for the amplitude and phase of the cyclic pattern, γ_k accounts for the attenuation in the amplitude, and ϕ_k is a time stretching factor for varying cell-cycle length.

We assume the covariance matrix is also characterized by partition specific parameter(s) so that $\boldsymbol{S}_i = \boldsymbol{S}(\boldsymbol{\xi}_k)$ if $z_i = k$. If $T_i = T$ for all i, and there is no restriction on the covariance structure, we can assume $\boldsymbol{S}_i = \boldsymbol{\Sigma}_k$, e.g., $\sigma_k^2 \boldsymbol{I}$ given $z_i = k$.

- **Stage 3**: We assume the partition label z_i's are independent and identically distributed, conditional on the total number of partitions K and mixing proportion $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$,

$$\Pr(z_1, \cdots, z_n) = \prod_{i=1}^{n} \Pr(z_i \,|\, K, \boldsymbol{\pi}), \tag{8.9}$$

with

$$\Pr(z_i = k \,|\, K, \boldsymbol{\pi}) = \pi_k, \tag{8.10}$$

for $k = 1, \cdots, K$, and $i = 1, \cdots, n$.

- **Stage 4**: At this stage, we specify the prior distributions for the partition specific parameters. Assume

$$\boldsymbol{\mu}_k \,|\, K, \boldsymbol{m}, \boldsymbol{V} \sim_{i.i.d.} \mathrm{N}_q(\boldsymbol{m}, \boldsymbol{V}), \tag{8.11}$$

$$\boldsymbol{\Sigma}_k^{-1} \,|\, K, g, \boldsymbol{R} \sim_{i.i.d.} \mathrm{Wishart}(g, (g\boldsymbol{R})^{-1}), \tag{8.12}$$

$$\boldsymbol{\pi} \,|\, K, \boldsymbol{\delta} \sim \mathrm{Dirichlet}(\boldsymbol{\delta}), \tag{8.13}$$

with priors on $\{\boldsymbol{\xi}_k\}$ if they are present in the model. We also include a "zero" partition with $A_k = B_k = 0$. Genes showing no cyclic pattern will be included in this partition.

- **Stage 5**: The hierarchy is completed with specification of prior constants and hyper-priors. Throughout the analysis, we choose $\boldsymbol{\delta}$ to be a K-vector of 1's for the Dirichlet prior. We assume the total number of partitions K follows a Poisson distribution with parameter λ if it is considered unknown. We choose $g = p$, the dimension of $\boldsymbol{\Sigma}_k$, for it is the least informative in the sense that the distribution is the flattest while being proper (Wakefield, et al., 1994 [25]).

When K is known, this hierarchical model has a partitioning-by-features interpretation, and posterior computations can be carried out using standard Markov chain Monte Carlo (MCMC) software such as WinBUGS (Spiegelhalter et al.,2002 [18]). When K is unknown, it can be treated as a random variable and inferred from the data. More sophisticated techniques such as reversible-jump MCMC (Richardson and Green, 1997 [17]) or birth-death MCMC (Stephens, 2000 [20]) are required to deal with the changing dimension. For more details on computation, see Wakefield, et al. (2003) and Zhou and Wakefield (2005).

8.4.4 Results

We now report the results from applying our enhanced hierarchical mixture model to the cell-cycle expression data.

Among all 6309 genes (including controls) on each of the 24 microarrays ($t = 105$ was dropped due to mRNA contamination), 6141 had no missing data across all chips, 75 had one missing value, 25 had two missing values, and 68 had three or more. A close inspection reveals genes with many missing values tend to be highly unreliable thus genes with three or more missing values were dropped. Some of the measurements were flagged as unreliable at the data processing stage, we still decided to include them in subsequent analysis because of the ad hoc nature of flagging.

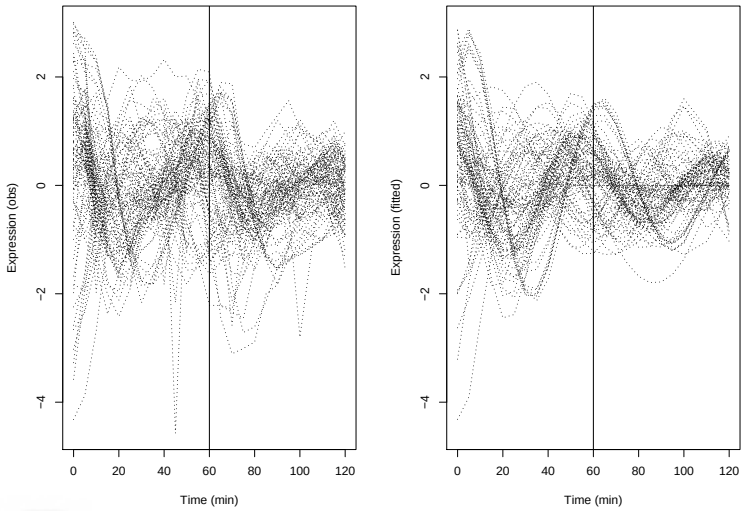


Fig. 8.8. Observed expression of the 100 known cell-cycle regulated genes, and their fitted values based on non-linear least squares estimates using Model (8.8).

We first identified 1680 genes as candidates for cell cycle regulation using the filter described in Section 8.4.2. Next we evaluate the extension to the mean structure. Figure 8.8 shows the observed curves and the fitted curves based on non-linear least square estimates from Model (8.8). Compared to Figure 8.3, the improvement in the attenuation adjustment and time stretching is clear.

We have found that the number of partitions K is highly sensitive to the prior specification, not only the Poisson prior, but also other priors on the variance parameters which could affect the size and shape of partitions. This

is in agreement with Stephens (2000) [20]. In addition, our enhanced model allows genes to be classified at a finer scale (with more features), which led to a large number of partitions. Given there is no clear definition regarding the underlying regulation pathways during the cell cycle, we found this number hard to interpret and highly variable depending on the prior choices, so we decided to restrict our attention to the analyses with K fixed. Figure 8.9 displays the classification and estimated mean profiles from fitting the enhanced model to the 38wt data with K fixed at 16. There is an inherit unidentifiability problem with Bayesian mixture modeling so that re-labeling needs to be carried out, see Stephens (2000) for discussion. Here we re-labelled the partitions on the basis of time to the first peak. This decision is based on the fact that the cell cycle events are regulated in an orderly fashion, the early activation or deactivation of transcription factors are often responsible for the next wave of gene expression, so this re-labelling has an appealing biological interpretation.

Our model was able to identify some interesting cell-cycle gene partitions, and the effect of model enhancement is obvious. From Figure 8.9, we can see that partitions 3, 6, 8, 13 and 16 are partitions with strong cyclic signals, and they all show the dissipation of synchrony over time. In particular, partition 3 has a greatly heightened first peak, which is large enough to obscure the later cyclic pattern. Without the improvement to the model, we may not be able to identify this group of genes. We suspect these genes are related to the mating process, so their expression is induced by the pheromone. Several partitions appear to have shortened first cycles, such as partition 2, 3 and 11. These are G_1 or G_2 phase genes, confirming our speculation that the synchronization may shorten the growth phase. At least 9 out of the 13 genes classified into partition 8 are the S-phase histone coding genes. The products of these genes form a single complex that is used for DNA condensation. These genes are coordinately regulated and have been well characterized. A closer inspection reveals that many genes in partition 2 are M–G_1 genes and share a promoter element called ECB; many genes in partition 5 and 6 are late G_1 genes and share MCB and/or SCB promoter elements; partition 9 consists of G_2-phase genes and many of them also share the MCB/SCB promoter elements; and many genes in partition 13 appear to share MCM1 and FKH sites. Partition 11 contains many genes involved in ribosome biogenesis. Their promoters are enriched for two sequence motifs referred to as PAC and RRPE (Wade et al., 2001 [23]; Hughes et al., 2000 [8]). Our data indicate that these transcripts are modestly periodic and peak ten minutes after the histones peak.

Note that the time to first peak in partition 16 is larger than 58 minutes,

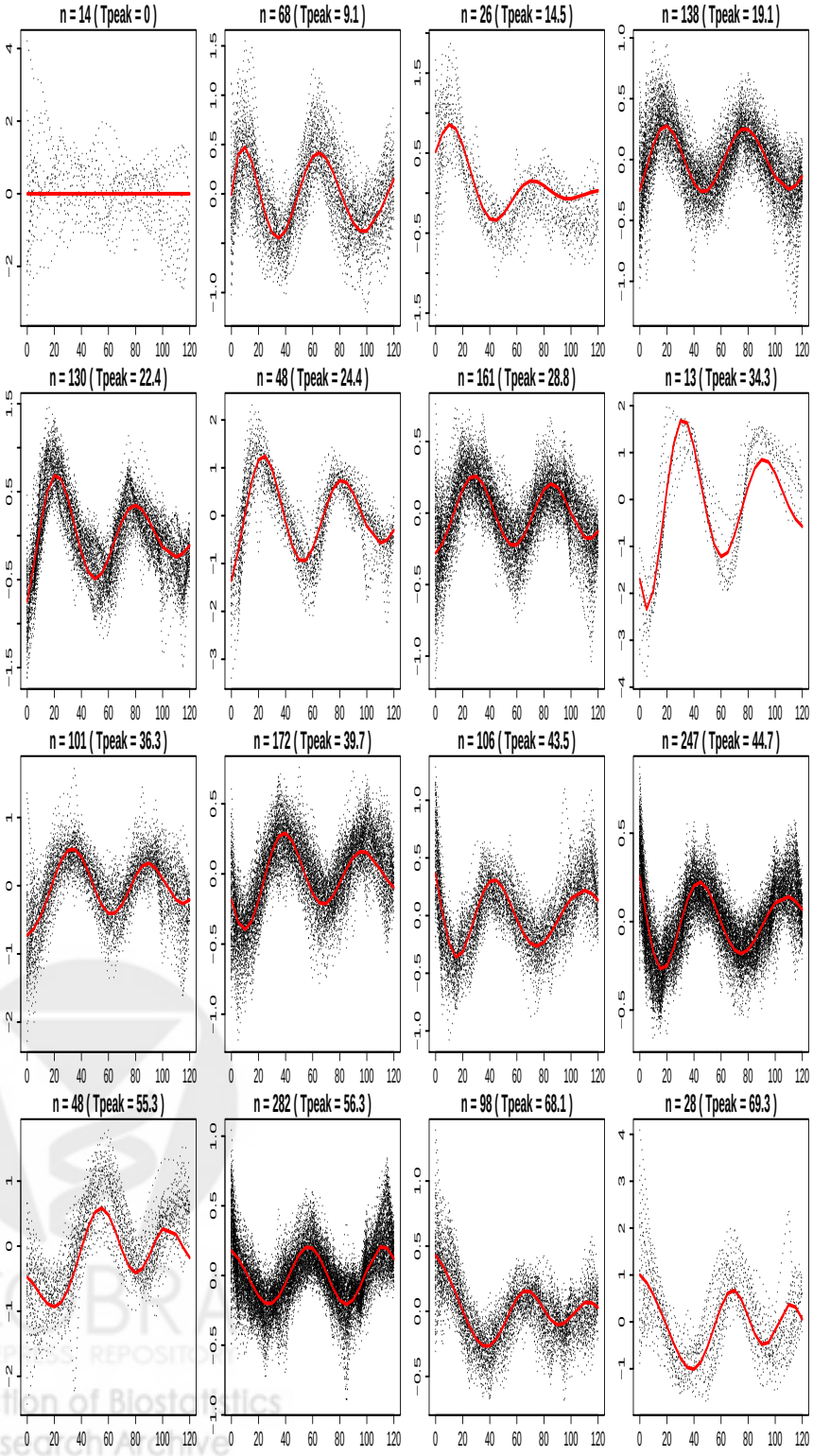
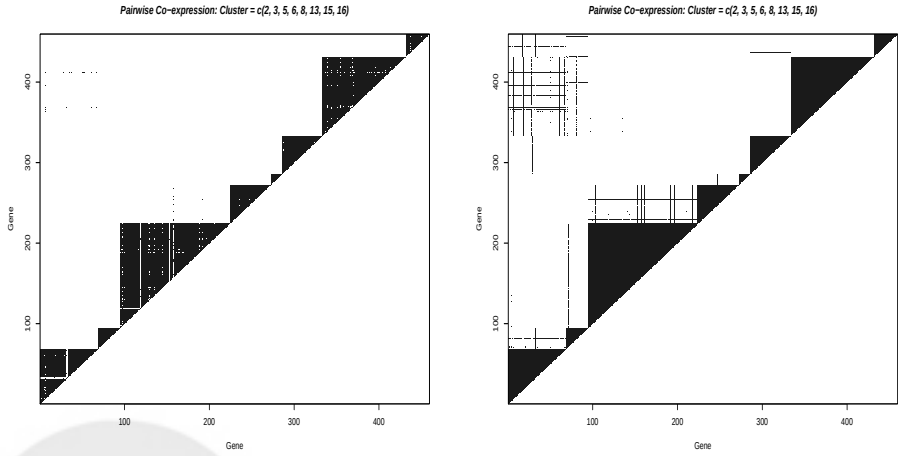


Fig. 8.9. Final partitioning with $K = 16$ fixed, note different vertical scales.

the normal cell cycle span we used. This is because the attenuation at the beginning of the experiment is so large that the first peak of this partition is obscured. If we shift the time to peak by 58 minutes, we can see that this group actually coincide with partition 2, except with much larger amplitude.

Under the Bayesian mixture models, specific partitions are susceptible to the re-labelling problem. But as suggested in Wakefield et al. (2003) [24], we can examine the probabilities of *co-expression* $p(z_i = z_{i'} \mid \boldsymbol{y})$, which are invariant to re-labelling. A good visual display of co-expression is the heat-map. Due to space limitation, we select a sub-sample of the partitions to display. Figure 8.10 shows the co-expression, with dark areas indicating high co-expression, and as expected, shaded areas are close to the diagonal, suggesting strong co-expression within partitions. There is some overlap between partition 1 and 2, which is not surprising given our previous discussion.



(a) Cutoff= 0.5 (b) Cutoff= 0.1

Fig. 8.10. Heat-map of probabilities that two genes share a common label, for partitions 2, 3, 5, 6, 8, 13, 15, and 16. Shaded blocks correspond to pairwise probabilities larger than the chosen cutoff.

The posterior classification probability of each gene $p(z_i = k \mid \boldsymbol{y})$ provides a natural measure of uncertainty concerning the partitioning of each individual gene. However, it is also of interest to measure the strength of the partitions, such as how tight genes are *within* a partition, and how much overlap there is *between* different partitions. So we examine the sensitivity and specificity

of the partitions where, *sensitivity* is the probability of co-expression, given labelling in the same partition, and *specificity* is the probability of non-co-expression, given labelling in different partitions. Such functions cannot be evaluated with traditional partitioning approaches.

The sensitivity of partition k is estimated by

$$\text{sensitivity} = \sum_{i,i' \in C_k} p(z_i = z_{i'} = k \,|\, \boldsymbol{y})/N_{k1}, \tag{8.14}$$

where C_k denotes partition k, and N_{k1} is the number of distinct gene pairs classified into C_k. The specificity of partition k is estimated by

$$\text{specificity} = \sum_{i \in C_k, i' \in C_{k'}, k' \neq k} p(z_i = k, z_{i'} = k' \,|\, \boldsymbol{y})/N_{k2}, \tag{8.15}$$

where C_k and $C_{k'}$ are different partitions, and N_{k2} is the number of distinct gene pairs with only one gene classified into C_k. The sensitivity and specificity of the 16 partitions are shown in Figure 8.11. Partition 1 is the "zero" partition for non-cyclic genes, so it is not surprising to see it has the lowest sensitivity. Partitions 11 and 12 only have weak signals and there is overlap between genes in these partitions, hence their sensitivity and specificity are low. Partition 8 contains a tight group of histone genes which have strong cyclic signals, and it is ranked the highest in terms of sensitivity and specificity. Other high quality partitions include partitions 3, 6, 13 and 16, as was evident from Figure 8.9. The sensitivity and specificity estimates provide a natural quantitative measure of the quality of partitions, based on which we can focus on the high quality partitions, and proceed with validation or more sophisticated analysis such as motif discovery.

Studying the co-expression can also provide important information about relationship between partitions. For example, Figure 8.12 shows several genes identified from the heat-map which had high co-expression with genes in partition 16 though they were classified into partition 2. Examination of the mean trajectories reveals the peaks of one trajectory appear to coincide with the other, suggesting these two partitions could be co-regulated, although the magnitude of the signals differs. Some would argue that these genes should be considered co-regulated as long as the peaks and troughs of their oscillations concur, regardless of their magnitude. Here we distinguish these genes, for we speculate that genes with higher amplitude may contain more promoting elements, or some other element(s) responsible for increased expression levels, or the low amplitude profiles may be from genes with unstable mRNAs. In fact, a sequence search reveals that partition 16 and partition 2 do share common MCM1 elements. The relevant motif

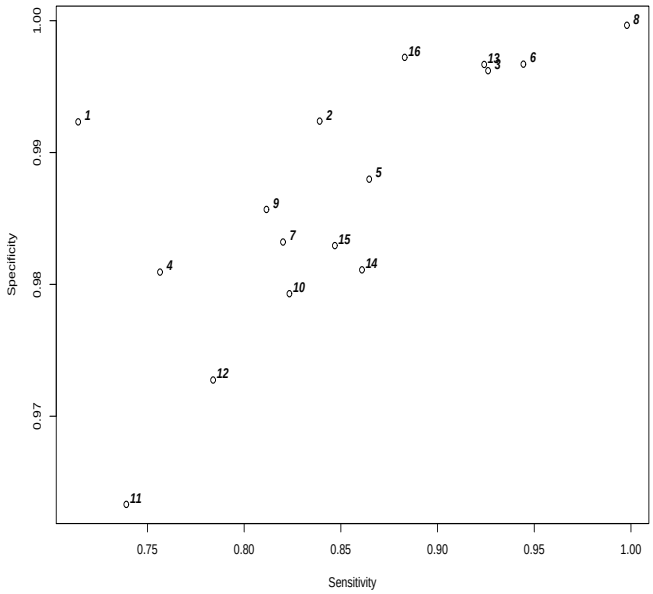


Fig. 8.11. Strength of co-expression within and between partitions, measured through sensitivity and specificity.

is `TTTCCNNNNNNGGAAA`, a flanking palindrome to which two MCM1 proteins bind (`N`=`A` or `C` or `G` or `T`). Such binding is required for transcriptional activation at the M/G_1 boundary. And as we thought, the partition 16 genes have multiple elements and a larger consensus sequence, and the partition 2 genes have only one site. Many partition 2 genes do not have the MCM1 site at all. This causes us to suspect that there may be new element(s) in partition 2 genes which have similar properties as MCM1. We will continue investigation of these speculatives.

8.5 Discussion

As explained above, the changing cell-cycle span and magnitude of signals are systematic and correspond to actual biological phenomena. Although a large number of research papers have been published on the topic of cell-cycle gene expression, few have taken these systematic variations into account. Zhao et al. (2001) [27] considered the issue of decreasing signals in their single-pulse model (SPM), in which they allowed the precision to decrease over time. Bar-Joseph (2002) [2] mentioned both issues, but used semi-

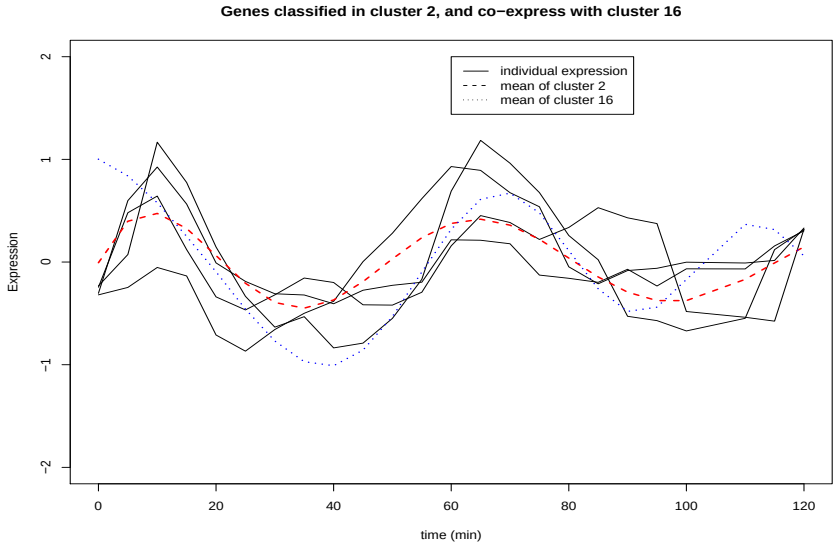


Fig. 8.12. Selected genes partitioned into group 2, but with co-expression with partition 16.

parametric models instead of directly modeling the phenomena. Here we advocate a science motivated, model-based approach towards cell-cycle gene expression analysis. We believe that it is less appropriate to rely totally on data-driven approaches, regardless of the biological context and scientific questions waiting to be addressed.

Because every synchronization protocol has its limitations, a prudent strategy for determining if a specific process is cell cycle regulated is to employ at least two different synchrony methods. If the oscillation can be observed through two or more mitotic cycles in two different synchrony experiments, it is unlikely the oscillation is induced by the arrest (Breeden, 1997 [3]). But combining analyses from different experiments is a difficult task, and has not been fully addressed by researchers. We leave it as future research, and do not attempt this problem here.

Our approach of assuming a mixture model with flexible mean structures is crucially different from the "model-based" clustering approach of [26], who analyzed similar data but simply assumed that the data arose from a mixture of T-dimensional normal distributions and hence did not acknowledge the time-ordering of the data (the analysis would be unchanged if the time ordering were permuted). In particular it would be desirable to allow serial dependence, within such an approach, but the MCLUST software [7] that

is used by Yeung et al. (2001) [26] does not allow for this possibility, and it does not perform well when the dimension T gets large. In their approach, missing data and unbalanced design also cause complications whereas in our model no such problems arise. Medvedovic and Sivaganesan (2002) [14] also proposed a Bayesian hierarchical model for clustering microarray data, but again they failed to take the time ordering into account in their approach.

We have demonstrated that our enhanced model can provide further insight into our understanding of cell-cycle transcription programs. In our enhanced model, each partition is characterized by a set of four parameters. Intuitively speaking, the finer we characterize the mean model, the easier to distinguish different features and we see more partitions. So we were not surprised to find that a large number of partitions were being identified under our refined model. Although many numerical methods for detecting underlying clusters based on gene expression data have been published, none of them are satisfactory. From our experience we have found that without plausible interpretation and biological validation, the number of partitions produced by numerical analysis is highly unreliable, and sometimes even misleading. The partitions are defined by the model, which in turn is motivated by the biology. The ultimate validation of the partitioning should be based on scientific investigation, with data analysis providing numerical support and further hypotheses. In other words, the conclusion should be based on science, not just on data analysis.

References

Axon Instruments, Inc. (2005) GenePix Pro. 6.0 User's Guide & Tutorial – Microarray Acquisition and Analysis Software for GenePix Microarray Scanners. *Axon Instruments, Inc.*

Bar-Joseph, Z., Gerber, G., Gifford, D. K., and Jaakkola, T. S. (2002). A new approach to analyzing gene expression time series data. *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology*, 39–48.

Breeden, L. L. (1997). α-factor synchronization of budding yeast. *Methods in Enzymology*, **283**:332–341.

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

Do, K., Müller, P. and Tang, F. (2005) A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society C*, **54**, 627–644.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14863–14868.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *Computer Journal*, **41**, 578–588.

Jason D. Hughes, Preston W. Estep, Saeed Tavazoie, George M. Church (2000).

Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology*, **296**, 1205–1214.

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

Kelly, T. J. and Brown, G. W. (2000). Regulation of chromosome replication. *Annu. Rev. Biochem.*, **69**, 829–880.

de Lichtenberg, U., Jensen, J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, **21**, 1164–1171.

Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. (2004). Statistical resychronization and bayesian detection of periodically expressed genes. *Nucleic Acids Research*, **32**, 447–455.

Luan, Y. and Li, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, **20**, 332–339.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, **18**, 1194–1206.

Morgan, D. O. (1997). Cyclin-dependent kinases: engines, clocks, and microprocesses. *Annu. Rev. Cell Dev. Biol.*, **13**, 261–291.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics Supplement*, **32**, 496–501.

Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2002) WinBUGS User Mannual, Version 1.4. *Medical Research Council Biostatistics Unit, Institute of Public Health*, Cambridge University.

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

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–498.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q. amd Kitareewan, S., Dmitrovsky, E., Lander, E., and Golub, T. (1999). Interpreting patterns of gene expression with self–organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**, 2907–2912.

Christopher Wade, Kathleen A. Shea, Roderick V. Jensen and Michael A. McAlear (2001). EBP2 Is a Member of the Yeast RRB Regulon, a Transcriptionally Coregulated Set of Genes That Are Required for Ribosome and rRNA Biosynthesis. *Molecular and Cellular Biology*, **21**, 8638–8650.

Wakefield, J., Zhou, C., and Self, S. (2003). Modelling gene expression over time: curve clustering with informative prior distributions. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meet-*

ing, Clarendon Press, Oxford, 721–733.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, **43**, 201–221.

Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–87.

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

Zhou, C. and Wakefield, J. (2005). A Bayesian Mixture Model for Partitioning Gene Expression Data. *Biometrics*, in press.