Statistica Sinica **12**(2002), 141-158

A SIMPLE STATISTICAL MODEL FOR DEPICTING THE CDC15-SYNCHRONIZED YEAST CELL-CYCLE REGULATED GENE EXPRESSION DATA

Ker-Chau Li, Ming Yan and Shinsheng Yuan

University of California, Los Angeles

Abstract: Microarrays are used for identifying cell-cycle regulated genes by Spellman et al. (1998). In one experiment, a strain of yeast (cdc15-2) was incubated at a high temperature (35^oC) for a long time, causing cdc15 arrest. Cells were then shifted back to a low temperature (23^oC) and the monitoring of gene expression is taken every 10 min for 300 min, using cDNA microarrays. The data are available from their web site (http://cellcycle-www.standford.edu). We find a simple statistical model that can be used to describe most of the expression curves. Three ideas are involved in our analysis: (1) the use of principal component analysis to suggest basis curves; (2) the use of nested models for organizing gene expression patterns; (3) the construction of a compass plot using known cycle-regulated genes for phase determination.

The first two ideas are mainly statistical in nature, but some biological discretion is necessary for successful application. On the other hand, the third idea uses biological information subject to some statistical discretion. The agreement and the difference between our results and the 800 genes identified by Spellman et al. are discussed. A rather unexpected finding is the existence of over 500 genes whose expression levels oscillate regularly every 10 min from time 70 min to time 250 min like a biological pendulum. Extension of our analysis to other cell-cycle experiments is briefly discussed.

Key words and phrases: Analysis of variance, cell-cycle, gene expression, microarray, nested models, principal component analysis.

1. Introduction

Microarray technology enables massive measurement of mRNA amount in cells (Lee and Lee (2000)). Because the mRNA level is a key regulation checking point for gene expression, this technology allows biologists to monitor the biosystem activities at the whole genome level. However, microarray experiments are known to suffer from high level noises due to many factors, including the intrinsic variation of gene expression levels in the cells, systematic or random errors generated by the equipment (robotic arrayer, fluorescence scanner, unspecific cross-hybridization, etc.,) and human inconsistency in the preparation of microarray samples. For most cases, proper analysis of the microarray data remains extremely challenging. Fortunately, many gene expression data are now made public-accessible through the internet. This allows researchers of different disciplines to explore the wealth of information contained in the same data from different perspectives.

In this article, we present a statistical approach for analyzing expression data with a temporal component. While we expect our methodology to have wide application, we illustrate how it works using primarily the data collected in the cdc15 experiment of Spellman et al. (1998). This is one of several experiments conducted in order to find genes that are cell-cycle-regulated in the yeast S. Cerevisiae. Application of our method to other experiments will only be discussed briefly.

Cell cycle is a repetitive life process through which cells grow and divide as the environment allows. Roughly speaking, a mother cell goes through four phases: $G1$, S, $G2$ and M, generating two daughter cells and returning to the $G1$ phase. Generally, $G1$ is the phase of cell growth in preparation to enter the next cycle, S is the phase when all the DNA in cell is replicated, $G2$ is the preparation phase for cell division, M is when the cell separates the genetic material and physically divides into two daughter cells. Researchers are interested in creating a comprehensive catalog of yeast genes whose transcript levels vary periodically within the cell cycle.

To produce useful data, yeast cultures must be synchronized first. Different experiments used different synchronization methods. Once this is done, cell samples were taken at fixed time intervals and the mRNA levels for all yeast genes were measured by microarrays. Combining data from three experiments, α factor arrest, cdc15 arrest, and an earlier cdc28 experiment done by Cho et al. (1998), Spellman et al. (1998) selected 800 genes whose expression levels met a criterion of periodicity. These 800 genes become the focus for further promoter region and other innovative studies.

The criterion used by Spellman et al. for detecting periodic patterns in the expression level curves is based on a combination of Fourier analysis and some biological judgment. The use of Fourier analysis appears logically natural because sine and cosine functions are known to be useful first order approximation for describing periodic phenomena in physical/engineering systems. On a second thought, we feel they may be too simplistic for biological systems. For example, their use suggests that an ideal gene expression curve should have evenly spaced peaks and valleys because this is the characteristic geometric property of sine and cosine functions. We suspect that such simplification may not reflect the biology well. Spellman et al.'s criterion also calls for some subjective determination on the range of periodicity for each experiment. The weight assignment for combining different experiments can also change the outcome of gene selection.

The approach we take is completely different. There are three ideas involved: let the data themselves decide what are the most appropriate basis curves to use; construct a nested statistical model for global partition of genes into homogeneous classes; construct a compass plot from known cycle-regulated genes for phase determination.

The first idea is statistical in nature. But to apply it successfully, some biological discretion is needed. Recall that our aim is to recognize genes with cycle expression patterns by statistical modeling methods. It turns out that basis curve searching can be performed by principal component analysis (PCA). However, we have to trim out the first few, and the last few, time points from each gene expression curve before applying PCA. The statistical and biological justifications for trimming are discussed later.

Using PCA alone is not enough to sort out different curve patterns. There are two reasons: not all curves are expected to be fitted equally well with a selected set of basis functions; the basis curves may or may not show cyclic patterns. It turns out that cyclic patterns only appear in the second and the third basis curves, not in the first. This leads to our second idea. We use the first three basis curves as the full model and partition genes into different groups, according to whether or not they comply to the full model or its submodels. Within groups, genes become more homogeneous in the sense that they share similar patterns which can be parametrized by submodels. From our perspective, nested statistical modeling can be used as a powerful data organization tool.

The third idea relies mostly on a set of 104 known cycle-regulated genes cited in Spellman et al. But we find that the group of SCB regulated genes (late $G1$ phase) do not exhibit consistent cycle patterns as compared with others. Our phase assignment is made without using this group.

The agreement and the difference between our results and the 800 genes identified by Spellman et al. are discussed. We also find a large group of genes which follow the first basis curve very well. Each gene in this group oscillates regularly every 10 min from time 70 min to time 250 min, like a pendulum. Biological implication of this finding is currently under investigation.

2. Data Sets

The expression data used in this paper are from the paper of Spellman et al. We downloaded the data from their web page at http://cellcycle-www.stanford .edu. The cdc15 experiment used cdc15-2(DBY8728), a temperature sensitive mutant, which was first grown to 2.5×10^6 cells/ml in YEP glucose medium at 23^oC. The culture was then shifted to an air incubator at 37^oC and held at that temperature for $3.5h$, causing a cdc15 arrest. Measurement of mRNA began after releasing the cells from cdc15 arrest by shifting the culture back to a 23^oC water bath.

According to Spellman et al., the samples were taken every 10 min for 300 min. However, several time points are missing from the data posted on the web. In fact, it appears that samples were taken every 20 minutes from 10 min to 70 min first, then every 10 min from 70 min to 250 min, and back to every 20 min from 250 min to 290 min.

In our analysis, we use only the 19 consecutive time points from 70 min to 250 min. There are four reasons for this choice. First, this portion of data is taken regularly every 10 min. Second, the tail portion of the data may suffer from what we call a marathon effect: because the pace of growing varies from cell to cell, synchronization has to decay as the time goes by. The situation is similar to runners in a marathon race, in fact Spellman et al. reported that "the third round of small buds appeared at 270 min, although by this time synchrony was decaying." Third, the early portion of data may suffer from what we call the recovery effect. The analogy is that cells, like patients returning home from hospitals, need time to recover. Thus the expressed gene activities in the early portion of data may be confounded with some transient biological phenomena. Lastly, we tried to use all time points but found the results hard to interpret.

Another step of simplification is taken: all ORFs (Open Reading Frames) with missing values during this time interval are removed. There are 4530 ORFs left. So our starting data is a 4530 by 19 matrix, each row representing the expression levels of a gene at the 19 time points in the log scale with base 2. This matrix is designated as M.

The functional grouping information of the genes reported in this paper is based on the Saccharomyces Genome Database (http://genome-www.stanford .edu/Saccharomyces/) and the functional catalog from the web site of Munich Information Center for Protein Sequences (http://www.mips.biochem.mpg.de/proj /yeast/catalogues/funcat/index.html).

3. Analysis Tools

In this section, we describe the tools to be used in our data analysis.

3.1. Two-way ANOVA

Two-way analysis of variance deals with the simplest statistical model on a matrix of data:

$$M_{i,j} = \mu + \alpha_i + \beta_j.$$

For our application, i represents the index for row (Gene) and j for column (Time), $i = 1, \ldots, I$, $j = 1, \ldots, J$, α_i is called the row effect and β_j is called the column effect. This model is of course too simple to be true for our 4530 by 19 matrix M. Nevertheless, it serves as a starting point for bringing out more complicated models. In fact, the more interesting quantities are the residuals from this model. The 4530 by 19 matrix of residuals will be denoted by Y and our curve modeling procedure is applied to it.

3.2. PCA and curve basis selection

One strategy in curve analysis is to fit the observed curve with a parametric family of analytic functions. Let Y_{ij} be the observed value for the i curve, at the jth time point t_j. The model can be written as

$$Y_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(t_j) + \epsilon_{ij}. \tag{1}$$

Common choices of the basis functions $\phi_k(t)$ include polynomials, Fourier series, splines, or wavelets. Theoretically speaking, by allowing for an arbitrarily large K, one can achieve a very high degree of accuracy in approximating any well-behaved function. Practically, a proper choice of the basis system and the number of terms is essential. To increase interpretability of the analysis, the number of terms K must be kept as small as possible. But for our application, none of these pre-determined basis systems appear flexible enough to resolve this model parsimony issue. To alleviate this difficulty, instead of using any preset basis functions, our approach is to leave them completely unspecified. Our rationale is that the most appropriate basis curves should be determined by the data themselves.

It turns out that one can use principal component analysis (PCA) to find basis functions. PCA is easy to describe. For our application, Y_{ij} is the ij entry of the matrix Y. PCA amounts to an eigenvalue decomposition of the p by p covariance matrix $\hat{\Sigma} = n^{-1} * Y'Y$, where $n = 4530$ and $p = 19$ here.

Let $v_1, \ldots, v_p$ be the eigenvectors of $\hat{\Sigma}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$: $\hat{\Sigma}v_k = \lambda_k v_k, \quad k = 1, \ldots, p$. We use the eigenvector v_k to estimate the k basis curve $\phi_k(t_j), j = 1, \ldots, p$. In vector form, we have

$$Y_i = \sum_{k=1}^{K} \hat{c}_{ik} v_k + e_i,$$

where $Y_i = (Y_{i1}, \ldots, Y_{ip})'$, $\hat{c}_{ik}$ is obtained by a least squares fit, separately for each curve. The residual e_i is obtained by taking the difference between Y_i and the fitted curve $f_i = \sum_{k=1}^{K} \hat{c}_{ik} v_k$.

In (1), error terms ϵ_{ij} are assumed to be uncorrelated, with mean 0 and variance σ_i^2, the variance being allowed to vary from gene to gene.

Under these assumptions, it can be proved that the basis vectors found by PCA are consistent (asymptotically). This means that as one has more and more curves in the data set, the estimates become more and more accurate in estimating the correct basis vectors. It is worth noting that the eigenvalue decomposition of $\hat{\Sigma}$ is equivalent to the singular value decomposition of Y.

3.3. Nested models and goodness of fit

To continue our analysis, we take $k = 3$ in the curve fitting model (1) of the previous subsection. The basis vectors are the first three eigenvectors of PCA; see Figure 1, panel $A-C$. For each gene expression curve we ask three questions: does the curve follow this three parameter model well ?; does the curve contain cycle component ?; is the curve smooth ?

Each question is conveniently translated into a hypothesis testing problem so that suitable testing statistics can be applied. In particular, the three hypotheses and their testing procedures are as follows.

(1) Compliance Check. Hypothesis H_0: The three-basis full model holds. Reject if R_i^2 and $rss_i > 7.25$.

(2) Cycle Component Check. Hypothesis H_0: $c_{2i} = c_{3i} = 0$. Reject if $\frac{(\hat{c}_{2i}^2+\hat{c}_{3i}^2)/2}{rss_i/15} > F_{2,15} = 3.68$.

(3) Smoothness Check. Hypothesis H_0: $c_{1i} = 0$. Reject if $\frac{\hat{c}_{1i}}{|\sqrt{rss_i/15}|} \geq t_{15}(0.975) = 2.131$.

Here R_i^2 is the R-squared value and rss_i is the sum of squared residuals from the least squares fit of the full model for the ith gene. The reason that there are only 15 degrees of freedom in (2) and (3) is that one degree of freedom is already lost when the intercept is subtracted out from the beginning. Further discussion of the cut-off values is to be given later.

4. Results

4.1. Two-way ANOVA of the cdc15 dataset

As a preliminary step in our analysis, we apply the standard two-way analysis of variance (ANOVA) to M. The purpose is to check the degree of constancy of the average expression level over time for each gene, and the constancy of the average expression level over all genes at each time point. The result is given in Table 1.

For better interpretation of these results, recall that the expression level records the logarithm (base 2) of the noise-adjusted light intensity ratio for the red channel to the green channel. The red channel uses mRNA sample from the synchronized cells at each time point. The green channel uses the mRNA sample from unsynchronized cells. The insignificance of row effects shows that the unsynchronized cell sample can be viewed as mixtures of synchronized cells from various time points. This is consistent with biological reality.

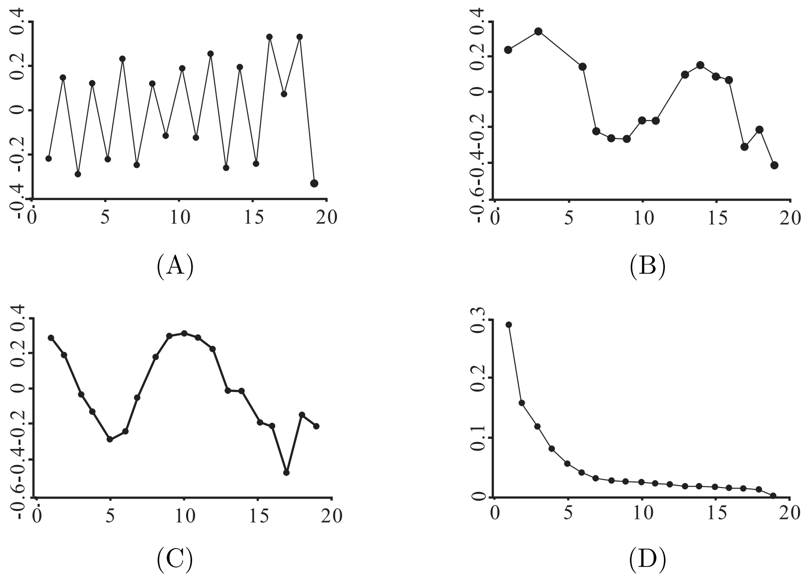
Table 1. Two Way ANOVA on the data matrix.

Factor	df	SS	MS	F
gene	4529	5.24E+2	1.16E−1	6.42E−1
time	18	2.97E+2	1.65E+1	9.16E+1
residual	81522	1.47E+4	1.80E−1	
total	86069	1.55E+4		

Column effects are seen to be statistically significant. However, since different DNA chips are used at each time point, this variation is confounded with chip to chip bias. From now on, we remove both the column effects and row effects from each expression curve. This amounts to considering the residual matrix Y as defined earlier. Such adjustment is minor. For instance, the maximum adjustment is less than 0.15 which represents only about one-tenth fold increase ($2^{0.15} = 1.11$).

4.2. Bases for cdc15 gene expression curves

The PCA analysis is applied to Y as described before. The first three basis functions obtained by PCA are shown in Figure 1, panel $A - C$. Panel D shows how rapidly the eigenvalues decay. It is seen that the first three components account for 56% of the total variance. As we shall soon see, these three bases already provide an adequate fit for the majority of gene expression curves.



(A)

(B)

(C)

(D)

Figure 1. Basis curve searching by PCA **(A)**: The first PCA direction. **(B)**: The second PCA Direction. **(C)**: The third PCA direction. **(D)**: The first three eigenvalues contains more than 56% of the total variance.

The first basis curve oscillates in an extremely regular fashion, moving up and down every 10 minutes (Figure 1, (A)). So far, we do not have a good explanation for the presence of this highly regular, high frequency component; see **Discussion** section.

The interesting patterns of cell cycle occur in the second and the third basis curves (Figure 1, (B) and (C)). Based on a visual inspection of these two curves, the period for the cycle is about 90 minutes, which is in good accordance with the yeast cell cycle period as determined by bud count. A closer examination shows that the second basis appears to trail the third basis by about 30 minutes. The correlation coefficient is 0.88 between the two basis vectors after shifting the second basis to the left by 3 time points.

4.3. Nested models for organizing gene expression patterns

The three-basis full model serves as an approximation to the true mRNA level curve. We need a mechanism to judge how well the gene expression curves comply to this model. For this purpose, we rely on two byproducts from fitting each curve with the full model. To qualify as a compliance member of this three parameter model, the R-squared value must be bigger than 0.56 (equivalently, the correlation coefficient between the fitted values and the observed values must be greater than 0.75, the square root of 0.56) or the sum of squared residuals must be less than 7.5 (equivalently, the standard deviation for residuals must be smaller than 0.68). Genes that do not meet this criterion are called non-compliance genes. To see if the value of 0.68 is reasonably small in terms of the original ratio scale, note that $2^{0.68} = 1.6$ and $2^{-0.68} = 0.62$. So in terms of fold change, one standard deviation of error amounts to 0.49 ($= [(1.6 - 1) + (1 - 0.62)]/2)$), or about 0.5, fold change. This is well within the range of precision in current Micro-Array technology.

Once the non-compliance genes are separated out, we can use the loading coefficients $\hat{c}_{i1}, \hat{c}_{i2}, \hat{c}_{i3}$ to organize genes according to whether they have significant cycle patterns or not. The non-cycle genes are those with small values for $\hat{c}_{i2}$ and $\hat{c}_{i3}$. The criterion is based on an F-test for $c_{i2} = c_{i3} = 0$. For those with significant cycle patterns, we further separate them into two groups, the smooth cycle group and the non-smooth cycle group. The smooth cycle group is determined by accepting the null hypothesis that $c_{i1} = 0$ via a t-test.

Figure 2 shows the partition tree. For the 4530 genes without missing values, there are only 41 non-compliance genes. The patterns within this group vary a lot. Some of them appear nearly constant except for a single spike at a single time point. Some of them exhibit periodic patterns. For the 4489 compliance genes, 2824 of them show no non-cycle patterns. The remaining 1665 genes have significant cycle coefficients, among which 714 have no bumpy components.

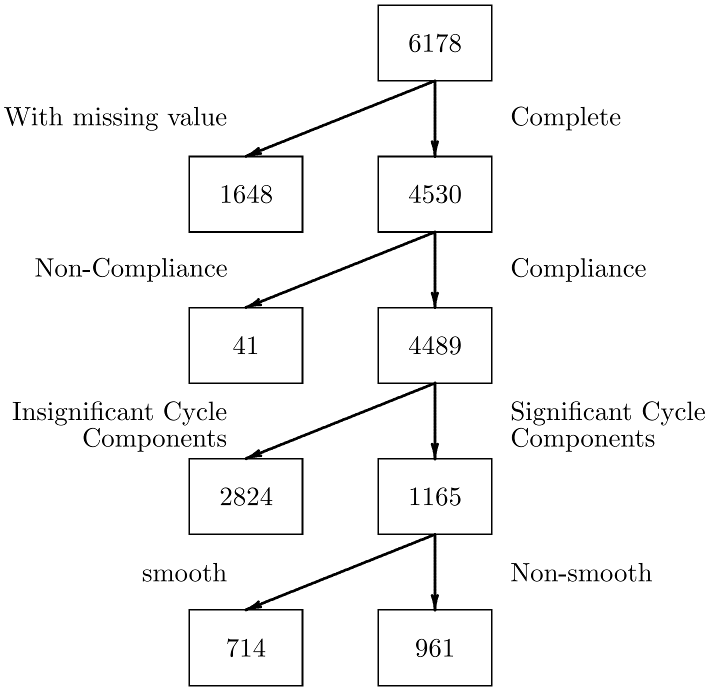


Figure 2. The partition tree for all genes.

4.4. Phase assignment with a compass plot

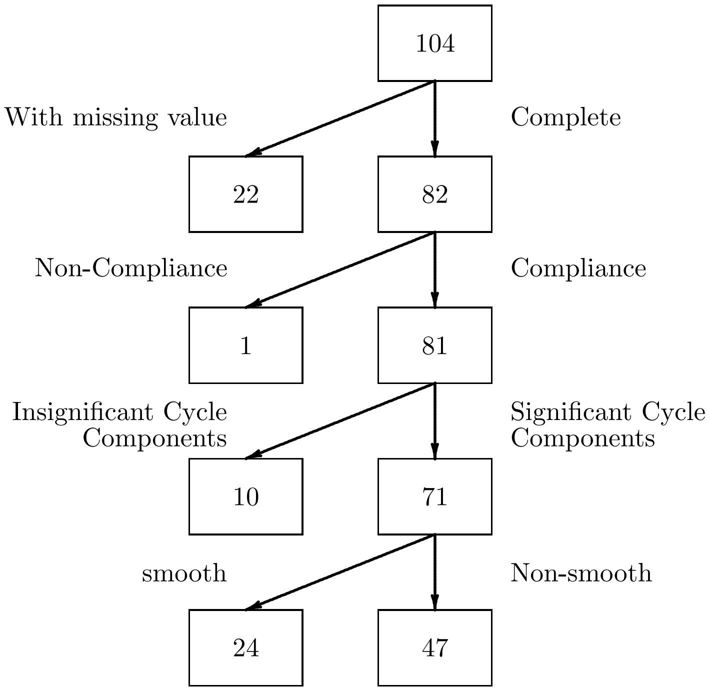
A compass plot based on the combination of the second and third PCA component is constructed for assigning phases to cell cycle regulated genes.

As shown above, the second and the third basis curves have clear cycle patterns. The loading coefficients $\hat{c}_{i2}, \hat{c}_{i3}$ for these two basis curves , after normalization to have length equal to one, plays the same role as the phase parameter in Fourier analysis. In other words, if we represent $(\hat{c}_{i2}, \hat{c}_{i3})$ in polar coordinates as $(\hat{r}_i \sin(\hat{\theta}_i), \hat{r}_i \cos(\hat{\theta}_i))$, where $\hat{r}_i = \sqrt{\hat{c}_{i2}^2 + \hat{c}_{i3}^2}$ is the length and $\hat{\theta}_i$ is the angle, then we can represent each cycle-gene by a point in a circle, using its numerical phase value $\hat{\theta}_i$. If these numerical phase values are calibrated to match the biological cell cycle phases of known cycle-regulated genes, then we can use them to assign biological phases to other genes in both the smooth cycle group and the non-smooth cycle group.

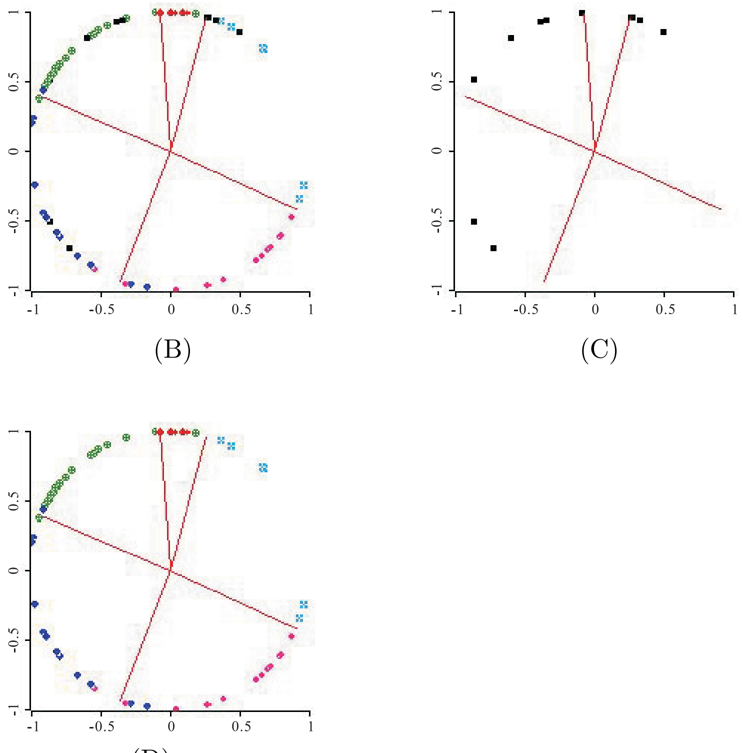
To achieve this, we rely on the 104 known phase genes reported in Spellman et al. The phases of these genes were determined by traditional experimental approaches. They were grouped into 6 categories by Spellman et al. : SCB ($G1$ phase), MCB ($G1$ phase), S phase, $S/G2$ phase, $G2/M$ phase, and $M/G1$ phase.

We first apply our partition rules to these 104 genes. The result is given in Figure 3(A). Genes with missing-values, noncompliant, or without significant cycle component are excluded. The phases of those genes passing the tests are plotted in the compass plot in Figure 3(B).

From this plot, we see that mathematically genes from the same biological phase category tend to group together. This is in accordance with the notion that the phases obtained with our method do reflect the phase of gene expression in the cell cycle. However, although both the MCB group and the SCB group belong to the $G1$ phase, the SCB group appears to be more scattered (Figure 3(C)). Visual inspection of each curve in this group shows that curve patterns are rather diverse for these genes. This can be further confirmed by a two-way ANOVA. In addition, some biological findings may be linked to the diversity of the peak expression times for the SCB group genes. For example, Spellman et al. wrote “It is now apparent that SBF is not as specific for SCB’s as was originally thought, but rather can bind, at least in some cases, to motifs more closely matching the MCB consensus (Partridge et al. (1997))”. Note that SBF is supposed to regulate the SCB group of genes, not the MCB group of genes. Other related evidence comes from Figure 2 of Spellman et al. The binding site



(A)



(B)

(C)

(D)

Figure 3. **(A)**: The partition tree for 104 genes with known phases. **(B)**: Compass plot: different biological phase groups are coded with different shapes and colors; Green for G1, Red for S, Cyan for S/G2, Magenta for G2/M, and Blue for M/G1. **(C)**: SCB group on compass plot. **(D)**: Remove SCB genes, final compass plot for determining the phases of genes.

frequencies for SCB appear much more divergent than MCB. For these reasons, we believe that MCB group of genes are more dependable for determining the $G1$ phase than the SCB group. We exclude the SCB group from the final compass plot (Figure 3(D)).

We can use the phase partition from the compass plot to assign phases for all genes in the smooth cycle-regulated group and in the non-smooth group separately.

4.4.1. Grouping of the 800 cell cycle regulated genes

Spellman et al. have identified 800 cell-cycle regulated genes. We apply our partition rules (Section 4.3) to these genes. The result is shown in Figure 4.

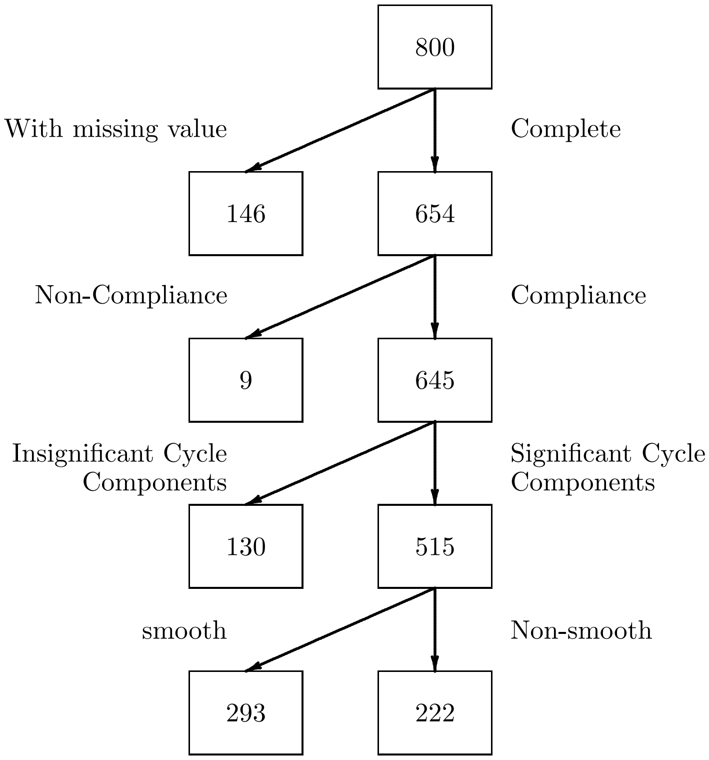
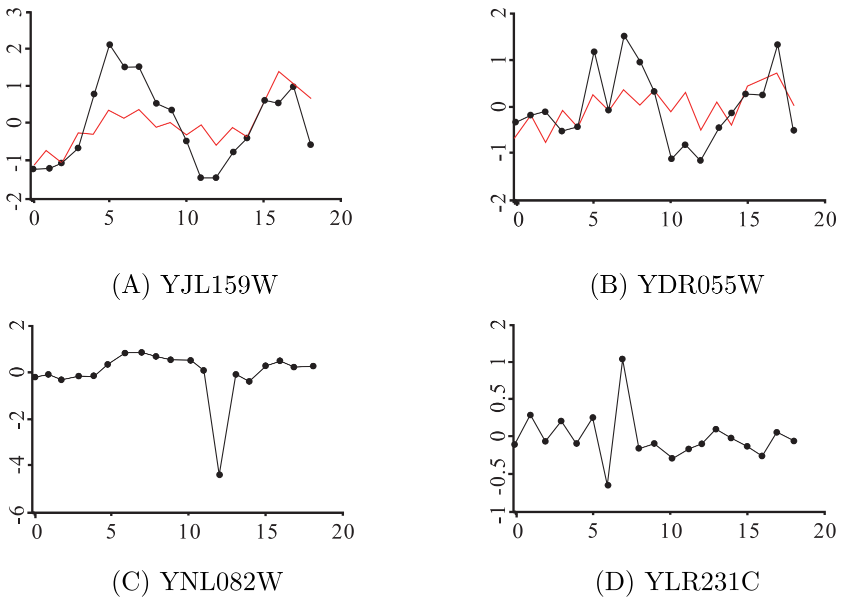


Figure 4. The partition tree for all 800 cycle-regulated genes from Spellman's paper.

For the 654 genes without missing values, only 9 of them are classified as non-compliant. Visual inspection on these genes show that some of them have a cyclic pattern but others do not. See Figure 5 (A)(B)(C).

For the 645 complete genes, 130 did not pass our test for significant cycling components. Many of them do not appear to be cyclic even if they receive high scores for Spellman et al. One such curve is given in Figure 5, (D).

There are 515 genes falling into our significant cyclic component group. Of them, 293 are smooth. We can compare our phase assignment with Spellman et al.'s. This is summarized in Table 2 (for the nonsmooth group) and Table 3 (for the smooth group). We see that most genes are assigned the same phases (the diagonal entries). A few genes are assigned to adjacent phases. Thus, overall the phase assignment by the two methods is fairly consistent. However, there are 6 cases in the Smooth group and 1 case in the Non-smooth group where the two assignments differ by two phases (bold-faced cases in Table 2 and 3). By visual inspection of these cases, we find that the signals are either low or confusing because of multiple peaks.



(A) YJL159W

(B) YDR055W

(C) YNL082W

(D) YLR231C

Figure 5. Some examples from Spellman et al.'s 800 cell-regulated genes - all are supposed to show cycle patterns.

(A): YJL159W is in the non-compliance group. It has a clear cycle pattern as shown by the black curve. The red one is the fitted curve by the three-basis model. The fit is apparently poor.

(B): YDR055W is another gene from the non-compliance group, the cycle pattern is less clear.

(C): YNL082W is from the non-compliance group, no visible cyclic pattern can be seen.

(D): YLR231C is from the insignificant cycle component group, no visible cyclic pattern can be seen.

Table 2. Contingency table for phase assignments of non-smooth group of genes.

Spellman's Our	$G1$	S	$S/G2$	$G2/M$	$M/G1$	Total
$G1$	59	6	0	0	0	65
S	4	3	0	0	0	7
$S/G2$	**1**	7	31	17	0	56
$G2/M$	0	0	3	47	1	51
$M/G1$	18	0	0	4	21	43
Total	82	16	34	68	22	222

Table 3. Contingency table for phase assignments of the smooth group of genes.

Spellman's Our	$G1$	S	$S/G2$	$G2/M$	$M/G1$	Total
$G1$	74	8	0	0	1	83
S	7	10	1	0	0	18
$S/G2$	**5**	11	43	17	**1**	77
$G2/M$	0	0	1	39	1	41
$M/G1$	43	0	0	3	28	74
Total	129	29	45	59	31	293

5. Discussion

We discuss some key methodological differences between our approach and several other works related to gene expression with a temporal component first, before turning to related biological issues.

5.1. Other related works and methodological differences

Our analysis hinges critically on the combined use of two statistical methods : PCA and nested-modeling. PCA is used to find important basis curves for parametrizing expression data. Nested-modeling is used to partition the genes into organized groups. One method alone does not work, it is the synergy that brings meaningful results.

PCA by itself has become a widely used technique for gene expression analysis. It is often described as a dimension reduction procedure. The procedure finds a small number of directions for projecting high dimension data. The projected points have the largest possible variance and visual inspection after projection is often recommended for detecting the presence of clusters or other nonlinear patterns. For instance, in Wen et al. (1998), PCA was applied to confirm gene clusters found independently by a cluster method FITCH (Felsenstein (1993)). But this does not work for the cdc15 data. The entire plot is crowded with data points after projection with the first three PCA directions. It is not easy to identify clusters by visual inspection.

Holter et al. (2000) applied singular value decomposition (SVD) for finding fundamental patterns underlying gene expression profiles with temporal components. As noted earlier, SVD is algebraically equivalent to PCA. Holter et al. viewed SVD as a numerical analysis tool which approximates an n by p data matrix by a matrix with a smaller rank. With some selected examples, they illustrated how the simplified data matrix can preserve the entire set of gene expression data. However, they ignored the possible discrepancy between the simplified data and the original data. This discrepancy is explicitly accounted for in our curve fitting model (1). The error term in (1) plays an essential role, especially when the PCA(or SVD) result is combined with compliance checking and other nested-modeling ideas.

One of Holter et al.'s examples also used cdc15 data, but they did not apply SVD to the entire data set. In fact, they only considered the 800 cycle-regulated genes and found the first two basis curves exhibiting sine and cosine shapes. From our perspective, this application only confirms that Fourier transform did play a key role when Spellman et al. defined the score of cell-cycle-regulation for each gene. For reasons unexplained in their paper, Holter et al.'s analysis excluded all the even time points from the data set. As requested by a referee, we applied PCA to the 800 genes using all time point between 70 min and 250 min. The first three basis curves are shown in Fig 6. The oscillating pattern now appears in the third basis.

More advanced use of SVD is illustrated in Alter, Brown and Botstein (2000). By example, they showed how SVD can be applied to preprocess the gene-expression data in order to filter out “eigengenes” or “eigenarrays” that are inferred to represent noise or experimental artifact. However, there does not appear to be any model for generating data sets that can be systematically analyzed by their approach. Their examples include α-factor arrest data and the elutriation data, but not the cdc15 data.

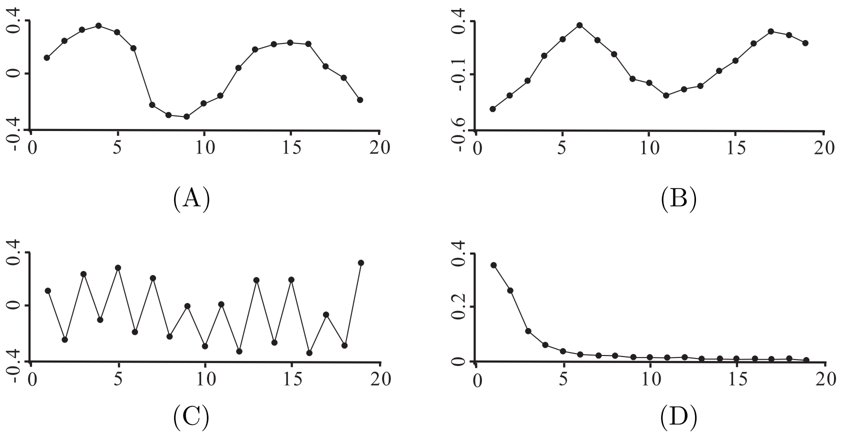
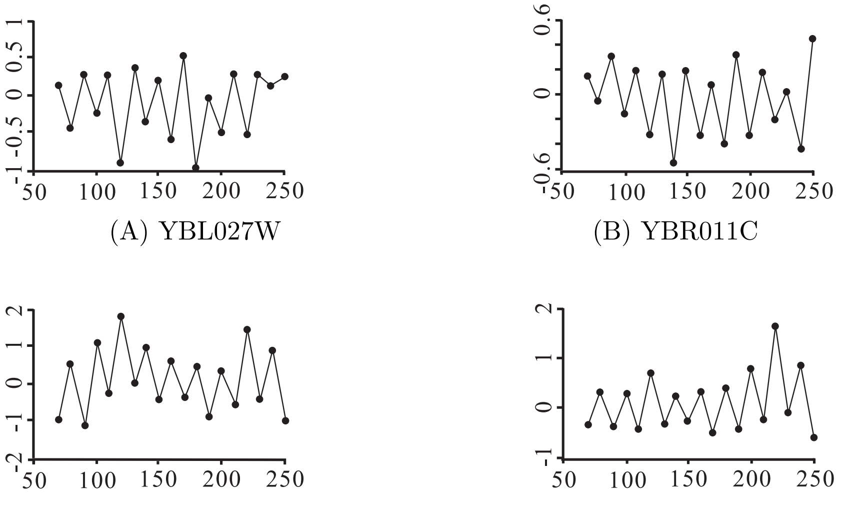


Figure 6. Basis curve searching by PCA on 800 genes. **(A)**: The first PCA direction. **(B)**: The second PCA Direction. **(C)**: The third PCA direction. **(D)**: The first three eigenvalues contains more than 72% of the total variance.

5.2. Oscillating genes

Our way of applying PCA to cdc15 data suggests three basis curves for model fitting. The first PCA direction reveals a high frequency component. A surprising feature about this basis is that the perceived fluctuation pattern does not appear to be random (Figure 1,(A))— it moves up and down alternately. We find that there are over 500 genes that show such regular oscillating patterns; see Figure 7 for four such curves. In fact, they can be further divided into two classes: 190 genes move up-down-up-down while 363 genes go down-up-down-up. Their protein products are involved in all kind of biological activities, including transportation, cytokinesis, RNA processing, transcription and translation. If the fluctuation pattern were completely random, so that there is a 0.5 probability to move up (or move down), then there is only 2^{-18} probability for a gene to follow the up-down-up-down (or down-up-down-up) pattern. Consequently, for the entire Yeast genome (with less than 2^{13} genes), on the average we anticipate to see less than one such gene ($2^{13} \times 2^{-18} = 2^{-5}$ in fact). This indicates that the coherent behavior from so many oscillating genes cannot be explained by chance error. Of course, this may well be due to a systematic fluctuation intrinsic to the Microarray experimental design. If this is the case, then it would be interesting to know why such systematic bias did not affect all genes. Indeed, an anonymous referee from an earlier version of this article reported that the experiment was done on different days, a possible source of artifact. If this is true, then some statistical effort should be made to remove this effect from the data. Our approach can be the first attempt in this direction. It can be done by substracting the first basis from the original data, for example.



(A) YBL027W

(B) YBR011C

(C) YFL011W

(D) YFL039C

Figure 7. Some oscillating genes. **(A)** YBL027W: RPL19B, 60 S Large subunit ribosomal protein L19.E. **(B)** YBR011C: IPP1, inorganic pyrophosphatase, cytoplasmic. **(C)** YFL011W: HXT10, hexose transporter. **(D)** YFL039C: ACT1, actin.

5.3. Further works and limitations

It is easy to apply our methodology to the other three gene expression experiments (α-factor, cdc28 and elutriation). It is interesting to find that, for each experiment, we still need only the first three basis curves from PCA. However, unlike the cdc15 data, the bases meaningful for extracting the cycle components do not occur in the second or the third PCA directions. Neither do they occur in the first PCA direction. Rather they are hidden in some combinations of the first three PCA directions. The extra basis exhibited a monotonic trend. But this additional complexity does not make the follow-up nested modeling based partition more difficult to carry out. Details will be reported elsewhere.

The basis functions we empirically identified for depicting cell-cycle regulated gene expression profiles serve as promising alternatives to Fourier basis functions. Although the Fourier transform is often used in studying periodic functions, it should be noted that this modeling technique can be problematic when the series is short. One nice property of sines and cosines is that the curve shape around the peak is a mirror reflection of that around the valley. Yet, this type of symmetry may not have a strong biological footing because the durations of different cell phases vary substatially.

Mathematically and statistically, it remains an unsettled issue regarding which linear combinations of the three basis functions should be used to represent the cyclic components. Recall that PCA produces orthogonal baisis. But orthogonality is not required in (1). For the cdc15 data, we have to rely on the supporting evidence about buddings from Spellman et al. (1998). This leads to our belief that there are no more than 2 cell-cycles within the 70-250 min interval. We have visually inspected combinations of our two chosen bases and verified that this assumption is not severely violated.

Model (1) can be improved by introducing correlation between errors. The ultimate gain will depend on many factors, including the proper assumption of correlation patterns and efficient ways of estimating the basis functions and correlation parameters simultaneously. This requires substantial further study. One can also consider the option of using more than 3 basis functions. Additional parameters would increase the goodness of fit, but the trade-off is stability of estimation and complexity in interpreting the results.

What we have done with the microarray data is a long way from understanding the biology behind how genes are cell-cycle regulated. There are intrinsic limitations on what can be learned from microarray data alone. First, genes that play critical roles in regulating cell cycles do not necessarily express cyclic patterns in their mRNA abundance levels. For instance, Cdc28 does not belong to the 800 cycle-cell-regulated genes, nor did it have a significant cycle component in our analysis. This is because Cdc28 is needed in all phases for forming complexes with different cyclin proteins. Second, it is the balance between making and degrading that determines the measured mRNA amounts. Thus the peaks in the expression curve may not honestly reflect the correct timing of gene activation. Third, current miroarray technologies are not sensitive enough at low expression levels to pick up the subtle gene activity which may lead eventually to some cascade effect.

Despite such limitations, we feel that there is still a lot of information that can be extracted from microarray data. We have presented a novel way of organizing genes so that their expression patterns can be described parsimoniously with only three parameters. Within our partition system, every gene without missing values has been accounted for. This helps biologists gain a global picture of expression patterns at the full genomic scale. For simplicity of presentation, the genes with missing values do not participate in the process of finding curve bases. However, they can be incooperated into the analysis without much further work.

Acknowledgements

Li's research is supported in part by NSF grants DMS-9803459 and DMS-0104038.

References

Alter, O., Brown, P. O. and Botstein, D. (2000). *Proc. Natl. Acad. Sci.* **97**, 10101-10106.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W. (1998). *Mol. Cell* **2**, 65-73.

Eisen, M. B., Spellman, P. T., Brown P. O. and Botstein, D. (1998). *Genetics* **95**, 14863-14868.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package), Version 3.5c, Department of Genetics, University of Washington, Seattle.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. and Fedoroff, N. (2000) *Proc., Natl, Acad. Sci.* **97**, 8409-8414.

Lee, P. S. and Lee, K. H. (2000). *Curr. Opin. Biotechol.* **11**, 171-175.

Partridge, J. F., Mikesell, G. E. and Breeden, L. L. (1997). *J. Biol. Chem.* **272**, 9071-9077.

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). *Mol. Biol. Cell* **9**, 3273-3297.

Wen, W., Stefanie Fuhrman, George S. Michaels, Daniel Carr, Susan Smith, Jeffery L. Barker and Roland Somogyi. (1998) *Proc. Natl. Acad. Sci.* **95**, 334-339.

Department of Statistics, 8130 Math. Science Building, Box 951554, University of California, Los Angeles, CA90095-1554.
E-mail: kcli@stat.ucla.edu

Department of Biochemistry and Computer Science, University of California, Los Angeles, U.S.A.

Department of Statistics, University of California, Los Angeles, U.S.A.
E-mail: syuan@stat.ucla.edu

(Received March 2001; accepted October 2001)