Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives

Article *in* EURASIP Journal on Bioinformatics and Systems Biology · February 2007

DOI: 10.1155/2007/70561 · Source: PubMed

Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 70561, 10 pages
doi:10.1155/2007/70561

Research Article

Clustering Time-Series Gene Expression Data Using Smoothing Spline Derivatives

S. Déjean,[1] P. G. P. Martin,[2] A. Baccini,[1] and P. Besse[1]

[1] *Laboratoire de Statistique et Probabilités, UMR 5583, Université Paul Sabatier, 31062 Toulouse Cedex 9, France*
[2] *Laboratoire de Pharmacologie et Toxicologie, UR 66, Institut National de la Recherche Agronomique (INRA),*
180 Chemin de Tournefeuille, BP 3, 31931 Toulouse Cedex 9, France

Received 14 December 2006; Revised 6 March 2007; Accepted 16 May 2007

Recommended by Stéphane Robin

Microarray data acquired during time-course experiments allow the temporal variations in gene expression to be monitored. An original postprandial fasting experiment was conducted in the mouse and the expression of 200 genes was monitored with a dedicated macroarray at 11 time points between 0 and 72 hours of fasting. The aim of this study was to provide a relevant clustering of gene expression temporal profiles. This was achieved by focusing on the shapes of the curves rather than on the absolute level of expression. Actually, we combined spline smoothing and first derivative computation with hierarchical and partitioning clustering. A heuristic approach was proposed to tune the spline smoothing parameter using both statistical and biological considerations. Clusters are illustrated a posteriori through principal component analysis and heatmap visualization. Most results were found to be in agreement with the literature on the effects of fasting on the mouse liver and provide promising directions for future biological investigations.

1. INTRODUCTION

In the context of microarray experiments, we focused on the analysis of time-series gene expression data. Our original data were hepatic gene expression profiles acquired during a fasting period in the mouse. Two hundred selected genes were studied through 11 time points between 0 and 72 hours, using a dedicated macroarray.

The literature concerning the analysis of time-series gene expression data mainly addresses two problems: identification of differentially expressed genes over time [1–4] and temporal profile clustering to identify genes which are coordinately regulated during the time course experiment [5–8]. Methods developed to propose solutions to the first problem can be viewed as a preliminary step that filters genes to which a clustering procedure can then be applied [9]. However, since we used a dedicated macroarray with a limited number of genes, we focused directly on the clustering of temporal profiles. In the above-mentioned articles that address the second problem, clustering is based on a set of predefined model profiles. This could be relevant when dealing with short time-series, but with 11 time points, we assumed that the information contained in the data was sufficient and that we did not require such prior information.

Since the aim of this paper is not prediction but curve clustering, the approach considered here does not refer to parametric statistical models (such ARMA) used to fit time-series. Furthermore, as mice differ from one point in time to another, models for longitudinal data are not relevant in the present context.

The purpose of the present study was to identify homogeneous clusters of genes. Nevertheless, a relevant clustering method must take into account the data specificity and, in particular, should integrate the temporal aspect. In this context, the absolute level of expression is generally of little interest, mainly because the probes on the microarray can have a significant influence on the measured intensities (see, e.g., [10]). Instead, the shapes of the curves may provide meaningful information on coordinate gene regulations. The suitable mathematical tool to describe this information is the derivative. Therefore, a preliminary stage consists in smoothing the temporal profiles in order to get regular and differentiable functions. The study of functional data is addressed in the statistical literature (see [11], for a survey). In the context of microarray data, Bar-Joseph et al. [12] use splines to provide continuous representations of time-series gene expression profiles, and thus to permit the interpolation of missing values and dataset alignment. We used the same mathematical tool to propose a methodology for curve clustering.

Our approach is in the framework of functional data analysis [11]. Its main originality lies in its focus on the first derivative of curves by means of a priori spline smoothing. The approach was composed of two steps. The first one can be viewed as a signal extraction method: assuming that gene expression profiles are regular curves, spline smoothing is performed. Tuning the smoothing parameter is a core problem that could not be achieved by the usual cross-validation method because of the poor quality of clustering results. Thus, we propose a heuristic approach that takes into account both statistical and biological considerations. The second step consisted in clustering the derivatives of the smoothed curves after discretization; hierarchical clustering and the *k*-means algorithm were used successively in order to obtain robust clusters.

Details of the biological experiment are given in the second section of the paper. Then, statistical methodology is developed with a focus on tuning the smoothing parameter. In the fourth section, clustering results are interpreted, then illustrated *a posteriori* through principal component analysis (PCA) and heatmap visualization of simultaneous clustering of curves and time points. Finally, some elements of discussion about the analysis of times-series gene expression data are given to conclude the paper.

2. BIOLOGICAL EXPERIMENT

2.1. Experimental design

Ten-week-old male *C57BL/6J* mice (wild-type) were obtained from Charles River France (Les Oncins, France) and were acclimatized to local animal facility conditions for two weeks prior to the fasting experiment. Mice were housed in groups of four in plastic cages at a temperature of 22°C (±2°C) with a 12/12 hours light/dark cycle. Mice were randomly assigned to the experimental groups. A total of 44 mice (11 cages × 4 mice/cage) were subjected to 11 different fasting periods ranging from 0 to 72 hours. All mice were moved into clean cages without food at 5 a.m. (2 hours prior to the beginning of the light phase). Since mice mainly eat during the night, this experimental setting corresponded to postprandial fasting. At each of the selected time points (0, 3, 6, 9, 12, 18, 24, 36, 48, 60, and 72 hours), 4 mice were euthanized. The liver was dissected, snap-frozen in liquid nitrogen, and stored at −80°C until RNA extraction.

The sampling rate in time-course experiments is discussed in [13]. In our case, gene expression was measured at 11 time points from 0 to 72 hours of fasting with a decreasing sampling rate. It was assumed that most of the gene expression changes would occur at the beginning of fasting. Nevertheless, the number of time points was determined to be able to observe fluctuations in the gene expression profiles, that is, changes in the sign of their derivatives, until the 72nd hour of fasting.

2.2. Production of INRArray 01.3

Selection, cloning, amplification, and spotting of the cDNA fragments onto nylon membranes have been previously described for version 01.2 of INRArray [14, 15]. The same procedure was followed for INRArray 01.3. Eighty genes were added to the panel of 120 genes present on INRArray 01.2, leading to a total of 200 genes. They were mainly genes involved in energy and xenobiotic metabolism. Furthermore, we developed a set of 13 probes and corresponding *in vitro* transcribed polyA-RNAs from yeast to be used as internal controls for normalization purposes (spiked-in RNAs). The full list of clones present on INRArray 01.3 can be found in [16]. Additionally, the spotting buffer (50% DMSO) was spotted on the macroarray at 200 different locations for the analysis of the background.

2.3. RNA extraction and labeling

Total RNA was extracted with TRIzol reagent (Invitrogen, Cergy Pontoise, France) according to the manufacturer's instructions. The integrity of the RNAs was evaluated on a Bioanalyzer 2100 (Agilent Technologies, Massy, France). For each sample, 3 μg of total RNA along with a fixed amount of the 13 spiked-in yeast RNAs were labelled by reverse transcription with Superscript II RT (Invitrogen) in the presence of 40 μCi of $[\alpha^{-33}\mathrm{P}]$dCTP (ICN, Orsay, France). The clean-up of the labelled cDNAs and the hybridization, washing, scanning, and image analysis of INRArray have been described previously [14].

2.4. Data preprocessing

All data were log-transformed. The normality of the background intensities was verified using the Kolmogorov-Smirnov test. Four macroarrays out of 44 exhibited *P*-values lower than 0.05. Each gene on each array was declared "present" when its intensity exceeded the mean plus twice the standard deviation of the background intensities. Only the genes declared "present" on a minimum of six macroarrays were retained for further analysis. This procedure yielded a total of 130 genes selected for further analysis. Data were normalized using the average signal of the 13 spiked-in yeast RNAs. Boxplots for the 44 macroarrays led us to declare 4 macroarrays as outliers, which were removed from the dataset. Thus the dataset studied in this paper consists of a matrix of log-transformed normalized intensities for 130 genes × 40 samples (40 mice).

3. STATISTICAL METHODOLOGY

Let us recall that our purpose consisted in clustering temporal profiles according to their shape. In this context, the mathematical tool to be used is the first derivative of the curve. Therefore, the first step aimed at getting one regular curve modeling the evolution of each gene.

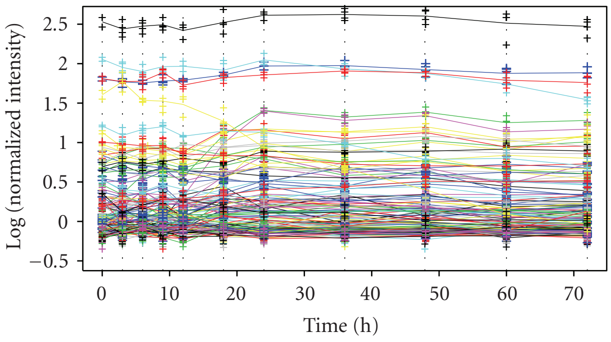


Figure 1: Log-normalised intensity versus time for 130 genes. For each gene, the line joins the average value at each time point. Vertical dashed lines indicate time points.

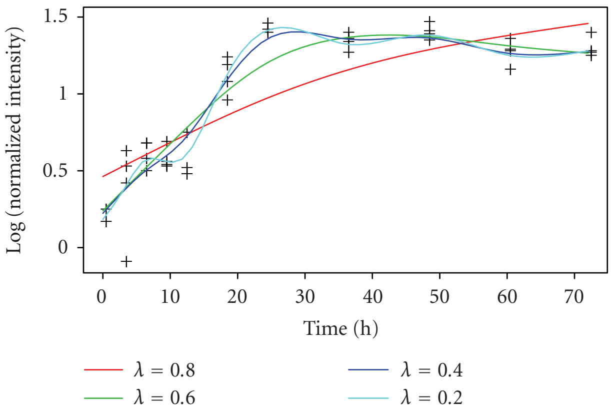


Figure 2: Smoothed curves obtained for the gene *Cyp4a10* with $\lambda = 0.2, 0.4, 0.6$, and 0.8.

3.1. Signal extraction

Rather than directly computing means of the observed values as in Figure 1, we tried a somewhat more realistic approach based on two essential assumptions:

(i) the values at each time point are noisy observations of the "true" value (obviously unknown),

(ii) this type of biological phenomenon should be a regular, and so differentiable, function of time. This means for us without singularities or any chaotic behavior. This is a sensible assumption when data are acquired at a macroscopic level; it may be false at a molecular or a single-cell scale. Furthermore, in this study, fasting is typically a progressive stimulus where hormonal changes take place progressively and should not imply biological thresholds.

This led us to consider the following nonparametric model for each gene expression:

$$y_i^j = f(t_j) + \varepsilon_{ij}, \quad i = 1, \dots, 4, \ j = 1, \dots, 11, \qquad (1)$$

where y_i^j denotes the observation for the ith mouse ($i = 1, \dots, 4$) at time t_j, f is a continuous and differentiable function, and ε_{ij} are independent and identically distributed random variables satisfying classical assumptions:

$$E(\varepsilon_{ij}) = 0, \qquad \mathrm{Var}(\varepsilon_{ij}) = \sigma^2. \qquad (2)$$

This problem is classically solved by a nonparametric estimation of f. Kernel smoothing or spline smoothing both achieve this objective, but we naturally preferred spline smoothing since we needed to estimate both the function and its derivative. This is quite easy using cubic spline smoothing. The estimation of any gene expression curve according to this model is then the solution to the following optimization problem [17]:

$$\min_{f\in H_1} \frac{1}{4\times 11} \sum_{i=1,4;j=1,11} \left[y_i^j - f(t_j)\right]^2 + \lambda \int_{t_1}^{t_{11}} \left[f''(u)\right]^2 du, \qquad (3)$$

where f belongs to H_1, the Sobolev space of continuous functions with integrable squared second derivative, and λ is the smoothing parameter. This parameter balances the influence between the left-hand term of (3), which forces solutions to be close to mean values, and the right-hand one, which controls the regularity of the function.

The solution $\hat{f}$ of (3) is a piecewise function which is defined on the basis of cubic polynomials. The solution shape and its smoothness depend directly on λ. On the one hand, as λ grows, the solution converges to a trivial linear regression since the integral in the right-hand term of (3) tends to zero (with the second derivative). On the other hand, if λ decreases towards zero, the solution becomes a piecewise polynomial interpolating function of the means of the four values at each time point since the left-hand term reaches its minimum value.

3.2. Tuning the smoothing parameter

The estimation of the function f in model (1) according to formula (3) clearly raises the central problem of how to tune the smoothing parameter λ in order to correctly extract the informative part of the signal. The influence of λ is illustrated with the *Cyp4a10* gene in Figure 2. Depending on the λ value, smoothed profiles exhibit more or fewer fluctuations along the time axis.

We first performed λ tuning by minimizing a generalized cross-validation estimation of a prediction error. Each gene was thus allocated one λ value. Results were disappointing: heterogeneous profiles were clustered together and biological interpretation was very difficult.

Therefore, we adopted another strategy: a unique λ value for all genes. We propose a heuristic approach combining two levels of reflection: eigenelements of the PCA performed *a posteriori* and biological interpretations of results.

Scree graph of eigenvalues and eigenvectors smoothness

The PCA computation requires the number of principal components (PC), that is, the projection space dimension, to be chosen. Some subspace stability argumentation is given

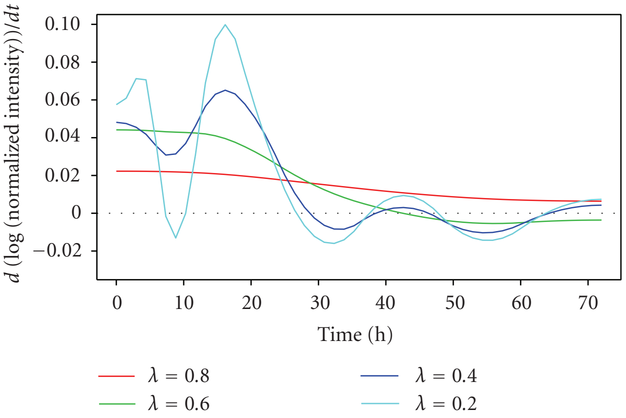


Figure 3: First derivatives of smoothed curves obtained for the gene *Cyp4a10* (Figure 2). Horizontal dotted line locates zero.

in [18] to point out the importance of the difference of values between the last eigenvalue kept and the first that is dropped out.

Practically, let us consider the following steps:

(i) each gene expression profile is smoothed according to the same λ value (Figure 2),

(ii) first derivatives (Figure 3) are computed and discretized, thus giving a new data matrix on which

(iii) a PCA is computed, leading to a scree graph (Figure 4) together with eigenvectors (Figure 5) that are also discretized time functions.

These graphs were plotted for different values of λ (Figures 4 and 5). When λ was large, each expression profile was fitted by a linear regression, and so the derivative was constant, equal to the slope. Obviously, a PCA gave only one large eigenvalue (Figure 4(a)) since the data matrix was of rank one. The same computations were run for different decreasing values of λ until a second eigenvalue arose from noise (Figure 4(b)). The eigenvectors associated with the two largest eigenvalues looked regular and led to easy interpretations of approximations of gene profiles which were projected onto the eigenbasis (Figures 5(a) and 5(b)). But as λ continued to decrease, a third eigenvalue arose from noise (Figures 4(c) and 4(d)) and the first two eigenvectors became much more irregular (Figures 5(c) and 5(d)), and thus much more difficult to interpret, with the risk of giving sense to a noise component.

Biological interpretation

A second consideration which should be addressed is the consistency with biological relevance. For higher λ values, the phenomena highlighted were mainly based on the opposition between the beginning and the end of the experiment. Then, clustering or factorial methods could highlight globally increasing, stationary or decreasing genes without any information about the intermediary period of fasting; two or three time points would have led to the same interpretation. As λ decreased, intermediary time points were integrated (through the second PC) but eigenvectors had to be checked to be smooth enough. Too many oscillations in the eigenvectors could be irrelevant and potentially lead to misinterpretation.

Synthesis

The two levels of consideration yielded approximately the same value for the parameter $\lambda \approx 0.6$. For this value, the detail level of curves was consistent with the number of observations; there were clearly two separate eigenvalues; the corresponding eigenvectors were smooth enough and led to simple and interpretable projection spaces for graphical displays.

3.3. Clustering

The aim of the analysis of these data was to identify some characteristic evolutions of gene regulation occurring during fasting. More precisely, we intended to obtain a few homogeneous clusters of curves, the curves being summarized by the values of the derivative of smoothed expression profiles at some discretization points. We chose 20 points equally spaced between 0 and 72 hours. This value roughly corresponds to the thinnest interval between two real measurements (3 hours) applied all along the 3-days fasting. Furthermore, let us note that when the smoothing is tuned through a penalization parameter, the number and the positions of the points are not very important; practically, results obtained with values from 10 to 50 discretization points were found to be very stable.

The data to be analyzed can be presented in a table with 130 individuals genes in rows and 20 variables dates in columns. The values are the discretized values of the derivative of smoothed curves.

In the context of microarray data analysis, hierarchical clustering is often performed. It was used here in an initial stage. Note that the distance chosen between two curves was the standard Euclidean distance computed between the 20 pairs of coordinates (correlation-based distance would be redundant with the use of the derivative). On the other hand, the criterion chosen to agglomerate two clusters was the Ward criterion, generally advocated by statisticians. It consists in fusing the two clusters that minimize the increase in the total within-cluster sum of squares [19]. We also performed clustering with the information summarized by the first two principal components but, as mentioned by [20], it did not improve the results.

A major weakness of the hierarchical algorithm is that an improper fusion at an early stage cannot be corrected later. In order to correct this weakness, at least partially, we performed a partitioning method (also described as k-means) in which initialization is given by the k centroids of the clusters obtained through hierarchical clustering. See, for example, [21] for a survey of k-means in the context of microarray data.

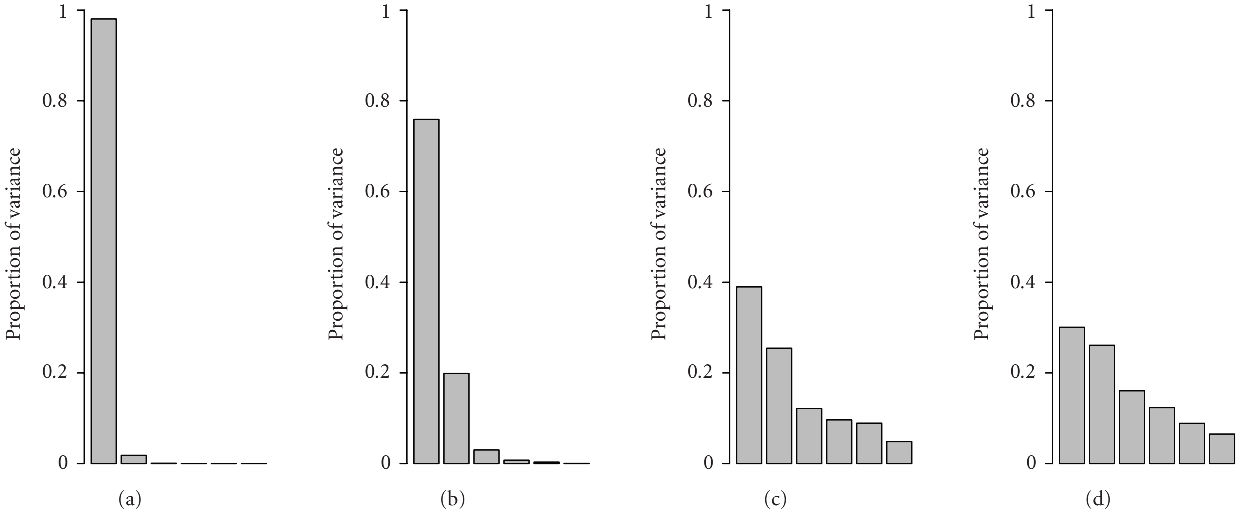


Figure 4: Influence of the smoothing parameter λ on the proportion of variance explained by the first six PCs. From left to right, λ equals (a) 0.8, (b) 0.6, (c) 0.4, and (d) 0.2.

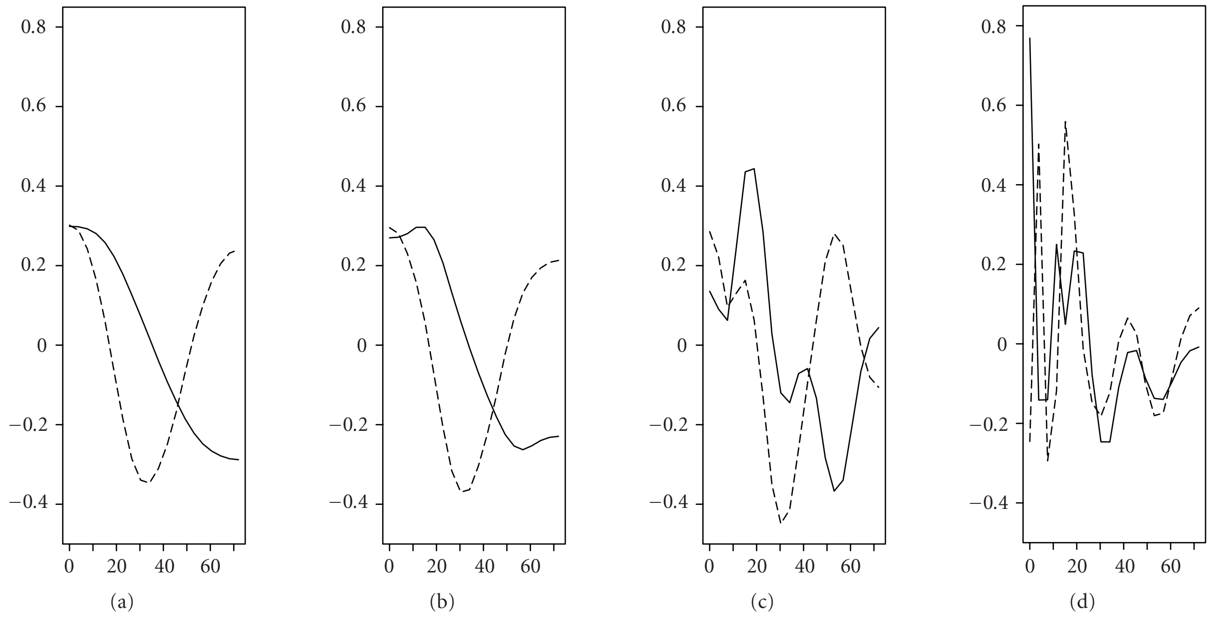


Figure 5: Influence of the smoothing parameter λ on the two first eigenvectors (first: full line, second: dashed line) of the PCA. From left to right, λ equals (a) 0.8, (b) 0.6, (c) 0.4, and (d) 0.2.

4. RESULTS

4.1. *Hierarchical clustering*

Hierarchical clustering produced a dendrogram (Figure 6) that led to arguable choices between 3 and 8 clusters. Four clusters were considered because they led to a relevant and easily perceived biological interpretation. Analysis of more than 4 clusters provides more precise information to the biologist studying gene expression changes during fasting and will be described elsewhere.

Let us note that the four clusters defined by the dendrogram globally correspond to four temporal expression profiles: decreasing (`hc3`), stationary (`hc2`), weakly increasing (`hc1`), strongly increasing (`hc4`).

4.2. *k-means partitioning*

To make the clustering more robust, we performed the k-means algorithm, specifying the initial centers as the centers of the classes obtained when cutting the dendrogram

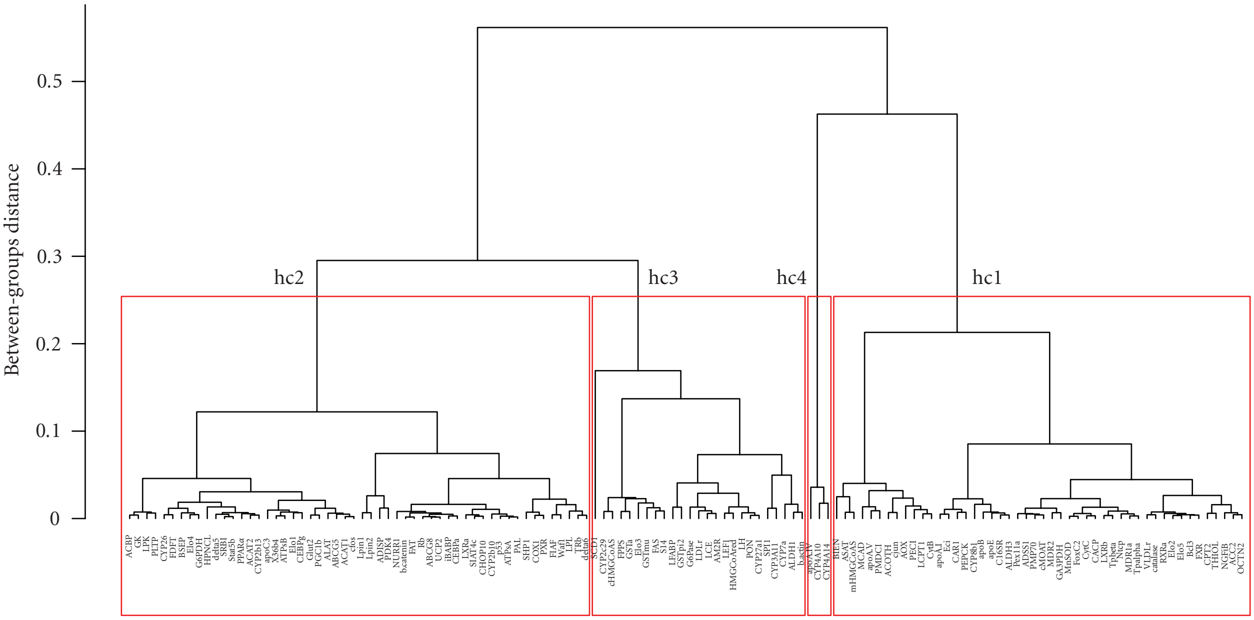


FIGURE 6: Dendrogram representing the result of the hierarchical clustering performed on the value of the first derivative smoothed curves using Euclidean distance and Ward criterion. The horizontal lines locate the cut level identifying 4 clusters (hc1, . . . , hc4).

TABLE 1: Changes between hierarchical clustering and k-means clusters.

Clusters	hc1	hc2	hc3	hc4	Sum
km1	26	3	0	0	29
km2	22	48	4	0	74
km3	0	3	21	0	24
km4	0	0	0	3	3
Sum	48	54	25	3	130

(see Figure 6). Changes that occurred during k-means are summarized in Table 1.

The main event lies in the 22 genes that change from hc1 (low increasing) to km2 (stationary). Other changes are minor and the three-gene cluster (hc4) remains unchanged (km4).

The four clusters of curves obtained after k-means partitioning are displayed in Figure 7; their interpretation is given below.

km1: the expression of the 29 genes which belong to the first cluster increases during the first half of fasting and then tends to decrease slightly or to stabilize. Most of these genes are involved in lipid catabolism. In particular, this cluster contains the genes encoding the three enzymes involved in fatty acid β-oxidation (Acyl-CoA oxidase, BIfunctional ENzyme, and 3-ketoacyl-CoA thiolase) and the enzyme involved in the rate-limiting step of ketogenesis (mitochondrial HMG-CoA synthase). During fasting, lipids stored in the adipose tissue are mobilized and the liver plays a major role in catabolizing these lipids to provide energy and appropriate substrates to peripheral organs. Peroxisome proliferator-activated receptor alpha (PPARα) is an important hepatic transcriptional modulator of lipid catabolism which is activated during fasting [22]. We noticed that most genes in km1 are well-described PPARα targets (reviewed in [23]). PPARα activation and subsequent coordinate induction of km1 genes likely provide a molecular interpretation of their clustering.

km2: the second cluster (74 genes) reveals quasi-constant curves. These genes are not regulated during fasting.

km3: the third one (24 genes) is characterized by a decrease of the gene expression with time. This cluster is mostly composed of genes which are involved in xenobiotic metabolism (the cytochromes P450 3a11, 2c29, and the glutathione-S-transferases α, μ, and π), lipogenesis (FAS, S14, SCD1), cholesterol metabolism (FPP synthase, Cyp7a, cytosolic HMG-CoA synthase, and reductase), and glucose metabolism (glucokinase, pyruvate kinase, and glucose 6-phosphatase). Since large amounts of lipids accumulate in mouse liver during fasting (data not shown), it is likely that the activity of the sterol regulatory element binding proteins (SREBP1 and SREBP2) is reduced. These transcription factors regulate numerous genes involved in lipid synthesis. Their reduced activity may provide a rationale for the decreased expression of lipogenesis and cholesterol synthesis genes. One striking observation is that the liver fatty acid-binding protein (L-FABP), a known PPARα target gene, was also repressed, and is thus found in this third cluster. This result is consistent with a previous report [22] and is currently being investigated.

km4: the fourth cluster is composed of the most strongly induced genes during fasting: *Cyp4a10* and *Cyp4a14*, the two

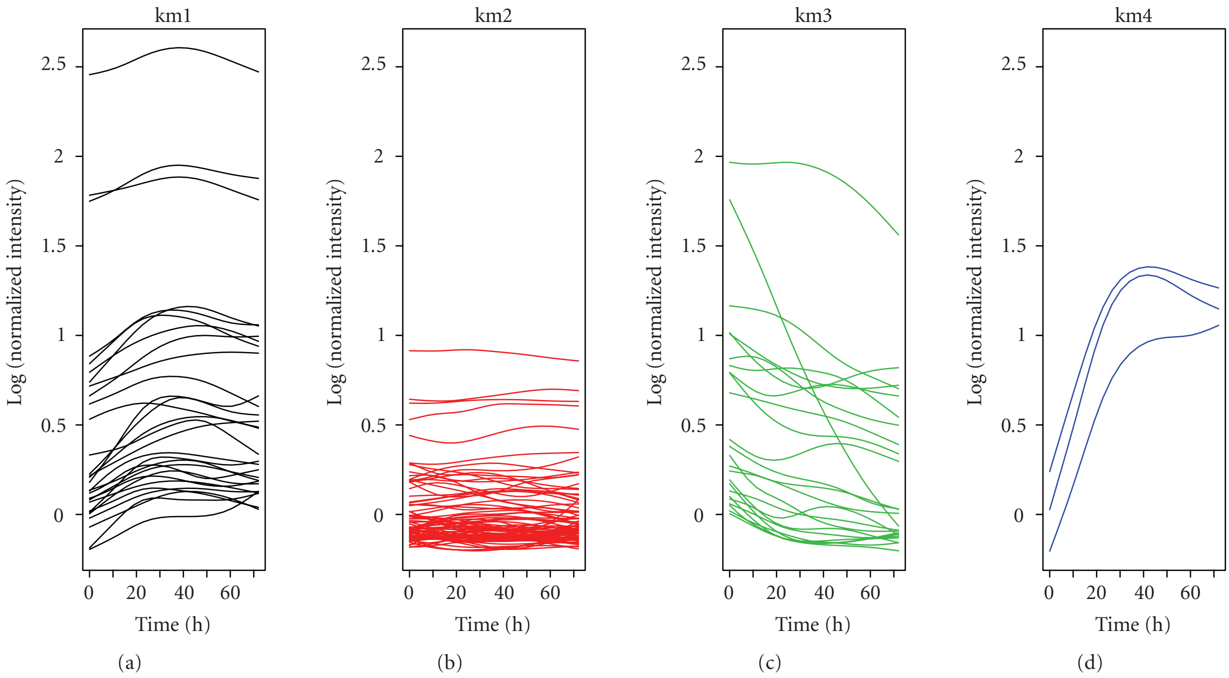


Figure 7: Representation of the smooth curves distributed in 4 clusters determined after hierarchical and k-means classification.

most responsive PPARα target genes and apoA-IV. Their expression strongly increases until the 40th hour of fasting and then stabilizes.

Overall, these results are consistent with the known hepatic gene expression modulations induced by fasting [24]. Hepatic fatty acid oxidation and fatty acid transport and trafficking are induced (mostly through induction of PPARα target genes) and allow the liver to manage, at least partially, the large amounts of lipids which are mobilized from the adipose tissue. On the other hand, lipogenesis and cholesterogenesis are decreased, probably due to reduced SREBP activity. Glucose metabolism genes are decreased, probably in parallel with the decrease in plasma glucose (data not shown). Additionally, some novel hypotheses were drawn from this clustering results and are subject to further experimental investigation.

4.3. Graphical display

We used two methods to give graphical evidence of clusters relevance: PCA and heatmap visualization of simultaneous clustering for genes and time points.

Principal component analysis

We performed a PCA that checked the relevance of the four clusters. The proportion of variance explained by the first two PCs reached about 96% (85% for the first PC), and thus justified a two-dimensional representation (Figure 8).

Genes are shown with different colors according to their cluster (Figure 8 right). The four clusters are distributed along the first (horizontal) axis in a specific order: from left to right, gene expression profiles go from a sharply increasing curve (`km4`, in blue) to a weakly increasing curve (`km1`, genes in black), then stationary profiles (`km2`, genes in red), and finally a decreasing curve (`km3`, genes in green). The second (vertical) axis highlights gene regulations occurring around the 30th hour of fasting. Analysis of more than 4 clusters helps in identifying groups of genes regulated during this intermediary phase of the fasting experiment (data not shown).

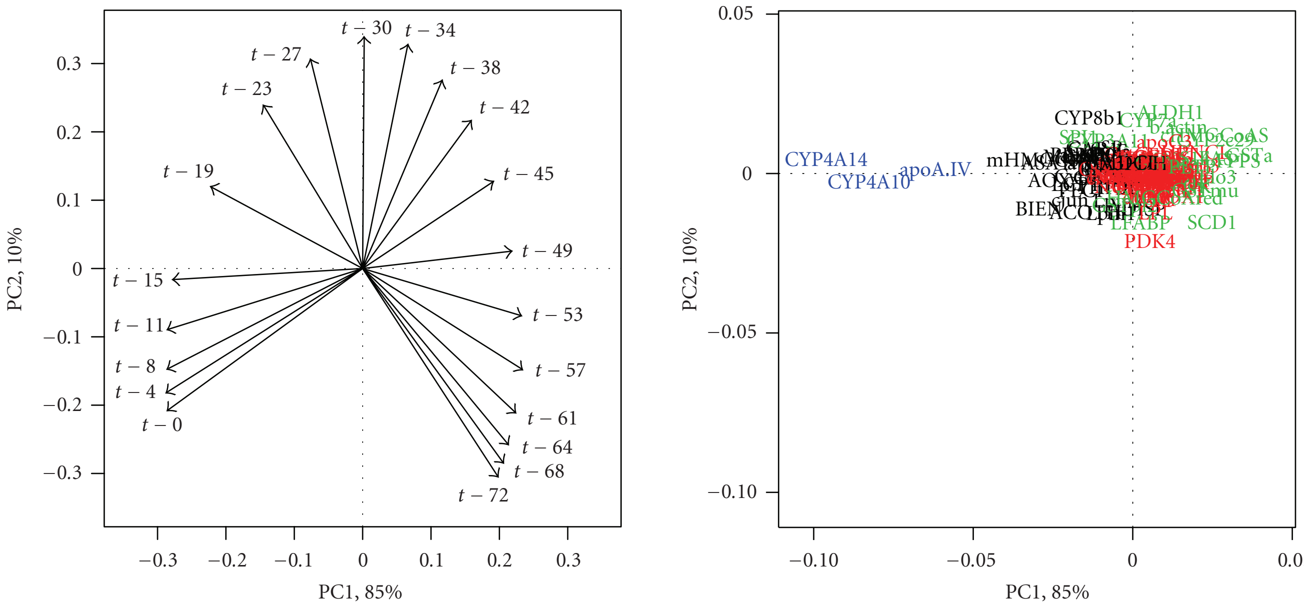
The times of discretization are also shown in Figure 8. Their regular pattern indicates the consistency of the smoothed and discretized data. The sort of inverted U formed by the times of discretization recalls well-known situations of variables connected with time.

Heatmap visualization

Heatmaps are widely used to graphically represent multidimensional gene expression data which have been subjected to clustering algorithms.

We first compared heatmaps obtained on two different data matrices: the matrix of discretized smoothed gene expression profiles; the matrix of discretized derivatives of the smoothed gene expression profiles. In both cases, we forced a reordering of the time points to follow, as much as the dendrogram allows it, their increase from left to right. Perfectly ordered time points were obtained. Genes were systematically reallocated to four clusters using k-means algorithm. This explains why a dendrogram cannot be drawn on the left side of the heatmap. Horizontal lines separated the four clusters obtained following k-means reallocation.

The comparison of the heatmaps obtained (not shown here) clearly highlighted a major advantage of color coding



(a)

(b)

Figure 8: Representation of variables (discretized time points, on the left) and individuals (genes, on the right) on the first two principal components. Genes are differentially displayed according to their cluster after *k*-means.

the derivatives instead of the profiles themselves. When color coding the profiles themselves, the eye needs to integrate the changes of colors along the ordered time points to extract the direction and the amplitude of the changes in gene expression. Conversely, color coding the derivatives allows a direct extraction of gene expression changes direction and amplitude at the different time points. Consequently, it becomes much easier to identify both the causes of the clustering and the time points at which major transcriptional changes occur.

Here, we present two heatmaps computed on the matrix of discretized derivatives of the smoothed gene expression profiles. The clustering of the gene expression profile derivatives was performed as described in the previous paragraphs. Similarly, the hierarchical clustering of the time points was done with the Euclidean distance and the Ward criterion. The first heatmap was computed with all 130 genes (Figure 9). The most strongly regulated genes are easily visualized: `km4` genes at the uppermost and *SCD1* which appears as a green line in the lower quarter of the heatmap. While `km4` genes appear most strongly upregulated until the 30th hour of fasting, *SCD1* is negatively regulated in a constant way during all the fasting periods. Thus, by contrast to `km4` genes, *SCD1* expression profile could have been equally well modelled by a straight line since its derivative appears constant with fasting time. One obvious drawback of this representation (Figure 9) is that the representation of `km4` and *SCD1* gene profile derivatives tend to strongly narrow the color range used to represent the other profile derivatives due to their extreme regulations in mouse liver during fasting. Once interpreted, `km4` and *SCD1* genes were thus removed from the dataset and a new heatmap was computed (Figure 10). Genes belonging to `km1` all display a clear increase in their expression until up to 30 hours of fasting. Their expression is stable from 30 to 45 hours. After 45 hours, divergent regulations are observed (stable, increased, or decreased expression) which could have been highlighted through the analysis of more than 4 clusters. A similar interpretation can be drawn for downregulated `km3` genes located in the lower part of the heatmap.

Interestingly, time points clustering highlighted that most gene expression changes occur during the first 30 hours of fasting although subtle gene expression modulations are still observed after this time point.

5. DISCUSSION

This paper presents an integrated use of statistical tools that provides a framework for the study of time-series data obtained with microarray technology. Before the usual clustering step, we perform spline smoothing as a denoising method. In this context, the quality of the results depends highly on the core problem of tuning the smoothing parameter. For this purpose, we propose an original strategy using both statistical and biological considerations. The procedure is completed by clustering the derivatives of the continuous curves resulting from smoothing, which actually represent the temporal variations of mRNA concentrations.

The main results obtained are clearly in accordance with previous studies on the effects of fasting on hepatic gene expression in the mouse. This study provides a novel time-dependent view of fasting effects on gene expression which are usually studied through 2 or 3 time points only (including a fed state corresponding to time 0). It may thus help in

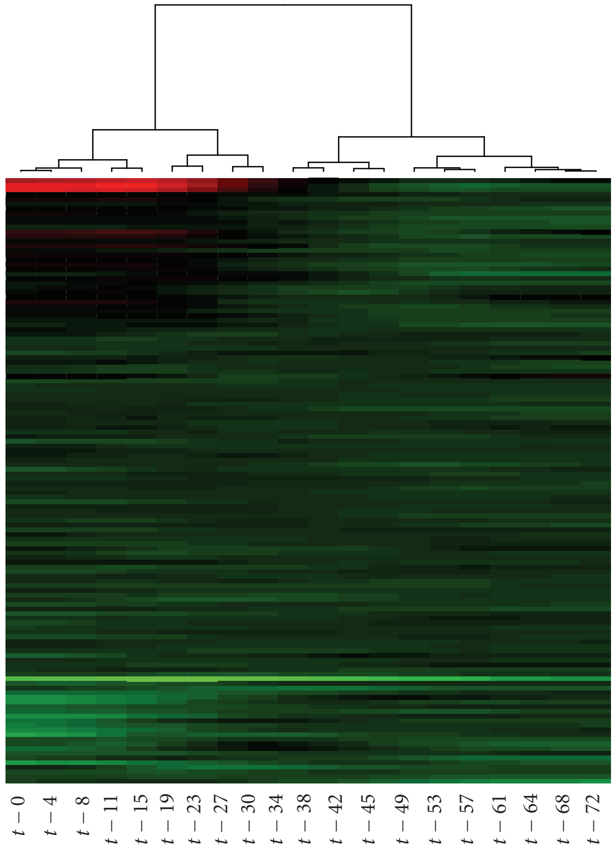


Figure 9: Heatmap of smoothed gene expression profiles for the whole dataset. Genes are ordered according to their cluster determined by the *k*-means algorithm. Horizontal blue lines separate the 4 clusters. Values increase from green to red *via* black.

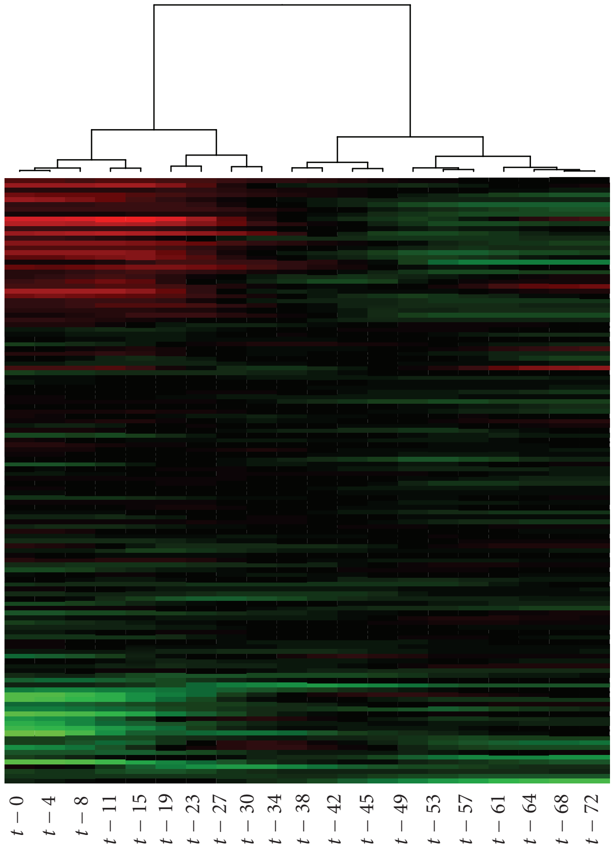


Figure 10: Heatmap of smoothed gene expression profiles without SCD1 and km4-genes. Graphical features are the same as Figure 9.

setting up future experiments where time points can be chosen more adequately depending on the scientific aims. Additionally, this work is the starting point of future investigations aiming at delineating the role of various transcription factors such as PPARα or SREBP in the observed gene expression regulations.

The statistical methodology proposed in the present paper was clearly developed for this specific dataset and its associated scientific aims. Other microarray time-course experiments may benefit from this methodology provided that sufficiently large sample sizes are considered. It is likely that the decreasing cost of microarray technology and the increasing development of cheaper dedicated macroarrays will rapidly yield several suitable time-course datasets.

The dataset studied in this paper and the *R* functions used to perform its analysis are available upon request from the authors.

ACKNOWLEDGMENTS

The authors are grateful to Thierry Pineau, Romain Barnouin, and Henrik Laurell for interesting discussions about biological interpretation of the results. They thank Dominique Haughton for critical review of the manuscript and Alice Vigneron for complementary works on this dataset. This work was partially supported by a grant from ACI IMPBio.

REFERENCES

[1] T. Park, S.-G. Yi, S. Lee, et al., “Statistical tests for identifying differentially expressed genes in time-course microarray experiments,” *Bioinformatics*, vol. 19, no. 6, pp. 694–703, 2003.

[2] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach, “Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference,” *Bioinformatics*, vol. 19, no. 7, pp. 834–841, 2003.

[3] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, “Significance analysis of time course microarray experiments,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12837–12842, 2005.

[4] Y. C. Tai and T. P. Speed, “A multivariate empirical Bayes statistic for replicated microarray time course data,” *The Annals of Statistics*, vol. 34, no. 5, pp. 2387–2412, 2006.

[5] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 14, pp. 9121–9126, 2002.

[6] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21, supplement 1, pp. i159–i168, 2005.

[7] C. D. Giurcăneanu, I. Tăbuş, and J. Astola, "Clustering time series gene expression data based on sum-of-exponentials fitting," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 8, pp. 1159–1173, 2005.

[8] N. A. Heard, C. C. Holmes, D. A. Stephens, D. J. Hand, and G. Dimopoulos, "Bayesian coclustering of Anopheles gene expression time series: study of immune defense response to multiple experimental challenges," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 47, pp. 16939–16944, 2005.

[9] A. Conesa, M. J. Nueda, A. Ferrer, and M. Talón, "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments," *Bioinformatics*, vol. 22, no. 9, pp. 1096–1102, 2006.

[10] J. Letowski, R. Brousseau, and L. Masson, "Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays," *Journal of Microbiological Methods*, vol. 57, no. 2, pp. 269–278, 2004.

[11] J. Ramsay and B. Silverman, *Functional Data Analysis*, Springer, New York, NY, USA, 2nd edition, 2005.

[12] Z. Bar-Joseph, G. K. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon, "Continuous representations of time-series gene expression data," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 341–356, 2003.

[13] Z. Bar-Joseph, "Analyzing time series gene expression data," *Bioinformatics*, vol. 20, no. 16, pp. 2493–2503, 2004.

[14] P. G. P. Martin, F. Lasserre, C. Calleja, et al., "Transcriptional modulations by RXR agonists are only partially subordinated to PPARα signaling and attest additional, organ-specific, molecular cross-talks," *Gene Expression*, vol. 12, no. 3, pp. 177–192, 2005.

[15] P. G. P. Martin, H. Guillou, F. Lasserre, et al., "Novel aspects of PPARα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study," *Hepatology*, vol. 45, no. 3, pp. 767–777, 2007.

[16] INRArray, Laboratoire de Pharmacologie et Toxicologie, INRA, 2005, http://www.inra.fr/internet/Centres/toulouse/pharmacologie/lpt.htm.

[17] B. Silverman, "Some aspects of the spline smoothing approach to non-parametric regression curve fitting," *Journal of the Royal Statistical Society: Series B*, vol. 47, no. 1, pp. 1–52, 1985.

[18] P. Besse, H. Cardot, and F. Ferraty, "Simultaneous nonparametric regressions of unbalanced longitudinal data," *Computational Statistics & Data Analysis*, vol. 24, no. 3, pp. 255–270, 1997.

[19] G. A. F. Seber, *Multivariate Observations*, John Wiley & Sons, New York, NY, USA, 1984.

[20] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[21] H. Chipman, T. J. Hastie, and T. Tibshirani, "Clustering microarray data," in *Statistical Analysis of Gene Expression Microarray Data*, T. Speed, Ed., pp. 159–200, Chapmann & Hall/CRC Press, Boca Raton, Fla, USA, 2003.

[22] S. Kersten, J. Seydoux, J. M. Peters, F. J. Gonzalez, B. Desvergne, and W. Wahli, "Peroxisome proliferator-activated receptor α mediates the adaptive response to fasting," *Journal of Clinical Investigation*, vol. 103, no. 11, pp. 1489–1498, 1999.

[23] S. Mandard, M. Müller, and S. Kersten, "Peroxisome proliferator-activated receptor α target genes," *Cellular and Molecular Life Sciences*, vol. 61, no. 4, pp. 393–416, 2004.

[24] M. Bauer, A. C. Hamm, M. Bonaus, et al., "Starvation response in mouse liver shows strong correlation with life-span-prolonging processes," *Physiological Genomics*, vol. 17, no. 2, pp. 230–244, 2004.