BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 14 2008, pages 1619–1624
doi:10.1093/bioinformatics/btn246

Systems biology

Estimating dynamic models for gene regulation networks

Jiguo Cao[1] and Hongyu Zhao[2,*]

[1]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada and [2]Department of Epidemiology and Public Health, Yale University, School of Medicine, 60 College Street, New Haven, CT 06520-8034, USA

Received on April 25, 2008; revised on April 26, 2008; accepted on May 23, 2008

Advanced Access publication May 27, 2008

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Transcription regulation is a fundamental process in biology, and it is important to model the dynamic behavior of gene regulation networks. Many approaches have been proposed to specify the network structure. However, finding the network connectivity is not sufficient to understand the network dynamics. Instead, one needs to model the regulation reactions, usually with a set of ordinary differential equations (ODEs). Because some of the parameters involved in these ODEs are unknown, their values need to be inferred from the observed data.

Results: In this article, we introduce the generalized profiling method to estimate ODE parameters in a gene regulation network from microarray gene expression data which can be rather noisy. Because numerically solving ODEs is computationally expensive, we apply the penalized smoothing technique, a fast and stable computational method to approximate ODE solutions. The ODE solutions with our parameter estimates fit the data well. A goodness-of-fit test of dynamic models is developed to identify gene regulation networks.

Contact: hongyu.zhao@yale.edu

1 INTRODUCTION

Gene expression is a highly regulated and fundamental biological process. Transcription is directed by a set of transcription factors, which may interact with each other for proper activation or inhibition of genes. A transcriptional regulatory network refers to the collection of genes, their regulators and their interactions. With recent advances in genomics technologies, there have been extensive research efforts on elucidating regulatory networks. See Alon (2007) for an excellent review on this topic. One of the first discoveries in genomic-level analysis of networks is that certain regulation patterns occur much more often than by chance, and these patterns are called network motifs. Among these motifs, the feed forward loop (FFL) involves three Genes X, Y, Z in which X regulates the expressions of Y and Z, and Y regulates the expression of Z. Gasch *et al.* (2000) collected time course gene expression data in the yeast *Saccharomyces cerevisiae* under different environmental changes and found that a large set of genes responded to almost all environmental transitions they made. Figure 1 shows the expression profiles of three genes (X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5) after the temperature is increased from 25°C to 37°C. These three genes compose a so-called Coherent Type 1 FFL, a type of FFL where X activates the expressions of Y and Z, and Y activates the expression of Z (Mangan and Alon, 2003).

An FFL can be modeled by a set of simple first order ODEs (Barkai and Leibler, 1997; Mangan and Alon, 2003; Savageau, 1976; Shen-Orr *et al.*, 2002). For example, the following ODEs have been used to model a Coherent Type 1 FFL:

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t}=-\alpha_y Y(t)+\beta_y f(X(t),K_{xy}), \tag{1}$$

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t}=-\alpha_z Z(t)+\beta_z g(X(t),Y(t),K_{xz},K_{yz}), \tag{2}$$

where $X(t)$, $Y(t)$ and $Z(t)$ represent the expression levels of Genes X, Y and Z, respectively, at time t. The activation function $f(u,K)=(u/K)^H/(1+(u/K)^H)$ has two parameters H and K. Parameter H controls the steepness of $f(u,K)$, and we choose $H=2$ in our following analysis. Parameter K defines the expression of Gene X required to significantly activate expression of other genes. When $u=K$, $f(u,K)=0.5$. Assuming Genes X and Y regulate Gene Z independently, the activation function from Genes X and Y to Gene Z is $g(t)=f(X(t),K_{xz})f(Y(t),K_{yz})$. Our objective is to estimate the vector of ODE parameters $\boldsymbol{\theta}=(\beta_y,\beta_z,\alpha_y,\alpha_z,K_{xy},K_{xz},K_{yz})$ from the noisy measurements for gene expression.

ODEs are popular tools to model dynamic processes in engineering, biology and many other areas. For example, in classical mechanics, the motion of a body is described in terms of its position, velocity (the first derivative of the position function) and acceleration (the second derivative of the position function) as time varies. Newton's Laws relate the position, velocity, acceleration and forces acting on the body with differential equations. How to estimate ODE parameters is an old but difficult problem. When ODEs have analytical solutions, this is essentially a non-linear regression problem (Bates and Watts, 1988). Unfortunately, most ODEs do not have analytical solutions.

Many methods have been proposed for estimating ODEs which cannot be solved analytically. Bock (1981) proposed a multiple shooting method, which partitioned the whole time interval into segments and applied the non-linear optimization procedure over each segment. Himmelblau *et al.* (1967) converted ODEs to non-linear equations by integrating ODEs with numerical quadrature. de Boor and Swartz (1973) approximated ODE solutions with piecewise polynomial functions by collocation. Ramsay and Silverman (2005) estimated the derivatives of the underlying curves by non-parametric smoothing, and then estimated ODE parameters with standard non-linear regression. Huang *et al.* (2005) proposed

[*]To whom correspondence should be addressed.

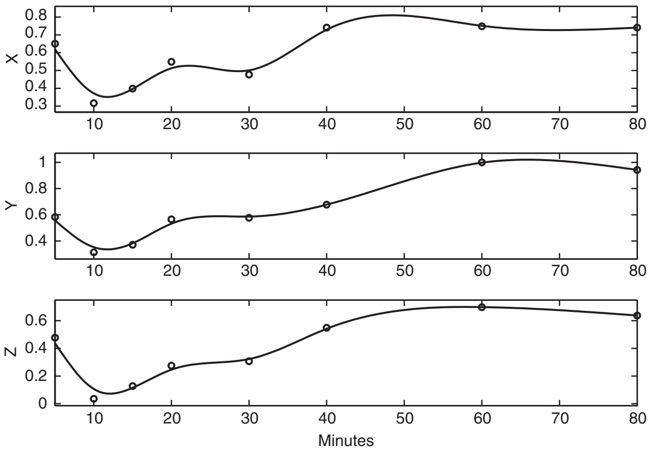

Fig. 1. The expression profiles of three genes (X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5) measured at 5, 10, 15, 20, 30, 40, 60, 80 min. The data were collected by DNA microarrays from yeast after the temperature was increased from 25°C to 37°C (Gasch *et al.*, 2000). The solid lines are the smooth curves estimated by penalized spline smoothing (The basis functions are cubic B-splines with 40 equal-spaced knots, and the value of the smoothing parameter is 10).

a Bayesian approach by numerically solving ODEs when updating ODE parameters.

Most recently, Ramsay *et al.* (2007) proposed a generalized profiling method to estimate ODE parameters. The ODE solution is approximated with a linear combination of basis functions. The basis coefficients are estimated with penalized smoothing with an ODE-defined penalty. The smoothing parameter controls the trade off between fitting the data and fidelity to the ODEs. Their method has several unique aspects. For example, the computation load is much lower than the other methods because it avoids the computational expense to obtain numerical ODE solutions. The method can estimate some ODE components when they are missing. It can also estimate initial values of ODE components by evaluating the fitted curves at the first time points. Cao and Ramsay (2007) extended this method and estimated the smoothing parameter in three nested levels of optimization.

Our article is organized as follows. Section 2 introduces the generalized profiling method in details. Section 3 applies the generalized profiling method to estimate the ODE parameters in (1) and (2) from the noisy measurements for gene expression. Section 4 validates the generalized profiling method by simulations. Section 5 introduces a goodness-of-fit test for the adequacy of the FFL to model the observed expression data. Section 6 gives the conclusions and discussion.

2 THE GENERALIZED PROFILING METHOD

We start with a simple dynamic system composed of one single component, and then extend to more general cases. Let $X(t)$ be a process modeled by one ODE $\mathrm{d}X/\mathrm{d}t = f(X|\boldsymbol{\theta})$, where $f(X|\boldsymbol{\theta})$ is known. We have n observations $y(t_j)$ at a number of time points with mean $X(t_j)$ and SD σ_j, $j=1,\ldots,n$, where $X(t)$ is the solution to the ODE over t. The parameter vector $\boldsymbol{\theta}$ is unknown and has to be estimated from data $\mathbf{y}=(y(t_1),\ldots,y(t_n))$. Two nested levels of optimization are implemented. In the inner level, we approximate $X(t)$ with a smooth curve $x(t)$ by penalized smoothing with the ODE-defined penalty, conditional on the ODE parameter vector $\boldsymbol{\theta}$. So the fitted curve $\hat{x}(t)$ can be treated as a function of $\boldsymbol{\theta}$. In the outer level, the ODE parameter vector is estimated by minimizing the likelihood function (the likelihood function is a function of the fitted curve, and thus is also a function of $\boldsymbol{\theta}$). The standard error for $\hat{\boldsymbol{\theta}}$ is estimated by the modified delta method, which includes the uncertainty coming from the fitted curve.

2.1 The point estimate for the ODE parameter

The ODE solution is approximated by a linear expansion of K basis functions $\phi_k(t), k=1,\ldots,K$, as follows:

$$x(t)=\sum_{k}^{K} c_k\phi_k(t)=\mathbf{c}'\boldsymbol{\phi}(t).$$

The basis system must have the flexibility to approximate ODE solutions and their derivatives. Many dynamic systems have sharp features, such as peaks, valleys, high frequency oscillations and discontinuities in derivatives. Ramsay and Silverman (2005) showed that the B-spline basis system can accommodate the discontinuities by assigning multiple knots to the critical locations. The B-spline basis functions are only non-zero in local intervals, and the computation can be made efficient using this property of local supports. In practice, the cubic B-splines are often used as the basis system. Usually we put one knot on each point with observations, but when data are sparse or the dynamic systems have sharp features, more knots are required to make the cubic B-spline flexible enough.

The basis coefficient $\mathbf{c}$ is estimated by optimizing the penalized likelihood function:

$$J(\mathbf{c}|\boldsymbol{\theta},\lambda,\mathbf{y})=-l(\mathbf{c}|\mathbf{y})+\lambda\int[\mathbf{c}'\frac{\mathrm{d}\boldsymbol{\phi}(t)}{\mathrm{d}t}-f(\mathbf{c}'\boldsymbol{\phi}(t)|\boldsymbol{\theta})]^2\mathrm{d}t,$$

where $l(\mathbf{c}|\mathbf{y})$ is the log likelihood function for $\mathbf{c}$. The fit of the smoothing curve $x(t)$ to the ODE is measured in the second term, which is the integrated squared difference between the two sides of the ODE. The smoothing parameter controls the tradeoff between fitting the data and fidelity to the ODE.

In general, suppose the dynamic system has G ODEs and T components:

$$\frac{\mathrm{d}x_g(t)}{\mathrm{d}t}=f_g(x_1(t),x_2(t),\ldots,x_T(t)|\boldsymbol{\theta}),g=1,\ldots,G,$$

where $f_g(x_1(t),x_2(t),\ldots,x_T(t)|\boldsymbol{\theta})$ is known, and $\boldsymbol{\theta}$ is the ODE parameter that has to be estimated. Each component is approximated by a linear combination of basis functions: $x_i(t)=\mathbf{c}_i\boldsymbol{\phi}_i(t), i=1,\ldots,T$. Suppose we have observations for only $M\leq T$ of these components: $\mathbf{y}_m=(y_m(t_{m1}),\ldots,y_m(t_{mn_m})), m=1,\ldots,M$. Let $\mathbf{c}=(\mathbf{c}_1',\ldots,\mathbf{c}_T')'$ and $\mathbf{y}=(\mathbf{y}_1',\ldots,\mathbf{y}_M')'$, then the fitting criterion can be generalized to be:

$$J(\mathbf{c}|\boldsymbol{\theta},\lambda,\mathbf{y})=-\sum_{j=1}^{M}\omega_j l(\mathbf{c}_j|\mathbf{y}_j)+$$

$$\sum_{\ell=1}^{G}\lambda_\ell\omega_\ell\int[\mathbf{c}_\ell'\frac{\mathrm{d}\boldsymbol{\phi}_\ell(t)}{\mathrm{d}t}-f(\mathbf{c}_1'\boldsymbol{\phi}_1(t),\ldots,\mathbf{c}_T'\boldsymbol{\phi}_T(t)|\boldsymbol{\theta})]^2\mathrm{d}t, \qquad (3)$$

where ω_j is the normalizing weight, which can be used to keep different components having comparable scales. For example, in this study we set the values of ω_j as the reciprocals of the variances taken over observations for the j-th component. There are sometimes some unobservable components, that is, $M < T$, but the basis coefficients for missing components are involved in the second term and can still be estimated. When $f_g(x_1(t), x_2(t), \ldots, x_T(t)|\boldsymbol{\theta})$ is a non-linear function, the integration terms can be approximated by numerical quadrature. For each given value of $\boldsymbol{\theta}$, we obtain one estimate $\hat{\mathbf{c}}$ by optimizing (3), so $\hat{\mathbf{c}}$ can be viewed as a function $\boldsymbol{\theta}$.

In the outer level, we estimate the ODE parameter $\boldsymbol{\theta}$ by optimizing the log likelihood function:

$$H(\boldsymbol{\theta}|\lambda, \mathbf{y}) = -\sum_{j=1}^{M} \omega_j l(\hat{\mathbf{c}}_j(\boldsymbol{\theta})|\mathbf{y}_j)$$

We can obtain the gradients and Hessian matrices analytically, so the optimization can be fast and stable. When $f_g(x_1(t), x_2(t), \ldots, x_T(t)|\boldsymbol{\theta})$ is a non-linear function, $\hat{\mathbf{c}}$ is an implicit function of $\boldsymbol{\theta}$, and the Implicit Function Theorem can be used to get the analytical gradients and Hessian matrices.

2.2 Statistical inference for the ODE parameter

Let Σ be the variance–covariance matrix for the data $\mathbf{y}$. Ramsay *et al.* (2007) show that the variance estimate for $\hat{\boldsymbol{\theta}}$ can be obtained with a modified delta method:

$$\widehat{\mathtt{Var}}(\hat{\boldsymbol{\theta}}) = \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right] \Sigma \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]',$$

where $d\hat{\boldsymbol{\theta}}/d\mathbf{y}$ can be derived with the implicit function theorem:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = -\left[\left.\frac{d^2H}{d\boldsymbol{\theta}^2}\right|_{\hat{\boldsymbol{\theta}}}\right]^{-1} \left[\left.\frac{d^2H}{d\boldsymbol{\theta} d\mathbf{y}}\right|_{\hat{\boldsymbol{\theta}}}\right],$$

where

$$\frac{d^2H}{d\boldsymbol{\theta}^2} = \frac{\partial^2 H}{\partial\boldsymbol{\theta}^2} + 2\frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\boldsymbol{\theta}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)' \frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}^2},$$

and

$$\frac{d^2H}{d\boldsymbol{\theta} d\mathbf{y}} = \frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\mathbf{y}} + \left[\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right]' \frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left[\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right]' \frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$$
$$+ \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\mathbf{y}}.$$

The variance for $\hat{\mathbf{c}}$ can also be estimated with the modified delta method:

$$\widehat{\mathtt{Var}}(\hat{\mathbf{c}}) = \left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right] \Sigma \left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right]',$$

where

$$\frac{d\hat{\mathbf{c}}}{d\mathbf{y}} = \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\frac{d\hat{\boldsymbol{\theta}}}{\partial\mathbf{y}}. \tag{4}$$

In (4), $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$ denotes the partial derivative of $\hat{\mathbf{c}}$ with respect to $\mathbf{y}$, and $d\hat{\mathbf{c}}/d\mathbf{y}$ denotes the full derivative of $\hat{\mathbf{c}}$ with respect to $\mathbf{y}$. When $\hat{\mathbf{c}}$ is an implicit function of $\boldsymbol{\theta}$, $\partial\hat{\mathbf{c}}/\partial\boldsymbol{\theta}$ can be attained with the Implicit Function Theorem. This method considers the functional relationship between $\mathbf{c}$ and $\boldsymbol{\theta}$, so the estimated variance for a parameter includes the uncertainty of other parameter estimates.

3 APPLICATIONS

To estimate the model parameters, we fix the value of K_{xy}, vary values for α_y and β_y to solve the ODE (1) with these parameter values, and calculate the sum squared differences between the ODE solution and the measured expression of Gene Y. Figure 2 displays the contour plot of the logarithms of these sums of squared differences. The optimal values of the two parameters α_y and β_y show strong collinearity, and most of them are located around the line $\alpha_y = 0.11 + 0.15\beta_y$. Similar conclusion can be found for α_z and β_z. So in the following, we fix the two parameters $\beta_y = 1$ and $\beta_z = 1$, and estimate parameters $\alpha_y, \alpha_z, K_{xy}, K_{xz}, K_{yz}$ in ODEs (1) and (2).

We estimate dynamic models for three different FFLs—FFL 1 is composed of X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5; FFL 2 is composed of X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV1; FFL 3 is composed of X: Gene PDR1; Y: Gene PDR3; Z: Gene PDR5. The expression function for Gene X, $X(t)$, is estimated by penalized spline smoothing. We estimate parameters $\alpha_y, \alpha_z, K_{xy}, K_{xz}, K_{yz}$ with the generalized profiling method from the measured gene expression of Gene Y and Z. Each component is approximated by cubic B-splines with 40 equally spaced knots. The smoothing parameter $\lambda = 1000$. Table 1 shows the parameter estimates and their standard errors. FFL 1 and FFL 2 have the same Genes X and Y, and they are measured together in the same environmental changes (the temperature is increased from 25°C to 37°C), so the parameters for Gene Y to regulate Gene X, α_y and K_{xy}, have the same values. The self-regulation parameter α_z for Gene Z have different values, which means Gene Z in FFL 2 is more self-repressed than Gene Z in FFL 1. The parameter K_{yz} has a larger value in FFL 2 than FFL 1, so Gene Y in FFL 2 has a higher level of threshold required to significantly activate the expression of Gene Z. For FFL 3, K_{xy} and K_{xz} are relative high, which indicates that Gene X in FFL 3 has a high threshold to significantly activate the expression of Genes Y and Z.

In order to validate the estimates of dynamic models, it is useful to compare gene expressions to ODE solutions with the parameter estimates. To solve ODEs it numerically requires the initial values for Genes Y and Z. We estimate the initial values for Genes Y and Z

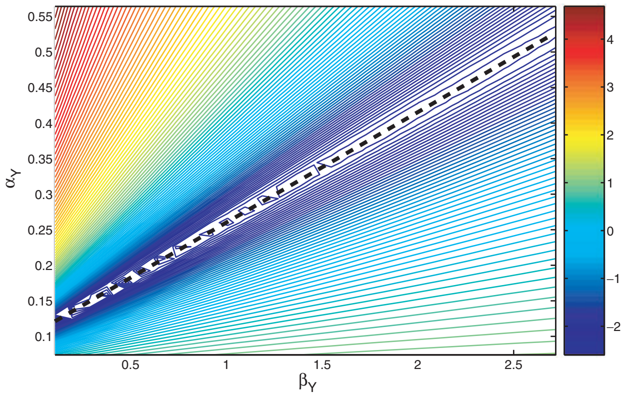

Fig. 2. The contour plot of the logarithm of the sums of squared differences between the measured expression of Gene Y shown in Figure 1 and the ODE (1) solution with different values of α_y and β_y. The value of K_{xy} is fixed as 0.93. The dashed line is $\alpha_y = 0.11 + 0.15 * \beta_y$.

Table 1. Parameter estimates and the standard errors for ODEs (1) and (2) from the measured expressions of Genes Y and Z

Parameters	α_y	α_z	K_{xy}	K_{xz}	K_{yz}
FFL 1: X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5					
Estimates	0.44	0.69	0.90	0.60	0.56
Standard Errors	0.22	0.18	0.33	0.06	0.15
FFL 2: X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV1					
Estimates	0.44	0.90	0.90	0.75	1.21
Standard Errors	0.22	0.01	0.33	0.44	0.74
FFL 3: X: Gene PDR1; Y: Gene PDR3; Z: Gene PDR5					
Estimates	0.32	0.56	2.11	1.06	0.76
Standard Errors	0.15	0.12	0.74	0.32	0.21

Each component is approximated by cubic B-splines with 40 equal spaced knots. The smoothing parameter $\lambda = 1000$.

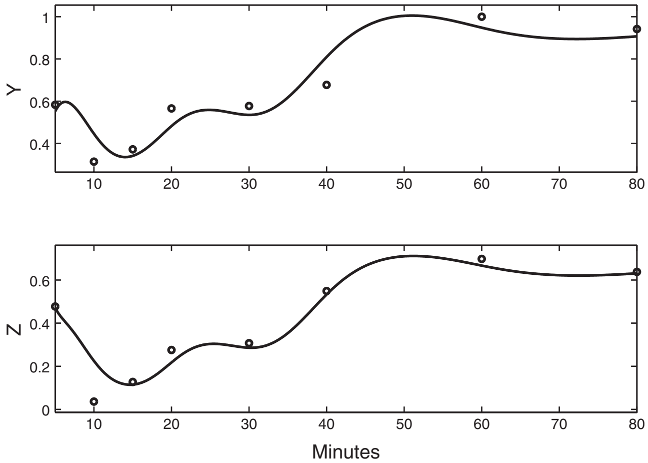

Fig. 3. The dynamic models for FFL 1 (X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV5). The circles are the real expression profiles of three genes, and the solid lines are the numerical solutions to ODEs (1) and (2) with the ODE parameter estimates $\alpha_y = 0.44, \alpha_z = 0.69, K_{xy} = 0.90, K_{xz} = 0.60, K_{yz} = 0.56$ and the estimated initial values $Y(t_0) = 0.55$ and $Z(t_0) = 0.47$.

by evaluating the smoothing curves at the start time point $t_0 = 5$, where the smoothing curves are estimated by minimizing penalized smoothing criterion (3). Figures 3–5 shows the numerical solutions to ODEs (1) and (2) with the ODE parameter estimates and the estimated initial values for the three FFLs. The ODE solutions are all close to the expressions of Genes Y and Z, which suggests ODEs (1) and (2) are good dynamic models for the FFL regulation network.

4 SIMULATIONS

We construct simulated data by adding Gaussian error with SD 0.1 to the solutions of ODEs (1) and (2) (shown in Fig. 4) at n equally spaced time points in $[5, 80]$. The true values of ODE parameters are chosen as the parameter estimates from the real data, and initial values of $Y(t)$ and $Z(t)$ are $Y(t_0) = 0.55$ and $Z(t_0) = 0.47$. The smoothing parameter $\lambda = 1000$. Table 2 shows the summary of parameter estimates from 100 simulations. We choose $n = 10$ or $n = 76$ to explore the effect of the number of observations on parameter estimations. When we have 10 observations for each gene, the biases for parameter estimates are around 15% of the real parameter values, and the coverages of the 95% confidence intervals are above 86% for each parameters. After the number of observations is increased to 76, the biases for parameter estimates are decreased to below 4% of the real parameter values except around 10% for α_z and K_{xz}. The estimated 95% confidence intervals have coverage near 95%.

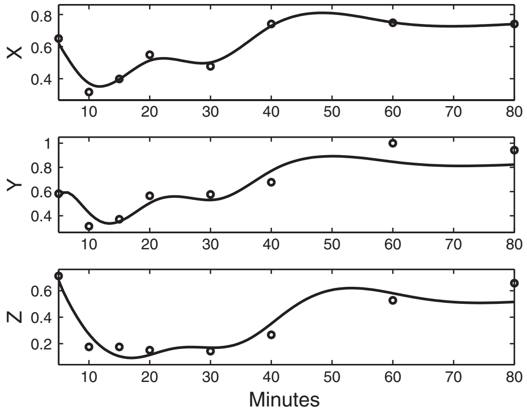

Fig. 4. The dynamic models for FFL 2 (X: Gene GCN4; Y: Gene LEU3; Z: Gene ILV1). The circles are the real gene expression profiles of three genes. The solid lines in the top panel is the estimated $\hat{X}(t)$, and the solid lines in the bottom panels are the ODE solutions to ODEs (1) and (2) with the ODE parameter estimates $\alpha_y = 0.44, \alpha_z = 0.90, K_{xy} = 0.90, K_{xz} = 0.75, K_{yz} = 1.21$ and the estimated initial values $Y(t_0) = 0.55$ and $Z(t_0) = 0.70$.

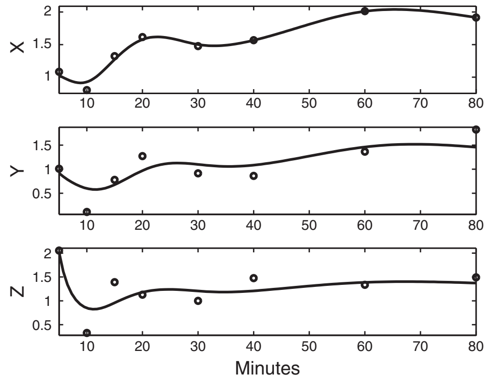

Fig. 5. The dynamic models for FFL 3 (X: Gene PDR1; Y: Gene PDR3; Z: Gene PDR5). The circles are the real gene expression profiles of three genes. The solid lines in the top panel is the estimated $\hat{X}(t)$, and the solid lines in the bottom panels are the ODE solutions to ODEs (1) and (2) with the ODE parameter estimates $\alpha_y = 0.32, \alpha_z = 0.56, K_{xy} = 2.11, K_{xz} = 1.06, K_{yz} = 0.76$ and the estimated initial values $Y(t_0) = 0.92$ and $Z(t_0) = 2.02$.

5 TEST THE GOODNESS OF FIT OF DYNAMIC MODELS

The goodness-of-fit test of dynamic models can be used to identify whether three genes of interest compose a FFL. Since the generalized

Table 2. Summary of parameter estimates from 100 simulated datasets

n	Parameter	α_y	α_z	K_{xy}	K_{xz}	K_{yz}
10	True	0.44	0.69	0.90	0.60	0.56
	Mean	0.38	0.60	1.06	0.81	0.45
	Bias	−0.06	−0.09	0.16	0.10	−0.10
	STD	0.10	0.25	0.24	0.36	0.29
	Coverage	90%	87%	98%	86%	89%
76	True	0.44	0.69	0.90	0.60	0.56
	Mean	0.44	0.63	0.91	0.68	0.54
	Bias	0.00	−0.06	0.01	0.08	−0.02
	STD	0.054	0.098	0.089	0.18	0.12
	Coverage	93%	93%	93%	96%	97%

n is the number of observations for each gene.

profiling method is computationally efficient (<20 s for parameter estimation for our problem), parametric bootstrap is used to test the goodness of fit of dynamic models, which is described as following. The ODE parameters $\boldsymbol{\theta}$ in ODEs (1) and (2) are estimated from the expressions of Genes X, Y and Z, then we calculate the sum of squared errors:

$$\mathtt{SSE}(y,s_y,z,s_z)=\sum_{i=1}^{n_y}[y(t_i)-s_y(t_i|\hat{\boldsymbol{\theta}})]^2+\sum_{i=1}^{n_z}[z(t_i)-s_z(t_i|\hat{\boldsymbol{\theta}})]^2,$$

where $y(t_i)$, $z(t_i)$ are expression profiles of Genes Y and Z at t_i, and $s_y(t_i|\hat{\boldsymbol{\theta}})$, $s_z(t_i|\hat{\boldsymbol{\theta}})$ are solutions of ODEs (1) and (2) with parameter values $\hat{\boldsymbol{\theta}}$ at t_i. The variance for residuals is estimated as $\hat{\sigma}^2=\mathtt{SSE}(y,s_y,z,s_z)/(n_y+n_z-p)$, where p is the number of ODE parameters. Then 1000 simulated datasets are generated, and each simulated dataset, $y^{(j)}(t_i)$, $z^{(j)}(t_i)$, $j=1,\dots,1000$, is generated by adding Gaussian noise with variance $\hat{\sigma}^2$ to ODE solutions $s_y(t_i|\hat{\boldsymbol{\theta}})$, $s_z(t_i|\hat{\boldsymbol{\theta}})$. The generalized profiling method obtain the ODE parameter estimate, $\hat{\boldsymbol{\theta}}^{(j)}$, from each simulated dataset. Let $s_y^{(j)}$, $s_z^{(j)}$ be the ODE solutions with the parameter value $\hat{\boldsymbol{\theta}}^{(j)}$, then we can calculate the sum of squared errors, $\mathtt{SSE}(y^{(j)},s_y^{(j)},z^{(j)},s_z^{(j)})$, for each simulated dataset. The empirical P-value can then be obtained.

As one example, three genes are randomly chosen from 6152 genes in yeast (X: YLL044W; Y: YER096W; Z: YDR279W). We use the time course gene expression data measured when the temperature is increased from 25°C to 37°C (Gasch *et al.*, 2000), which is displayed in Figure 6. The sum of squared errors from the real data to ODE solutions, $\mathtt{SSE}(y,s_y,z,s_z)$, is equal to 0.84. The histogram for $\mathtt{SSE}(y^{(j)},s_y^{(j)},z^{(j)},s_z^{(j)})$ is displayed in Figure 7. The parametric boostrap gives the empirical P-value 0.046, which indicates that the three genes do not compose a Coherent Type 1 FFL.

In contrast, the real data for the three sets of genes used in Section 3 are applied to test if they compose FFLs. The sum of squared errors from the real data to ODE solutions, $\mathtt{SSE}(y,s_y,z,s_z)$, are displayed in Table 3. The empirical P-values from the goodness-of-fit test are all much larger than 0.05, which verify that the three sets of that genes do compose FFLs. The parametric boostrap for the goodness-of-fit test is computionally intensive, which takes 11 h to finish the above example on a standard computer. Formal test statistics are required to find out all FFLs in thousands of genes, which will be addressed in the future study.

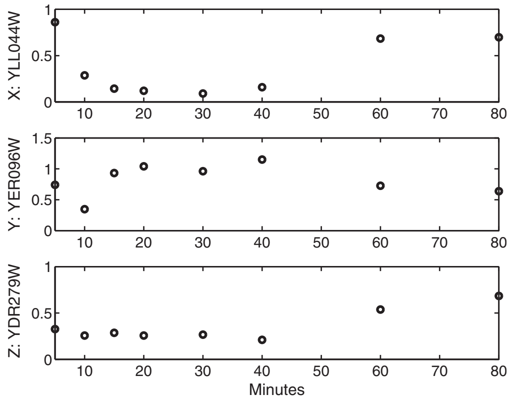

Fig. 6. The expression profiles of three genes measured at 5, 10, 15, 20, 30, 40, 60, 80 min. The data were collected by DNA microarrays from yeast after the temperature was increased from 25°C to 37°C (Gasch *et al.*, 2000).

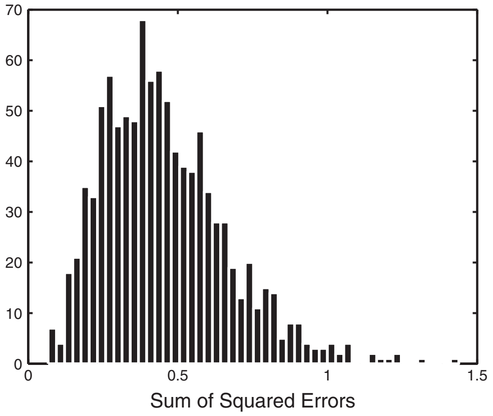

Fig. 7. The histogram for the sum of squared errors from the simulated data to ODE solutions.

Table 3. Identify whether three genes can compose a FFL with the goodness-of-fit test of dynamic models

Gene X	Gene Y	Gene Z	SSE	P-values
GCN4	LEU3	ILV5	0.090	0.25
PDR1	PDR3	PDR5	1.17	0.33
GCN4	LEU3	ILV1	0.092	0.34
YLL044W	YER096W	YDR279W	0.84	0.046

SSE is the sum of squared errors from the real data to ODE solutions.

6 DISCUSSION AND CONCLUSIONS

ODEs are widely used for modeling dynamic processes in engineering, biology, medicine, economics and many other areas. In this article, we propose to apply the generalized profiling method to estimate parameters in a set of non-linear ODEs for modeling gene regulation networks. The initial values for the gene expression are estimated by evaluating the fitted curves at the start time points. We show that the ODE solutions found with our estimated parameter values and initial values fit the data well. This is a good validation to show that the dynamic model can describe the observed behavior of the regulation system well. We also find two pairs of parameters show strong collinearity, an issue that can be alleviated with more observations.

Most differential equations used to model real systems are non-linear and do not have analytic solutions. Many methods for estimating ODEs have to solve ODEs numerically when searching for optimized ODE parameter values, which is computationally expensive and requires knowing the initial values of the ODE components. On the other hand, the generalized profiling method approximates ODE solutions with penalized smoothing splines, which requires a much lower computational load. A modified Delta method is developed to estimate the standard errors of the ODE parameter estimates, which takes into account the uncertainty of other parameter estimates.

Although technologies for gene expression analysis are becoming less expensive, analysis of such complex systems is still limited by the constraints on the number of microarray experiments that can be performed due to array cost and limitations of biological sample collection. We have found that some ODE parameters cannot be reliably estimated from sparse data routinely collected in microarray experiments. It is interesting to determine the frequency requirement for data collection in order to estimate ODE parameters of interest precisely. At the same time, the locations of measurement points also play an important role in the parameter estimations. This experimental design issue will be addressed in future research.

In conclusion, ODEs provide elegant models for gene regulation networks. The generalized profiling method can estimate ODE parameters quickly from noisy observations. The resulting ODE solutions using the estimated parameter values can fit the data well, which can lend evidence to the validity of the proposed ODE models.

ACKNOWLEDGEMENTS

We thank Dr Olga Troyanskaya and three reviewers for their constructive comments.

Funding: The first author was partially supported by NSERC Discovery Grant of Canada. The second author was supported in part by NIH grant GM59507 and NSF grant DMS 0714817.

Conflict of Interest: none declared.

REFERENCES

Alon,U. (2007) *An introduction to Systems Biology*. Chapman & Hall/CRC, London.

Barkai,N. and Leibler,S. (1997) Robustness in simple biochemical networks. *Nature*, **387**, 913–917.

Bates,D.M. and Watts,D.B. (1988) *Nonlinear Regression Analysis and its Applications*. Wiley, New York.

Bock,H.G. (1981) Numerical treatment of inverse problems in chemical reaction kinetics. In Ebert, K. *et al*. (eds), *Modelling of Chemical Reaction Systems*, Springer, New York, pp. 102–125.

Cao,J. and Ramsay,J.O. (2007) Parameter cascades and profiling in functiona data analysis. *Comput. Stat.*, **22**, 335–351.

de Boor,C. and Swartz,B. (1973) Collocation at gaussian points. *SIAM J. Numer. Anal.*, **10**, 582–606.

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

Himmelblau,D. *et al*. (1967) Determination of rate constants for complex kinetics models. *Ind. Eng. Chem. Fund.*, **6**, 539.

Huang,Y. *et al*. (2005) Hierachical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, **62**, 413–423.

Mangan,S. and Alon,U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Nat. Acad. Sci.*, **100**, 11980–11985.

Ramsay,J.O. and Silverman,B.W. (2005) *Functional Data Analysis*. 2nd edn. Springer, New York.

Ramsay,J.O. *et al*. (2007) Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. R. Stat. Soc. B*, **69**, 741–796.

Savageau,M. (1976) *Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology*. Addison–Wesley, Reading, MA.

Shen-Orr,S. *et al*. (2002) Network motifs in the transcriptional regulation network of escherichia coli. *J. Anim. Ecol.*, **31**, 64–68.