BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 1 2006, pages 68–76
doi:10.1093/bioinformatics/bti742

Gene expression

Classification using functional data analysis for temporal gene expression data

Xiaoyan Leng[1,*] and Hans-Georg Müller[2]

[1]Wake Forest University School of Medicine, Public Health Sciences, Section on Biostatistics, Medical Center Blvd., MRI-3, Winston-Salem, NC 27157, USA and [2]Department of Statistics, University of California, One Shields Avenue, Davis, CA 95616, USA

Received on September 1, 2005; revised on October 17, 2005; accepted on October 21, 2005
Advance Access publication October 27, 2005
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Temporal gene expression profiles provide an important characterization of gene function, as biological systems are predominantly developmental and dynamic. We propose a method of classifying collections of temporal gene expression curves in which individual expression profiles are modeled as independent realizations of a stochastic process. The method uses a recently developed functional logistic regression tool based on functional principal components, aimed at classifying gene expression curves into known gene groups. The number of eigenfunctions in the classifier can be chosen by leave-one-out cross-validation with the aim of minimizing the classification error.

Results: We demonstrate that this methodology provides low-error-rate classification for both yeast cell-cycle gene expression profiles and *Dictyostelium* cell-type specific gene expression patterns. It also works well in simulations. We compare our functional principal components approach with a B-spline implementation of functional discriminant analysis for the yeast cell-cycle data and simulations. This indicates comparative advantages of our approach which uses fewer eigenfunctions/base functions. The proposed methodology is promising for the analysis of temporal gene expression data and beyond.

Availability: MATLAB programs are available upon request.

Contact: ileng@wfubmc.edu

Supplementary information: Supplementary materials are available on the journal's website.

1 INTRODUCTION

Since cDNA and oligonucleotide microarray techniques were developed to monitor the expression of many genes in parallel (Schena *et al*., 1995, 1996), this high-capacity system has been applied routinely for identifying and analyzing genes involved in various biological processes in different organisms (Spellman *et al*., 1998; Cho *et al*., 1998, 2001; Eisen *et al*., 1998; Wen *et al*., 1998; Golub *et al*., 1999; Iyer *et al*., 1999; White *et al*., 1999; Hill *et al*., 2000; Laub *et al*., 2000; Iranfar *et al*., 2001; Breyne and Zabeau, 2001). Recently, microarray experiments have been widely used to collect large-scale temporal data to monitor gene expression underlying development or other dynamic processes in many organisms. For example, a precise regulation of gene activity probably controls the molecular processes of DNA replication, chromosome segregation and mitosis during the cell cycle, which makes the study of cell-cycle dependent genome-wide expression an attractive system for genetic analysis. The first genome-wide expression analyses of cell-cycle regulating genes were performed in budding yeast by Spellman *et al*. (1998). More recently, several other genome-wide expression studies of cell-cycle regulated genes have been completed in bacteria (Laub *et al*., 2000), fission yeast (Rustici *et al*., 2004; Peng *et al*., 2005), plants (Breyne and Zabeau, 2001) and humans (Cho *et al*., 2001).

Another widely studied organism is the amoeba, *Dictyostelium discoideum*, which provides opportunities for studying fundamental cellular processes, including aspects of development such as cell-type determination. Recent work on *Dictyostelium* includes a review by Mohanty and Firtel (1999) on mechanisms controlling spatial patterning and cell-type proportioning, and a study of gene expression patterns with microarrays by Shaulsky and Loomis (2002). Cell-type specific gene expression patterns were studied in *Dictyostelium* by Iranfar *et al*. (2001). Other types of gene expression data were generated in large-scale temporal gene expression studies in the mapping of development of the mouse central nervous system (Wen *et al*., 1998), physiological response of human fibroblasts to serum (Iyer *et al*., 1999), and development of *Caenorhabditis elegans* (Hill *et al*., 2000) and *Drosophila* (White *et al*., 1999; Arbeitman *et al*., 2002). Information gleaned from the analysis of temporal gene expression profiles will provide an added dimension to insights into the characterization of gene function.

For these large-scale data, classifying genes into different functional groups is a first step in order to gain more sophisticated knowledge of different biological pathways and/or functions. Many classification analyses have been performed for such temporal gene expression profiles. Hierarchical clustering (Spellman *et al*., 1998; Eisen *et al*., 1998; Wen *et al*., 1998, Iyer *et al*., 1999; Gasch *et al*., 2000; Qin *et al*., 2003), *k*-means clustering (Tavazoie *et al*., 1999; Wu *et al*., 2003), principal component analysis (PCA) and singular value decomposition (SVD) (Alter *et al*., 2000, 2003; Raychaudhuri *et al*., 2000; Li *et al*., 2002; Holter *et al*., 2000), self-organizing maps or its variants (Tamayo *et al*., 1999; Nikkila *et al*., 2002; Resson *et al*., 2003), correlation analysis (Kruglyak and Tang, 2001) and independent component analysis (ICA) (Liebermeister, 2002; Lee and Batzoglou, 2003), as well as simulated annealing (Lukashin and Fuchs, 2001), and support vector machines (SVM) (Brown *et al*., 2000) have been used.

These statistical and computational methods belong to the general framework of multivariate analysis, i.e. data are treated as vectors of

[*]To whom correspondence should be addressed.

discrete samples and permutation of components will not affect the analysis results, hence the timing of the biological processes is irrelevant in these analyses. A more efficient way to look at such data is to incorporate the information that is inherent in time order and smoothness of processes over time. The tools for such an approach are provided by the recently developed methodology of functional data analysis (FDA; Ramsay and Silverman, 2005), especially discrimination and classification methods based on FDA (Hall *et al*., 2001; James and Hastie, 2001; Müller, 2005), dynamic time warping (Aach and Church, 2001; Liu and Müller, 2003) and periodicity analysis (Zhao *et al*., 2004). Recent non-parametric applications for the analysis of temporal gene expression data include work by Klevecz and Murray (2001), Luan and Li (2003) and Bar-Joseph *et al*. (2003). In the latter two papers, B-spline approaches to cluster genes based on mixed effects and mixture models were emphasized.

We view the observed gene expression profiles as independent realizations of a smooth stochastic process. The covariance function of the process is then also smooth and can be expanded into smooth orthogonal eigenfunctions (functional principal components), leading to the *Karhunen-Loève* representation of the observed sample paths as a sum of a smooth mean trend and an expansion of the random part in terms of these eigenfunctions. A truncated version of the random part of this representation serves as a statistical approximation of the random process (Rice and Silverman, 1991). In this paper, we consider functional discrimination through logistic regression based on functional principal components. We demonstrate the usefulness of this approach in a simulation study and for the analysis of yeast cell-cycle temporal data, as well as for data on the differential expression patterns of *Dictyostelium* cell-type specific genes. Although our methods do not require a regular time design, these two datasets happen to have equally spaced time points. For more details on FPCA methods for irregular and/or sparse data, see Yao *et al*. (2005).

An alternative way to model random curves is provided by B-splines, which have been previously used for clustering problems. Rice and Wu (2001) proposed a non-parametric mixed effect model based on B-splines (see also Shi *et al*., 1996; Luan and Li, 2003), emphasized cluster analysis derived from these approaches. We compared our approach based on functional principal components with a B-spline implementation of functional discriminant analysis for the yeast cell-cycle data and in a simulation study. This indicates comparative advantages of our approach, which unlike the B-spline based model does not rely on Gaussian assumptions.

2 MODELS AND METHODS

2.1 Functional Principal Component Analysis (FPCA)

We model the sample curves as independent realizations of a square integrable stochastic process $X(t)$ on $[0,\ T]$, with mean $E\{X(t)\} = \mu(t)$ and covariance function $\operatorname{cov}\{X(s),\ X(t)\} = G(s,t)$ (Rice and Silverman, 1991; Capra and Müller, 1997). By *Mercer*'s Theorem, $G(s,\ t)$ has an orthogonal expansion in $L^2([0,\ T])$:

$$G(s,t) = \sum_m \lambda_m \rho_m(s)\rho_m(t) \quad m = 1, 2, \ldots, \pm \tag{1}$$

where ρ_m and λ_m are eigenfunctions and eigenvalues ordered by size, $\lambda_1 \geq \lambda_2 \geq \ldots$.

A random curve from the population then has the following *Karhunen–Loève* representation:

$$X(t) = \mu(t) + \sum_m \varepsilon_m \rho_m(t) \quad 0 \leq t \leq T, \tag{2}$$

where

$$\varepsilon_m = \int_0^T (X(t) - \mu(t))\rho_m(t)dt \tag{3}$$

are uncorrelated random variables with $E(\varepsilon_{\mathrm{m}}) = 0$, $E(\varepsilon_m^2) = \lambda_m$ and $\sum \lambda_m < \infty$. The eigenfunctions ρ_m are referred to as functional principal components (FPCs) with FPC scores ε_m.

The deviation of each sample function from the mean is thus represented as a sum of orthogonal curves with uncorrelated random coefficients. We shall suppose that the mean curve and the FPCs are smooth and that the random part can be sufficiently well approximated by the first M FPCs, for an $M < \infty$; we discuss methods how to choose M data-adaptively in Supplement (S3).

For a sample of n random curves observed on a closed interval $[0, T]$, let $X_i = (X_i(t_{i1}), X_i(t_{i2}), \ldots, X_i(t_{in_i}))^T$ be the vector of observations for the random curve $X_i(\cdot)$ at time points t_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. An estimate $\hat{\mu}(t)$ of the mean function $\mu(t)$ can be obtained by any linear scatterplot smoother (see Supplement (S1); compare Fan and Gijbels, 1996).

Forming a dense grid s_k of $[0, T]$, e.g. $s_k = (k-1)/(S-1)\,T$, $k = 1, \ldots, S$, for a suitable large S, the estimation of the covariance function $G(s,t)$ proceeds via the empirical covariances

$$C_n(s_k, s_l) = \frac{1}{n}\sum_{i=1}^{n}\{(X_i(s_k) - \hat{\mu}(s_k))(X_i(s_l) - \hat{\mu}(s_l))\} \tag{4}$$

for all pairs of times $(s_k,\ s_l)$, $k, l = 1, \ldots, S$, $k \neq l$. For the case of irregular time grids, a pre-smoothing step may be included. The empirical covariances are then smoothed, using a 2D scatterplot smoother on the dense grid of points $(s_k,\ s_l)$, $k, l = 1, \ldots, S$ (S1). A spectral analysis is performed on the resulting $S \times S$-matrix $\hat{G} = (\hat{G}(s_k, s_l))$, yielding the first M eigenvectors/eigenvalues for $\hat{G}$. The m-th eigenvector is $(\hat{\rho}_m(s_1), \ldots, \hat{\rho}_m(s_S))'$ with the corresponding eigenvalue $\hat{\lambda}_m$ for $m = 1, \ldots, M$. The FPC scores ε_{im} for the i-th gene and the m-th FPC are obtained numerically by

$$\hat{\varepsilon}_{im} = \sum_{k=1}^{S}(X_i(s_k) - \hat{\mu}(s_k))\hat{\rho}_m(s_k). \tag{5}$$

Alternative shrinkage estimators are described in Yao *et al*. (2003).

Individual temporal gene expression profiles can then be predicted, using their FPC scores, by

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{m=1}^{M}\hat{\varepsilon}_{im}\hat{\rho}_m(t) \quad 0 \leq t \leq T. \tag{6}$$

The FPCA was performed on the combined dataset. The FPC scores can then be used to describe both between-group variability and between-group mean differences that may be of relevance for classification. The FPCA methods can be easily extended to cover the case of missing or highly irregular and sparse data (Yao *et al*., 2005).

2.2 Functional Logistic Regression

Generalized linear models are extensions of classical linear models with the following three components (McCullagh and Nelder, 1989): a random component where for the responses, $Y \sim$ exponential family, with means $E(Y) = \mu$; linear predictors, $\eta = \sum x_p \beta_p$, where x_p is the p-th predictor variable; and a monotone link function, $g(\mu) = \eta$. When Y is binomial, this is a binomial regression model. A special case is logistic regression where the link function $g(\cdot)$ is the logit function, i.e. $\operatorname{logit}(x) = \log\{x/(1-x)\}$, so that $g^{-1}(x) = \mathrm{e}^x/(1 + \mathrm{e}^x)$.

In the framework of the classification problem, the response Y denotes membership in one of two groups, say G_0 and G_1, coded as a binary random variable, where $Y = 1$ if the observation comes from G_1 and $Y = 0$ if it comes from G_0 (Efron *et al*., 1975; Press and Wilson, 1978). The predictor function *X(t)*, $t \in [0, T]$ from now on is assumed to be a centered random curve, i.e. $\mu(t) \equiv 0$. For an i.i.d. sample $X_i(t)$, for $i = 1, \dots, n$ the linear predictors are defined by $\eta_i = \alpha + \int \beta(t)\, X_i(t) dt$, leading to the functional generalized linear model (James, 2002; Müller and Stadtmüller, 2005):

$$Y_i = g^{-1}(\eta_i) + e_i, \quad i = 1, \dots, n, \tag{7}$$

where $g(\cdot)$ is a link function as before, α is a constant and $\beta(\cdot)$ is the parameter function. The errors e_i are assumed to be independent, $E(e_i) = 0$, $\mathrm{var}(e_i) < C < \infty$. The *M*-truncated model (see S2) becomes

$$Y_i = g^{-1}\left(\alpha + \sum_{m=1}^{M} \beta_m \varepsilon_{im}\right) + e_i, \quad i = 1, 2, \dots, n. \tag{8}$$

For fixed M, $\beta^T = (\alpha, \beta_1, \beta_2, \dots, \beta_M)$, the unknown parameter vector, can be estimated by solving the estimating or score equation

$$U(\beta) = \sum_{i=1}^{n} (Y_i - \mu_i) g'(\eta_i) \varepsilon_i / \sigma^2(\mu_i). \tag{9}$$

Denote the solution by $\hat{\beta}^T = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_M)$.

For functional binomial regression, as in classical binomial regression for discriminant analysis, set $\pi_i = P(Y_i = 1)$ and prior probabilities p_1 and p_0 for the groups G_1 and G_0, respectively. We estimate π_i by $\hat{\pi}_i = \hat{P}(Y_i = 1 | X_i(t)) = g^{-1}(\hat{\alpha} + \sum_{m=1}^{M} \hat{\beta}_m \hat{\varepsilon}_{im})$. Then we classify the *i*-th observation into G_1 if $\hat{\pi}_i \geq p_1$, otherwise into G_0.

3 RESULTS

3.1 Application to temporal gene expression data for yeast cell cycle

3.1.1 Functional discriminant analysis Temporal gene expression data (α factor synchronized) for the yeast cell cycle were obtained by Spellman *et al*. (1998). There are 6178 genes in total, and each gene expression time-course consists of 18 data points, measured every 7 min between 0 and 119 min. Of 90 genes, which were identified by traditional methods and have data available, 44 are known to be related to G_1 phase regulation and 46 to non-G_1 phase regulation (i.e. S, S/G_2, G_2/M and M/G_1 phases) of the yeast cell cycle; these serve as the training set. The expression profiles for these 90 genes are depicted in Figure 1, differentiated into phase-specific groups of gene expression profiles in Figure 2. The estimated covariance surface for these 90 genes in Figure 3 illustrates the pattern of time-dependence of gene expression and provides the basis for constructing the eigenfunctions by spectral decomposition. The diagonal elements $\hat{C}(s_k, s_k)$ were not used when constructing this surface estimate, as these elements may reflect additional measurement errors.

The number *M* of FPCs is chosen by minimizing the leave-one-gene-out cross-validation classification error rate. For each gene in the training set, the FPC scores are estimated from the data of the other 89 genes. Then a functional logistic regression model is fitted for these 89 genes, and group membership for the left-out gene is predicted; this procedure is iterated over all 90 genes, providing the cross-validated predictions.

The solid line in Figure 4 displays the cross-validation classification error rate as a function of *M*, the number of FPCs. This plot indicates that when the first five FPCs are used, the overall cross-validation classification error rate is at a minimum 10.00%, with the misclassification rate for G_1 genes estimated at 11.36% and for non-G1 genes at 8.70%.

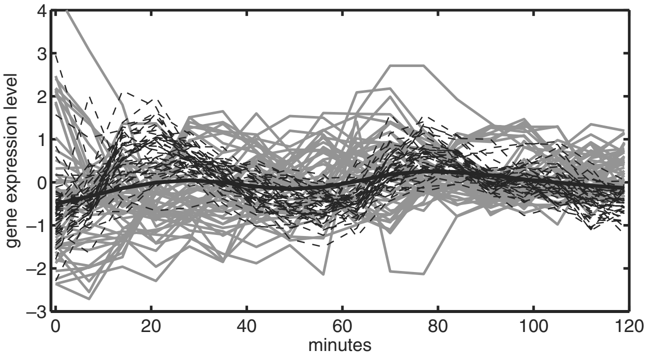

Fig. 1. Temporal gene expression profiles of yeast cell cycle. Dashed lines: G_1 phase; Gray solid lines: non-G_1 phases; Black solid line: overall mean curve.

The first five FPCs for the gene expression curves in the training set are depicted in Figure 5. Plotting the FPC scores for the second FPC versus the first FPC reveals interesting patterns for genes of different phases (Figure 6, left panel). We find that although both G_1 and S genes tend to have positive second FPC scores, all the S genes have positive first FPC scores, whereas most G_1 genes are on the negative side. However, most S/G2, G2/M and M/G1 genes have negative second scores; S/G_2 and G_2/M genes also tend to have positive first FPC scores. The discrimination between S and G_1 genes and also between G_1/S and non-G_1/S genes based on the first two FPC scores is seen to be relatively clear-cut. Similar plots can be produced for each pair of the first five FPC scores. The pairwise scatterplots also highlight the order of the phases. This feature appears to be most evident in the scatterplot of the third versus second FPC scores (Figure 6, right panel). The clockwise order of the genes is G_1→S→S/G_2→G_2/M→M/G_1→G_1.

A closer look at the misclassified genes showed that there were five genes in the G_1 group that were misclassified into the non-G_1 group. The left panel of Figure 7 displays four of these five genes, overlaid with the trajectories of G_1 genes and S genes. It appears that the trajectories of these genes are in fact close to those of the S genes. The right panel in Figure 7 displays the fifth misclassified gene in the G_1 group, overlaid with trajectories of G_1 genes and M/G_1 genes. This gene's trajectory is seen to be close to the M/G_1 trajectories. It appears likely that these five genes are actually non-G_1 genes, but somehow were erroneously identified as G_1 genes using traditional methods. There are four misclassified genes in the non-G_1 group (data not shown). Upon close inspection, we find that the trajectories of two of these four genes are closer to those of the G_1 group than to those of the non-G_1 group. The trajectory of one gene cannot be clearly associated with either group and the fourth gene lies on the boundary between the two groups.

3.1.2 Comparison with B-spline based method Rice and Wu (2001) and Shi *et al*. (1996) proposed a mixed effects model for unequally sampled noisy curves. Let $X_i = (X_i(t_{i1}), X_i(t_{i2}), \dots, X_i(t_{in_i}))^T$ be the vector of observations for the *i*th curve for $i = 1, 2, \dots, n$,

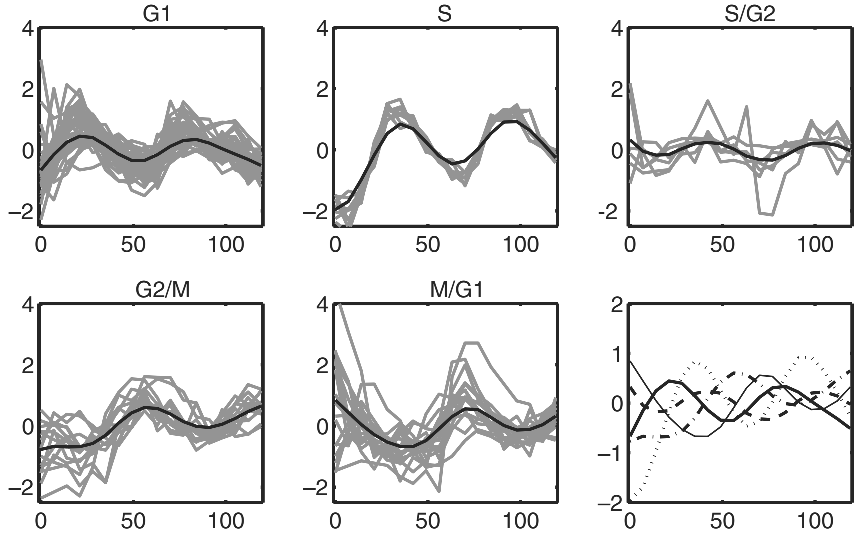

Fig. 2. Yeast cell-cycle gene expression profiles sorted by phases. The first five panels provide the expression profiles for G_1, S, S/G_2, G_2/M, and M/G_1 phases, with mean functions indicated by the black solid lines. The lower right panel contains the mean curves of all phases overlaid: G_1: thick solid line; S: dotted line; S/G_2: dashed line; G_2/M: dash–dot line; and M/G_1: thin solid line.

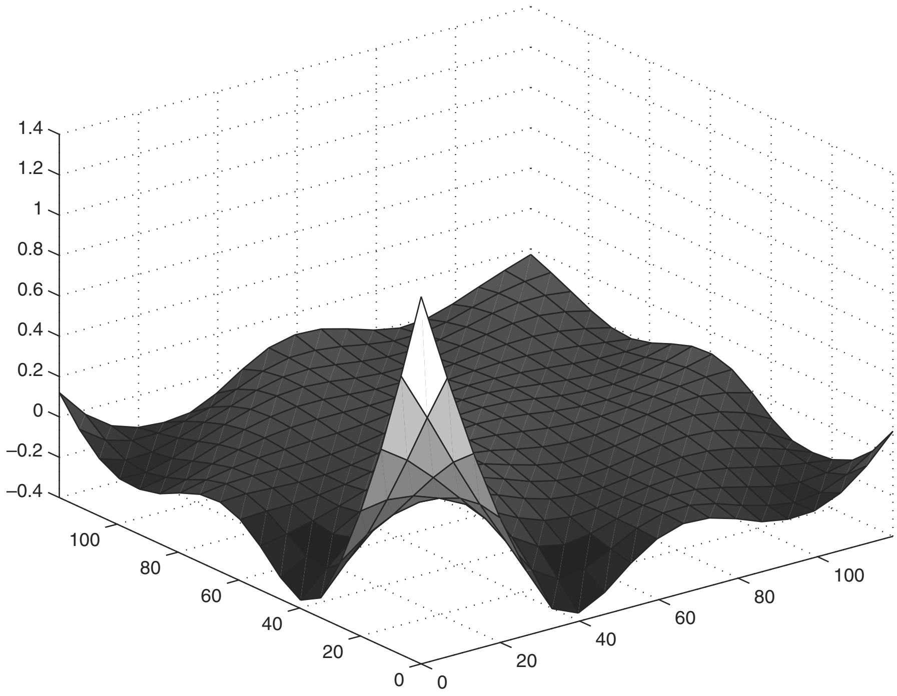

Fig. 3. Estimated covariance surface for the 90 known yeast cell-cycle genes.

where $X_{ij} = X_i(t_{ij})$ is the observed function value at time t_{ij}, $j = 1, \dots, n_i$. Note that the setup is the same as described above.

The approximating model of Rice and Wu is

$$X_{ij} = \sum_{k=1}^{p} \beta_k \bar{B}_k(t_{ij}) + \sum_{l=1}^{q} \gamma_{il} B_l(t_{ij}) + \varepsilon_{ij},$$

where the mean function is $E(Y_i(t)) = \mu(t) = \sum_{k=1}^{p} \beta_k \bar{B}_k(t)$ and $\bar{B}_k(\cdot)$, $B_l(\cdot)$ are possibly different B-spline bases on [0, T]. The γ_{il} are random effects, with $E(\gamma_{il}) = 0$ and $\mathrm{cov}(\gamma_{il}) = \Gamma$. The corresponding estimate of an individual trajectory is the smooth curve

$$\hat{X}_i(t) = \sum_{k=1}^{p} \hat{\beta}_k \bar{B}_k(t) + \sum_{l=1}^{q} \hat{\gamma}_{il} B_l(t),$$

where the estimates are obtained by least squares. Classification can be based on the random coefficients γ_i, and the dashed line in

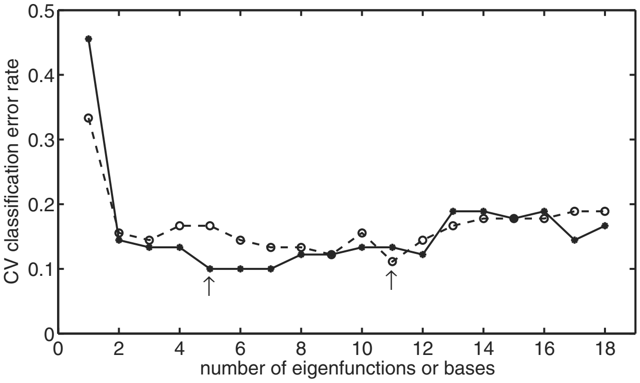

Fig. 4. Choosing M, the number of eigenfuntion for yeast cell-cycle data. Leave-one-out cross-validation estimates of the misclassification rate, in dependency on M: FPCA: solid line; *B*-spline: dashed line.

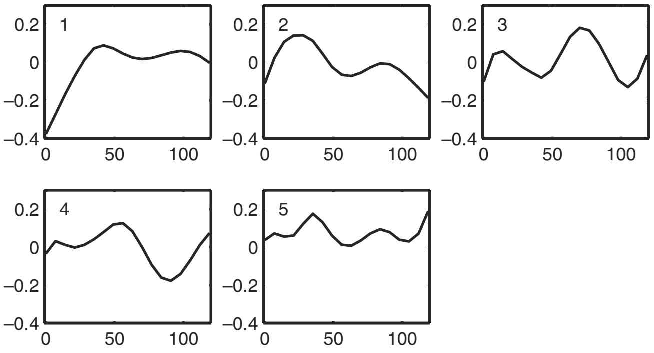

Fig. 5. The first five FPCs for the known 90 genes for yeast cell-cycle data. These five FPCs account for 98.9% of the total variation, with the first FPC accounting for 66.5%, the second for 21.9%, the third for 4.7%, the fourth for 3.1% and the fifth for 2.7%.

Figure 4 shows the overall misclassification error rates. Using B-splines, the misclassification error rate attains its lowest value 11.1% for 11 bases, whereas the minimum error rate using FPCA is 10% and is achieved with only 5 FPCs, as described above. Thus FPCA is seen to be advantageous in this example by employing fewer components than the B-spline approach, while simultaneously yielding a slightly lower misclassification error rate.

3.2 Application to expression patterns of cell-type specific genes in *Dictyostelium*

Iranfar *et al*. (2001) studied expression patterns of cell-type specific gene fragments in *Dictyostelium discoideum*. Such studies are of particular interest for *Dictyostelium*, since only prestalk and prespore cells are differentiated during development. DNA microarrays carrying 690 targets were used to determine expression profiles during development. Fitting a biologically based kinetic equation to extract the times of transcription onset and cessation, the authors recognized 35 cell-type specific genes, including 17 newly identified ones, which were confirmed by Northern blots. We used these 35 genes, with 14 prestalk genes and 21 prespore genes, as our training set to explore other potential cell-type specific genes. Figure 8 shows the relative intensity of the signals for prestalk and prespore genes. A considerable number of prestalk genes peaked between 8 and 10 h of development and then decreased significantly, whereas most prespore genes were not expressed until 10 h of development and continued to be expressed thereafter.

We used these genes as training set for functional discriminant analysis. Cross-validation error rates indicated that using the first three FPCs yields the lowest overall misclassification rate of 22.86%, with 28.57% for prestalk genes and 19.05% for prespore genes. Misclassified prestalk genes are highlighted in the left panel of Figure 9. Besides *emcA* and *emcB*, which were screened out by Iranfar *et al.* (2001) owing to their 'prespore-like' profiles, we found that *ostB* and *mitA* might not be correctly classified either. These two genes show no cell-type specific features yet were grouped into prestalk genes. For another gene *rasD*, mentioned by Iranfar (Iranfar *et al*., 2001), the estimated probability for classification into prespore genes with 0.4376 only slightly exceeds the threshold of 0.4, which is the prior probability according to the training set. The right panel of Figure 9 highlights the four misclassified prespore-specific genes. Gene *cbpB* shows an early peak at 8 h, and follows the pattern of prestalk genes. Genes *sodA* and *thfA* did not show obvious cell-specific features in their expression patterns. Gene *cprF* did not start to express until 20 h, which may have contributed to its misclassification.

We then used the model fitted to the training set to classify the rest of the genes, and chose ranges of estimated probabilities for a gene to be classified into the prestalk group of [0.95, 1], [0.85, 0.95) and [0.75, 0.85), in order to identify subgroups of prestalk genes. Analogous probability ranges of [0, 0.05], (0.05, 0.15] and (0.15, 0.25] were used to identify subgroups of prespore genes. The results are shown in Figure 10. With these three classification probability ranges, 40 prestalk-specific and 36 prespore-specific genes were identified, displaying reasonably homogeneous patterns within each identified subgroup. Especially the genes in the left upper panel show very typical cell-specific patterns.

3.3 A simulation study

3.3.1 Functional discriminant analysis A data-based simulation study was performed based on the first five estimated FPCs from the yeast cell-cycle data, where we assume that these correspond to the real underlying FPCs. Then five random coefficients ε_m, $m = 1,\ldots,5$, were generated for each subject from normal distributions with means 0.6, 0.5, 0.4, 0.3, and 0.2 for group 1 and −0.6, −0.5, −0.4, −0.3, and −0.2 for group 2, with variances $\sigma_m^2 = 2.6950$, 0.8850, 0.1957, 0.1266 and 0.1079 for both groups. These variances correspond to the estimated eigenvalues from the yeast cell-cycle data. The priors for the two groups were chosen equal, i.e. $\pi_1 = \pi_2 = \frac{1}{2}$, so that the generated samples have overall mean 0. For all subjects, 18 equally spaced data points were taken, just as is the case for the yeast cell-cycle data. We generated 100 training and test datasets. Each dataset was composed of 200 samples, where the first 100 samples formed the training set and the remaining 100 samples the test set.

For each of the 100 simulated datasets, classification error rates were calculated for the test data based on FPCA and B-spline methods, respectively. The simulation classification error rates based on FPCA and B-splines are compared in Table 1 (Monte

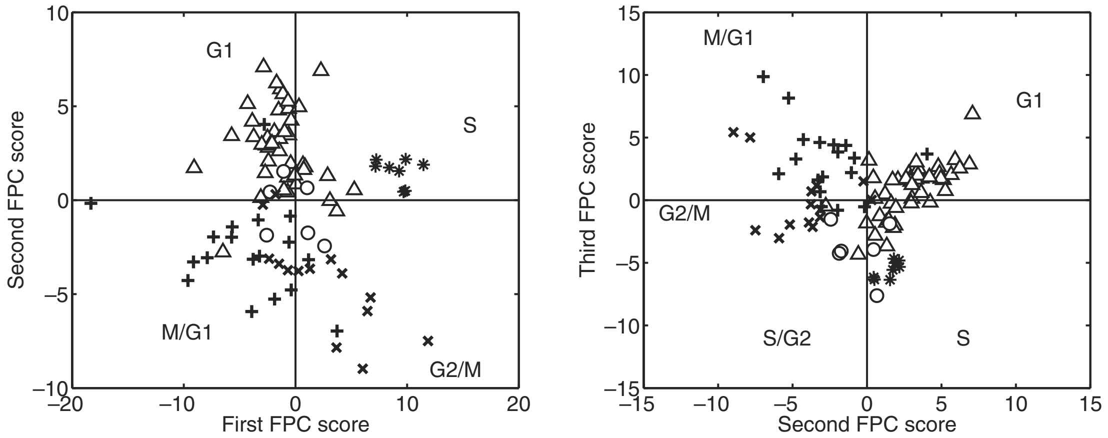

Fig. 6. Scatterplot of pairwise FPC scores for genes in the five cell-cycle phases. G_1: triangles; S: stars; S/G_2: circles; G_2/M: x-marks; M/G_1: plus signs. Left panel: Second versus first FPC score; Right panel: Third versus second FPC score, for yeast cell-cycle data.

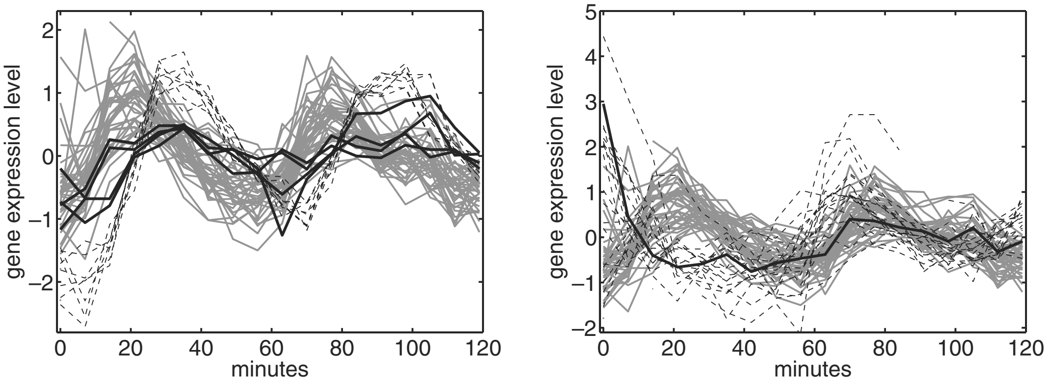

Fig. 7. Profiles of misclassified yeast cell cycle genes. Left panel: black solid lines: misclassified G_1 genes; gray sold lines: G_1 genes; dashed lines: S genes; note that the misclassified G1 genes (YDL055C, YDR113C, YDR356W and YJL092W) are actually close to the trajectories of the S genes. Right panel: black solid line: misclassified G_1 gene; gray solid lines: G_1 genes; dashed lines: M/G_1 genes; note that the misclassified G_1 gene (YCL055W) is actually close to the trajectories of the M/G_1 genes. The genes YJL092W and YCL0055W were also pointed out by Spellman *et al*. (1998).

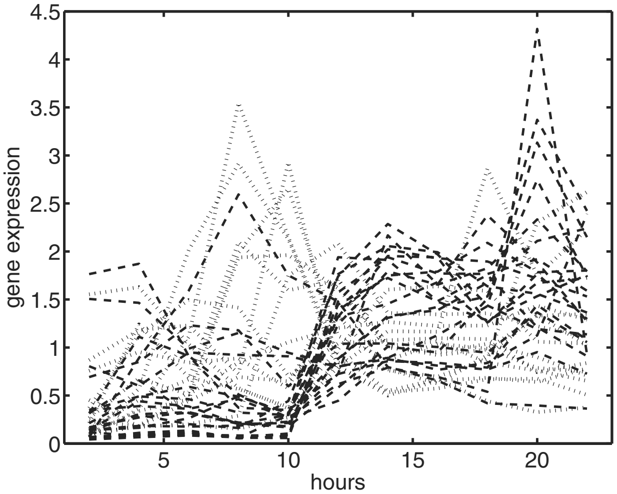

Fig. 8. Developmental temporal patterns of *Dictyostelium* gene expression. Dotted lines: prestalk-specific genes; dashed lines: prespore-specific genes.

Carlo standard errors are in parentheses). The average classification error rate over the 100 datasets indicates that all five FPCs should be used for optimal classification with either method. Except for the case with one eigenfunction/base function, the overall classification error rates based on FPCA are always slightly lower than those observed for B-splines.

4 DISCUSSION AND CONCLUSIONS

Owing to the dynamic nature of biological systems, temporal gene-expression data play a critical role in exploring the regulation of gene expression, in particular, in highlighting genes that are time critical for the regulation of certain biological processes such as the cell cycle for different organisms (Spellman *et al*., 1998; Laub *et al*., 2000; Breyne and Zabeau, 2001; Cho *et al*., 2001), the central nervous system development (Wen *et al*., 1998), *Drosophila* development (White *et al*., 1999; Arbeitman *et al*., 2002) and *Dictyostelium* cell differentiation (Iranfar *et al*., 2001). With rapidly accumulating amounts of temporal microarray gene expression data, developing adequate models to analyze such data is urgent. In this paper, we propose a functional discriminant analysis method,

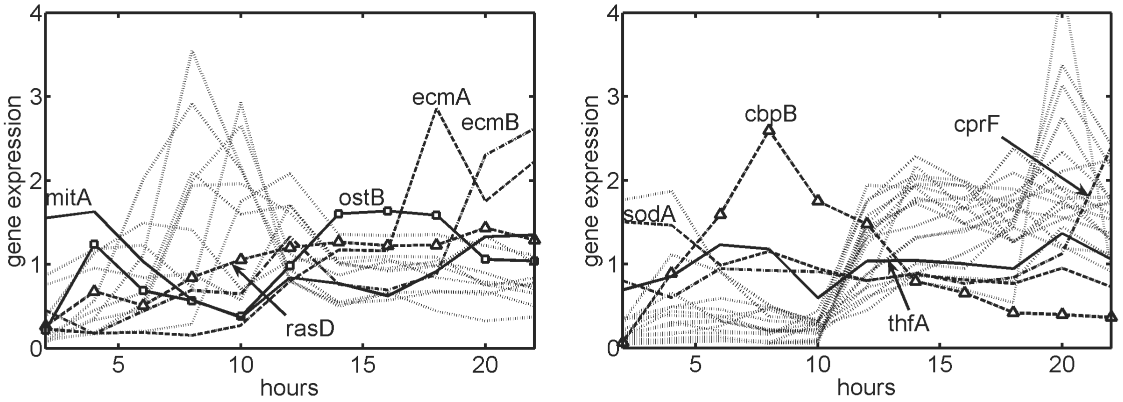

Fig. 9. Misclassified cell-type specific genes for *Dictyostelium*. Left panel: Prestalk genes are shown in light gray. The four misclassified prestalk-specific genes are labelled and highlighted in thick lines; gene *ras*D is also indicated (see text). Right panel: Prespore genes are shown in light gray. The four misclassified prespore-specific genes are lablelled and highlighted in thick lines.

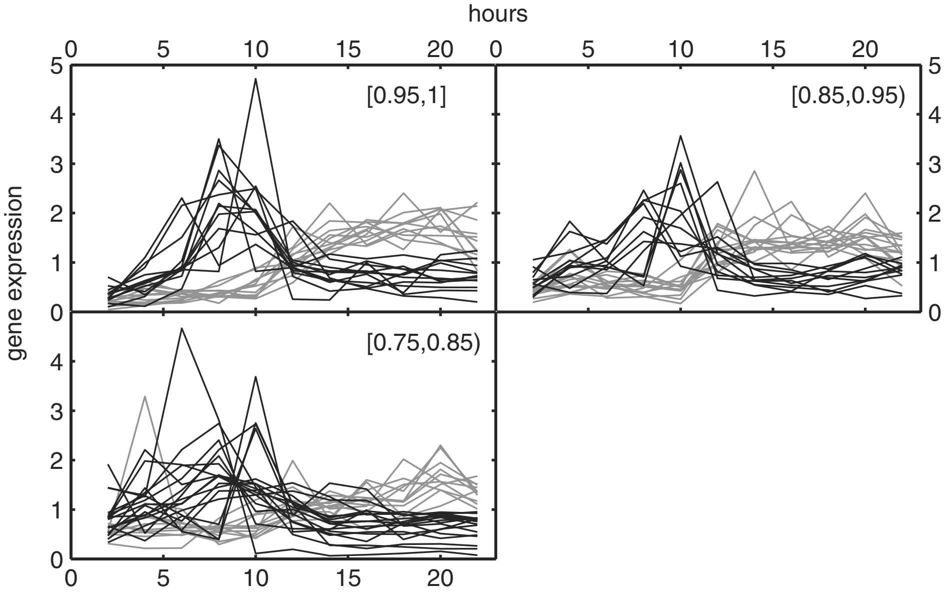

Fig. 10. Subgroups of *Dictyostelium* cell-type specific genes according to different posterior probabilities. Black solid lines: prestalk-specific genes corresponding to the indicated range of posterior probabilities; gray solid lines: prespore-specific genes corresponding to posterior probabilities of [0.0,05] (upper left panel), (0.05, 0.15] (upper right panel) and (0.15, 0.25] (lower panel).

Table 1. Classification error rates based on FPCA and B-Splines (B-S)

No. of FPCs or base functions	Group 1		Group 2		Overall	
	FPCA	B-S	FPCA	B-S	FPCA	B-S
1	32.7 (0.07)	30.5 (0.07)	33.0 (0.08)	29.8 (0.07)	32.8 (0.05)	30.1 (0.04)
2	27.8 (0.07)	36.8 (0.09)	26.1 (0.07)	37.5 (0.08)	27.0 (0.04)	37.2 (0.05)
3	11.4 (0.05)	14.6 (0.05)	11.9 (0.05)	15.0 (0.05)	11.7 (0.03)	14.8 (0.03)
4	10.8 (0.05)	11.2 (0.05)	10.3 (0.05)	11.2 (0.05)	10.6 (0.03)	11.2 (0.03)
5	10.3 (0.04)	10.8 (0.05)	10.3 (0.05)	10.7 (0.05)	10.3 (0.03)	10.8 (0.03)

using a functional version of logistic regression and functional principal components for the temporal gene expression data.

Temporal gene expression data provide valuable functional information about temporal patterns of gene expression and also interactions between genes. For example, a typical yeast mitotic cell cycle is commonly broken down into the four standard phases: G_1, S, G_2, and M. When the daughter cell breaks away from the mother cell, it is typically smaller than the mother cell. During the G_1 phase, the daughter cell will grow until it is of a large enough size to enter the cell cycle. The G_1 phase of the cell cycle is important for determining the fate of the cell. Statistically identifying genes that regulate the G1 phase will be helpful for studies of genetic cell-cycle regulation. Since most biological processes are in fact continuous, temporal gene expression data can be viewed as discretized samples from smooth random gene expression trajectories over time, naturally leading to a functional data analysis approach. The proposed method provides low-error rate (10%) classification for the yeast cell-cycle gene expression data and also in simulations. Differentiating cell-cycle regulated genes from non-cell-cycle regulated genes is another important goal for cell-cycle studies. Extensions of the functional methods proposed here will be of interest in approaching this problem.

In comparisons with the B-spline approach, both yeast cell-cycle data analysis and simulations demonstrate overall lower-error rates with fewer eigenfunctions/base functions when using the FPCA method. The proposed FPCA methods allow the identification of genes that were probably misclassified by traditional biological classification methods. The phase order displayed by scatterplots of pairwise FPC scores suggests that FPCA has potential for time ordination analysis of temporal gene expression.

The NIH has designated *Dictyostelium discoideum* as a model organism for the functional analysis of sequenced genes. Applying our methods, we screened out previously misclassified cell-type specific genes. Furthermore, we identified 76 genes falling into subgroups that show cell-type specific features of gene expression. Extending the proposed algorithm to functional cluster analysis is feasible and useful in the common situation where group membership is unknown, as is often the case in biological applications.

ACKNOWLEDGEMENTS

The comments of three reviewers led to numerous improvements and are gratefully acknowledged. Research supported in part by NSF grants DMS03-54448 and DMS05-05537.

Conflict of Interest: none declared.

REFERENCES

Aach,J. and Church,G.M. (2001) Alignment gene expression time series with time warping algorithms. *Bioinformatics*, **17**, 495–508.

Alter,O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modelling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Alter,O. *et al*. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc. Natl Acad. Sci. USA*, **100**, 3351–3356.

Arbeitman,M.N. *et al*. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science*, **297**, 2270–2275.

Bar-Joseph,Z. *et al*. (2003) Continuous representation of time-series gene expression data. *J. Comput. Biol*., **10**, 341–356.

Breyne,P. and Zabeau,M. (2001) Genome-wide expression analysis of plant cell cycle modulated genes. *Curr. Opin. Plant Biol*., **4**, 136–142.

Brown,M.P.S. *et al*. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Capra,W.B. and Müller,H.G. (1997) An accelerated-time model for response curves. *J. Am. Statist. Ass*., **92**, 72–83.

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

Cho,R.J. *et al*. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet*., **27**, 48–54.

Efron,B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Ass*., **70**, 892–898.

Eisen,M.B. *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fan,J. and Gijbels,I. (1996) *Local Polynomial Modelling and its Applications*. Chapman & Hall, New York.

Gasch,A.P. *et al*. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Bio. Cell*, **11**, 4241–4257.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hall,P. *et al*. (2001) A functional data-analytic approach to signal discrimination. *Technometrics*, **43**, 1–9.

Hill,A.A. *et al*. (2000) Genomic analysis of Gene expression in *C. elegans*. *Science*, **290**, 809–812.

Holter,N.S. *et al*. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.

Iranfar,N. *et al*. (2001) Expression patterns of cell-type specific genes in *Dictyostelium*. *Mol. Bio. Cell*, **12**, 2590–2600.

Iyer,V.R. *et al*. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83–87.

James,G.M. and Hastie,T.J. (2001) Functional linear discriminant analysis for irregular sampled curves. *J. R. Statist. Soc. B*, **63**, 533–550.

James,G.M. (2002) Generalized linear models with functional predictors. *J. R. Statist. Soc. B*, **64**, 411–432.

Klevecz,R.R. and Murray,D.B. (2001) Genome wide oscillations in expression: wavelet analysis of time series data from yest expression arrays uncovers the dynamic architecture of phenotype. *Mol. Biol. Reports*, **28**, 73–82.

Kruglyak,S. and Tang,H. (2001) A new estimator of significance of correlation in time series data. *J. Comput. Biol*., **8**, 463–470.

Laub,M.T. *et al*. (2000) Global analysis of the genetic network controlling a bacterial cell cycle. *Science*, **290**, 2144–2148.

Lee,S.I. and Batzoglou,S. (2003) Application of independent component analysis to microarrays. *Genome Biol*., **4** (Art. R76).

Li,K.C. *et al*. (2002) A simple statistical model for depicting the cdc15-synchronized yeast cell-cycle regulated gene expression data. *Statistica Sinica*, **12**, 141–158.

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.

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

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

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*. Chapman & Hall, London.

Mohanty,S. and Firtel,R.A. (1999) Control of spatial patterning and cell-type proportioning in *Dictyostelium*. *Semin. Cell Dev. Biol*., **10**, 597–607.

Müller,H.G. (2005) Functional modelling and classification of longitudinal data. *Scand. J. Stat*., **32**, 223–240.

Müller,H.G. and Stadtmüller,U. (2005) Generalized functional linear models. *Annals Stat*., **33**, 774–805.

Nikkila,J. *et al*. (2002) Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, **15**, 953–966.

Press,S.J. and Wilson,S. (1978) Choosing between logistic regression and discriminant analysis. *J. Am. Statist. Ass*., **73**, 699–705.

Peng,X. *et al*. (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell*, **16**, 1026–1042.

Qin,J. *et al*. (2003) Kernel hierarchical gene clustering from microarray expression data. *Bioinformatics*, **19**, 2097–2104.

Ramsay,J.O. and Silverman,B.W. (2005) *Functional Data Analysis*. Springer, New York.

Raychaudhuri,S. *et al*. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput*., **2000**, 455–466.

Resson,H. *et al*. (2003) Clustering gene expression data using adaptive double self-organizing map. *Physiol. Genomics*, **14**, 35–46.

Rice,J.A. and Silverman,B.W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B*, **53**, 233–243.

Rice,J.A. and Wu,C.O. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Rustici,G. *et al*. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet*., **36**, 809–817.

Schena,M. *et al*. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schena,M. *et al*. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.

Shaulsky,G. and Loomis,W.F. (2002) Gene expression patterns in *Dictyostelium* using microarrays. *Protist*, **153**, 93–8.

Shi,M.G. *et al*. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J. R. Statist. Soc. C*, **45**, 151–163.

Spellman,P.T. *et al*. (1998) Comprehensive identification of cell cycle-regulated gene of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamayo,P. *et al*. (1999) Interpreting pattern of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tavazoie,S. *et al*. (1999) Systematic determination of genetic network architecture. *Nat. Genet*., **22**, 281–285.

Wen,X. *et al*. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 34–339.

White,K.P. *et al*. (1999) Microarray analysis of *Drosophila* development during metamorphosis. *Science*, **286**, 2179–2184.

Wu,F.X. *et al*. (2003) A genetic *K*-means clustering algorithm applied to gene expression data. *Lecture in Artificial Intelligence*, **2671**, 520–526.

Yao,F. *et al*. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, **59**, 676–685.

Yao,F. *et al*. (2005) Functional Data Analysis for Sparse Longitudinal Data. *J. Am. Statist. Ass*., **100**, 577–590.

Zhao,X. *et al*. (2004) The functional data analysis view of longitudinal data. *Statistica Sinica*, **14**, 789–808.