

Continuous-Time Identification of Gene Expression Models

DANIEL E. ZAK,¹ RONALD K. PEARSON,² RAJANIKANTH VADIGEPALLI,²
GREGORY E. GONYE,² JAMES S. SCHWABER,² and FRANCIS J. DOYLE III³

ABSTRACT

One objective of systems biology is to create predictive, quantitative models of the transcriptional regulation networks that govern numerous cellular processes. Gene expression measurements, as provided by microarrays, are commonly used in studies that attempt to infer the regulation underlying these processes. At present, most gene expression models that have been derived from microarray data are based in discrete-time, which have limited applicability to common biological data sets, and may impede the integration of gene expression models with other models of biological processes that are formulated as ordinary differential equations (ODEs). To overcome these difficulties, a continuous-time approach for process identification to identify gene expression models based in ODEs was developed. The approach utilizes the modulating functions method of parameter identification. The method was applied to three simulated systems: (1) a linear gene expression model, (2) an autoregulatory gene expression model, and (3) simulated microarray data from a nonlinear transcriptional network. In general, the approach was well suited for identifying models of gene expression dynamics, capable of accurately identifying parameters for small numbers of data samples in the presence of modest experimental noise. Additionally, numerous insights about gene expression modeling were revealed by the case studies.

INTRODUCTION

A FUNDAMENTAL GOAL OF SYSTEMS BIOLOGY is to understand quantitatively the logic and behavior of the transcriptional regulatory networks that govern diverse cellular functions. This process involves both the construction and analysis of system-wide quantitative models of gene expression dynamics. These models of gene expression dynamics describe how *inputs* (e.g., transcription factor [TF] levels, perturbations) influence transcriptional *outputs* (mRNA levels of various genes) over time. The present work is concerned with a key step in the process of constructing these models: the estimation of their parameters. This process is referred to in the present work as the *identification* of the models, rather than *modeling*, to make refer-

¹Department of Chemical Engineering, University of Delaware, Newark, Delaware.

²Daniel Baugh Institute for Functional Genomics and Computational Biology, Department of Pathology, Cell Biology and Anatomy, Thomas Jefferson University, Philadelphia, Pennsylvania.

³Department of Chemical Engineering, University of California, Santa Barbara, California.

Online supplementary material available at: www.dbi.tju.edu/dbi/publications/omics03.

ence to *process identification*. Process identification is distinct from process modeling in that the former involves the construction of empirical black-box models from the experimental data whereas the latter is concerned with the development of models from fundamental physical principles (Ogunnaike and Ray, 1994; Ljung 1999). The present work is concerned with models that can be feasibly identified from system-wide measures of gene expression, such as microarray data (cDNA and oligonucleotide arrays). Given that such data is characteristically limited to a few measurements per gene and is often corrupted by experimental noise, the models to be identified must be relatively simple. An effort is made, however, to include key features, such as saturating nonlinearities and transcript degradation. For this reason, the models of the present work can be considered *gray box* models (Pearson and Pottman, 2000), in that fundamental understanding of gene expression is combined with empirical aspects to yield models that are well suited to the data.

Numerous groups have studied the dynamical properties of transcriptional regulatory networks, and many of these modeled gene expression as a continuous-time, biochemical process, expressed in ordinary differential equations (ODEs) (Hargrove et al., 1991; Goldbeter, 1996; Smolen et al., 1998; Cherry and Adler, 2000; De Jong, 2002; Isaacs et al., 2003). Similarly, models of signal transduction pathways, that may ultimately be linked to gene expression models, have been routinely formulated as sets of ODEs (Chen et al., 2000; Kholodenko et al., 1999). ODEs provide a convenient framework to represent gene expression and signaling, given the wide array of analysis tools that are available, such as DASSL/DASPK (Maly and Petzold, 1996), and MATLAB (Shampine and Reichelt, 1997). A contrary trend, however, has been seen in studies that identify gene expression models from microarray data. In these studies, the models, with few exceptions (Chen et al., 1999; Yeung et al., 2002; Wahde and Hertz, 2000), have been formulated often as discrete-time gray box models, where time is not a continuous variable but rather one that moves forward in finite steps (D'Haeseleer et al., 1999; Weaver et al., 1999; Wessels et al., 2001; Hartemink et al., 2002). The use of a distinct framework for microarray-derived gene expression models disconnects gene expression from other cellular processes that may ultimately lead to difficulty in integrating whole-cell models. Discrete-time models of gene expression suffer from other difficulties. Exponentially spaced sampling, rather than uniform sampling, is often employed in biological experiments in an attempt to determine the order of magnitude of the time required to exhibit a significant response. Discrete-time models may not be readily identified from exponentially sampled data. Additionally, computational models of biochemical processes often are assembled from multiple data sets that may not have been collected at a single sampling rate, further complicating discrete approaches for creating integrated models. Given these tendencies for biological data to be asynchronous, and for biological models to be constructed from multiple data sources, continuous-time models that are robust to how the data is collected are preferable to discrete-time models that are not. Other arguments for continuous-time models of gene expression are that, with some exceptions (McAdams and Arkin, 1999), the physical principles that govern gene expression are most commonly formulated in continuous-time; that discrete-time models may exhibit undue sensitivity to parameters in comparison to continuous-time models (Unbehauen and Rao, 1998); and that nonlinear discrete-time models may have very different qualitative character (e.g., chaos, multiplicity) from continuous-time systems (Unbehauen and Rao, 1998; Pearson, 1999).

Following the above arguments, the present work is specifically concerned with the identification of continuous-time models of gene expression from gene expression time course data. The present work differs from previous continuous-time gene expression model identification efforts in that the method that is employed neither involves the direct estimation of derivatives from data (Ronen et al., 2002; Yeung et al., 2002), which can be undesirable due to noise corruption (Unbehauen and Rao, 1998), nor complex nonlinear optimization techniques (Wahde and Hertz, 2000), and is not limited to linear models (Chen et al., 1999). Rather, through the use of the modulating functions technique (Shinbrot, 1957; Patra and Unbehauen, 1995), parameters in nonlinear continuous-time ODE models of gene expression are identified using linear regression.

The modulating functions approach was developed by Shinbrot (1957) as a means to estimate parameters in continuous-time nonlinear systems by converting sets of differential equations into a linear regression. It has been employed in diverse process identification studies in the literature (Pearson and Lee, 1985; Co and Ydstie, 1990; Patra and Unbehauen, 1995; Daniel-Berhe and Unbehauen, 1999; Balestrino et al.,

2000). The advantages of the modulating functions approach, in addition to the fact that parameter identification is accomplished by linear regression, are that it removes any need for approximating derivatives from data, and there is no requirement for uniformly sampled data. Disadvantages of the modulating functions approach are that it cannot be used to estimate parameters for all nonlinear model types, and for some nonlinear models, it cannot guarantee bias-free estimates (Niethammer et al., 2001). In the present work, the Hartley modulating function (HMF) approach (Patra and Unbehauen, 1995) is employed. The HMF approach is advantageous over other modulating function approaches in that HMFs are entirely real valued (Unbehauen and Rao, 1998), which can give greater computational efficiency.

In the present work, the HMF method is used to identify several gene expression models from simulated data. Details about the implementation of the HMF method are provided in the online supplementary material. In the first section, the essential elements of modulating functions are illustrated with a simple linear gene expression system. In the second section, a more complex nonlinear model for an autoregulatory gene, capable of multiple steady states, is considered. In the final section, a case study in identifying gene expression models from microarray data using data generated by a previously described transcriptional network simulator (Zak et al., 2001, 2003) is presented. The examples demonstrate that the HMF method is well suited for identifying models from gene expression data.

Before proceeding it must be noted that the problem of simultaneously identifying network structure and network parameters is not the focus of the present work. For the problem of network structure identification, the reader is referred to techniques in promoter bioinformatics (for example, Tavazoie et al., 1999; Vadigepalli et al., 2003) and genome-wide location analysis (Ren et al., 2000).

APPLICATIONS

Linear gene expression model

In the present section, the problem of identifying the parameters a and d , given a time course of $x(t)$ and $u(t)$, is considered for the following simple linear gene expression model:

$$\dot{x}(t) = au(t) - dx(t) \tag{1}$$

where $x(t)$ is the mRNA level over time, $a \times u(t)$ is the transcription rate of x , a is the transcription rate constant, $u(t)$ is the activity of a TF that binds to the promoter of the gene, and d is the linear degradation rate constant for the transcript.

The first step is to multiply both sides of equation 1 by the known modulating function, $\phi_m(t)$, and integrate from $t = 0$ (the first time of the measurements) to $t = T$ (the final measurement time):

$$\int_0^T \phi_m \dot{x} dt = a \int_0^T \phi_m u dt - d \int_0^T \phi_m x dt \tag{2}$$

By applying integration by parts and requiring that the modulating function be differentiable and have the property: $\phi_m(0) = \phi_m(T) = 0$, equation 2 is simplified to:

$$- \int_0^T \dot{\phi}_m x dt = a \int_0^T \phi_m u dt - d \int_0^T \phi_m x dt \tag{3}$$

In equation 3, the derivative has been shifted from the data ($x(t)$) to the known smooth function, $\phi_m(t)$, thereby avoiding the need to estimate derivatives from the data. By using M different modulating functions, a linear expression for the parameters a and d is constructed that can be solved by linear regression:

$$- \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_M \end{bmatrix} = \begin{bmatrix} X_{1a} & -X_{1d} \\ X_{2a} & -X_{2d} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ X_{Ma} & -X_{Md} \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} \tag{4}$$

where:

$$\begin{aligned}
 Y_m &= \int_0^T \phi_m x dt \\
 X_{m_a} &= \int_0^T \phi_m u dt \\
 X_{m_d} &= \int_0^T \phi_m x dt
 \end{aligned}
 \tag{5}$$

The integrals in equation 5 can be evaluated using one of several techniques for numerical quadrature (Heath, 1997). The particular quadrature approach employed in the present work, as well as the approach taken to select M , is discussed in the online supplementary material. Unlike discrete-time formalisms, the data does not need to be sampled uniformly for the parameters to be estimated. In fact, for the simple model above, $u(t)$ and $x(t)$ do not even need to be obtained at the same time points. This flexibility with respect to sampling makes the modulating functions approach especially suitable for biological problems.

The system in equation 1 was used to explore how the accuracy of the identification results from the HMF method depend on the number of samples and experimental noise. A step input in $u(t)$ from $u = 0$ to $u = 10$ at $t = 1$ min was used to excite the system, and the nominal parameter values were $a = 5 \text{ min}^{-1}$ and $d = 1 \text{ min}^{-1}$. There were five time points for $x(t)$ and $u(t)$, from $t = 0$ to $t = 50$ min, sampled asynchronously (exponential, with a minimum inter-sample time of 1 min). The accuracy of the identification was quantified by θ_{norm} , the summed squared relative error in the parameter estimates (online supplementary material). To simulate the experimental noise in microarray data, the log-normal multiplicative model for noise given by Rocke and Durbin (2001) was used. The HMF method was benchmarked against a method that identified the parameters through direct estimation (DE) of the derivatives from the data. Figure 1a shows the results as the number of samples is varied from 3 to 10, demonstrating clearly that, within a realistic range of samples, the parameter estimates from the HMF method are more accurate than those obtained by DE. Figure 1b shows the results for the two methods as a function of experimental noise (5 samples, exponential sampling). Median values of $\theta_{norm} \pm$ one median absolute deviation of the median (MAD) obtained from 200 Monte Carlo runs are plotted against increasing noise magnitude, σ_η . MAD was used because it provides an outlier-insensitive measure of spread in the data (Huber, 1981). For a wide range of noise levels, the HMF method gave more accurate estimates of the parameter values than DE, with less variability. It

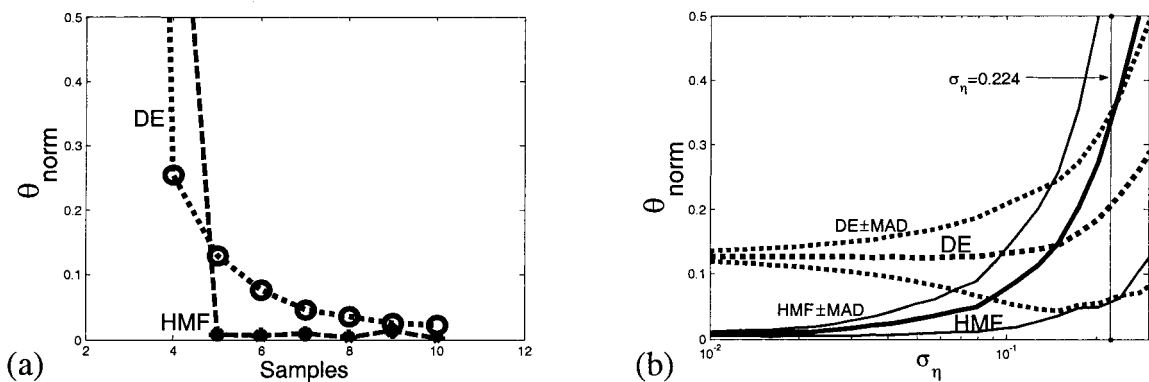


FIG. 1. Impact of number of samples and noise on parameter identification accuracy for HMF and DE approaches. Exponentially sampled step response. **(a)** Sum squared relative error in parameter estimates (θ_{norm}) versus the number of samples, demonstrating that the accuracy of the parameter identification using the HMF method is less sensitive to the number of samples than that using DE. Dashed line, stars, HMF; dotted line, circles, DE. **(b)** Median values of $\theta_{norm} \pm$ one MAD obtained from 200 Monte Carlo runs/data point are plotted against increasing noise magnitude, σ_η (log-normal model), five samples. Solid lines, HMF; dotted line, DE. Over a wide range of noise levels, the HMF method gives more accurate estimates of the parameter values, with less variability. At levels of noise similar to that observed in actual microarray studies ($\sigma_\eta \sim 0.224$), the accuracy and spread of the DE parameter estimates were more favorable than those obtained using HMF, however.

should be noted that when the estimates of a and d were considered independently, the variability was comparable for the two methods. At levels of noise similar to that observed in actual microarray studies ($\sigma_\eta \sim 0.224$), the DE parameter estimates were more favorable than those obtained from HMF. At this high level of noise, however, there was substantial error in the parameter estimates from both methods.

These results indicate that the HMF method may provide accurate parameter estimates for low numbers of samples and a wide range of experimental noise levels. At high levels of noise, such as that typically observed in microarray data, parameter identification through direct estimation of derivatives may actually be favorable, although the parameter estimates will be coarse. To obtain accurate parameter estimates from microarray data, it may be necessary to combine the HMF method with averaging over replicated measurements. The online supplementary material contains all of the details for the computations performed in this example, as well as an additional example that demonstrates how asynchronous sampling can be preferable to uniform sampling for this system.

Autoregulatory gene expression model

The modulating functions approach is extensible to systems that are more complex than equation 1. One example is the bistable autoregulatory circuit described by Smolen and colleagues (1998). Their model describes a TF that forms a homodimer and then activates its own transcription, with the maximal rate of transcription being influenced by its degree of phosphorylation. The objectives in studying this system were to demonstrate how the HMF method can be used to identify complex nonlinear models of gene expression and to explore how the accuracy of the parameter identification depends on the number of samples, the type of input used to excite the system, and the magnitude of measurement noise.

The Smolen model is given by:

$$\dot{y}(t) = av(t) \frac{y(t)^2}{b + y(t)^2} - cy(t) + d \quad (6)$$

where $y(t)$ is the TF concentration, $a \times v(t)$ is the maximal rate of transcription, $v(t)$ is the extent of phosphorylation of the TF (the external input), b is the dissociation constant of the TF dimer from its promoter, c is the linear degradation rate constant of the TF, and d is the basal synthesis rate of the TF. This system is interesting because, depending on the parameter values, it can be toggled between insensitive and excitable steady states by varying $v(t)$ (Smolen et al., 1998). In the present work, the nominal parameter values used are those given by Smolen and colleagues ($a = 10 \text{ min}^{-1}$, $b = 10 \text{ concentration}^2$, $c = 1 \text{ min}^{-1}$, $d = 0.1 \text{ concentration} \cdot \text{min}^{-1}$, $v(0) = 1$).

Using the modulating functions approach, the system in equation 6 was converted to a form similar to equation 4, allowing the parameters to be estimated using linear regression:

$$Y_m = X_{m_a}a + X_{m_b}b + X_{m_{bc}}bc + X_{m_c}c + S_{m_d}d \quad (7)$$

where the Y and X terms involve numerical quadrature and are similar to those in equation 5 (complete derivation given in the online supplementary material).

Using equation 7 and the HMF method, the impact of varying the number of samples on the accuracy of the parameter identification was explored. The input consisted of two pulses, with the first driving the system from its insensitive steady state to its excitable steady state, and the second providing further excitation. Results of the parameter identification for 20, 50, and 100 samples are shown in Figures 2a, 2b, and 2c, respectively (uniform sampling, $x(t) = y(t) - y(0)$, $u(t) = v(t) - v(0)$). With as few as 20 samples, the general bistable character of the system is captured, although there is significant overshoot in the response to the first pulse, and the error in the parameter estimates is significant ($\theta_{norm} = 2.37$). Increasing the number of samples to 50 only slightly reduces the overshoot, but the parameter estimates are greatly improved ($\theta_{norm} = 0.12$). Finally, increasing the number of samples to 100 gives excellent parameter estimates and captures the system dynamics very well. This example demonstrates how, for the more complex gene expression model in equation 6, there can be significant disagreement between the observed and predicted behavior, even when there is reasonably good agreement between the estimated and actual parameter values.

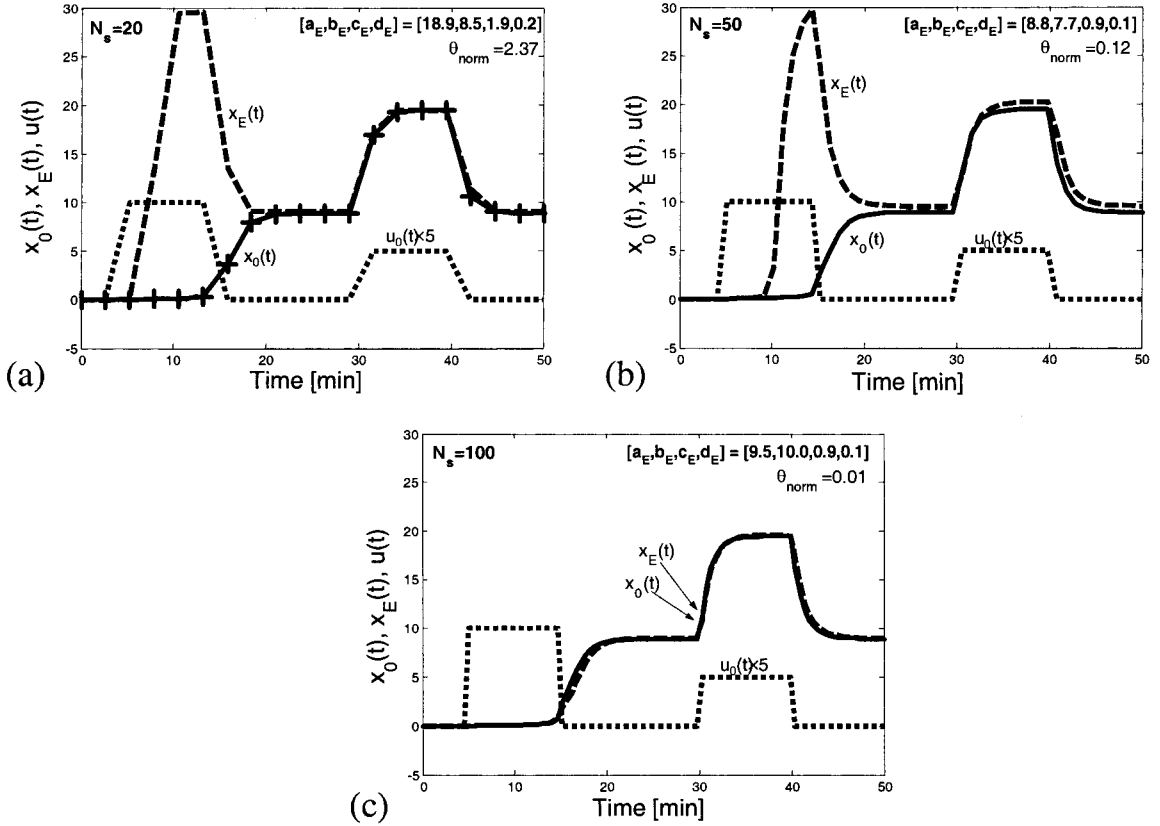


FIG. 2. Influence of the number of samples on parameter identification accuracy for the autoregulatory model. (a) 20 samples. (b) 50 samples. (c) 100 samples. $x(t) = y(t) - y(0)$, $u(t) = v(t) - v(0)$. Solid line, $x_0(t)$; dashed line, $x_E(t)$; dashed-dotted line, $u_0(t) \times 5$. The smallest number of samples captures the essential bistable nature of the system, but displays significant overshoot and inaccurate parameter estimates. At 50 samples, the parameter estimates are reasonably accurate, but the system still displays overshoot. At 100 samples, the parameter estimates are highly accurate, and the dynamics of the system are captured.

It also demonstrates how a greater number of samples is required to obtain accurate parameter estimates as compared to the simpler model (equation 1).

Equation 7 and the HMF method were also used to explore how the accuracy of the parameter identification was influenced by experimental noise and the type of input ($v(t)$) used to excite the system (number of samples fixed at 50). The log-normal noise model was applied as above. Three different inputs were considered. Two of the inputs were 10-min pulses, with one being strong enough to drive the system to the higher steady state (super-threshold pulse), and one too weak to do so (sub-threshold pulse). The third was the double pulse input from the previous example. Median values from 50 Monte Carlo runs are shown in Figure 3, where it is clear that parameter identification accuracy is sensitive to both the input and the amount of noise in the data. For the sub-threshold pulse, the parameter identification results were the least sensitive to noise, but were generally poor ($\theta_{norm} \sim 3.4$, where $\theta_{norm} = 4$ roughly corresponds to 100% error in the parameter estimates for this system). The single pulse that drove the system to the higher steady state was highly sensitive to noise, giving accurate parameter estimates for only very low noise levels. The double pulse input gave accurate parameter estimates for the widest noise range. Beyond $\sigma_\eta \sim 0.05$, however, the parameter estimates became worse than those obtained using the sub-threshold pulse. None of the inputs gave accurate parameter estimates for the experimentally-identified level of noise in microarray data ($\sigma_\eta = 0.224$), indicating that the noise in microarray data may be a significant impediment to its use in deriving complex models of transcriptional regulation. Rather, alternative sources of gene expression data,

GENE EXPRESSION MODEL IDENTIFICATION

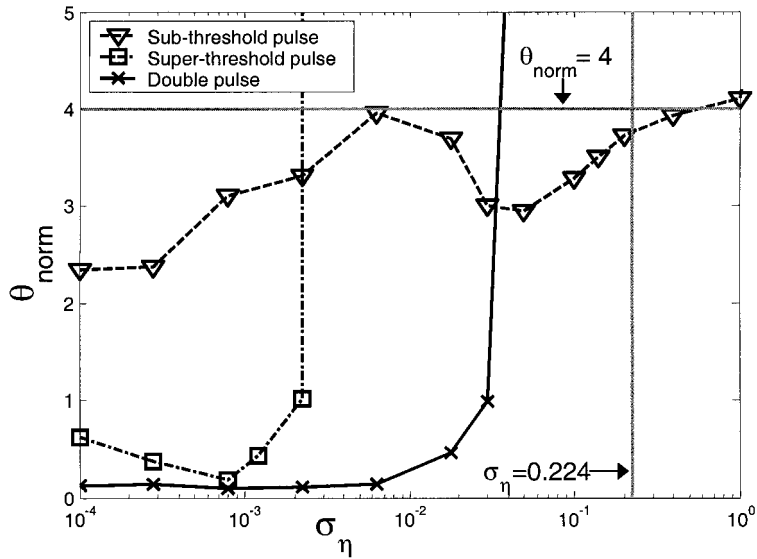


FIG. 3. Impact of noise and input on parameter identification for the autoregulatory model. Median sum squared relative error in parameter estimates (θ_{norm}) from 50 Monte Carlo runs/data point versus noise magnitude, σ_η (log-normal model), 50 samples. Only the double pulse input gave accurate parameter estimates, but for relatively low noise levels. None of the inputs gave accurate parameter estimates for an experimentally observed noise level in microarray data ($\sigma_\eta = 0.224$). Parameters that were estimated using the sub-threshold pulse were generally not sensitive to noise, but were generally inaccurate ($\theta_{norm} \sim 3.4$).

such as that used in (Ronen et al., 2002), may be required. The results for the sub-threshold and double pulse inputs are also suggestive of previous results from the literature that showed that complex inputs are required for accurate system identification (Kauffman et al., 2003) and that simpler inputs can be favorable at significant levels of noise (Zak et al., 2003).

Identifying gene expression models from microarray data

There are several aspects of microarray data and transcriptional regulation that make the problem of identifying dynamic gene expression models from microarray data particularly interesting. As has already been discussed in the present work, microarray time series often consist of a small number of asynchronously sampled data points that are corrupted by significant amounts of noise. In addition, it must be recognized that the time scale on which any mRNA level may respond to changes in transcription is determined by that mRNA's half-life (Hargrove and Schmidt, 1989). Also, TF mRNAs do not directly interact with their target genes. At the very least, translation of the TF must occur first, which introduces a delay. Finally, TFs may influence transcription rates of their target genes in a manner that depends nonlinearly on the active TF concentration, with the transcriptional regulation of a single gene being governed by the coordinate activity of multiple TFs. In the present section, a model structure for gene expression is developed that, when combined with the HMF approach to model identification, begins to address many of the above issues. This model structure is applied to a case study using simulated microarray data generated from a previously described transcriptional network simulator (Zak et al., 2001, 2003).

Model structure. The simplest formulation for a gene expression model is similar to equation 1 and is readily identified using the HMF approach:

$$\dot{x}(t) = af(u(t)) - dx(t) \quad (8)$$

where $x(t)$ is the scaled ($-1 \leq x(t) \leq 1$, $x(0) = 0$) concentration of the mRNA from the target gene as measured with the microarray, and $u(t)$ is the scaled ($0 \leq u(t) \leq 1$, $u(0) \neq 0$) concentration of the TF mRNA

concentration, also measured with the microarray. The details of the scaling procedures are given in the on-line supplementary material. The function $f()$ is a nonlinear function that describes how the transcription rate of the target gene depends on the TF concentration. While equation 8 is appealing for its simplicity, it may not be realistic because it suggests that a change in the mRNA of the TF can instantaneously affect the transcription rates of its targets, without any delay for its conversion to functional protein. An alternative formulation can address this delay through an intermediate state p :

$$\begin{aligned} \dot{x}(t) &= ap(t) - dx(t) \\ \dot{p}(t) &= f(u(t)) - ep(t) \end{aligned} \quad (9)$$

where e is the first order degradation rate constant for the intermediate state p . By differentiating the top expression in equation 9, and solving for p and its derivative in terms of $x(t)$, the derivatives of $x(t)$, and $u(t)$, the following expression is obtained:

$$\dot{x}(t) + (d + e)\dot{x}(t) + dex(t) = af(u(t)) \quad (10)$$

The parameters in equation 10 can be readily identified using the HMF approach given $x(t)$ and $u(t)$.

The introduction of the lag between TF mRNA levels and the effect of the TF on the transcription rates of its target genes could also have been accomplished with a discrete delay, a technique that has been employed in several models of gene expression (Smolen et al., 1998; Lema et al., 2000). Techniques have been developed to estimate such discrete delay systems using the modulating functions approach (Balestrino et al., 2000). It is the opinion of the present authors, however, that the type of delay introduced in equation 9 is more realistic because it is more closely related to one of biophysical processes that is responsible for the delay: translation (first order) of transcript into protein. Should the delay in equation 9 not be sufficient to describe the data, due to the presence of other intermediate biochemical steps in addition to translation, it is possible to include a longer delay by adding more intermediate states.

The final step is to define a functional form for the nonlinearity $f()$. Given that increasing concentration of active TF generally has a saturating effect on the transcription rate of its target genes, a natural functional form is:

$$g(u) = \frac{(K + 1)u}{K + u} \quad (11)$$

where $f(u)$ is related to $g(u)$ by: $f(u(t)) = g(u(t)) - g(u(0))$. Three values of K are considered in the present work, corresponding to approximate linearity ($K = 10$), weak nonlinearity ($K = 0.5$), and strong nonlinearity ($K = 0.1$). A plot of $g(u)$ for the three values of K is given in Figure 4.

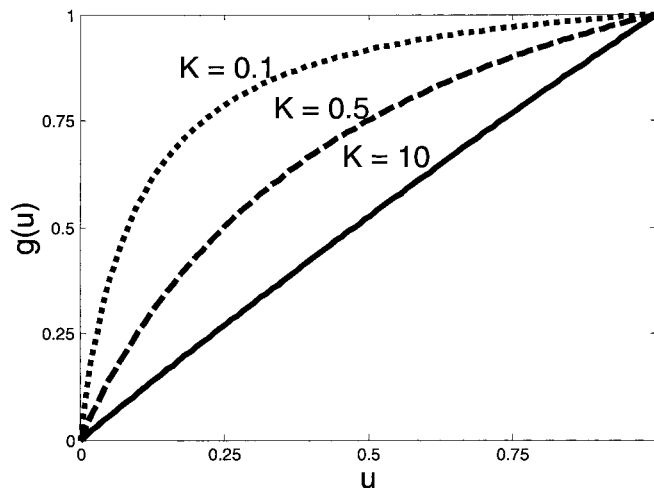


FIG. 4. $g(u)$ versus u for three different values of K . As K is increased, the extent of the nonlinearity is decreased. $K = 10$ corresponds to approximate linearity, $K = 0.5$ corresponds to weak nonlinearity, and $K = 0.1$ corresponds to strong nonlinearity.

The approach to identifying gene expression models from microarray data proposed in the present work can now be summarized:

- i. Collect microarray time course data for TFs and their target genes.
- ii. Scale the TF microarray data as $u(t)$ and scale the target gene microarray data as $x(t)$ (online supplementary material) to give the scaled experimental data $x_0(t)$ and $u_0(t)$.
- iii. Identification of the parameters a and d in the delay-free model (equation 8) and identification of parameters a , $(d + e)$, and (ed) for the model with delay (equation 10) using the HMF method for all three values of K (a total of six sets of parameters are estimated).
- iv. For cases where mRNA half-life data is already available, the parameter d for the models is estimated by $d = \ln(2)/t_{1/2}$, where $t_{1/2}$ is the mRNA half-life. In this case, only parameter a is estimated in the delay free model, and only parameters a and e are estimated in the delay model.
- v. For all six models, the sum of squared errors (SSE) between the scaled microarray data for the target genes, $x_0(t)$ and the expression data predicted using the modeling results (assuming that $u(t)$ varies linearly between samples) is calculated: $SSE(model) = \sum_{i=1}^n (x_0(t_i) - x_E(t_i))^2$. The delay-non-linearity combination that gives the smallest $SSE(model)$ is selected. The $SSE(model)$ for this combination is retained as a measure of how well the best model could describe the data.
- vi. Additionally, the SSE between $x_0(t)$ and $c(u(t) - u(0))$ is calculated, where c minimizes: $\sum_{i=1}^n (x_0(t_i) - c(u_0(t_i) - u_0(0)))^2$. This SSE, called $SSE(correlation)$, gives an estimate of how correlated $x_0(t)$ and $u_0(t)$ are that is useful in evaluating the identification results.

Simulated microarray data. The convenience of using simulated genetic regulatory networks to test and benchmark gene network modeling techniques has been recognized by several authors (Wahde and Hertz, 2000; Zak et al., 2001, 2003; Smith et al., 2002). In the present section, a simulated genetic regulatory network is used to explore the approach to identifying gene expression models from microarray data outlined above. The simulator has been described previously (Zak et al., 2001, 2003). It is a 10-gene network with a receptor that responds to ligand input, with transcription rates that depend nonlinearly on TF levels, and biochemical delays between the appearance of TF transcripts and functional TFs. It is described mathematically by 44 coupled nonlinear ODEs. Additional information about the model is provided in the online supplementary material. The response of the network to a pulse of ligand was used as the source of simulated microarray data. Ten data points (exponentially sampled, minimum inter-sample time of 3 min) up to 24 h after the injection of ligand were compiled.

Representative results are shown in Figure 5, where C and E have been modeled as targets of the transcriptional repressor D. The agreement between the simulation data and the predicted models was excel-

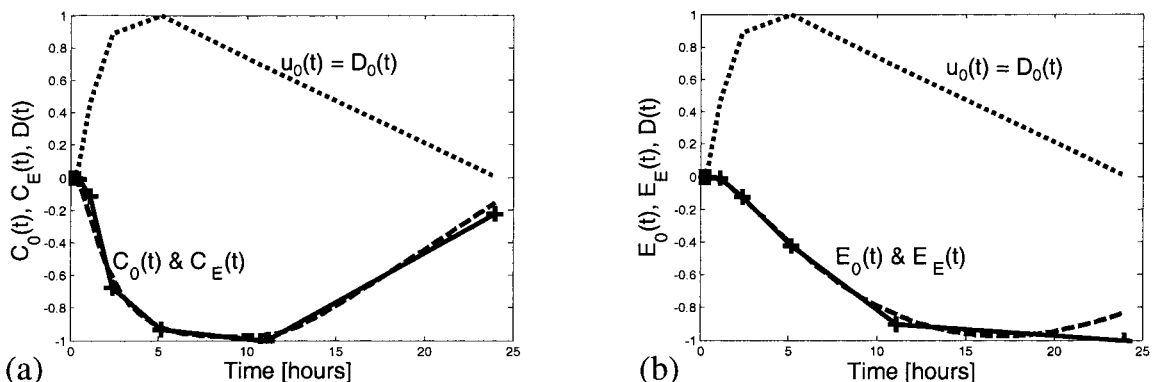


FIG. 5. Modeling results for transcriptional targets of D, showing excellent agreement. (a) Gene C ($SSE(model)_{DC} = 0.027$, strongly nonlinear, delay-free model). (b) Gene D ($SSE(model)_{DE} = 0.032$, linear, delay-free model).

lent ($SSE(model)_{DC} = 0.027$, $SSE(model)_{DE} = 0.032$), although this is expected given that C and E are targets of D in the actual underlying simulator network. Interestingly, the best fit model for regulation of C by D was strongly nonlinear ($K = 0.1$), while the best fit model for regulation of E by D was approximately linear ($K = 10$). For both cases, the delay-free model outperformed the delayed model.

Another set of results is shown for gene B in Figure 6. B has been modeled as a target of the TFs F, C, A, or D. In the true underlying network, B is a target of A and F. B was modeled reasonably well as a target of its true regulator F ($SSE(model)_{FB} = 0.19$, linear, delay-free model), poorly modeled as a target of C (that does not regulate it in the model) ($SSE(model)_{CB} = 1.89$, weakly nonlinear, delayed model) and poorly modeled as a target of its other true regulator A ($SSE(model)_{AB} = 2.0$, linear, delayed model). The poor results for A may be due to the fact that B is actually a target of both A and F in the underlying network, and modeling it as a target of either TF individually may not be sufficient. They may be also due to the fact that there are additional protein-protein interactions between A and B in the simulator that are not accounted for in the model structure. Interestingly, B is best modeled as a target of D, a gene that is co-regulated with B (D is also a target of F) ($SSE(model)_{DB} = 0.004$, linear, delay-free model). The fact that B was more successfully modeled as a target of a co-regulated gene than the correct TF provides a message of caution for gene network identification studies in which the network structure is not known *a priori*. When the structure is not known, care must be taken to avoid confusing correlation with causation. The

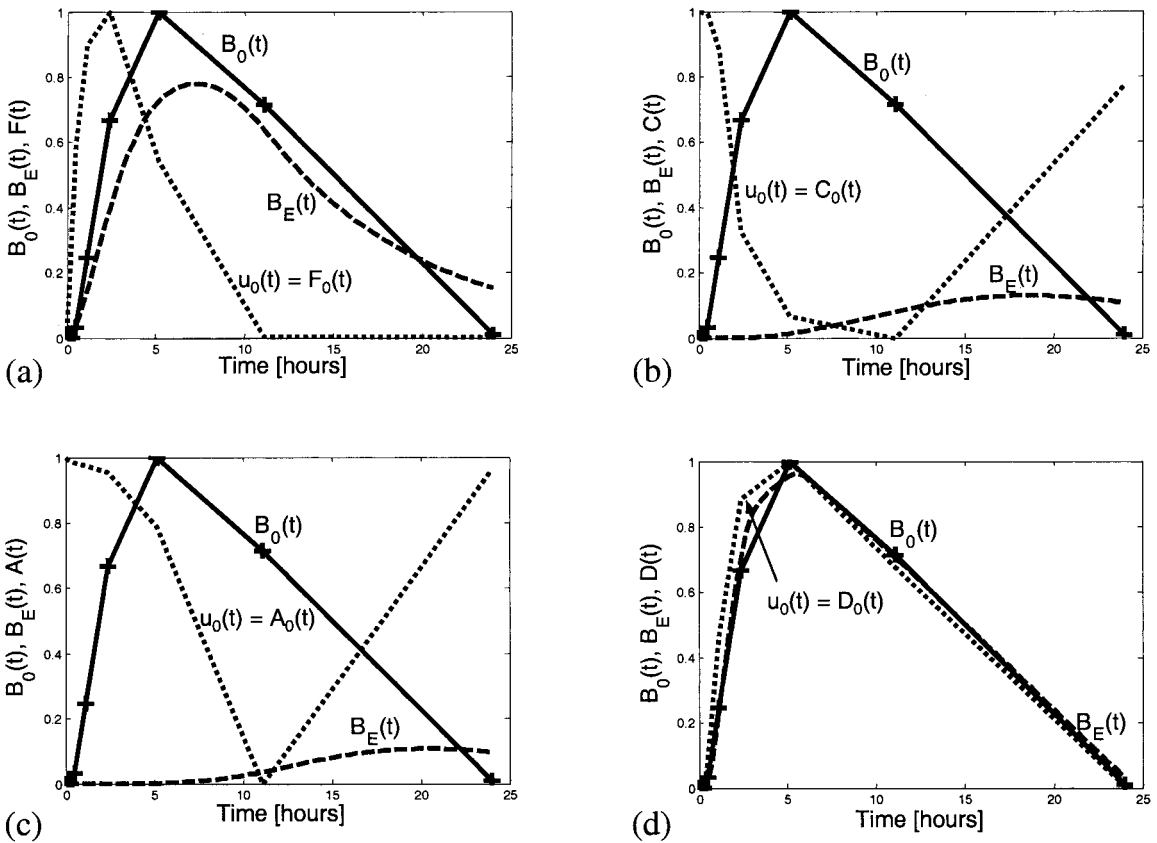


FIG. 6. Modeling B as a target of F, C, A, or D. **(a)** B as a target of F, one of its true regulators. The agreement is reasonable ($SSE(model)_{FB} = 0.19$, linear, delay-free model). **(b)** B as a target of C, a TF it is not regulated by. The agreement is poor ($SSE(model)_{CB} = 1.89$, weakly nonlinear, delayed model). **(c)** B as a target of A, one of its true regulators. The agreement is poor ($SSE(model)_{AB} = 2.0$, linear, delayed model). **(d)** B as a target of D, a gene that it is co-regulated with. There is strong agreement between the modeling result and the data, as well as strong correlation between B and D ($SSE(model)_{DB} = 0.004$, linear, delay-free model; $SSE(correlation)_{DB} = 0.07$).

correlation between D and B was also revealed by $SSE(correlation)_{DB}$, which is also much smaller than the $SSE(model)$ for either of the true regulators ($SSE(correlation)_{DB} = 0.07$).

In the final example, H was modeled as a target of G (its true regulator), C, D, or F. Surprisingly, H is modeled well as a target of G ($SSE(model)_{GH} = 0.002$, strongly nonlinear, delayed model), as a target of C (the upstream regulator of G) ($SSE(model)_{CH} = 0.002$, weakly nonlinear, delayed model), as a target of D ($SSE(model)_{DH} = 0.006$, linear, delayed model), or as a target of F ($SSE(model)_{FH} = 0.01$, strongly nonlinear, delayed model). Interestingly, the delayed model structure gave the best results for all combinations. If this were an experimental system, and no data in addition to the microarray data were available, it would be impossible to determine which TF actually regulates H. A figure with these results is provided in the online supplementary material (www.dki.tju.edu/dbi/publications/omics03).

DISCUSSION

In the present work, an argument was made for the use of continuous-time identification methods, particularly the Hartley modulating functions (HMF) method, for the identification of dynamic models of gene expression. The HMF method was applied to three example gene expression modeling problems of varying complexity, demonstrating the method's applicability and providing insights into the specific systems.

Using the simplest gene expression model, it was demonstrated that the HMF approach is well suited for parameter identification from asynchronous data. It was also observed that the HMF method may provide accurate parameter identification from a few samples and over a range of noise levels. At high levels of noise, comparable to that observed in microarray data, however, the preferred method for parameter identification may be through direct estimation of derivatives, although the parameter estimates will be coarse. Accurate parameter estimates from microarray data can be obtained using the HMF method, however, by reducing experimental noise through averaging over replicated measurements and other means. In the online supplementary material, it also was demonstrated how asynchronous sampling can lead to improved process identification, even though it makes discrete-time modeling difficult.

It was demonstrated, using the bistable autoregulatory gene expression model, how the HMF method is readily applicable to more complex models of transcriptional regulation. It was also shown how, for the more complex model, more samples must be collected for accurate parameter estimation. Additionally, more complex perturbations are required to excite the system, a result that parallels those observed in another study (Kauffman et al., 2003). This result was only applicable to the case of moderate noise, however, as a simpler input was preferable in cases of extreme noise, a result also paralleled by previous results (Zak et al., 2003). These requirements may preclude microarray data from being a sole data source to identify such complex models. Alternative types of gene expression data, such as that obtained in Ronen et al. (2002), may be required.

Using a network simulator, it was observed that the HMF method can be used to identify gene expression models using data of the type that is obtained in microarray studies. Without information that may constrain possible networks in addition to gene expression, however, it may be impossible to determine which TFs actually regulate each gene. The expression profiles of some downstream genes can be equally well modeled as targets of numerous transcription factors (TFs) in the system, making it impossible, without additional information, to uncover the true underlying network. Additional information could come in the form of additional time courses, with different dynamic responses, promoter bioinformatics (Tavazoie et al., 1999; Vadigepalli et al., 2003), or alternative experimental techniques such as genome-wide location analysis (Ren et al., 2000). The results also have shown that flexibility in model structure is important for gene expression modeling. Even though the true underlying network contained delays and nonlinearities, whether or not there were delays or nonlinearities in the model that best explained the data was dependent on the TF–target gene combination.

The successes or failures of the gene expression model identification techniques considered in the present work may or may not be attributed to the particular functional forms chosen and the data considered. The ability of the continuous-time HMF approach to identify models from asynchronously sampled data that is very common in biology, however, is a feature of the approach itself. The results of this modeling

approach may readily be integrated with models of upstream signaling networks that are formulated as ODEs. Similarly, the approach can be adapted to include efforts for inferring network structures through promoter bioinformatics (Tavazoie et al., 1999; Vadigepalli et al., 2003), or approaches that attempt to infer network connectivity through imposing constraints on interconnections (Yeung et al., 2002). Given the ability of the approach to integrate existing models of related cellular processes and varied analysis methods, it can play an important role in integrative environments like BioSPICE.

The present techniques fit well with the work of several other BioSPICE investigators. The autoregulatory gene results are directly relevant to efforts in the Byrne and Collins labs, which have an interest in nonlinear dynamical models of gene expression (Smolen et al., 2000; Isaacs et al., 2003). The ODE gene expression models of the present work parallel the hybrid approaches the Kumar and Rubin labs use to model prokaryotic genetic regulation. The dynamic microarray-based modeling approaches described above are complementary to the Liao lab's efforts in combining microarray data and operon information (Sabatti et al., 2002) with Bayesian techniques to predict prokaryotic genetic regulatory networks, and to the Collins lab's efforts in predicting genetic regulatory networks using gene expression data and steady state models (Gardner et al., 2003). The present modeling approaches may also be used with the dynamic gene expression data from the Jett lab to generate hypothesis about the regulation underlying host responses to pathogens. Finally, the modeling approaches of the present work may be used to link the core circadian clock models emerging from the Bryne lab and Jewett, Weaver, and collaborators (Forger et al., 1999; Smolen et al., 2001) with emerging circadian gene expression profiles in *Drosophila* (Etter and Ramaswami, 2002) and mouse (Delaunay and Laudet, 2002) into models that predict how the core clock regulates downstream genes and ultimately drives circadian changes in physiology.

Future work will involve the extension of the modeling approach to include genes that are targets of more than one transcription factor, and evaluation of additional nonlinear functional forms that describe how target gene transcription rates depend on TF concentration.

ACKNOWLEDGMENTS

This work is supported by the DARPA BioComp Initiative under contract number DE-AC03-76SF00098 (Francis J. Doyle III, PI) and contract number F30602-01-2-0578 (James Schwaber, PI). F.J.D. also acknowledges support from the Alexander von Humboldt Foundation and DEZ acknowledges the University of Delaware Department of Chemical Engineering for funding.

REFERENCES

- BALESTRINO, A., LANDI, A., and SANI, L. (2000). Identification of Hammerstein systems with input/output time delay via modulating functions. *Proc. IFAC. Linear Time Delay Systems* **2000**, 168–172.
- CHEN, T., HE, H.L., and CHURCH, G.M. (1999). Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput.* **4**, 29–40.
- CHEN, K.C., CSIKASZ-NAGY, A., GYORFFY, B., et al. (2000). Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell.* **11**, 369–391.
- CHERRY, J.L., and ADLER, F.R. (2000). How to make a biological switch. *J. Theor. Biol.* **203**, 117–133.
- CO, T.B., and YDSTIE, B.E. (1990). System identification using modulating functions and fast Fourier transforms. *Comput. Chem. Eng.* **14**, 1051–1066.
- DANIEL-BERHE, S., and UNBEHAUEN, H. (1999). Physical parameters estimation of the nonlinear continuous-time dynamics of a DC motor using Hartley modulating functions method. *J. Franklin I* **336**, 481–501.
- DELAUNAY, F., and LAUDET, V. (2002). Circadian clock and microarrays: mammalian genome gets rhythm. *Trends Genet.* **18**, 595–597.
- DE JOHG, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **9**, 67–103.
- D'HAESELEER, P., WEN, X., FUHRMAN, S., et al. (1999). Linear modeling of mRNA expression levels during CNS development and injury. *Proc. Pac. Symp. Biocomput.* **4**, 41–52.

- ETTER, P.D., and RAMASWAMI, M. (2002). The ups and downs of daily life: profiling circadian gene expression in *Drosophila*. *Bioessays* **24**, 494–498.
- FORGER, D.B., JEWETT, M.E., and KRONAUER, R.E. (1999). A simpler model of the human circadian pacemaker. *J. Biol. Rhythms* **14**, 532–537.
- GARDNER, T.S., DI BERNARDO, D., LORENZ, D., et al. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105.
- GOLDBETER, A. (1996). *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour*, 2nd ed. (Cambridge, University Press, Cambridge).
- HARGROVE, J.L., and SCHMIDT, F.H. (1989). The role of mRNA and protein stability in gene expression. *FASEB J.* **3**, 2360–2370.
- HARGROVE, J.L., HULSEY, M.G., and BEALE, E.G. (1991). The kinetics of mammalian gene expression. *Bioessays* **13**, 667–674.
- HARTEMINK, A.J., GIFFORD, D.K., JAAKOLA, T.S., et al. (2002). Combining location and expression data for principled discovery of genetic regulatory network models. *Proc. Pac Symp Biocomput* **7**, 437–449.
- HEATH, M.T. (1997). *Scientific Computing* (McGraw-Hill, New York).
- HUBER, P.J. (1981). *Robust Statistics* (John Wiley and Sons, New York).
- ISAACS, F.J., HASTY, J., CANTOR, C.R., et al. (2003) Prediction and measurement of an autoregulatory genetic module. *Proc. Natl. Acad. Sci. USA* **100**, 7714–7719.
- KAUFFMAN, K.J., OGUNNAIKE, B.A., and EDWARDS, J.S. (2003). Design of high-throughput profiling experiments: a mathematical analysis of network identification (submitted).
- KHOLODENKO, B.N., DEMIN, O.V., MOEHREN, G., et al. (1999). Quantification of short-term signaling by the epidermal growth factor receptor. *J. Biol. Chem.* **274**, 30169–30181.
- LEMA, M.A., GOLOMBEK, D.A., and ECHAVE, J. (2000). Delay model of the circadian pacemaker. *J. Theor. Biol* **204**, 565–573.
- LJUNG, L. (1999). *System Identification: Theory for the User*, 2nd ed. (Prentice Hall PTR, Upper Saddle River, NJ).
- MALY, T., and PETZOLD, L.R. (1996). Numerical methods and software for sensitivity analysis of differential-algebraic systems. *Appl. Numer. Math.* **20**, 57–79.
- McADAMS, H.H., and ARKIN, A. (1999). It's a noisy business: genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69.
- NIETHAMMER, M.N., MENOLD, P.H., and ALLGOWER, F. (2001). Parameter and derivative estimation for nonlinear continuous-time system identification. Presented at the 5th IFAC Symposium on Nonlinear Systems.
- OGUNNAIKE, B.A., and RAY, W.H. (1994). *Process Dynamics, Modeling, and Control* (Oxford University Press, New York).
- PATRA, A., and UNBEHAUSEN, H. (1995). Identification of a class of nonlinear continuous-time systems using Hartley modulating functions. *Int. J. Control* **62**, 1431–1451.
- PEARSON, R.K. (1999). *Discrete-time dynamic models* (Oxford University Press, New York).
- PEARSON, A.E., and LEE, F.C. (1985). On the identification of polynomial input-output differential systems. *IEEE Trans. Automat. Contr.* **AC-30**, 778–782.
- PEARSON, R.K., and POTTMANN, M. (2000). Gray-box identification of block-oriented nonlinear models. *J. Process Contr.* **10**, 301–315.
- REN, B., ROBERT, F., WYRICK, J.J., et al. (2000). Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309.
- ROCKE, D.M., and DURBIN, B. (2001). A model for measurement error for gene expression arrays. *J. Comput. Biol.* **8**, 557–569.
- RONEN, M., ROSENBERG, R., SHRAIMAN, B.I., et al. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* **99**, 10555–10560.
- SABATTI, C., ROHLIN, L., OH, M.K., et al. (2002). Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.* **30**, 2866–2893.
- SHAMPINE, L.F., and REICHEL, M.W. (1997). The MATLAB ODE suite. *SIAM J. Sci. Comput.* **18**, 1–22.
- SHINBROT, M. (1957). On the analysis of linear and nonlinear systems. *Trans. Am. Soc. Mech. Eng.* **79**, 547–552.
- SMITH, V.A., JARVIS, E.D., and HARTEMINK, A.J. (2002). Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics* **18**, S216–S224.
- SMOLEN, P., BAXTER, D.A., and BYRNE, J.H. (1998). Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. *Am. J. Physiol.* **274**, C531–C542.
- SMOLEN, P., BAXTER, D.A., and BYRNE, J.H. (2000). Mathematical modeling of gene networks. *Neuron* **26**, 567–580.

- SMOLEN, P., BAXTER, D.A., and BYRNE, J.H. (2001). Modeling circadian oscillations with interlocking positive and negative feedback loops. *J. Neurosci.* **21**, 6644–6656.
- TAVAZOIE, S., HUGHES, J.D., CAMPBELL, M.J., et al. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285.
- UNBEHAUEN, H., and RAO, G.P. (1998). A review of identification in continuous-time systems. *Annu. Rev. Control* **22**, 145–171.
- VADIGEPALLI, R., CHAKRAVARTHULA, P., ZAK, D.E., et al. (2003). PAINT: A promoter analysis and interaction network generation tool for genetic regulatory network identification. *OMICS* **7**, 235–252.
- WAHDE, M., and HERTZ, J. (2000). Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems* **55**, 129–136.
- WEAVER, D.C., WORKMAN, C.T., and STORMO, G.D. (1999). Modeling regulatory networks with weight matrices. *Proc. Pac. Symp. on Biocomput.* **4**, 102–111.
- WESSELS, L.F.A., VAN SOMEREN, E.P., and REINDERS, M.J.T. (2001). A comparison of genetic network models. *Proc. Pac. Symp. Biocomput.* **6**, 508–519.
- YEUNG, M.K., TEGNER, J., and COLLINS, J.J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA* **99**, 6163–6168.
- ZAK, D.E., DOYLE, F.J., III, GONYE, G.E., et al. (2001). Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. Presented at the 2nd International Conference on Systems Biology.
- ZAK, D.E., GONYE, G.E., SCHWABER, J.S., et al. (2003). Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network. *Genome Res.* **13**, 2396–2405.

Address reprint requests to:

Dr. Francis J. Doyle III
Department of Chemical Engineering
University of California
Santa Barbara, CA 93106

E-mail: doyle@engineering.ucsb.edu