Biostatistics (2009), **10**, 1, *pp.* 32–45
doi:10.1093/biostatistics/kxn011
Advance Access publication on May 22, 2008

Time-synchronized clustering of gene expression trajectories

RONG TANG*

Division of Biostatistics, Center for Devices and Radiological Health,
Food and Drug Administration, Rockville, MD 20850, USA
rong.tang@fda.hhs.gov

HANS-GEORG MÜLLER

Department of Statistics, University of California–Davis,
One Shields Avenue, Davis, CA 95616, USA

Summary

Current clustering methods are routinely applied to gene expression time course data to find genes with similar activation patterns and ultimately to understand the dynamics of biological processes. As the dynamic unfolding of a biological process often involves the activation of genes at different rates, successful clustering in this context requires dealing with varying time and shape patterns simultaneously. This motivates the combination of a novel pairwise warping with a suitable clustering method to discover expression shape clusters. We develop a novel clustering method that combines an initial pairwise curve alignment to adjust for time variation within likely clusters. The cluster-specific time synchronization method shows excellent performance over standard clustering methods in terms of cluster quality measures in simulations and for yeast and human fibroblast data sets. In the yeast example, the discovered clusters have high concordance with the known biological processes.

Keywords: Clustering; Gene expression analysis; Microarray; Time warping.

1. Introduction

DNA microarray data are collected through microscopic DNA probes attached to a solid surface, forming an array for the purpose of expression profiling. In a temporal microarray experiment, the arrays are collected over a period of time (Storey *and others*, 2005). These experiments expose thousands of genes to an experimental condition simultaneously, and the dimensions of such studies pose great challenges. Cluster analysis has become a popular dimension reduction tool in microarray studies, and recently there has been interest in classification and clustering especially of gene expression time course data (Luan and Li, 2003; Ma *and others*, 2006). Hierarchical clustering (Johnson, 1967) and K-means clustering (Marriott, 1982) are widely used to identify functionally related gene groups. These nonmodel-based

*To whom correspondence should be addressed.

cluster algorithms have 2 important components: linkage and distance measures (Jain *and others*, 1999). We use average linkage throughout this study and propose a new distance measure for situations where expression trajectories exhibit patterns of time variation.

Time variation or warping is addressed in functional data analysis, an area of statistical research that focuses on data in the form of continuous functions or curves. Even though gene expressions can only be observed at discrete time points, they can be viewed as functional data because the underlying biological process is usually thought to be continuous. Curve alignment or time warping (Sakoe and Chiba, 1978; Gasser and Kneip, 1995; Ramsay and Li, 1998; Gervini and Gasser, 2004, 2005) is often used as a preliminary step in functional data analysis to reduce time variation. Versions of time warping have been applied to gene expression data (Leng and Müller, 2006). However, in the presence of multiple shape patterns, time warping is often unidentifiable due to the confounding of time and shape variation. This motivates the notion of "shape clusters" to represent groups of genes that share similar patterns, regardless of their individual time dynamics. The dilemma is that we cannot identify the shape clusters without first removing time variation; on the other hand, we cannot remove time variation without causing some degree of shape distortion if the cluster structure is unknown.

There are a number of methods that incorporate time delay into a similarity measure such as the event method (Kwon *and others*, 2003) or the time-delayed correlation method (Li *and others*, 2006). In these approaches, the continuous biological processes are discretized, which may lead to potential loss of important information. The time delay is restricted to be a multiple of the time interval between observed time points. Moreover, after adjusting for time variation, the distance between any 2 genes decreases, regardless of whether they have similar shapes. This overall decrease of pairwise distance will affect the ability of an algorithm to identify groups with different shapes. A mixed-effects model that contains random parameters for scaling and translation in time was proposed by Chudova *and others* (2003). Assumptions include a multivariate normal distribution and that there is no correlation between 2 different measurements taken from the same subject, which is unrealistic for the time course data. Possible misalignment of time intervals is another problem.

The time-synchronized clustering method proposed here successfully combines the competing goals of curve alignment and shape clustering, by identifying clusters that contain gene profiles with similar shapes and a variety of time dynamic patterns within a cluster but not across clusters. This is biologically sensible as we shall demonstrate with gene expression time courses from yeast and human fibroblasts. The proposed approach treats the expression profiles as continuous functions of time, and the individual time dynamics are modeled by a monotone increasing function, the time warping function. A cluster-specific warping procedure is proposed which implements time warping only within small neighborhoods, where the neighborhood relation is determined by the L^2 distance of trajectories after a pairwise warping step. The performance of this method in both simulations and real time course gene expression data indicates that it can successfully identify the true shape clusters and produce tighter and better defined clusters than do K-means and hierarchical clustering, both of which use conventional distance measures.

2. Models and methods

In this section, we introduce a model that accommodates data sets with multiple shape patterns (clusters) and allows for time heterogeneity among curves within the same cluster. Let $Y_1,\ldots, Y_n$ denote n continuous random functions defined on a closed real interval $\mathcal{T} = [0, T]$. These curves are observed at time points $t_j = \frac{j-1}{m-1}T$, $j = 1, \ldots, m$, under additional random errors, with measurements

$$y_{ij} = Y_i(t_j) + \epsilon_{ij} = X_i(h_i^{-1}(t_j)) + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{2.1}$$

where the random errors ϵ_{ij} are assumed to be independent and to satisfy $E\epsilon = 0$ and $E\epsilon^2 = \sigma^2 < \infty$. In this model, X_i are underlying random functions that define amplitude variation of the curves and h_i are random warping functions, that is, transformations of the time axis, where $h_i\colon \mathcal{T} \to \mathcal{T}$. The inverse warping functions h_i^{-1} are assumed to exist and define the actual timescale of Y_i.

The amplitude variation functions X_i follow the model

$$X_i(t) = V_i(t) + \delta Z_i(t), \quad i = 1, \dots, n, \ \ t \in \mathcal{T}, \tag{2.2}$$

where $V_i = \mu_l$ for any Y_i in cluster l and μ_l are fixed twice continuously differentiable cluster-specific shape functions that satisfy $\int_{\mathcal{T}_o} \mu_l'(t)^2\,\mathrm{d}t > 0$ for any nondegenerate interval $\mathcal{T}_o \subseteq \mathcal{T}$, and there exists $C \gg \delta$ such that $\int_{\mathcal{T}} (\mu_{l_1}(t) - \mu_{l_2}(t))^2\,\mathrm{d}t \geqslant C > 0$ for clusters $l_1 \neq l_2$. The coefficient δ is a small positive constant, while Z_i are twice continuously differentiable independent random processes that satisfy $EZ_i(t) = 0$, $EZ_i^2(t) < \infty$, $EZ_i'(t)^2 < \infty$ and $EZ_i''(t)^2 < \infty$, for $t \in \mathcal{T}$. We may normalize these processes by requiring $\int_{\mathcal{T}} E(Z_i^2(t))\,\mathrm{d}t = 1$, thus also providing a unique specification of δ. Processes Z_i are assumed to be independent of h_i. Note that δZ_i plays the role of a small additional amplitude variation process. We refer to Section 5 for more discussion. For the practically important case where one has discrete measurements as in (2.1), we include a presmoothing step (see, e.g. Leng and Müller, 2006).

Denoting the true cluster membership of Y_i by M_i, we have $E(X_i(t)|M_i = l) = \mu_l(t)$. Additionally, for identifiability, we assume

$$E(h_i(t)|M_i = l) = t, \quad 1 \leqslant i \leqslant n, \ \ 1 \leqslant l \leqslant L. \tag{2.3}$$

The size of cluster l is given by

$$n_l = \sum_{i=1}^{n} \mathbf{1}_{\{M_i = l\}}, \quad l = 1, \dots, L, \tag{2.4}$$

with $0 < n_l < n$ and $\sum_{l=1}^{L} n_l = n$. As mappings between timescales, the warping functions h_i should satisfy the common endpoints condition, that is, $h_i(0) = 0, h_i(T) = T$, and the strict monotonicity condition, that is, $h_i(t_{j1}) < h_i(t_{j2})$, for $0 \leqslant t_{j1} < t_{j2} \leqslant T$. Illustrating these concepts, Figure 1 displays simulated data with 4 different shape patterns, distorted by a considerable amount of time variation within each cluster.

2.1 *Pairwise warping*

Pairwise warping is a curve synchronization method that utilizes relative time dynamics of every curve pair to arrive at a global warping function (Tang and Müller, 2008). For any 2 curves Y_i, Y_k in the same cluster, the underlying pairwise warping function g_{ik} is a composite of the 2 individual warping functions, $g_{ik}(t) = h_i(h_k^{-1}(t))$. Estimates of pairwise warping functions are obtained by minimizing a target function $C(Y_i, Y_k, g)$, subject to the monotonicity and endpoint constraints

$$\widetilde{g}_{ik}(t) = \arg\min_{g} C(Y_i, Y_k, g), \tag{2.5}$$

where g is required to satisfy $g_{ik}(0) = 0, g_{ik}(T) = T$, and $g_{ik}(t_1) < g_{ik}(t_2), 0 < t_1 < t_2 < T$, where entire trajectories $Y_i, i = 1, \dots, n$, are assumed to be either directly observed or obtained after a smoothing step, by applying a standard smoothing method. The target function to be minimized is

$$C_\lambda(Y_i, Y_k, g) = E\left\{ \int_{\mathcal{T}} (Y_i(g(t)) - Y_k(t))^2 + \lambda(g(t) - t)^2\,\mathrm{d}t | Y_i, Y_k \right\},$$

where $\lambda \geqslant 0$.

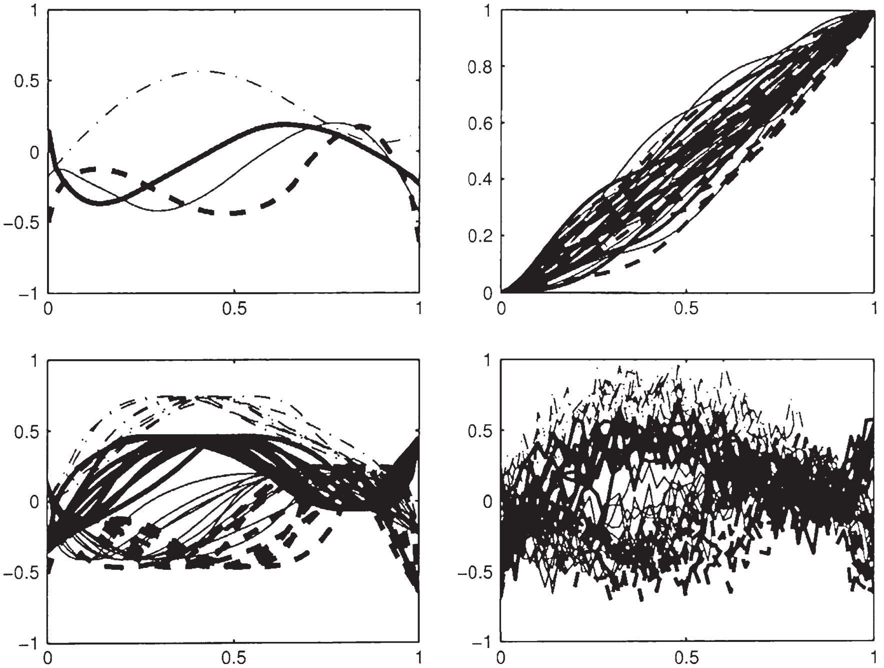

Fig. 1. An example of simulated clustered random functions. Upper left: mean functions of 4 clusters; upper right: 40 generated time warping functions; lower left: 40 simulated clustered random functions without noise; lower right: 40 simulated clustered random functions with noise. Mean functions μ_1 for cluster 1 (dashed–dotted line), μ_2 for cluster 2 (bold dashed line), μ_3 for cluster 3 (solid line), and μ_4 for cluster 4 (bold solid line).

We note that $C_\lambda(Y_i, Y_k, g)$ and therefore $\widetilde{g}_{ik}$ are not symmetric in i and k, as $\widetilde{g}_{ik}$ and $\widetilde{g}_{ki}$ have different reference trajectories. A measure of proximity between the 2 trajectories Y_i and Y_k is the L^2 distance after aligning Y_i to the reference Y_k,

$$\widetilde{d}_{\mathrm{pw}}(i,k) = \left\{ \int_{\mathcal{T}} (Y_i(\widetilde{g}_{ik}(t)) - Y_k(t))^2 \, \mathrm{d}t \right\}^{\frac{1}{2}}. \tag{2.6}$$

As long as the 2 curves being aligned are from the same cluster, their distances after aligning one to the other are expected to be small because the main source of variation is time variation. However, if they come from different clusters, the shape variation between them cannot be accounted for with monotone transformations $\widetilde{g}_{ik}$, due to the fact that $\int_{\mathcal{T}} \left(\mu_{l_1}(t) - \mu_{l_2}(t)\right)^2 \mathrm{d}t \geqslant C > 0$ for $l_1 \neq l_2$. This results in much larger distances $\widetilde{d}_{\mathrm{pw}}$ for pairs of curves that belong to different clusters after their alignment. While $\widetilde{d}_{\mathrm{pw}}$ is not symmetric in i and j, a bona fide symmetric distance between time-synchronized trajectories is defined in (2.10) and provides the basis for the actual clustering step of our algorithm.

2.2 *Cluster-specific warping*

Clusters in gene expression trajectories usually display multiple shape patterns, and directly applying time warping may not succeed in synchronizing the curves (upper left panel of Figure 2) and may even cause shape distortions. To address this problem, we propose a "cluster-specific warping" procedure.

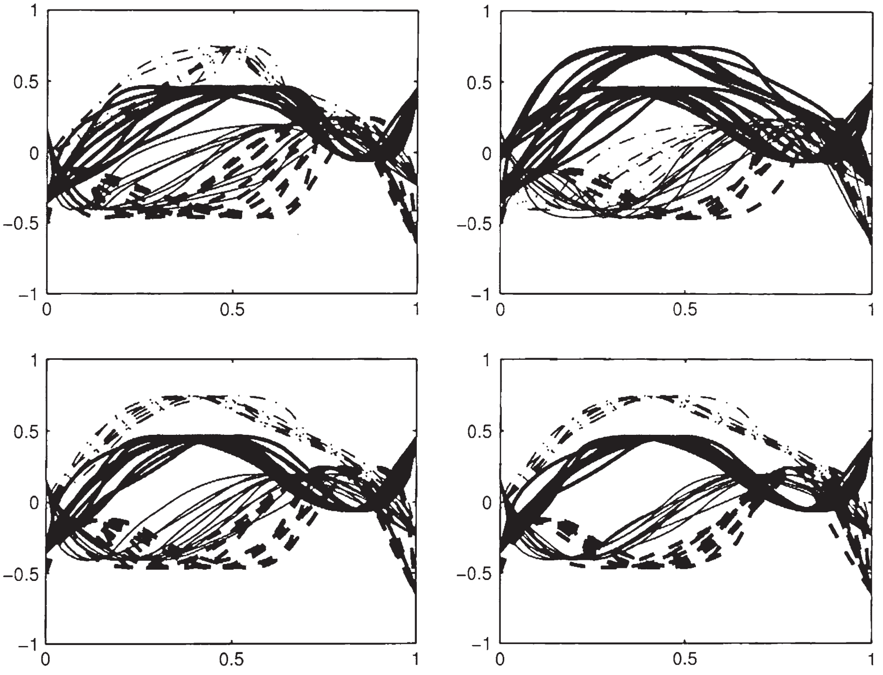

Fig. 2. Results for simulations, using as input the simulated data from the lower right panel of Figure 1. Upper left: application of pairwise warping to all data; upper right: K-means clustering with L^2 distance; lower left: cluster-specific warping with threshold $d_o = 1$; lower right: pairwise warping within true clusters—cluster 1 (dashed–dotted lines), cluster 2 (solid dashed lines); cluster 3 (solid lines); cluster 4 (bold solid lines).

If cluster membership information is available, according to (2.3), $E(g_{ik}(t)|h_i, h_k, M_i = M_k) = t$ becomes $E(h_i(h_k^{-1}(t))|h_i, h_k, M_i = M_k) = h_k^{-1}(t)$ after replacing t with $h_k^{-1}(t)$. This motivates to estimate the individual cluster-specific warping function for trajectory Y_k through

$$\widetilde{h}_k^{-1}(t) = \frac{1}{n_l}\sum_{i=1}^{n} \widetilde{g}_{ik}(t)\mathbf{1}_{\{M_i=M_k=l\}}. \tag{2.7}$$

After numerical inversion to obtain estimated cluster-specific warping functions $\widetilde{h}_k$, we achieve very good synchronization and removal of time variation within clusters (lower right panel of Figure 2). Implementing the estimates in (2.7) requires knowledge of true cluster membership, which is the objective of the clustering procedure in the first place. The L^2 distances after pairwise warping $d_{\text{pw}}(i,k)$ are expected to behave quite differently for pairs from the same cluster as compared to pairs from different clusters. This motivates to estimate cluster-specific warping functions h_k based exclusively on pairs for which $d_{\text{pw}}(i,k) < d_o$, for a preselected threshold d_o, leading to

$$\widehat{h}_k^{-1}(t) = \frac{1}{\sum_{i=1}^{n}\mathbf{1}_{\{d_{\text{pw}}(i,k)<d_o\}}}\sum_{i=1}^{n} \widetilde{g}_{ik}(t)\mathbf{1}_{\{d_{\text{pw}}(i,k)<d_o\}}, \tag{2.8}$$

where d_o serves as a cutoff point for distances after initial pairwise warping.

2.3 *Data synchronization and clustering*

Once we obtain cluster-specific warping functions, trajectories are time synchronized via

$$Y_k^*(t) = Y_k(\widetilde{h}_k^{-1}(t)), \quad k = 1, \ldots, n. \tag{2.9}$$

In the simulated example (lower right panel of Figure 1), we find that amplitude variation, that is, the variation caused by the presence of multiple shape patterns, is enhanced in the time-synchronized curves (lower left panel of Figure 2). Time-synchronized L^2 distances are defined as

$$d_t(i,k) = \left\{ \int_{\mathcal{T}} (Y_i^*(t) - Y_k^*(t))^2 \, \mathrm{d}t \right\}^{\frac{1}{2}}, \quad 1 \leqslant i, k \leqslant n. \tag{2.10}$$

Because the threshold in (2.8) is usually chosen conservatively in order to prevent shape distortion, the curves resulting from cluster-specific warping are not as tightly synchronized as those resulting from warping within clusters under known cluster structure (lower panels in Figure 2).

We note that after the time-synchronizing transformation, one can apply any preferred clustering method for the final determination of the clusters.

2.4 *Choice of threshold*

The selection of the threshold d_o is important as it determines which pairwise warping functions will be included in (2.8). A large threshold may lead to shape distortion, while a small threshold may not sufficiently adjust for time variability. In our experience, it is more important to prevent shape distortion, and therefore we advocate choosing a relatively small threshold d_o.

Assuming there are L clusters, the following consideration provides a lower bound for the number of pairs belonging to the same cluster. In the entire data set of n curves, there are $n(n-1)$ possible combinations of curve pairs. Pairs (Y_i, Y_k) and (Y_k, Y_i) count as different combinations because the outcome of pairwise warping between a curve pair is dependent on which of the 2 curves is used as reference. Among these $n(n-1)$ curve pairs, the total number of pairs belonging to the same cluster is $\sum_{l=1}^{L} n_l(n_l - 1) = \sum_{l=1}^{L} n_l^2 - n$.

One can show that if the data are partitioned into L clusters, then the number of curve pairs belonging to the same cluster is

$$\sum_{l=1}^{L} n_l^2 - n \geqslant L \sqrt[L]{n_1^2 n_2^2, \ldots, n_L^2} - n \geqslant \frac{n(n-L)}{L},$$

where the lower bound is achieved for $n_1 = n_2 = \cdots = n_L = \frac{n}{L}$.

As a consequence, regardless of cluster size distribution, one has at least $\frac{n(n-L)}{L}$ pairs of curves that belong to the same cluster. Let $\mathbf{D}$ be a $1 \times n(n-1)$ vector that contains the distances after pairwise warping between all possible curve pairs and $\mathbf{D}_k = (d_{\mathrm{pw}}(1,k), \ldots, d_{\mathrm{pw}}(n,k))$ be vectors of distances after pairwise warping of $Y_i, i = 1, \ldots, n$, against Y_k. Denote the qth quantile of a data sample $U = (u_1, \ldots, u_n)$ in the sense of Hyndman and Fan (1996) by $Q(U, q)$. The pairs (i,k) for which $d_{\mathrm{pw}}(i,k)$ is smaller than $Q\big(\mathbf{D}, \frac{(n-L)}{(n-1)L}\big)$ are likely members of the same cluster. This motivates the threshold value

$$d_o = Q\left(\mathbf{D}, \frac{(n-L)}{(n-1)L}\right),$$

which we implement in practice by

$$d_{ko} = \max\left(Q\left(\mathbf{D}_k, \frac{3}{n}\right), Q\left(\mathbf{D}, \zeta \frac{n-L}{(n-1)L}\right)\right), \quad k = 1, \ldots, n. \tag{2.11}$$

Here, ζ is a parameter that controls the threshold and $Q\big(\mathbf{D}_k, \frac{3}{n}\big)$ ensures that at least 3 pairwise warping functions are included in the cluster-specific warping. The optimal choice of ζ can be obtained by

$$\zeta^* = \underset{\zeta}{\arg\min}\; H/S, \tag{2.12}$$

where H measures the homogeneity and S the separation of the resulting clusters (Tang and Vemuri, 2005). In the context of discovering similar but potentially time-warped clusters, homogeneity H and separation S are computed for time-synchronized curves as follows:

$$H = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{\widehat{n}_l}\left\{\sum_{i=1}^{n}\mathbf{1}_{\{\widehat{M}_i=l\}}\int_{\mathcal{T}}(Y_i^*(t)-\bar{Y}_l^{\,*}(t))^2\,\mathrm{d}t\right\} \tag{2.13}$$

and

$$S = \frac{1}{\sum_{l_1\neq l_2}^{L}\widehat{n}_{l_1}\widehat{n}_{l_2}}\sum_{l_1\neq l_2}^{L}\widehat{n}_{l_1}\widehat{n}_{l_2}\int_{\mathcal{T}}(\bar{Y}_{l_1}^{\,*}(t)-\bar{Y}_{l_2}^{\,*}(t))^2\,\mathrm{d}t, \tag{2.14}$$

where $\bar{Y}_l^{\,*}(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i^*(t)\mathbf{1}_{\{\widehat{M}_i=l\}}$ are the average profiles of time-warped curves in cluster l, $\widehat{n}_l = \sum_{i=1}^{n}\mathbf{1}_{\{\widehat{M}_i=l\}}$ are the sizes of the estimated clusters, and $\widehat{M}_i$ are the estimated cluster memberships. A small H/S suggests that the resulting clusters are homogeneous and well separated. To accelerate computation, we usually search for the best ζ within the interval [0.5, 1]. In many cases, $\zeta = 1$ leads to very good results.

This time-synchronized clustering method is effective in identifying clusters of curves that share similar shape patterns in the presence of time variation. It can be used to remove time variation from the data, after which they can be entered into any clustering algorithm, which utilizes distances between data. We compare the performance of various clustering methods in a simulation study as well as applications. The code for the proposed time-synchronized clustering algorithm is available upon request.

3. Simulation study

To compare the performance of the proposed time-synchronized warping method with conventional hierarchical clustering and K-means clustering, we generated 100 samples, each with 40 curves falling in 4 different clusters. The mean functions of each cluster were constructed from linear combinations of cubic B-spline basis functions with random coefficients, determining the underlying pattern of each cluster. In the upper left panel of Figure 1 is an example of such mean functions for 4 clusters. For each sample, new spline coefficients were generated to obtain 4 new mean functions. Thus, we were able to study a wide range of shape patterns and evaluate the performance of the proposed method in a multitude of situations.

The time transformations h_i were generated using the area under a second set of random curves, which were generated similarly to the mean functions. Because the B-spline basis is positive on the domain spanned by the knots and zero elsewhere (Eilers and Marx, 1996), these random curves are positive and the area under the curves is monotone increasing. The resulting warping functions induce substantial time distortion (upper right panel of Figure 1). Each time transformation was randomly assigned to one of the 4 mean functions with equal probability. To avoid sparse clusters, the minimum cluster size was set to 6. An example of 40 curves generated from the described procedure is displayed in the lower right panel of Figure 1. The lower left panel of Figure 1 shows the structure of the same data before noise was added.

After applying time-synchronized warping, the results were compared with those obtained from K-means and hierarchical clustering based on the regular L^2 distance

$$d_{L^2}(i,k) = \left\{ \int_{\mathcal{T}} (Y_i(t) - Y_k(t))^2 \, \mathrm{d}t \right\}^{\frac{1}{2}}, \tag{3.1}$$

which does not take into account time variation. To assess the quality of the discovered clusters, we used the Rand index (Rand, 1971) and the Jaccard coefficient (Tan *and others*, 2005). The Rand index can be viewed as the ratio of pairs for which there is an agreement, while the Jaccard coefficient reflects the percentage of agreement between the estimated and the true clusters. When the Rand index is infinity or the Jaccard coefficient is 1, the estimated cluster structure is in perfect agreement with the true cluster structure.

The histograms of the Jaccard coefficients for the various clustering methods obtained from the simulations are shown in Figure 3 (Rand indices were not plotted, as they are infinity for perfect agreement). The parameter ζ that determines the threshold of cluster-specific warping in the proposed method was chosen as in (2.12). The clusters obtained using time-synchronized L^2 distance and optimal ζ were often in perfect agreement with the true cluster structures, regardless of the clustering techniques used, and there were only few incidences where K-means and hierarchical clustering with time-synchronized L^2 distance failed to reach more than 60% concordance with the true cluster structures. Overall, K-means with time-synchronized L^2 distance performed best and identified the true cluster structures 93 times among the 100 simulation samples. Hierarchical clustering based on time-synchronized L^2 distance also performed very

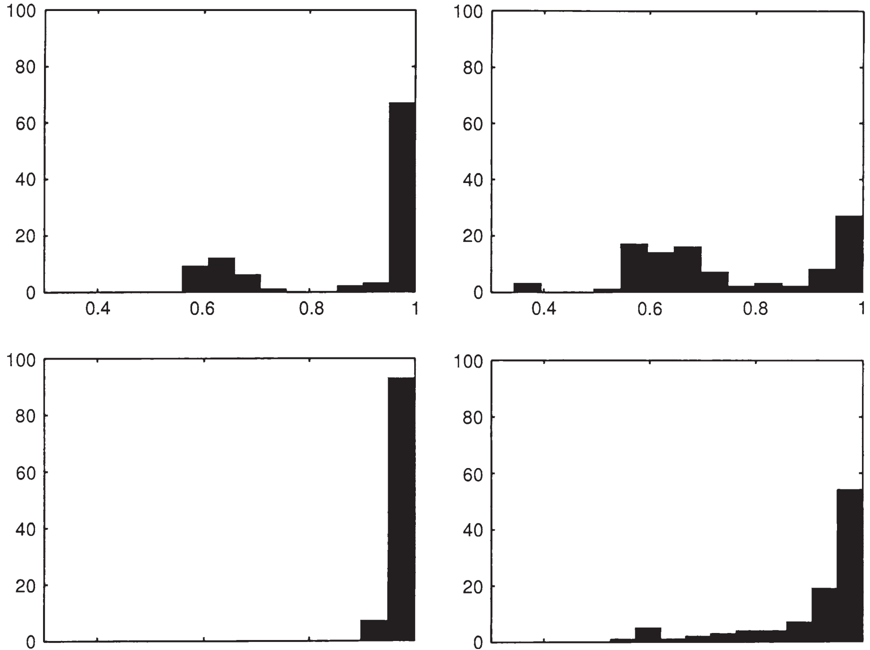

Fig. 3. Histogram of Jaccard coefficients obtained by simulating different clustering methods. Upper left panel: hierarchical clustering with time-synchronized L^2 distance (2.10), with ζ chosen by (2.12); upper right panel: hierarchical clustering with L^2 distance (3.1); lower left panel: *K*-means clustering method with time-synchronized L^2 distance (2.10), with ζ chosen by (2.12); lower right panel: K-means clustering with L^2 distance (3.1).

well and reached Jaccard $= 1$ in 63 cases. In contrast, K-means and hierarchical clustering with regular L^2 distance only successfully identified the true cluster structures 54 and 23 times, respectively.

4. Applications to time course gene expression data

The 3 organizing principles of gene ontology are cellular component, biological process, and molecular function (Ashburner *and others*, 2000). It is commonly believed that the relationship of biological processes with gene expression is guided by the "guilt by association" principle, that is, genes that share similar expression patterns also have similar functions (Brown *and others*, 2000; Lagreid *and others*, 2003). In the following examples, we first consider an application where biological annotation is not available, assessing the quantitative performance of the proposed method for a human fibroblast gene expression. In a second application, we use available information on biological function of genes for one of the most studied organisms, *Saccharomyces cerevisiae*, where 6 known biological processes define 6 natural clusters.

4.1 *Response of human fibroblasts to serum data*

The human fibroblast gene expression data contain complementary DNA microarray measurements of over 8000 genes at 12 different time points after the introduction of serum. For the purpose of clustering, we rescale the domains of these curves to [0, 1] and select the 517 genes studied by Iyer *and others* (1999). The full biological annotation is not yet available for these 517 genes. According to Khatri *and others* (2004), only 71 of these genes have been associated so far with 14 biological pathways.

In the absence of further biological annotation, as is the case for these data, 2 particularly useful features are the within-clusters homogeneity H (2.13) and the between-clusters separation S (2.14). Because the goal of the proposed method is to identify classes of comparable genes which exhibit similar behaviors, subject to transformation of the time axes, the quantities H and S are calculated based on the aligned curves (2.9). The values of H and S for alternative clustering methods without time adjustment are directly calculated using the original gene profiles. Small H and large S indicate smaller within-cluster variation and larger between-cluster variation, so that a low value of H/S indicates a desirable cluster structure. We use this measure to evaluate the overall performance of clustering with and without time adjustment, varying the number of clusters L. The threshold in the proposed method for cluster-specific warping was determined according to (2.11).

As shown in Figure 4, using the time-synchronized L^2 distance systematically produced tighter and better separated clusters regardless of the clustering algorithm applied. We also performed this analysis on other time course gene expression data that were measured during rat central nervous system development (Wen *and others*, 1998). Similar improvements in H and S were observed in this second application. These applications provide evidence that the clusters produced by the proposed method enjoy better properties than standard methods of clustering for temporal gene expression data.

4.2 *Reconstruction of yeast biological processes*

Six yeast biological processes (Hong *and others*, 2006) for which corresponding genes showed time-delayed co-expressed patterns in a cell cycle experiment (Spellman *and others*, 1998) are response to stress, protein biosynthesis, mitotic sister chromatid segregation, protein amino acid phosphorylation, cell wall organization and biogenesis, and conjugation with cellular fusion.

There are 51 genes involved in these 6 processes, and their gene expression values were taken from a microarray experiment in which relative messenger RNA abundances were measured every 7 min for 119 min after the yeast strain was arrested in the G1 phase by alpha factor. It is evident that the gene

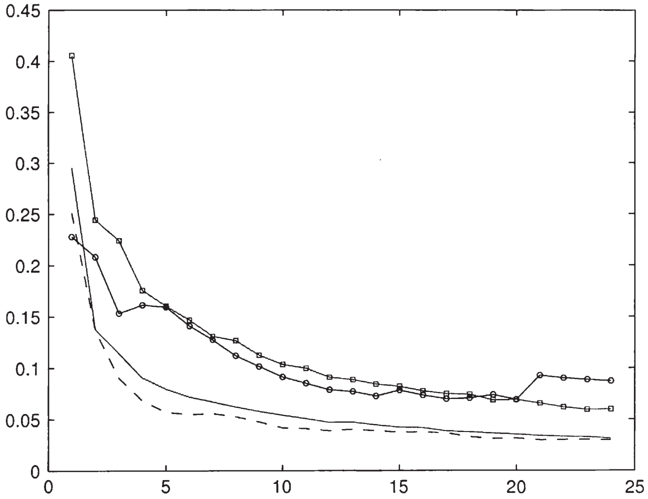

Fig. 4. H/S (2.13, 2.14) for different clustering methods for fibroblast gene profiles when the number of clusters ranges from 2 to 25: hierarchical clustering with time-synchronized L^2 distance (dashed line); K-means clustering with time-synchronized L^2 distance (solid line); K-means clustering (line with squares); hierarchical clustering (line with circles). Smaller H/S ratio indicates better clustering.

expression trajectories for the same biological process exhibit similar baseline patterns after sup-norm normalization, that is, replacing $f(x)$ by $\frac{f(x)}{\sup_y |f(y)|}$. There is, however, substantial time variation within each cluster (Figure 5). For this reason, genes in different biological processes may appear to have smaller L^2 distances (3.1) than those that belong to the same process, leading to inefficient clustering when using the regular L^2 distance.

Addressing the time variation, the proposed local cluster-specific time warping algorithm was applied with threshold value 0.75, chosen via (2.12), followed by the hierarchical clustering method based on the time-synchronized L^2 distance (2.10), aiming at 6 clusters. The resulting clusters are homogeneous (Figure 6). We also obtained alternative cluster estimates, using K-means and hierarchical clustering without time synchronization.

Because the biological processes form 6 known natural clusters, we are able to calculate Rand indices to compare different methods. This measure assesses the agreement between estimated and true clusters. The proposed method achieved a Rand index of 42.35, about 6–7 times higher than the Rand indices for the standard K-means and hierarchical clustering methods, which were only 7.85 and 6.11, respectively. Another quality measure for method clustering, homogeneity H (2.13), was also compared for time-synchronized and standard clustering. Small H indicates more homogeneous clusters, and it turned out that the proposed method had the smallest homogeneity, 0.033, while the homogeneities of standard K-means and hierarchical clustering were 0.042 and 0.041, respectively, even after time synchronizing within the estimated clusters.

As do virtually all clustering algorithms, the proposed method requires a prespecified number of clusters. It is therefore of interest to evaluate its robustness under misspecification of the number of clusters. When the number of clusters varied between 2 and 8, time-synchronized clustering consistently produced better results than K-means and hierarchical clustering with L^2 distance, in terms of the Rand index, as illustrated in Figure 7.

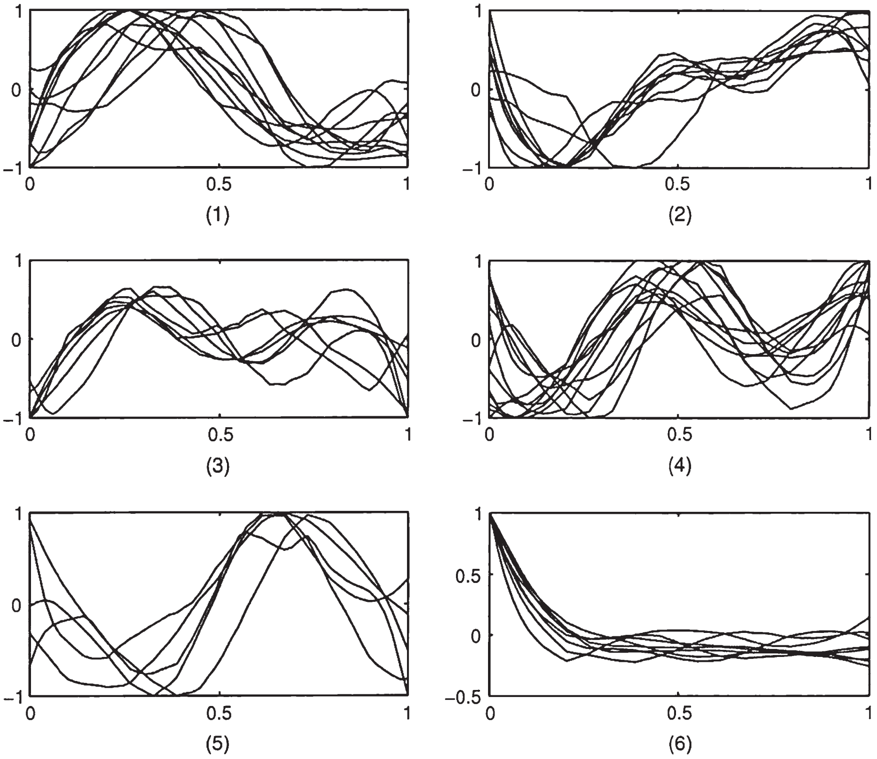

Fig. 5. Expression patterns of 51 yeast genes that are involved in 6 biological processes: (1) response to stress, (2) protein biosynthesis, (3) mitotic sister chromatid segregation, (4) protein amino acid phosphorylation, (5) cell wall organization and biogenesis, and (6) conjugation with cellular fusion.

5. CONCLUDING REMARKS

We note that the small random amplitude perturbation function δZ_i in model (2.2) is not a target of interest. This component plays no role in the clustering process and rather is a nuisance component that is included to allow for amplitude variation in the data, thus more realistically reflecting many functional data encountered in applications. Our theoretical arguments demonstrate that this component indeed can be ignored as long as δ remains small.

For the case of discrete measurements as reflected in (2.1), we include a presmoothing step, which requires to that a minimum number of repeated measurements per trajectory are available. How many measurements are actually needed to obtain a sufficiently accurate estimate of the underlying trajectory when implementing this smoothing step depends on several factors, including the design of the measurements and the signal-to-noise ratio. For commonly observed data, 10 or more repeated measurements per trajectory will often suffice. In the examples in Section 4, one has 12, respectively 17, measurements available.

Our results clearly demonstrate that time synchronization must be considered at the cluster estimation stage and not independent of it. The proposed time-synchronized clustering method produces more accurate and homogeneous clusters by coupling time synchronization with a suitable clustering algorithm by adopting this idea. We find that time variation adjustment in gene expression clustering can substantially improve upon standard clustering techniques for such data. With the aid of the proposed

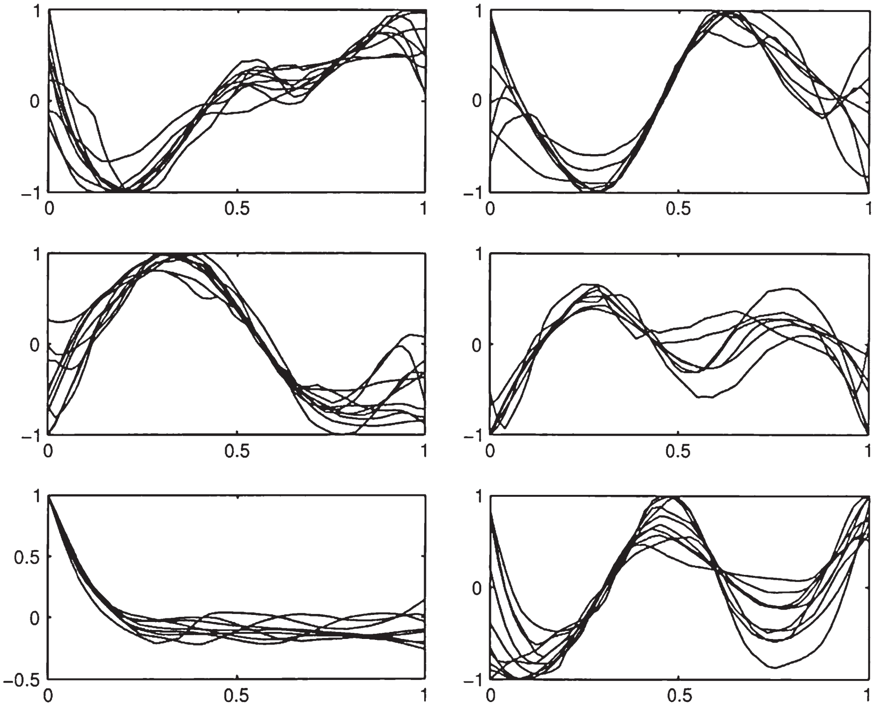

Fig. 6. Clusters obtained by the hierarchical clustering, based on time-synchronized L^2 distances, where the curves within the same clusters are synchronized, for the yeast gene expression data of Figure 5. Rand index = 42.35 and Jaccard coefficient = 0.8758.

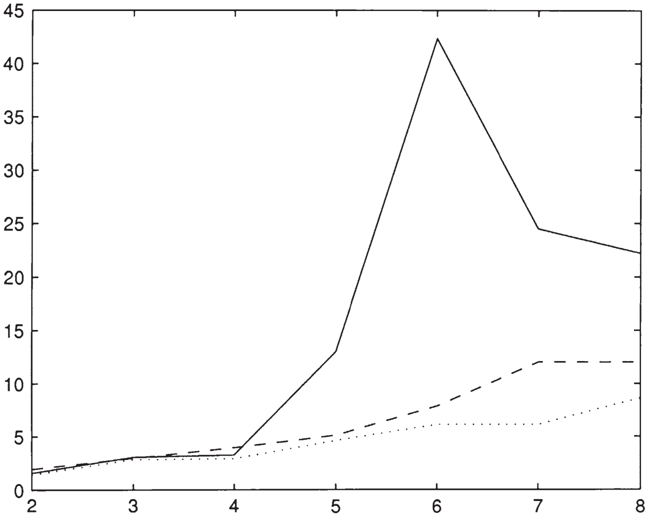

Fig. 7. Rand index profiles when the number of clusters varies from 2 to 8: Rand indices for time-synchronized clustering (solid line), hierarchical clustering (dotted line), and K-means clustering (dashed line).

time-synchronized clustering method, one can discern expression trajectories with similar shapes even under the interference of time variation, enabling a closer connection between the resulting clusters and the functionally related gene groups.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

National Science Foundation (DMS03-54448, DMS05-05537).

REFERENCES

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T. *and others* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25–29.

BROWN, M. P., GRUNDY, W. N., LIN, D., CRISTIANINI, N., SUGNET, C. W., FUREY, T. S., ARES, JR, M. AND HAUSSLER, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 262–267.

CHUDOVA, D., GAFFNEY, S., MJOLSNESS, E. AND SMYTH, P. (2003). Translation-invariant mixture models for curve clustering. *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 79–88.

EILERS, P. H. C. AND MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

GASSER, T. AND KNEIP, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association* **90**, 1179–1188.

GERVINI, D. AND GASSER, T. (2004). Self-modelling warping functions. *Journal of the Royal Statistical Society, Series B* **66**, 959–971.

GERVINI, D. AND GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92**, 801–820.

HONG, E. L., BALAKRISHNAN, R., CHRISTIE, K. R., COSTANZO, M. C., DWIGHT, S. S., ENGEL, S. R., FISK, D. G., HIRSCHMAN, J. E., LIVSTONE, M. S., NASH, R. *and others* (2006). *Saccharomyces Genome Database*. Available at: ftp://ftp.yeastgenome.org/yeast/. (Last accessed 16 March 2006).

HYNDMAN, R. J. AND FAN, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50**, 361–365.

IYER, V. R., EISEN, M. B., ROSS, D. T., SCHULER, G., MOORE, T., LEE, J. C. F., TRENT, J. M., STAUDT, L. M., HUDSON, JR, J., BOGUSKI, M. S. *and others* (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87.

JAIN, A. K., MUTRY, M. N. AND FLYNN, P. J. (1999). Data clustering: a review. *ACM Computing Surveys* **31**, 265–323.

JOHNSON, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **36**, 241–254.

KHATRI, P., BHAVSAR, P., BAWA, G. AND DRAGHICI, S. (2004). Onto-tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Research* **32**, 449–456.

Kwon, A. T., Hoos, H. H. and Ng, R. (2003). Inference of transcriptional regulation relationships from gene expression data. *Bioinformatics* **19**, 905–912.

Lagreid, A., Hvidsten, T. R., Midelfart, H., Komorowski, J. and Sandvik, A. K. (2003). Predicting gene ontology biological process from temporal gene expression patterns. *Genome Research* **13**, 965–979.

Leng, X. Y. and Müller, H. G. (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* **22**, 68–76.

Li, X., Rao, S. Q., Jiang, W., Li, C. X., Xiao, Y., Guo, Z., Zhang, Q. P., Wang, L. H., Du, L., Li, J. *and others* (2006). Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics* **7**, 7–26.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-spline. *Bioinformatics* **19**, 474–482.

Ma, P., Castillo-Davis, C. I., Zhong, W. and Liu, J. S. (2006). A data-driven clustering method for time course gene expression data. *Nucleic Acids Research* **34**, 1261–1269.

Marriott, F. H. C. (1982). Optimization method of cluster analysis. *Biometrika* **69**, 417–421.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B* **60**, 351–363.

Rand, W. M. (1971). Objective criterion for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Sakoe, H. and Chiba, C. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43–49.

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.

Storey, J. D., Xiao, W. Z., Leek, J. T., Tompkins, R. G. and Davis, R.W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12837–12842.

Tan, P. N., Steinbach, M. and Kumar, V. (2005). *Introduction to Data Mining*. Boston, MA: Addison Wesley.

Tang, N. and Vemuri, V. R. (2005). An artificial immune system approach to document clustering. *SAC '05: Proceedings of the 2005 ACM Symposium on Applied Computing*. ACM, pp. 918–922.

Tang, R. and Müller, H. G. (2008). Pairwise curve synchronization for functional data. *Biometrika* (in press).

Wen, X. L., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. S. and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 334–339.

[*Received November 29, 2007; revised March 24, 2008; accepted for publication April 8, 2008*]