Time Ordering of Gene Co-expression

Xiaoyan Leng

Department of Public Health Sciences

Section on Biostatistics

Wake Forest University School of Medicine

Medical Center Boulevard, MRI-3

Winston-Salem, NC 27157

Hans-Georg Müller

Department of Statistics

University of California

One Shields Avenue

Davis, CA 95616

Revised Version

February 14, 2006

---

*Corresponding author: Xiaoyan Leng, Department of Public Health Sciences, Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Boulevard, MRI-3, Winston-Salem, NC 27157. Tel: 336-716-4564; Fax: 336-716-5425/6427; email: ileng@wfubmc.edu

ABSTRACT

Temporal microarray gene expression profiles allow characterization of gene function through time dynamics of gene co-expression within the same genetic pathway. In this paper, we define and estimate a global time shift characteristic for each gene via least squares, inferred from pairwise curve alignments. These time shift characteristics of individual genes reflect a time ordering that is derived from observed temporal gene expression profiles. Once these time shift characteristics are obtained for each gene, they can be entered into further analyses, such as clustering. We illustrate the proposed methodology using *Drosophila* embryonic development and yeast cell cycle gene expression profiles, as well as simulations. Feasibility is demonstrated through the successful recovery of time ordering. Estimated time shifts for *Drosophila* maternal and zygotic genes provide excellent discrimination between these two categories and confirm known genetic pathways through the time order of gene expression. The application to yeast cell cycle data establishes a natural time order of genes that is in line with cell cycle phases. The method does not require periodicity of gene expression profiles. Asymptotic justifications are also provided.

Key words: Curve alignment; Functional data analysis; Gene expression profiles; Microarray; Time dynamics; Warping.

1 Introduction

The thousands of genes of an organism or a cell must be expressed in a regulated manner to enable the organism or the cell to utilize the biological information contained within the genome. One of the most important dimensions of gene regulation is temporal control of gene expression within genetic pathways or biological processes, through which groups of genes are thought to be co-expressed coherently across different time periods. These processes are related to the function of the proteins encoded by these genes. In general, the timing of mRNA expression for a given gene correlates well with the function of the resultant protein (Bozdech *et al.*, 2003; Bähler, 2005).

The temporal control of gene expression extends over many time scales, such as development and maturation over the entire life span on the organismal level or cell cycles on the cellular level. For example, embryogenesis is an example of a time-sensitive biological process, whereby the embryo forms and develops. A fertilized egg of the fruit fly (*Drosophila melanogaster*) undergoes cleavage, blastoderm, gastrulation, germ band elongation and retraction, head involution and dorsal closure, and differentiation over about 24 hours. The coordination in time and space of the expressions of maternal and zygotic genes guarantees normal embryogenesis (Lodish *et al.*, 2000; Weigmann *et al.*, 2003; Pollard and Earnshaw, 2004). Another example is provided by the timing characteristics of the yeast cell cycle, which can be modeled as "a line of dominoes" or dependent pathway, whereby a gene will

not start to express until certain other genes have already expressed (Pollard and Earnshaw, 2004).

Uncovering the time ordering of gene expression could therefore be expected to aid the elucidation of gene function, and to provide a context for interpreting organismal/cellular responses to drugs, growth conditions or environmental perturbations. This lends motivation to employ increasingly collected temporal microarray gene expression data, when aiming at detection of the time ordering of gene expression. The assumption that genes with similar temporal expression patterns are part of a common genetic pathway provides further motivation for global time ordering analysis of gene expression profiles.

Previous time ordering analyses have aimed at pairwise gene expression profile alignment. Butte *et al.* (2001) utilized digital signal-processing tools including power spectral densities, coherence, transfer gain, and phase shift, to find pairwise gene associations based on periodically expressed time-invariant gene profiles. Qian *et al.* (2001) adopted a local clustering approach to determine optimal pairwise alignment based on dynamic programming, while Aach *et al.* (2001) implemented both simple and interpolative time warping based on dynamic programming to identify an optimal time alignment of two gene expression time series. Expanding this approach, Liu and Müller (2003) proposed a non-parametric time warping technique to construct modes of temporal structure for a sample of gene expression profiles, adapting a time synchronization approach (Liu and Müller, 2004).

For general warping and curve alignment procedures we refer to Wang and Gasser (1997) and Gervini and Gasser (2004). Bar-Joseph *et al.* (2003a, 2003b) developed yet another approach and aligned temporal data sets under varying conditions, extracting shift and stretch parameters for each data set. In a very general approach, Arkin and Ross (1995) and Arkin *et al.* (1997) devised models with time shifts for chemical reaction pathways, and proposed to determine pairwise time shifts by maximizing correlation. A related global time shift model for functional data was considered by Silverman (1995).

In this paper, we define and infer global time shift characteristics for genes that are based on observations of optimal pairwise curve alignments, which then are symmetrized through the minimization of a functional distance via a least squares step. We show that a conditional L_2 distance between two curves is minimized near the true underlying pairwise time shift. The resulting global time shifts for genes lead to the proposed time order characteristics. Further analysis such as gene grouping/clustering within the same genetic pathway can then be based on the estimated time order characteristics in a subsequent step. The proposed methodology may play an auxiliary role in determining a genetic pathway.

The organization of the paper is as follows: We introduce pairwise alignment in section 2, including asymptotic consistency results for identifying underlying time shifts based on minimizing L_2 distance. The connection to global time shifts is made in section 3, followed by simulations (section 4) and the analysis of the

time ordering for *Drosophila* embryonic development (section 5) and yeast cell cycle (section 6). The paper ends with discussion and conclusions (section 7). Theoretical results and proofs can be found in the appendix which is posted at the journal's website.

2 Pairwise Curve Alignment

Assume we observe a collection of K gene expression profiles with expression levels Y_{im} measured at times t_{im}, $i = 1, \ldots, K$, $m = 1, \ldots, n$, $t_{im} \in [0, T]$, with

$$Y_{im} = X_i(t_{im}) + e_{im},$$

where X_i are the underlying expression profiles, which we view as realizations of a stochastic process, and e_{im} are i.i.d. errors with zero mean and finite variance. Furthermore, assume

$$X_i(t) = X_i(t, \tau_i) = \mu(t - \tau_i) + \delta Z_i(t - \tau_i),$$

where $\mu(t)$ is a non-random mean curve and $Z_i(\cdot)$ are i.i.d. realizations of a stochastic process Z, s.t. $E(Z(t)) = 0$ and $E(Z^2(t)) < \infty$. It is not necessary to specify the covariance function of Z. If, on the other hand it is known, some efficiency might be gained. The τ_i are random time shifts and δ is a small constant, so that $\delta Z_i(\cdot)$ might be viewed as a small random perturbation. For a pair of random curves $X_i(t)$ and $X_j(t)$, the relative time shift is $s^*_{ij} = \tau_i - \tau_j$, $i, j = 1, \ldots, K$. Note that this pairwise time shift is symmetric in the sense that $s^*_{ij} = -s^*_{ji}$.

We are interested in identifying the time shifts τ_i which are associated with specific gene profiles X_i, and also in the inherent ordering of the τ_i, $\tau_{(1)} < \tau_{(2)} < \ldots < \tau_{(K)}$. The time shift τ_i are characteristics of individual genes associated with their expression trajectories. They reflect the inherent time order of the ensemble of genes and therefore have a clear biological interpretation.

Given a random trajectory $X_i(t)$, we align other trajectories $X_j(t)$ against $X_i(t)$ on the interval $\mathcal{T} = [T_1, T_2]$ for some constants $T_1 < T_2$ by minimizing a distance $d(X_i(t), X_j(t-s))$ with regard to s. We aim at the minimizers

$$\begin{aligned} s_{ij} &= \arg\min_s E(d^2(X_j(t-s), X_i(t))|\tau_i, \tau_j), \\ d_{ij} &= d(X_j(t-s_{ij}), X_i(t)) \quad \text{for } t \in \mathcal{T}. \end{aligned} \tag{2.1}$$

Similarly, we define s_{ji} and d_{ji} by exchanging i and j. We assume implicitly that $X_i(t-s)$ is defined for all $s \in [-T_0, T_0]$, for a $T_0 > 0$, by extending the domain of the $X_i(\cdot)$ suitably beyond $\mathcal{T}$. For distance d in function space, we consider the L_2 pseudo-distance

$$d(f,g) = \left\{ \int (f(t) - g(t))^2 \, \mathrm{dt} \right\}^{\frac{1}{2}}. \tag{2.2}$$

All functions are pre-normalized by the transformation $f(t) \to \frac{f(t)}{(\int f(s)^2 \, \mathrm{ds})^{\frac{1}{2}}}$, aiming to de-emphasize differences in amplitude and to emphasize differences in horizontal shift. Other distances could also be used, such as correlation and rank correlation (Arkin and Ross, 1995; Arkin *et al.* 1997; Heckman and Zamar, 2000).

Theorem 1. *For two random functions X_i and X_j, let*

$$\Delta_{ij}(s) = E\left(\int_{\mathcal{T}} \Big(X_i(t,\tau_i) - X_j(t-s,\tau_j)\Big)^2 \,\mathrm{dt} \,\Big|\, \tau_i, \tau_j\right).$$

Under conditions (A1) - (A4) (see appendix), for sufficiently small δ,

$$s_{ij} = \arg\min_s \Delta_{ij}(s) = s^*_{ij} + O(\delta) = \tau_i - \tau_j + O(\delta).$$

Theorem 1 shows that, under suitable assumptions, s_{ij} asymptotically tracks the true value s^*_{ij}. The proof is in the appendix.

After pairwise alignment, we obtain two matrices: the minimum distance matrix $\mathbf{D}_{K\times K}$ and the relative time shift matrix $\mathbf{S}_{K\times K}$, where $\mathbf{D} = \{d_{ij}\}$ and $\mathbf{S} = \{s_{ij}\}$, $i, j = 1, \ldots, K$. We note that the matrices $\mathbf{S}$ and $\mathbf{D}$ are generally asymmetric. The elements of $\mathbf{S}$ consist of the pairwise relative time shifts s_{ij}, which serve as responses in a global time shift model that is discussed in the next section.

3 Global Time Shift Model and Inference for Time Shifts

For each pair of random curves $X_i(t)$, $X_j(t)$, $i, j = 1, \ldots, K$, the relative shift of X_j with respect to X_i is expected to be close to $s^*_{ij} = \tau_i - \tau_j$, if δ is relatively small, since $\mu(t - \tau_i) = \mu(t - \tau_j - s)$ for $s = s^*_{ij}$. We note that in general for the pairwise time shifts s_{ij} (2.1), $s_{ij} \neq s_{ij}$, so that a reasonable algorithm needs to include a symmetrization step. Without loss of generality, let $\tau_1 = 0$. For K gene expression

trajectories, the equations

$$s^*_{ij} \;=\; \tau_i - \tau_j, \quad i,j = 1,\ldots,K, \quad i \neq j, \tag{3.1}$$

correspond to a linear system

$$\boldsymbol{s^*} = \boldsymbol{A\tau}, \tag{3.2}$$

where $\boldsymbol{s^*}$ is a $K(K-1)$ vector of stacked pairwise relative time shifts, $\boldsymbol{A}$ the corresponding design matrix corresponding to equations (3.1), and $\boldsymbol{\tau} = (\tau_2,\ldots,\tau_K)'$ the shift parameter vector, where x' denotes the transpose of a (column or row) vector x; specifically, $\boldsymbol{s^*} = (s_1^{*'},\ldots,s_k^{*'})'$, where s_i^* is a $K-1$ vector of s_{ij}^*, $i,j = 1,\ldots,K$, $j \neq i$. Note that A is always of full rank by design.

Gene expression profiles X_i are often not continuously observed but are rather observed over a discrete grid of measurement times t_{im}, giving rise to discrete observations (t_{im}, Y_{im}) per profile. Since the measurement times can be irregular, there seems no obvious way to obtain the distance between any two observed profiles without making distributional assumptions, which we prefer to avoid; compare Yao *et al.* (2005) for the implementation of Gaussian assumptions for irregularly observed functional data. Smooth trajectories for profiles X_i can be obtained from the discrete measurements by applying a linear scatterplot smoother to the scatterplot (t_{im}, Y_{im}), denoted by $\hat{X}_i(t)$, when evaluated at point t. While various smoothing methods are available (Fan and Gijbels, 1996), we use a class of kernel smoothers (Gasser and Müller, 1984) that are well suited for our purpose; further details about these smoothers are provided in the appendix.

Theorem 2. *Using kernel smoothers (A-1, see appendix), for any two random functions X_i and X_j and their smoothed estimates $\hat{X}_i$ and $\hat{X}_j$, let*

$$\begin{aligned} \tilde{\Delta}_{ij}(s) &= E\left(\int_{\mathcal{I}} \left(\hat{X}_i(t,\tau_i) - \hat{X}_j(t-s,\tau_j)\right)^2 \mathrm{dt} \,\middle|\, \tau_i, \tau_j\right), \\ \tilde{s}_{ij} &= \arg\min_s \tilde{\Delta}_{ij}(s). \end{aligned} \tag{3.3}$$

Under assumptions (A1) - (A6) (see appendix), it holds that

$$|\tilde{s}_{ij} - s^*_{ij}| = O_p(n^{-\frac{1}{5}}) + O(\delta). \tag{3.4}$$

For the proof, we refer to the appendix. Theorem 2 says that the minimizers $\tilde{s}_{ij}$ asymptotically track the minimizing pairwise time shifts s_{ij} and therefore according to Theorem 1 also the true shifts s^*_{ij}, up to an asymptotically small error. Together with equations (3.1) and (3.2) this suggests a least squares approach.

With relative time shifts $\tilde{s}_{ij}$, model (3.2) becomes

$$\tilde{\boldsymbol{s}} = \boldsymbol{A}\boldsymbol{\tau} + \boldsymbol{\varepsilon}, \tag{3.5}$$

where $\tilde{\boldsymbol{s}}$ is the vector of estimated pairwise relative time shifts obtained from (3.3), and $\boldsymbol{\varepsilon}$ is the corresponding vector of errors. The least squares estimator $\hat{\boldsymbol{\tau}}$ of $\boldsymbol{\tau}$ is then

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\tilde{\boldsymbol{s}}. \tag{3.6}$$

Corollary 1. *The least squares time shifts $\hat{\tau}_i$ satisfy*

$$\hat{\tau}_i = \tau_i + O_p(n^{-\frac{1}{5}}) + O_p(\delta), \quad i = 2, \ldots, K. \tag{3.7}$$

The proof is in the appendix. Relative time shift estimates are practically implemented based on the aligned smoothed trajectories $\hat{X}_i$, $\hat{X}_j$ via

$$\hat{s}_{ij} = \arg\min_s \int_{\mathcal{T}} \left(\hat{X}_i(t) - \hat{X}_j(t-s)\right)^2 \mathrm{dt}, \tag{3.8}$$

which are entered on the right hand side of the least squares equation (3.5). Then the least square solution is $\hat{\boldsymbol{\tau}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\hat{\boldsymbol{s}}$, where $\hat{\boldsymbol{s}}$ is defined from (3.8) in analogy to $\boldsymbol{s}^*$ as defined in (3.2).

The gene-specific time shifts can alternatively be estimated through weighted least squares. Introducing the estimated squared distances $\hat{d}_{ij} = \int_{\mathcal{T}} \left(\hat{X}_i(t) - \hat{X}_j(t - \hat{s}_{ij})\right)^2 \mathrm{dt}$ resulting from (3.8), we define case weights $w_{ij} = 1/\hat{d}_{ij}$. Then the weighted least squares solution is

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A})^{-1}(\boldsymbol{A}'\boldsymbol{W}\hat{\boldsymbol{s}}), \tag{3.9}$$

where $\boldsymbol{W}$ is a diagonal matrix with diagonal elements w_{ij}. Updated estimates for pairwise time shifts $s^*_{ij} = \tau_i - \tau_j$ are then obtained by $\hat{s}^*_{ij} = \hat{\tau}_i - \hat{\tau}_j$. These estimates are symmetric, as $\hat{s}^*_{ij} = -\hat{s}^*_{ji}$.

In the analysis of temporal gene expression profiles, one may be interested in finding clusters of genes, which can be defined as genes with similar time shifts. Reaching high expression levels at about the same time might imply that such genes are related in function and may be involved in a common genetic pathway. Cluster analyses, based on various methods such as k-means, hierarchical or nonparametric density estimation based clustering (Fukunaga and Hostetler, 1975; Wong and Lane, 1983) may be performed in a second step, based on estimated time shifts $\hat{\tau}_i$.

4 Simulation Studies

To compare the performance of distance and other similarity criteria, we conducted two small simulation studies.

Simulation 1: We generated a set of curves ($K = 21$) with known non-random time shifts τ_i, $\tau_i = 0.02(i-1)$, $i = 1, 2, \ldots, 21$, on the interval $[-0.2, 1.2]$:

$$Y_i(t_{im}, \tau_i) = \xi_{1i} sin(\pi(t_{im} - \tau_i)) + \xi_{2i} sin(2\pi(t_{im} - \tau_i)) + e_{im},$$

where $\xi_{1i} \sim Lognormal(1, 0.3)$, $\xi_{2i} \sim Lognormal(0, 0.5)$, $e_{im} \sim N(0, 0.04)$ and data were sampled on an equi-spaced measurement grid t_{im} on $[-0.2, 1.2]$, $m = 1, \ldots, 71$. In this setup, the mean function is $\mu(t) = E(\xi_1)sin(\pi t) + E(\xi_2)sin(2\pi t)$, $Z(t) = (\xi_1 - E(\xi_1))sin(\pi t) + (\xi_2 - E(\xi_2))sin(2\pi t)$, where $E(\xi_1) = \exp(\frac{9}{200})$ and $E(\xi_2) = \exp(\frac{1}{8})$ and e_{im} are i.i.d. errors. The number of Monte Carlo runs was $M = 100$.

The normalized smoothed curves for one data set are shown in a time-ordered manner in the left panel of Figure 1. We aligned the normalized smoothed curves using distances d (2.2), and also distances based on correlation and rank correlation. For each criterion, we obtained the least squares solution for the time shifts, assuming $\tau_1 = 0$ in the model. The performance of each criterion was measured by the observed mean squared prediction errors for $K = 21$, $M = 100$, $MSPE = 1/(KM) \sum_{mc=1}^{M} \sum_{i=1}^{K} (\hat{\tau}_i - \tau_i)^2$.

Both correlation and rank correlation measures (Heckman and Zamar, 2000),

as alternative distances led to larger MSPEs as compared to the proposed distances d (2.2). We also compared the least squares estimates computed with and without case weights. The introduction of case weights reduced MSPE for all criteria, especially for the proposed L^2 distance d. The upper panel of Table 1 provides the comparisons for various similarity measures with regard to MSPE.

Table 1: Comparisons of different measures with or without case weights

Criteria	MSPE (w/o weights)	MSPE (w/ weights)
simulation 1		
d	0.0068	0.0035
correlation	0.0071	0.0062
rank correlation	0.0067	0.0064
simulation 2		
$Z(t)$: three points moving average		
d	0.0039	0.0015
correlation	0.0039	0.0029
rank correlation	0.0039	0.0037
$Z(t)$: nine points moving average		
d	0.0038	0.0014
correlation	0.0038	0.0029
rank correlation	0.0038	0.0036
$Z(t)$: fifteen points moving average		
d	0.0038	0.0014
correlation	0.0038	0.0029
rank correlation	0.0038	0.0036

The right panel of Figure 1 shows estimated time order using weighted least squares based on distance d for one set o simulated curves. It is evident that the original time order is well recovered.

Simulation 2: We also carried out a second simulation to investigate the behavior for less smooth trajectories Z than those considered in simulation 1, with Z constructed as moving averages of Gaussian white noise $N(0, 1)$ with varying span sizes. The mean curve $\mu(t)$ and errors were defined as in simulation 1. We report the results for three different levels of smoothness of Z in the second panel of Table 1. The results are quite similar as those of simulation 1. The distance d (2.2) again led to smaller MSPEs, as compared to correlation and rank correlation measures, and the introduction of case weights reduced MSPE for all criteria, especially for distance d, which emerged as the overall best distance for time alignment in both simulation studies and was used in the following applications. We conclude that our procedure is robust with regard to differing levels of smoothness of Z.

5 Time Ordering for *Drosophila* Embryonic Developmental Genes

Arbeitman and colleagues reported cDNA array gene expression patterns for nearly one-third of all 4028 *Drosophila melanogaster* genes during a complete time course of development (Arbeitman *et al.*, 2002), covering 66 sequential time periods beginning at fertilization and spanning the embryonic, larval, and pupal periods and the first 30 days of adulthood. In the first hours of embryonic development between fertilization and gastrulation (about 6-7 hours after fertilization), gene expression

is highly dynamic (Brody, 1996; Weigmann *et al.*, 2003). Two broad categories of transcripts are present at this time: those deposited into the egg during oogenesis (produced by maternal genes) and those that are expressed only after fertilization (produced by zygotic genes). To illustrate the proposed methodology, we time-ordered 27 strict maternal genes (including *swallow*) and 21 transiently expressed zygotic genes identified by Arbeitman *et al.* We also included the maternal anterior group genes, namely *bioid*, *swallow* and *exuperantia* (no data were available for *staufen* and *exuperantia-like*) (Brody, 1996). The gene expression patterns of these 50 genes for the first 10.5 hours of the fly embryo stage are displayed in Figure 2.

Here, it is natural to set the time shift of the first strict maternal gene as zero. The obtained time orders for the *Drosophila* genes are given in Table 2. The gene expression patterns are depicted in Figure 3, ordered by their estimated time shifts. The genes with complete peaks in the lower part of the figure are zygotic and the genes with "half" peaks in the upper half are maternal. We note that the time shift for zygotic genes starts at about 2 hours after fertilization, in accordance with the time when zygotic transcription initiates the switch from maternal to zygotic control of mitotic cycles (Foe *et al.*, 1993).

We also demonstrate another application of the proposed methodology by including the anterior genes in the alignment. The anterior system is one of the four maternal systems for assuring proper polarity of the oocyte prior to fertilization. *Bicoid* is the principal protein organizing anterior development in *Drosophila*. Di-

rectional action of *Exuperantia, Exuperantia-like, Swallow* and *staufen* are required during the process, in which *bicoid* mRNA is transported along the microtubule network of the oocyte to its anterior pole (Brody, 1996). The recovered early peakings of *swallow* and *exuperantia* prior to *bicoid* confirm the known pathway in terms of time order of expression.

The time order analysis also reveals that one gene (CG1624 or DPLD) previously identified as maternal, displays a relatively large lagged shift, which is a typical feature of zygotic genes. The pattern of this gene is very similar to that of zygotic genes and it likely has been mis-classified in previous analysis as a maternal gene.

6 Time ordering for Yeast Cell Cycle Genes

We illustrate the proposed methods for a set of time-course microarray expression profiles of 90 yeast genes (α factor synchronized), which have been identified as cell cycle regulating genes using traditional methods, such as Northern blot analysis (Spellman *et al.*, 1998). The expression level of each gene was measured during a period from 0 to 119 minutes with 7-minute intervals. The smoothed normalized gene expression profiles for these 90 genes are shown in Figure 4. Although the cell cycle is a continuous process resulting from a sequence of biochemical events, it has traditionally been divided into four phases: G1, S, G2, and M. Of the 90 genes, 18 were previously known to be related to M/G1 phase regulation, 44 to G1

phase regulation, 8 to S phase regulation, 6 to S/G2 phase regulation, and 14 to G2/M phase regulation of the yeast cell cycle.

Without employing knowledge of phase assignment, when applying our algorithm to obtain time order characteristics for these genes, the phases were successfully recovered in the natural time order: G2/M $\rightarrow$ M/G1 $\rightarrow$ G1 $\rightarrow$ S $\rightarrow$ S/G2 $\rightarrow$ G2/M, as illustrated by the dot plot in Figure 5. A complete list of time orders for the cell cycle genes is in Table 3. Furthermore, we would expect that relatively small time shifts between two genes indicate that they are closely co-regulated. The pair-wise time order characteristics are visualized in Figure 6. One may discern four time clusters (darker squares) corresponding to G2/M, M/G1, G1, and S-S/G2 genes.

The gene expression profiles can be aligned in their time order as shown in Figure 7. We note that no sharp cut-off occurs between genes in different phases, i.e., there is a continuum of timing of expression. Assignment of a "phase boundary" gene to a certain cycle phase therefore appears subjective (see also Cooper and Shedden, 2003). Further studies on these "boundary" genes might reveal gene regulating mechanisms during phase transitions.

Besides contributing to the issue of "boundary genes", the time order analysis points to a few genes as likely having been misclassified into underlying cycle phases by previous methods. These include $G1$ phase genes YGL225W, YDR113C, and YJL173C, for which the time shifts fall within the cluster of S-S/G2 phase

genes, and YDL102W, which according to its time shift belongs to the cluster of M/G1 genes. In addition, two M/G1 genes, namely, YER111C and YNR044W, fall into the cluster of G1 genes. Moreover, the time shifts of genes YDR150W and YPR141C, previously classified as S/G2, are found to be considerably closer to those of G2/M and S genes, respectively. This demonstrates the utility of the inferred time order for clustering and classification,

7 Discussion and Conclusions

The timing of individual gene expression offers insights into the function of the proteins they encode. A recent study on the transcriptome of the intraerythrocytic developmental cycle (IDC) of the malaria parasite *Plasmodium falciparum* led Bozdech *et al.* (2005) to conclude that the "expression profiles for developmentally regulated genes in the IDC transcriptome reveal an orderly timing of key cellular functions" to ensure the completion of the *P. falciparum* IDC, and "groups of functionally related genes share common expression profiles". The timing of gene expression suggests potential target genes for drug discovery and new vaccine therapies as also pointed out by Bozdech *et al.*. These findings indicate there is a broad scope for applications of the proposed time ordering methodology.

Recognizing the importance of timing characteristics of gene expression, in this paper we developed methods to uncover the underlying time ordering. We define and estimate a global time shift for each gene that is rooted in pairwise curve

alignments and is obtained in a symmetrization step through minimization of a functional distance via least squares. The proposed approach is easy to implement. In practice, the interval of integration $\mathcal{T} = [T_1, T_2]$ can be determined by searching over different ranges and choosing the range that leads to a "best global alignment", e.g., minimizing the sum of all pairwise distances $\sum_{i=1}^{K} \sum_{j=1}^{K} \int_{\mathcal{T}} (X_i(t - \hat{\tau}_i) - X_j(t - \hat{\tau}_j))^2/(T_2 - T_1)$. Although we sought to specifically emphasize the global temporal order of gene expression, finding time clusters for classifying sets of genes by their time dynamics might also be of interest in some studies and can be based at least in part on the proposed time ordering.

We demonstrated the methodology for *Drosophila* embryonic development and yeast cell cycle gene expression profiles, as well as in simulations. Feasibility of the approach was demonstrated by successful recovery of time ordering. While recovering time ordering of genes in itself is not sufficient to construct gene regulatory networks, it contains valuable information towards that goal and may serve as an informative guiding tool for biologists to further explore and discover relevant regulatory relationships within certain genetic pathways. For example, the timeline of early embyonic development of *Drosophila* has been well studied and estimated time shifts of participating genes will assign them into various embyonic sub-stages, thereby informing about their respective functions.

Acknowledgments

We wish to thank an associate editor and two referees for helpful comments, and Ms. Karen Klein at Office of Research, Wake Forest University School of Medicine, for assistance in text editing. This research was supported in part by NSF grants DMS03-5448 and DMS05-05537.

References

AACH, J. AND CHURCH, G.M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics* **17**, 495-508.

ARBEITMAN, M.N., FURLONG, E.E.M., IMAM, F., JOHNSON, E., NULL, B.H., BAKER, B.S., KRASNOW, M.A., SCOTT, M.P., DAVIS, R.W., AND WHITE, K.P. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270-2275.

ARKIN, A. AND ROSS, J. (1995). Statistical construction of chemical-reaction mechanisms from measured time-series. *J. Phys. Chem* **99**, 970-979.

ARKIN, A., SHEN, P., AND ROSS, J. (1997). A test case of correlation metric construction of a reaction pathway from measurements. *Science* **277**, 1275-1279.

BÄHLER J. (2005). Cell cycle control of gene expression in budding and fission yeast. *Annu. Rev. Genet.* **39**, 69-94.

BAR-JOSEPH, Z., GERBER, G., SIMON, I., GIFFORD, D.K. AND JAAKKOLA,

T.S. (2003a). Comparing the continuous representation of time-series expression profiles to identify differently expressed genes. *Proc. Natn. Acad. Aci. U.S.A.* **100**, 10146-10151.

BAR-JOSEPH, Z., GERBER, G., GIFFORD, D.K., JAAKKOLA, T.S., AND SIMON, I. (2003b). Continuous representation of time-series gene expression data. *Journal of Computational Biology* **10**, 2-4.

BOZDECH, Z., LLINAS, M., PULLIAM, B.L., WONG, E.D., ZHU, J. AND DERISI, J.L. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLOS Biology* **1**, 85-100. First published on 18 August 2003, DOI: 10.1371/journal.pbio.0000005.

BRODY, T. B. (1996). *The Interactive Fly: A Cyberspace Guide to Drosophila Genes and Their Roles in Development*. Society for Developmental Biology. *URL: http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/aimain /1aahome.htm*.

BUTTE, A.J., BAO, L., REIS, B.Y., WATKINS, T.W. AND KOHANE, I.S. (2001). Comparing the similarity of time-series gene expression using signal processing metrics. *J. Biomed. Inf.* **34**, 396-405.

COOPER, S. AND SHEDDEN K. (2003). Microarray analysis of gene expression during the cell cycle. *Cell & Chromosome* **2**, 1-12.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

FOE, V.E., ODELL, G.M. AND EDGAR, B.A. (1993). *Mitosis and Morphogenesis in the Drosophila Embryo*, Edited by M. Bate and A. Martinez-Arias, Cold Spring Harbor Press.

FUKUNAGA, K. AND HOSTETLER, L.D. (1975). Estimation of the gradient of a density-function, with applications in pattern recognition. *IEEE Transactions on Information Theory* **IT21**, 32-40.

GASSER, T. AND MÜLLER, H.G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11**, 171-185.

GERVINI, D. AND GASSER, T. (2004). Self-modelling warping functions. *J. Roy. Statist. Soc. B* **66**, 959-971.

HECKMAN, N.E. AND ZAMAR, R.H. (2000). Comparing the shapes of regression functions. *Biometrika* **87**, 135-144.

LIU, X.L., MÜLLER, H.G. (2003). Modes and clustering for time-warped gene expression profile data. *Bioinformatics* **19**, 1937-1944.

LIU, X.L., MÜLLER, H.G. (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Am. Statist. Assoc.* **99**, 687-699.

LODISH, H., BERK A., ZIPURSKY, L.S., MATSUDAIRA, P., BALTIMORE, D., AND DARNELL J. (2000). *Molecular Cell Biology.* 4th Edition. W. H. Freeman and Company, New York, USA.

POLLARD, T.D. AND EARNSHAW, W.C. (2004). *Cell Biology.* 1st Edition. W.B. Saunders, Philadelphia, USA.

QIAN, J., DOLLED-FILHART, M., LIN, J., YU, H. and Gerstein, M. (2001). Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identified new, biologically relevant interactions. *J. Mol. Biol.* **314**, 1053-1066.

SILVERMAN, B.W. (1995). Incorporating parametric effects into functional principal components-analysis. *J. Roy. Statist. Soc. B* **57**, 673-689.

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 gene of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273-3297.

WANG, F.M. AND GASSER, T. (1997). Alignment of curves by dynamic time warping. *Annals of Statistics* **25**, 1251-1276.

WEIGMANN, K., KLAPPER, R., STRASSER, T., RICKERT, C., TECHNAU, G. M., JÄCKLE, H., JANNING, W. AND KLÄMBT, C. (2003). FlyMove – a new way to look at development of *Drosophila*. *Trends in Genetics* **19**, 310-311.

WONG, M.A., LANE, T. (1983). A k-th nearest neighbor clustering procedure. *J. Roy. Statist. Soc. B* **45**, 362-368.

YAO, F., MÜLLER, H.G. AND WANG, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Assoc.* **100**, 577-590.

Table 2: Estimated relative time shifts for *Drosophila* maternal (m) and zygotic (z) genes.

CG number*	Category	Time shift (hrs)	CG number	Category	Time shift (hrs)
CG7627	m	0.000	CG15737	m	1.408
CG10308	m	0.145	CG4916	m	1.409
CG4790	m	0.165	CG9925	m	1.491
CG2913	m	0.296	CG15634	z	1.943
CG8180	m	0.319	CG1745	z	1.961
CG3800	m	0.363	CG9506	z	1.990
CG2707	m	0.368	CG14025	z	2.352
CG4735	m	0.375	CG1624	m	2.431
CG9523	m	0.419	CG1378	z	3.002
CG18543	m	0.440	CG1677	z	3.044
CG8994(exu)	m	0.513	CG7288	z	3.045
CG7108	m	0.522	CG12750	z	3.148
CG10387	m	0.541	CG7626	z	3.197
CG6319	m	0.556	CG1078	z	3.328
CG14444	m	0.619	CG16901	z	3.449
CG3429(swa)	m	0.851	CG10417	z	3.598
CG5940	m	0.857	CG2916	z	3.610
CG3510	m	0.895	CG5175	z	3.623
CG7660	m	0.968	CG7269	z	3.708
CG1034(bcd)	m	1.049	CG13096	z	3.955
CG17018	m	1.146	CG4602	z	4.083
CG5568	m	1.187	CG14722	z	4.212
CG14764	m	1.221	CG11988	z	4.942
CG7730	m	1.281	CG9839	z	5.387
CG5263	m	1.341	CG8606	z	5.493

* CG number stands for Computed Gene Identifier.

Table 3: Estimated relative time shifts for yeast cell cycle genes

CG number	Phase*	Time shift (min)	CG number	Phase	Time shift (min)
YIL106W	G2/M	-19.500	YDL102W	G1	-5.543
YDR146C	G2/M	-19.496	YLR079W	M/G1	-5.132
YMR001C	G2/M	-19.330	YNL192W	M/G1	-4.059
YPR119W	G2/M	-19.112	YBR083W	M/G1	-4.000
YLR131C	G2/M	-19.045	YDL127W	M/G1	-3.298
YGR108W	G2/M	-18.397	YCL055W	G1	-2.729
YOR058C	G2/M	-17.747	YCL027W	M/G1	-2.574
YHR152W	G2/M	-17.652	YLR452C	M/G1	-2.465
YGL116W	G2/M	-17.360	YNL327W	M/G1	-2.159
YAL040C	M/G1	-15.435	YKL045W	G1	-1.206
YGR092W	G2/M	-14.535	YKL042W	G1	-1.136
YDR150W	S/G2	-13.141	YGR109C	G1	-1.087
YDR033W	G2/M	-12.394	YNL082W	G1	-0.558
YBR202W	M/G1	-12.307	YER111C	M/G1	-0.355
YBR054W	G2/M	-11.051	YDR309C	G1	0
YJL194W	M/G1	-10.848	YBR067C	G1	0.051
YAR018C	G2/M	-10.801	YBL035C	G1	0.236
YDR077W	G2/M	-10.584	YDL197C	G1	0.268
YFL026W	M/G1	-9.542	YPR120C	G1	0.333
YLR274W	M/G1	-9.329	YDR097C	G1	0.798
YNL145W	M/G1	-7.888	YJL115W	G1	0.868
YJL157C	M/G1	-7.819	YAR007C	G1	1.028
YDL179W	M/G1	-7.294	YNL102W	G1	1.240
YKL185W	M/G1	-7.197	YNR044W	M/G1	1.395

* Previously identified by traditional methods (Spellman et al., 1998).

Table 3 (con't): Estimated relative time shifts for yeast cell cycle genes

CG number	Phase*	Time shift (min)	CG number	Phase	Time shift (min)
YLR103C	G1	1.752	YDL055C	G1	8.374
YKL113C	G1	2.172	YDR356W	G1	9.638
YPL153C	G1	2.175	YPR159W	G1	10.635
YDL003W	G1	2.235	YJL092W	G1	10.733
YPR175W	G1	2.317	YLR342W	G1	10.967
YBR088C	G1	2.465	YPR141C	S/G2	11.363
YJL187C	G1	2.500	YBL002W	S	11.807
YDL164C	G1	2.762	YBR010W	S	11.905
YER070W	G1	2.811	YNL031C	S	12.135
YPL256C	G1	3.119	YGL225W	G1	12.171
YER095W	G1	3.281	YBL003C	S	12.648
YER001W	G1	3.740	YNL030W	S	12.699
YIL066C	G1	3.944	YDR113C	G1	12.726
YMR199W	G1	4.097	YBR009C	S	12.890
YOL090W	G1	4.866	YDR224C	S	12.903
YNL262W	G1	5.057	YDR225W	S	13.352
YKL101W	G1	6.643	YMR198W	S/G2	14.782
YOR074C	G1	6.929	YKL096W-A	S/G2	14.938
YGL163C	G1	7.852	YJL173C	G1	15.239
YNL312W	G1	8.192	YKL096W	S/G2	16.936
YMR307W	G1	8.338	YLR210W	S/G2	17.225

* Previously identified by traditional methods (Spellman et al., 1998).

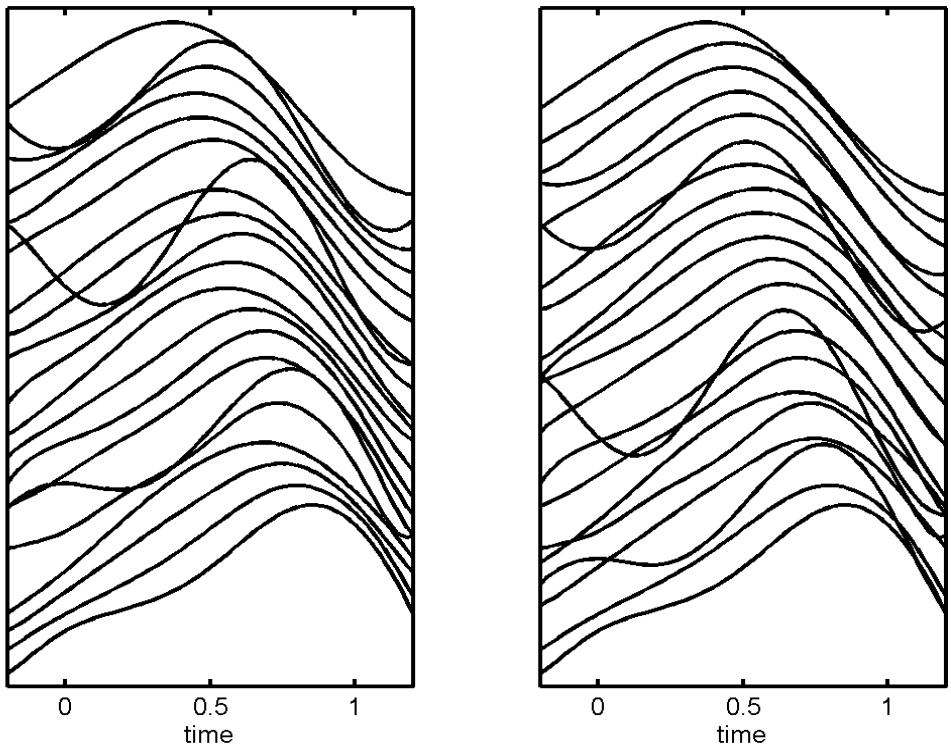

Figure 1: Illustrating time shifts for a set of simulated data. *Left panel: normalized smoothed curves plotted in a vertical sequence corresponding to increasing true time shifts τ_i from top to bottom, so that timewise leading profiles appear on the top and lagged profiles at the bottom; right panel: normalized smoothed curves plotted in the same way but with estimated time shifts $\hat{\tau}_i$ (3.9).*

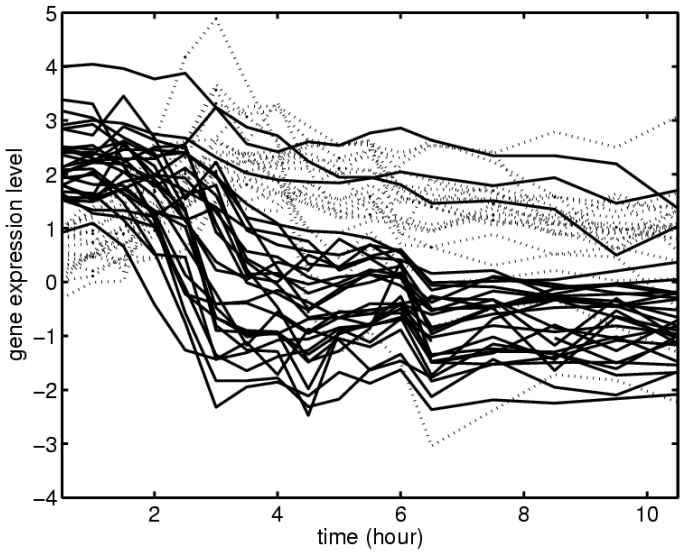

Figure 2: *Drosophila* gene expression profiles during the first 10.5 hours. *Solid: strict maternal genes (including maternal anterior genes); dotted: transient zygotic genes.*

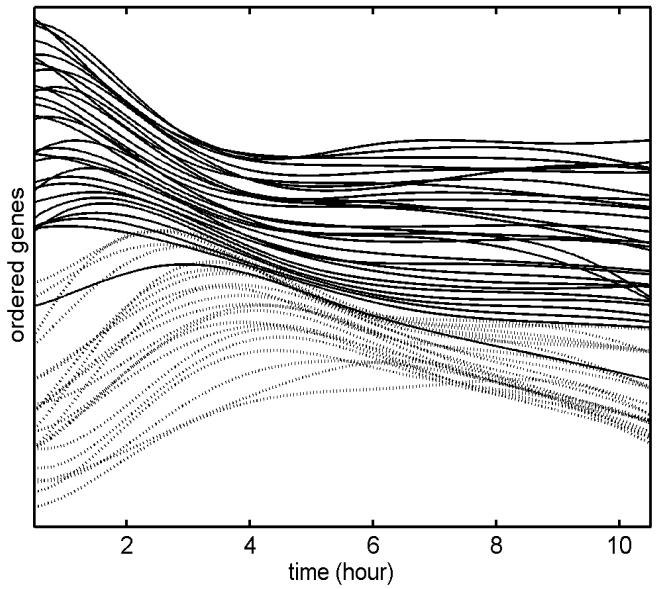

Figure 3: Time-ordered maternal and zygotic gene expression profiles of *Drosophila. The curves in the figure are normalized and smoothed gene expression profiles, ordered vertically from top to bottom according to increasing estimated time shifts. Profiles of early expressed genes appear near the top, those of late expressed genes near the bottom. Solid: maternal genes; dotted: zygotic genes.*

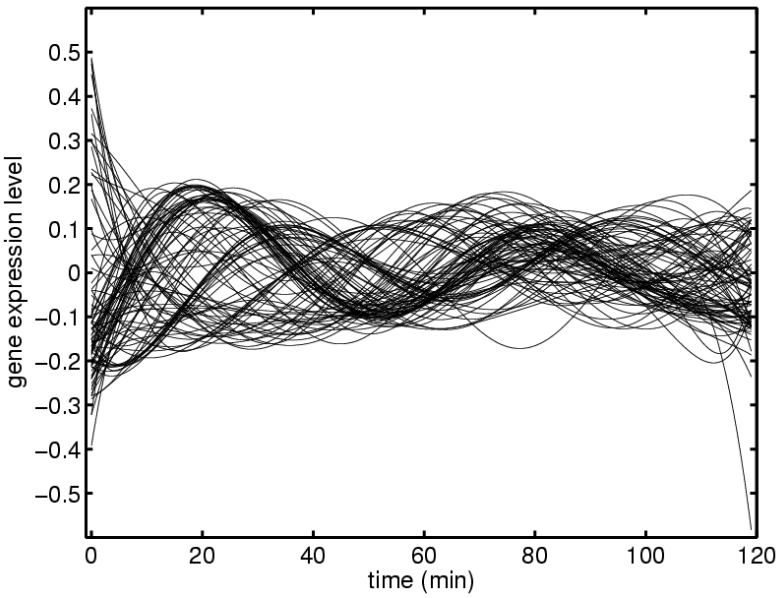

Figure 4: Normalized smoothed yeast cell cycle gene expression profiles.

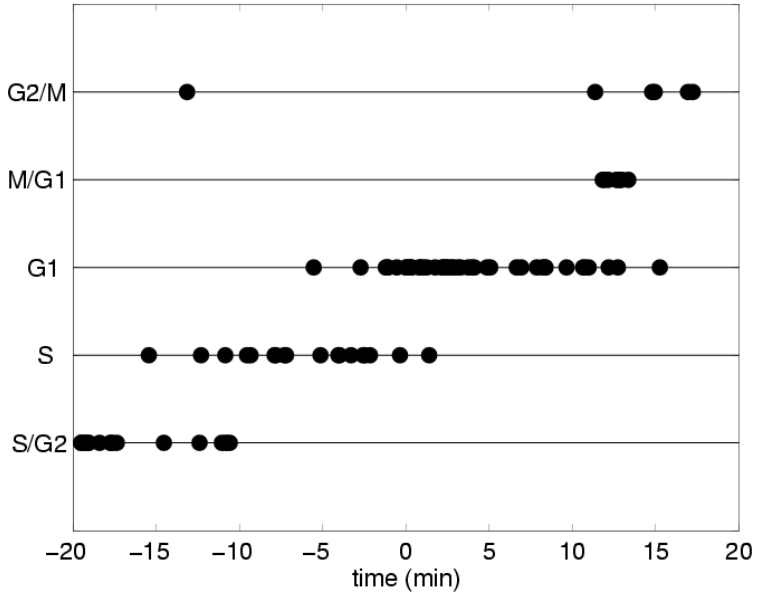

Figure 5: Dot plot of estimated time shifts $\hat{\tau}_i$ (3.9) ordered according to cell cycle phases. *The time order is G2/M → M/G1 → G1 → S → S/G2.*

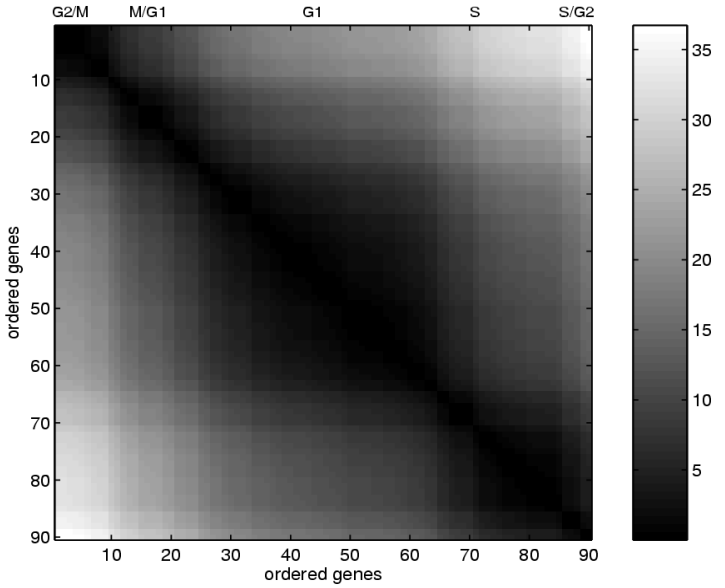

Figure 6: Relative time shift image plot for time-ordered yeast cell cycle gene expression profiles. *From top to bottom and from left to right, genes are time-ordered by increasing estimated gene-specific time shifts $\hat{\tau}_i$ (3.9). The plot shows estimated relative time shifts between pairs of genes as indicated by gray scale (darker indicates smaller relative time shifts, see gray scale bar on the right). The darker squares correspond to time clusters of genes, i.e., genes whose expression profiles have similar time shifts.*

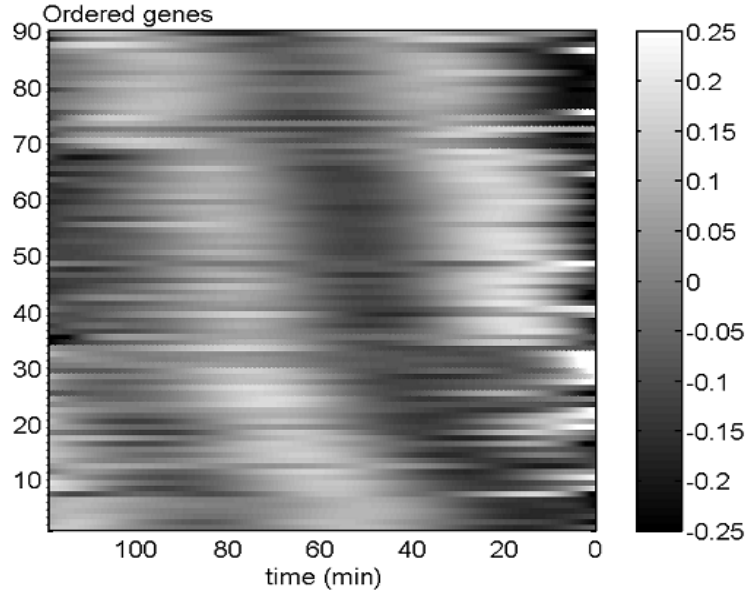

Figure 7: Time-ordered normalized smoothed yeast cell cycle gene expression profiles. *From bottom to top, gene expression profiles are ordered according to increasing estimated time shifts $\hat{\tau}_i$ (3.9). The gene expression level is gray-scale-coded (see color bar on the right). Note the shifting patterns in peak locations.*