Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2009, Article ID 357937, 5 pages
doi:10.1155/2009/357937

Research Article

Characterizing Gene Expressions Based on Their Temporal Observations

Jiuzhou Song,[1] Hong-Bin Fang,[2] and Kangmin Duan[3]

[1] *Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA*
[2] *Division of Biostatistics, University of Maryland Greenebaum Cancer Center (UMGCC), Baltimore, MD 21201, USA*
[3] *Department of Microbiology and Infectious Diseases, Health Sciences Centre, University of Calgary, Calgary, AB, Canada T2N 4N1*

Correspondence should be addressed to Jiuzhou Song, songj88@umd.edu

Received 31 October 2008; Revised 4 February 2009; Accepted 5 March 2009

Recommended by Zhenqiu Liu

Temporal gene expression data are of particular interest to researchers as they contain rich information in characterization of gene function and have been widely used in biomedical studies. However, extracting information and identifying efficient treatment effects without loss of temporal information are still in problem. In this paper, we propose a method of classifying temporal gene expression curves in which individual expression trajectory is modeled as longitudinal data with changeable variance and covariance structure. The method, mainly based on generalized mixed model, is illustrated by a dense temporal gene expression data in bacteria. We aimed at evaluating gene effects and treatments. The power and time points of measurements are also characterized via the longitudinal mixed model. The results indicated that the proposed methodology is promising for the analysis of temporal gene expression data, and that it could be generally applicable to other high-throughput temporal gene expression analyses.

1. Introduction

The high-throughput gene expression techniques, such as oligonucleotide and DNA microarray, serial analysis gene expression (SAGE) make it possible now to quickly generate huge amount of time series data on gene expression under various conditions [1–5], and have been widely applied in biomedical studies. The current temporal gene expressions usually have several main features: containing large scale of data set, having many genes, involving many procedure noises, and absenting statistical confidence, but few measuring time series levels. Using the difference at two or very few time points to understand changes has also some fundamental limitations. It tells us nothing about each gene's trajectory, and does not consider "overall" difference, nor does it allow studying evolution difference. For these such data with observations at very few time points, the current widely used analysis methods are various clustering methods, fold expression changes, ANOVA [6–9], and recently the hidden Markov chain models (Yuan and Kendziorski 2006). It is simple to interpret the results, and all the available data are analyzed when these methods are applied. However, there are problems associated with these methods which include merely qualifying characteristics of the gene behaviors and clearly absenting quantitative description, and it may take a risk of having false positive and false negative when looking strictly at fold change [9, 10]. Some genetic information may be lost using fold change analysis, and difficulties arise when genes having a bigger folds change in one expression experiment have different performance in multiple arrays or different experiments. It is even more problematic when multiple testing was carried out. For the widely used ANOVA or univariate method, it only analyzes difference between observed means and treats changes of individual gene profile as noise. The main limitation is that the data must be balanced, that is, all measurements occur at same times for all genes, no distinction between unequally spaced time points and equally spaced time points. The ANOVA does not produce a parameter that evaluates the rate of change over time for different treatment groups. Besides, it provides an oversimplification representation for the mean of a data set. The generalized linear models are also used in analyzing gene expression data, but they are based on analyzing the data at each time point separately. They do not take into

account the fact that the gene expression measurements are not independent and do not address the difference in how the mean changes over time. Both the "classical" univariate and multivariate procedures assume that covariance matrix of each data is the same for all measurements at different times, regardless of group or compound symmetry. This assumption implies a very pattern of correlation among observations taken on the same unit at different times which is quite unrealistic for longitudinal data [11]. The other characteristic shared both by the classical univariate and multivariate methods is that time itself does not appear explicitly in the model.

By characterizing the entire pattern of gene expression, and distinguishing the individual gene profile changes subgroup and population-average profile changes, precise estimates with good capability and excellent combination of gene and condition effects were achieved with observations at much more time points. A prospective cohort study where repeated measures are taken over time for each gene is usually designed to answer the following two questions. First, how many observation points are needed over time? Second, how are the variables of interest including genes and conditions associated with each other over time? Therefore, the longitudinal observations with enough time points are most appropriate for the investigation of individual gene changes over time and for the study of effects of other factors such as experimental conditions. In this paper, we illustrate the strategy with an example of a 15-gene set in *Pseudomonas aeruginosa* expressed in three conditions and measured at 48 time points. These 15 genes are either quorum-sensing (QS) genes or quorum sensing regulated genes. Quorum sensing system is a bacterial gene regulatory system that employs small secreted molecules called autoinducers as signaling molecules to coordinate gene expression in a population manner. The autoinducers synthesized and diffused into the growth medium by individual cells increase in amount when the cell number increases, and when the concentration of autoinducers reaches a threshold they bind to cognate transcription regulator to modulate transcriptions of the bacterial genes. So the cell behaves as a whole. The quorum sensing systems in *P. aeruginosa* play a central role in regulating virulence factor expression and in biofilms formation. It has been reported that the expression of one of the genes in QS systems, rhlI is regulated by the iron conditions of the growth medium. However, the extent that this gene is regulated by iron availability is rather small. It is hard to assess the importance of this effect of iron on the QS system in *P. aeruginosa*. Employing the strategy described in this paper, we are able to determine the definite effect of iron availability using a relative large dataset which includes 15 genes over 48 time points in three different conditions totaling 2160 data points.

To analyze such data of temporal gene expressions, the longitudinal mixed model is used. The linear mixed models are extensions of linear regression models used to analyze longitudinal (correlated) data. They accommodate both fixed effects and random effects where the random effects are used to model between-gene variation and the correlation induced by this variation. Linear mixed models are extremely flexible analysis tools, which are especially suitable for unbalanced data with unequally spaced time points and of emphasis on both individual gene level and population-level components. The longitudinal mixed model analysis we present provides a strategy to analyze more complex time series gene expression datasets. The gene expression longitudinal data is characterized by repeated observations over time on the same set of genes, and the repeated observations on the same gene tend to be correlated, therefore, any appropriate statistical analysis must take this correlation into account. The longitudinal mixed model analysis is useful to identify general trends within genes over time, to detect nonlinear changes over time, and also to provide information about the amount of interindividual gene variability. This analysis incorporates different subgroups on the same graph to explain interindividual gene variability.

Table 1: Culture media.

Condition treatments	Description
C1T13	TSBDC
C2T13	TSBDC + 400 ug/mL EDDA
C3T13	TSBDC + 50 ug/mL $FeCl_3$

2. Materials and Methods

2.1. Gene Expression Data in P. Aeruginosa. The promoter regions of selected *P. aeruginosa* virulence factors were amplified by PCR using oligonucleotide primers synthesized [12] according to the PAO1 genome data and PAO1 chromosomal DNA as the template. The PCR amplified promoter regions were then cloned into the *Xho*I-*Bam*HI sites of pMS402 and transformed into PAO1 by electroporation. PCR and DNA manipulation and transformation were performed following general procedures. The promoterless *luxCDABE* operon in pMS402 enables the activity of the promoter fused upstream of the operon to be measured as counts per second (CPS) of light production in a Victor^2 multilabelcounter [12].

TSBDC minimal medium supplemented with EDTA (400 ug/mL) and 50 ug/mL $FeCl_3$ was used in gene expression assays (Table 1). Overnight cultures of the reporter strains were diluted 1 : 200 in a 96-well microtiter plate and the promoteractivity of the virulence factors in different conditions was measured every 30 minutes for 24 hours. Bacterial growth was monitored at the same time by measuring the optical density at 620 nm (OD_{620}) in the Victor^2 multilabel counter.

2.2. Statistical Methods. To analyze these longitudinal data of temporal gene expressions, the mixed model

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad (1)$$

will be used, where Y_i is an $(n_i \times 1)$ vector of expression for the ith gene, $i = 1, \ldots, m$. X_i is an $(n_i \times p)$ design matrix that characterizes the systematic part of the gene expression, for example, depending on covariates and time. β is a $(p \times 1)$ vector of parameters usually referred to as fixed

effects, that complete the characterization of the systematic part of the gene expression. Z_i is an $(n_i \times k)$ design matrix that characterizes random variation in the response attributable to among genes. b_i is a $(k \times 1)$ vector of the random effects variables that completes the characterization of the among-gene variation. ε_i is an $(n_i \times 1)$ vector of within-gene errors characterizing variation due to the way in which the expression levels are measured on the ith gene.

The data vector Y_i has a multivariate normal distribution with $E(Y_i) = X_i\beta$, $\text{var}(Y_i) = Z_i D Z_i' + R_i = \Sigma_i$, and $Y_i \sim N(X_i\beta, \Sigma_i)$. Here, the usual assumptions are $b_i \sim N(0, D)$, D is a $(k \times k)$ covariance matrix that characterizes variation due to among-gene source, and the dimension of D corresponds to the number of among-gene random effects in model. $\varepsilon_i \sim N(0, R_i)$, R_i is an $(n_i \times n_i)$ covariance matrix that characterizes variance and correlation due to within-gene sources. The form of Σ_i implied by the model has two distinct components, the first having to do with variation solely from among-gene sources and the second having to do with variation solely from within-gene sources. We used maximum likelihood (ML), restricted maximum likelihood (REML), and minimum variance quadratic unbiased estimation (MIVQUE0) to estimate the covariance parameters of the G and R, respectively.

In order to check the influence of temporal measurements for longitudinal mixed analysis, we further constructed a dataset of the same dimension and with the same covariates and factor values for which power is to be calculated. With F-test statistics, we calculated noncentrality parameter (ϕ) and degrees of freedom ν_1 and ν_2, then power is calculated as $P(F_{\nu_1,\nu_2,0} > F_C)$, F_C is a critical value. All analyses were implemented by SAS package.

3. Results and Analysis

3.1. The Trajectories of the Longitudinal Gene Expression Data. To validate the models for our data set, we plotted the expression profiles for all genes under different conditions. The trajectories of the 15-gene set are shown in Figure 1. From the figure, we can see that there is high degree of variations between genes. There are also correlation genes at different time points, and the correlation structure cannot be ignored in analysis. The expression trajectories of the genes change over time for all of the genes, and at a certain time point, the change rate for each gene is different from other time point and from that of other genes. From Figure 2, we can see that the trajectories of experimental treats are also changing over time, and the change rate varies from conditions.

3.2. Choice of and Assessing the Goodness-of-Fit Covariance Structure. In the longitudinal data, there are three sources of error in the residual, including serial correlation, measurement error, and random component. In order to use longitudinal mixed-model methodology, it is assumed that the data has a linear mean and a reasonable covariance structure. The reasonable covariance structure is a parsimonious covariance just enough to be estimated with available current data and yet rich to capture probable covariance between

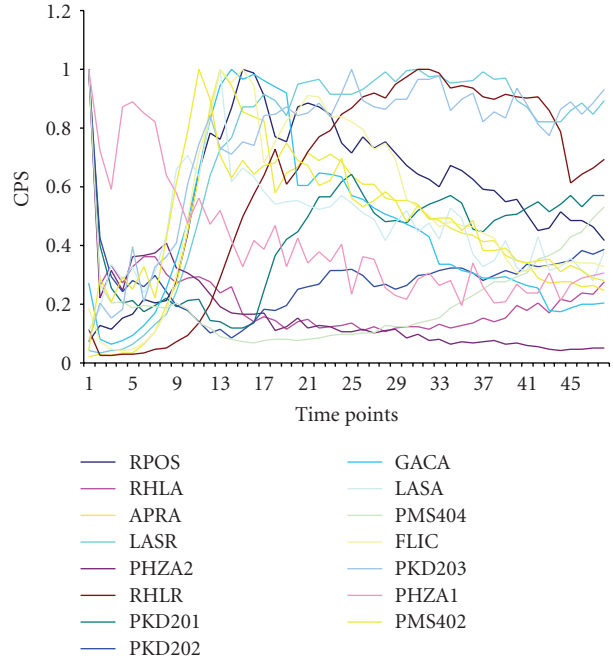

Figure 1: The trajectories of the 15 gene-set in C1T13: TSBDC condition.

Table 2: Covariance structures using ML.

Model	Description	AIC	BIC	−2 log likelihood
1	General linear model (GLM)	1811.8	1856.2	1798.8
2	Compound symmetry (CS)	1811.5	1856.0	1796.7
3	Variance components (VC)	1665.0	1651.3	1645.0
4	Heterogeneous CS (CSH)	1636.8	1618.0	1610.8
5	Spatial power (SP)	1689.2	1685.6	1600.2

AIC: Akaike's information criteria; BIC: Bayesian information criteria for each model selected.

gene expression observations. The fitting information shown in Table 2 provides some statistics about the estimated mixed model. The log likelihood supplies the estimation information of covariance G and R in the mixed models. Akaike's information criteria (AIC) can be used to compare models with the same fixed effects but different variance structures. Models having the smallest AIC are deemed the best. The Schwarz Bayesian criteria (BIC) are also computed, and models with smaller BIC are also preferred. The six models with different covariance structure were fitted, and preference was selected based on the AIC and BIC values. Inspection of AIC and BIC values for each of the six models revealed that the values of both the AIC and BIC in the assumed same covariance structure are larger than those of the assumed different ones. Both criteria are the smallest for the chosen separate spatial power (SP) structures for each treatment. The values of both AIC and BIC in SP are the minimum among the models. The log likelihood of the model is also the best for separate SP structures. As both criteria agree, it would be sensible to choose the

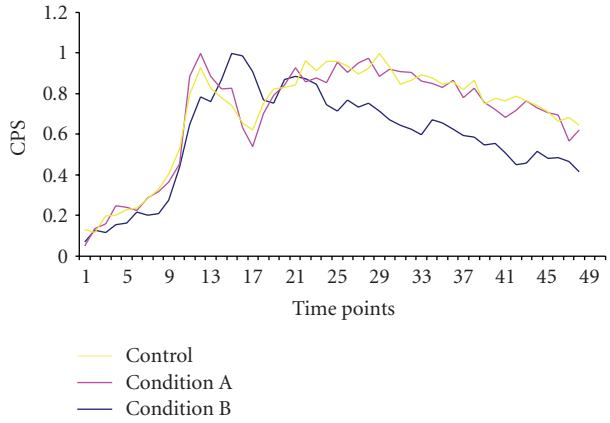

Figure 2: The trajectories of one gene in 3 conditions. Control: TSBDC, Condition A: TSBDC + 50 ug/mL $FeCl_3$ Condition B: TSBDC + 400 ug/mL EDDA.

model to represent the covariance structure that has different variance and covariance in different treatments. Interestingly, we found there were the almost same values AIC, BIC, and likelihood value between GLM and CS model, which indicated that univariate GLM calculations are identifical to MIXED estimates when using CS for the balanced data sets. The multivariate GLM cannot determine best fit when the data set is a longitudinal data.

3.3. Power and Sample Size Determination for Longitudinal Mixed Model. In statistical analysis, one typically expresses the belief that some effects exist in a population by specifying an alternative hypothesis to H_1, a null hypothesis H_0 as the assertion that effect does not exist and attempt to gather evidence to reject H_0 in favor of H_1. If H_0 is rejected but there is really no effect, this is called a Type I error, which is usually designated α; if there really is an effect in the population but H_0 is not rejected, then a Type II error has been made, which is usually designated β. The probability $1 - \beta$ of avoiding a Type II error, that is, correctly rejecting H_0 and achieving statistical significance, is called the power. We simulated our data structure and calculated the power of estimating condition effects via the longitudinal mixed model. As shown in Figure 3, we found the model can get maximum power while more than 7 or 8 measurements were taken. So the 48 temporal measurements of each gene in our research could have enough power to obtain the estimation of treatments and gene effects.

3.4. Estimation of the Effect of Iron Condition on QS Genes by the Mixed Model. We adopted the longitudinal mixed model with heterogeneous compound symmetry variance to estimate the effects of iron condition on QS genes expression. From Figure 2, the effects of the culture media TDBDC and TSBDC + 400 ug/mL EDDA are almost equal and higher than that of TDBDC + $FeCl_3$. Comparing with the TSBDC, the addition of TSBDC + 50 ug/mL $FeCl_3$ positively regulates the expression of these genes as shown in Figure 4. To check the detailed differences of the genes, the longitudinal mixed model was used to estimate the gene effects, as shown

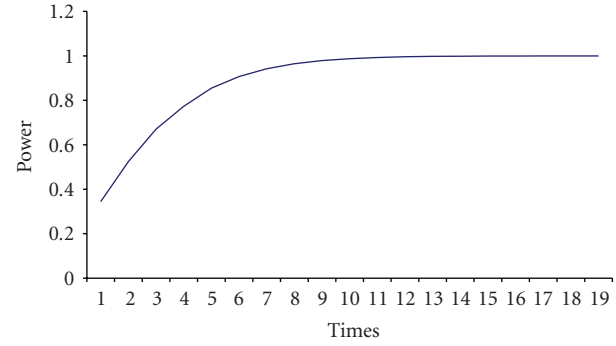

Figure 3: Power analysis under the longitudinal mixed model with heterogeneous compound symmetry variance structure.

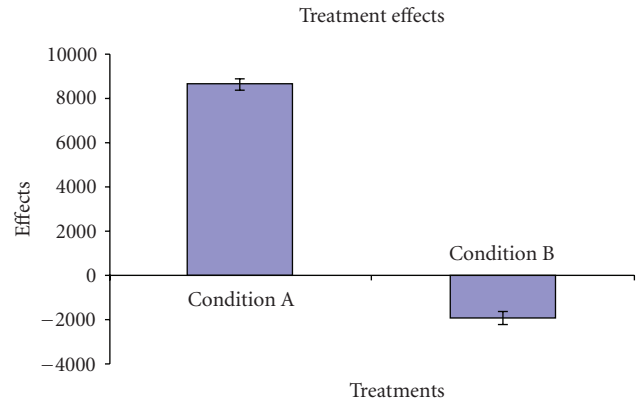

Figure 4: The estimation of condition effects. Condition A: TSBDC + 50 ug/mL $FeCl_3$,Condition B: TSBDC + 400 ug/mL EDDA.

in Figure 5. We found that most of genes, including FliC, LasR, PKD202, PKD203, and PhlR, demonstrate positive expression effects in condition of addition of 400 ug/mL EDDA, whereas PhlA shows opposite expression effect.

4. Discussion

The identification of genes that show changes in expression between varying biological conditions is a frequent goal in microarray experiments. Under different biological conditions, the patterns of gene expressions may be various. To obtain efficient information for temporal gene expression, the number of longitudinal observations should be enough for individual gene changes over time and the study of effects with biological conditions.

In longitudinal studies, time effect is the changes over time for each gene, and cohort effect is the 22 differences among genes in their baseline values. Longitudinal studies can distinguish these time and cohort effects while cross-sectional studies cannot. In this paper, we have considered mixed model with longitudinal covariates, the analysis of longitudinal data should take into account firstly, the within-subject correlation, secondly the measurements taken at unequal time intervals and finally the missing observations. Repeated measures analysis of variance can be used to

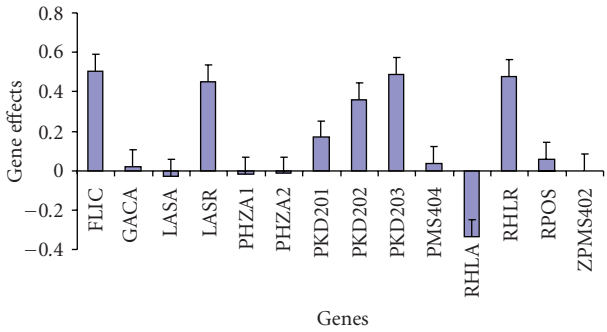

FIGURE 5: The estimation of gene effects under condition TSBDC + 50 ug/mL $FeCl_3$.

analyze longitudinal or repeated measures data for balanced study design, that is, when all genes are measured at equal time points and there are no missing data. In large scale of gene expression analysis, if having unbalanced datasets in longitudinal studies, it is necessary to use some alternative techniques which can handle unbalanced data. In this research, we confirmed that univariate GLM calculations are identifical to MIXED estimates when using CS for balanced data sets. The multivariate GLM cannot determine best fit when the data set is a longitudinal data. Therefore, the procedures of best fit mixed model include: (1) the choice of the model, (2) the choice of the variance-covariance structure (specifying the working correlation structure for each gene, e.g., independence, exchangeable, stationary, and autoregressive), (3) assessing the goodness-of-fit of the model, and (4) assessing the goodness-of-fit of the variance covariance structure.

Although the paper only analyzed the effects of three treatments and 15-gene effects, it proved that the longitudinal mixed model is a feasible method in dense temporal gene expression analysis. We found that the addition of TSBDC + 50 ug/mL $FeCl_3$ positively regulates the expression of these genes in our analysis. It has been reported that iron availability in the growth condition affects the expression of genes. However, the changes of expression are rather small. It is thus difficult to assess whether there is a pronounced effect of iron on the QS genes. Accordingly the current analysis method, using the mixed model described aforementioned a definite effect could be determined. A comprehensive understanding of biological processes requires the acquisition of expression data at different developmental stages, in different tissues and different treatment conditions with different organisms. The addition of time as a variable allows observation of the modulation of gene expression whether due to the regulation of development or the changing impact of a treatment condition. The expectation is that high-throughput gene expression analysis conducted in the higher dimensions of genes, conditions, tissues, and time as variables will help elucidate what the genes do, when, where, and how they are expressed as elements of an orchestrated system under the effects of perturbations and developmental processes, and we will explore the possibility of generalized mixed model in higher dimensions expression data [13–16].

References

[1] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[2] H. Zhu, Y. Tang, L. Ivanciu, et al., "Temporal dynamics of gene expression in the lung in a baboon model of E. coli sepsis," *BMC Genomics*, vol. 8, article 58, pp. 1–23, 2007.

[3] R. J. Cho, M. J. Campbell, E. A. Winzeler, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[4] P. T. Spellman, G. Sherlock, M. Q. Zhang, et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[5] J. Bjarnason, C. M. Southward, and M. G. Surette, "Genomic profiling of iron-responsive genes in *Salmonella enterica* serovar Typhimurium by high-throughput screening of a random promoter library," *Journal of Bacteriology*, vol. 185, no. 16, pp. 4973–4982, 2003.

[6] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

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

[8] H. Li, Y. Luan, F. Hong, and Y. Li, "Statistical methods for analysis of time course gene expression data," *Frontiers in Bioscience*, vol. 7, pp. a90–a98, 2002.

[9] S. Draghici, O. Kulaeva, B. Hoff, A. Petrov, S. Shams, and M. A. Tainsky, "Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays," *Bioinformatics*, vol. 19, no. 11, pp. 1348–1359, 2003.

[10] T. S. Tanaka, S. A. Jaradat, M. K. Lim, et al., "Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 16, pp. 9127–9132, 2000.

[11] S. L. Zeger and K.-Y. Liang, "An overview of methods for the analysis of longitudinal data," *Statistics in Medicine*, vol. 11, no. 14-15, pp. 1825–1839, 1992.

[12] K. Duan, C. Dammel, J. Stein, H. Rabin, and M. G. Surette, "Modulation of *Pseudomonas aeruginosa* gene expression by host microflora through interspecies communication," *Molecular Microbiology*, vol. 50, no. 5, pp. 1477–1491, 2003.

[13] J. Z. Song, K. M. Duan, T. Ware, and M. Surette, "The wavelet-based cluster analysis for temporal gene expression data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 39382, 7 pages, 2007.

[14] H. V. Westerhoff, E. Mosekilde, C. R. Noe, and A. M. Clemensen, "Integrating systems approaches into pharmaceutical sciences," *European Journal of Pharmaceutical Sciences*, vol. 35, no. 1-2, pp. 1–4, 2008.

[15] H. V. Westerhoff, A. Kolodkin, R. Conradie, et al., "Systems biology towards life in silico: mathematics of the control of living cells," *Journal of Mathematical Biology*, vol. 58, no. 1-2, pp. 7–34, 2009.

[16] I. P. Androulakis, E. Yang, and R. R. Almon, "Analysis of time-series gene expression data: methods, challenges, and opportunities," *Annual Review of Biomedical Engineering*, vol. 9, pp. 205–228, 2007.