JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 7, 2013

Pp. 524–539
DOI: 10.1089/cmb.2011.0222

Model Discrimination in Dynamic Molecular Systems: Application to Parotid De-differentiation Network

JAEJIK KIM,[1] JIAXU LI,[2] SRIRANGAPATNAM G. VENKATESH,[3]
DOUGLAS S. DARLING,[3] and GRZEGORZ A. REMPALA[1]

ABSTRACT

In modern systems biology the modeling of longitudinal data, such as changes in mRNA concentrations, is often of interest. Fully parametric, ordinary differential equations (ODE)-based models are typically developed for the purpose, but their lack of fit in some examples indicates that more flexible Bayesian models may be beneficial, particularly when there are relatively few data points available. However, under such sparse data scenarios it is often difficult to identify the most suitable model. The process of falsifying inappropriate candidate models is called model discrimination. We propose here a formal method of discrimination between competing Bayesian mixture-type longitudinal models that is both sensitive and sufficiently flexible to account for the complex variability of the longitudinal molecular data. The ideas from the field of Bayesian analysis of computer model validation are applied, along with modern Markov Chain Monte Carlo (MCMC) algorithms, in order to derive an appropriate Bayes discriminant rule. We restrict attention to the two-model comparison problem and present the application of the proposed rule to the mRNA data in the de-differentiation network of three mRNA concentrations in mammalian salivary glands as well as to a large synthetic dataset derived from the model used in the recent DREAM6 competition.

Key words: parotid dedifferentiation, ODE model, parameter estimation, Bayesian factor.

1. INTRODUCTION

THE PROCESS OF FALSIFYING INAPPROPRIATE CANDIDATE MODELS in order to find the best-suited model among different hypotheses is called model discrimination and is often based on discriminant analysis—a statistical technique going back to Fisher (1936) and applied in the fields of statistics, pattern recognition, and machine learning for the purpose of objects classification. The application of discriminant-like methods to modern biological data seems also useful in the current era of systems biology. Indeed, in most modern approaches to modeling biological phenomena via systems methods, one faces the challenge of evaluating multiple competing mathematical models of complicated reaction networks describing interactions among the biochemical species, such as various RNA molecules, proteins, enzymes, metabolites, and so on (see, e.g., Liao et al., 2005, or Carrillo et al., 2010, for specific examples in enzyme dynamics). Typically, before the

[1]Department of Biostatistics and Cancer Research Center, Georgia Regents University, Augusta, Georgia.
[2]Department of Mathematics, [3]Department of Oral Health and Systemic Disease, University of Louisville, Louisville, Kentucky.

correct model of a biochemical system is found, different hypothetical models might be reasonable and consistent with previous knowledge and available data. Whereas the full validation and fine-tuning of such models may require single cell-level data obtainable only through complex experiments, the initial model discrimination may sometimes be achieved by the analysis of aggregated data obtained from multiple cells (see, e.g., Hecker et al., 2009, for a recent review). Such data is often much more easily obtained with standard (hence cheaper) experimental methods, such as reverse transcriptase–polymerase chain reaction (RT-PCR). The dynamics of such aggregated data models are often amenable to analysis via a system of ordinary differential equations (ODEs) approximating the average stochastic effects of the cellular-level behavior (see, e.g., El Samad et al., 2005). Under the assumption that such a representation is meaningful, that is, the aggregated system indeed averages out the stochastic fluctuations around a single cellular-level steady state, the mathematical models are deterministic in nature and their sole sources of uncertainty come from the biased representation of reality and the data collection methodology (e.g., measurement bias and errors). This in turn allows one to invoke the general statistical methodology of (posterior) likelihood-based model discrimination, reminiscent of the classical linear and quadratic discriminant methods (LDA and QDA) (see, e.g., Hastie et al., 2001, Chapter 4), as well as the general methodology of the statistical validation of kinetic models (see, e.g., Zhan et al., 2011).

In a related problem of validating a given ODE model of a biological system with the aggregated time-course RT-PCR data, Rempala et al. (2007) proposed a likelihood-based procedure applicable to simple kinetic models, like, for instance, a simple transcription network (e.g., L1-gene retrotransposition). The procedure described in Rempala et al. (2007) was based on the ideas of Bayarri et al. (2007) on statistical validation in computer models, and was seen to work well for simple biological networks but not necessarily for more complicated ones such as complex signaling pathways. For these more complicated settings, a more comprehensive approach seems to be required, involving both the model validation and selection (see, e.g., Neff et al., 2011; Morton et al., 2010; Steimer et al., 2010, for applications to various molecular systems). For the latter, the ability to perform a discriminant-like analysis between competing models based on available time-course data is essential.

The aim of the current work is to propose a relatively general procedure for discriminating between different ODE-based kinetic models of a biological phenomenon under the assumption that the time-course data from the trajectory is available and that basic sources of both experimental and instrumental uncertainty and bias are accounted for. The proposed procedure extends Rempala et al.'s (2007) approach to multivariate time-course data and nonlinear interactions and is illustrated with two examples of data from time-course measurements. The first one is based on examining mRNA expression in acinar cells, obtained by RT-PCR with the purpose of analyzing transcription factor (TF) regulation of genes, which are indicators of differentiation. This was the original biological problem that motivated the methodological study. However, since the resulting network of the first example is relatively simple due to limited data availability, we have also considered another example in which two much larger synthetic models are compared. The second example is based on the ODE models considered in the recent Dialogue for Reverse Engineering Assessments and Methods (DREAM6) competition.

In order to motivate our biological example, we briefly recall some basic facts of the salivary glands biology relevant to our modeling efforts discussed below. (For a more general introduction to the topic, see, for instance, Harunaga et al., 2011 and Gorr et al., 2005.) Salivary glands are important for maintaining a healthy oral cavity in most mammals by production of saliva, which is necessary for lubrication and digestion initiation and which also contains proteins with antibacterial and antifungal properties. However, in humans salivary glands are often damaged or destroyed by radiation therapy, surgery for head and neck cancers, or by advanced Sjögren's syndrome. Secretion from salivary glands originates in clusters of acinar cells, which are classified as either serous (protein-secreting) or mucous (mucin-secreting). The parotid is the largest salivary gland, composed of serous acinar cells, and therefore is responsible for the secretion of the majority of salivary proteins. The most abundant proteins in parotid saliva are amylase and parotid secretory protein (PSP). The serous acinar cell is the salivary cell type that is most sensitive to radiation (Grundmann et al., 2009), and understanding its differentiation process is a necessary step in enabling the restoration or regeneration of diseased or destroyed parotid salivary tissue. Unfortunately, unlike in some differentiated tissues, the molecular mechanisms that drive differentiation of parotid acinar cells are not well characterized. Deciphering how the genes are transcriptionally regulated during de-differentiation is therefore important, as it might provide molecular insights into the reverse process of differentiation. Acinar cells, as with certain other cell types, lose their specialized characteristics rapidly when they are removed from the animal and placed in cell culture (Gorr et al., 2005). The synthesis and secretion of amylase and PSP

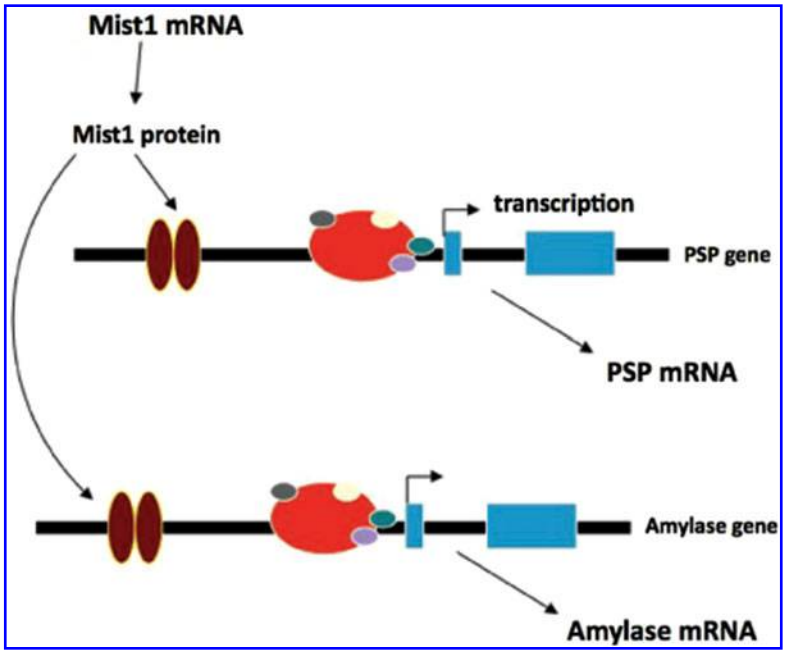

FIG. 1. Hypothesized interactions driving differentiation. A minimal hypothesized network consisting of the Mist1 transcription factor driving expression of both the parotid secretory protein (PSP) gene and the amylase gene. During differentiation of the cell, Mist1 would increase the amount of PSP mRNA and amylase mRNA. During the loss of differentiation *in vitro,* the decrease of Mist1 would lead to a decrease of PSP and amylase mRNAs.

are markers of terminal differentiation, which are rapidly lost under the conditions of culture (Venkatesh et al., 2012), providing for an excellent experimental system to study the loss of differentiation.

As the first step in analyzing the de-differentiation process, one would attempt to relate the mRNA levels of the terminal markers of differentiation (amylase and PSP) with that of exocrine-specific transcription factors such as Mist1, which is known to contribute to differentiation of pancreatic and salivary acinar cells (Luo et al., 2005). In Figure 1, we present a portion of a hypothesized parotid molecular differentiation model, expressing the idea that Mist1 mRNA generates Mist1 protein, which acts as a dimer at both the PSP gene and the amylase gene to stimulate the expression of both PSP and amylase mRNAs. On the basis of the hypothesized structure summarized in Figure 1, in order to help develop formal ODE-based models of their interactions, the longitudinal mRNA expression levels (measured via RT-PCR) were collected for the three genes. Similar ODE models have recently been used to characterize cell differentiation during hematopoiesis (Wang 2007; Soneji et al., 2007), but have been used to model differentiation of few if any other cell types. Nonetheless, they have been used successfully in other systems. For example, Gin et al. (2007) developed an ODE model to study the fluid secretion from adult parotid acinar cells. In addition, ODE modeling has been used extensively to study critical aspects of the cell cycle (Sveiczer et al., 2000; Tyson et al., 2001; Novak and Tyson, 2004) and the molecular interactions that cause circadian rhythms (Brown et al., 2000; Baker et al., 2011; Jolma et al., 2010).

2. METHODS AND RESULTS

We start by providing some further description of the time-course mRNA expression data and the RT-PCR procedure in Section 2.1 below. In the following Section 2.2, two possible ODE models for the given data are proposed; the procedures for validation and selection between them are presented in Section 2.3, with the results discussed in Section 2.4. The additional results from synthetic large-scale DREAM6 model comparisons are summarized in Section 2.5.

2.1. Time-course mRNA data

The procedure of collecting the data may be described as follows. For each experiment, parotid glands from 3–5 rats were collected and pooled, and the acinar cells isolated (see Venkatesh et al., 2012, for a detailed description of the experiment). This single pool of acinar cells was then split into eight cultures (in 6-well plates) and incubated for 0, 1, 6, or 24 hours. At each time point, the cells in two different wells were separately used to make RNA samples, therefore creating technical replicates at each time point. The

TABLE 1. TIME-COURSE mRNA DATA FOR AMYLASE, Mist1, AND PSP

	Amylase			*Mist1*			
Time (t_i)	*A1*	*A2*	*A3*	*M1*	*M2*	*M3*	*M4*
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	0.7020	2.1320	0.8440	0.7385	0.6083	0.7135	0.6322
6	0.0540	0.2420	0.2055	0.0565	0.2176	0.2400	0.3197
24	0.0060	0.0875	0.0375	0.0035	0.0202	0.1005	0.0444
	Mist1			*PSP*			
Time (t_i)	*M5*	*M6*	*M7*	*P1*	*P2*	*P3*	*P4*
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	0.8840	0.7435	0.7110	0.6310	1.0170	1.0745	1.0435
6	0.2645	0.1160	0.0640	0.0520	0.0825	0.1000	0.0995
24	0.0225	0.0105	0.0030	0.0040	0.0260	0.0095	0.0220

expression of amylase, Mist1, and PSP mRNAs were measured by TaqMan RT-PCR assays and normalized to glyceraldehydes 3-phosphate dehydrogenase (GAPDH) mRNA, which served as the internal experimental control. For each timepoint expression measured, the corresponding values of the two technical replicates were averaged and expressed relative to their initial value at the experiment inception (0-hour timepoint). The measurements from distinct wells were treated as independent. Mist1 and either PSP or amylase (due to limiting amounts of RNA) was measured in a total of seven independent experiments, as shown in Table 1.

The average mRNA expression data for each gene are plotted in Figure 2. For better visualization, in this as well as in all subsequent time-course plots presented in this article, we follow the convention of

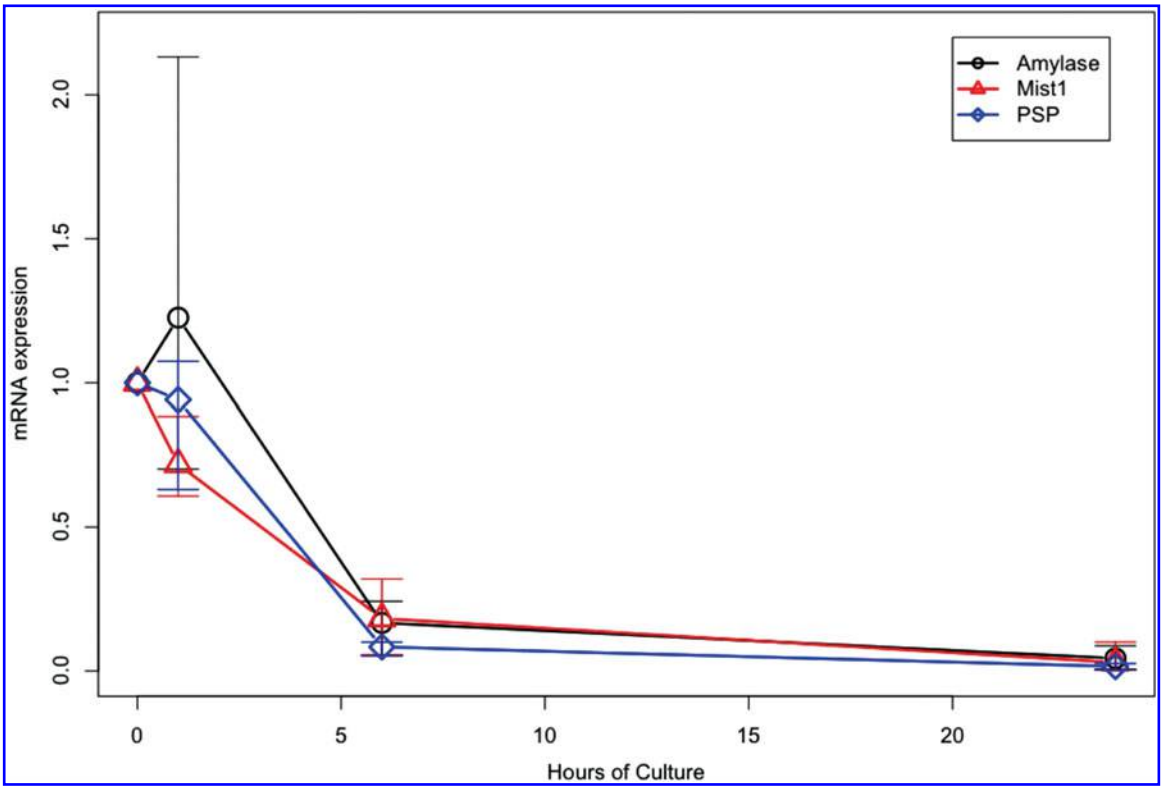

FIG. 2. mRNA expression time-course data. Colored solid lines are linearly interpolated average values for the three genes across time points. Each interval for each gene and time point shows the minimum and maximum mRNA expression obtained from the RT-PCR experiment.

extending all graphs beyond the observation time points via the linear interpolation. As seen from the plot in Figure 2, mRNA expression of amylase and PSP reach a saturation point at approximately 1 hour and then exponentially decrease. In contrast, Mist1 expression does not seem to have a saturation point and is seen to rapidly decrease in acinar cell cultures. As expected, the variations for each species also decrease with the time of culture.

2.2. ODE models

Based on the data presented in Figure 2, we may derive the following qualitative chemical ''mass transfer'' model (see, e.g., Chang et al., 2011; Likhoshvai and Ratushny, 2007) represented, in our case, by the coupled biochemical reactions

$$\begin{aligned} &Mist1\ (M) \longrightarrow PSP\ (P),\\ &Mist1\ (M) \longrightarrow amylase\ (A). \end{aligned} \tag{1}$$

While this is a very simple network, it represents a realistic, testable example, as it is well documented that Mist1 transcription factor helps drive terminal differentiation of acinar cells, and expression of both PSP and amylase are markers of acinar differentiation. To quantify the model above, it was hypothesized that the action of Mist1 was carried out by the dimerization of Mist1 molecules and that the saturation of PSP and amylase levels were attained while Mist1 was abundant. Two ODE representations of these interactions were considered. The first one (denoted here $\mathcal{M}_1$) employed the standard Hill-function ODE mass transfer model (see, e.g., Likhoshvai and Ratushny, 2007) as follows

$$\begin{aligned} \frac{dA(t)}{dt} &= \frac{v_1 M(t)^2}{v_2^2 + M(t)^2} - v_3 A(t)\\ \frac{dM(t)}{dt} &= -v_4 M(t)\\ \frac{dP(t)}{dt} &= \frac{v_5 M(t)^2}{v_6^2 + M(t)^2} - v_7 P(t). \end{aligned} \tag{2}$$

Here $A(t)$, $M(t)$, $P(t)$ are mRNA expressions of amylase, Mist1, and PSP, respectively, at time t, with $A(0) = 1$, $M(0) = 1$, and $P(0) = 1$, and $v_1, \ldots, v_7$ are nonnegative coefficients parametrizing $\mathcal{M}_1$. In particular, v_3, v_4, and v_7 are seen as the respective degradation rates, v_1 and v_5 as the maximum reaction rates, and v_2 and v_6 as the half-saturation points.

The particular form of the ODE model $\mathcal{M}_1$ is based on some earlier work (see, e.g., Venkatesh et al., 2012, and references therein) stipulating the presence of the trajectories extrema for all three species (maxima for amylase and PSP and a minimum for Mist1) at some $t_0 > 0$. The reason for the Hill-function form of $\mathcal{M}_1$ is the consideration of the saturation effect of the molecules of Mist1. However, the Hill–type growth does not take directly into account the ''mass transfer'' aspect of the reactions (1), which require that $M(t) \to 0$ as $t \to \infty$. The simplified version of the model $\mathcal{M}_1$ (denoted $\mathcal{M}_2$) may be obtained by employing the Taylor series approximation $x^2/(a+x^2) \approx x^2$ valid for small x, where ''small'' may be defined more precisely by considering non-zero intersection points of both curves and depends on the value of $a > 0$ (cf., also Sorribas et al., 2007). The new model is

$$\begin{aligned} \frac{dA(t)}{dt} &= v_1 M(t)^2 - v_2 A(t)\\ \frac{dM(t)}{dt} &= -v_3 M(t)\\ \frac{dP(t)}{dt} &= v_4 M(t)^2 - v_5 P(t), \end{aligned} \tag{3}$$

where $A(0) = 1$, $M(0) = 1$, and $P(0) = 1$ and $v_1, \ldots, v_5$ are, as before, assumed to be nonnegative system parameters. The main difference between $\mathcal{M}_1$ and $\mathcal{M}_2$ is the lack of the denominator terms in $\mathcal{M}_2$, which allows it to be parametrized by two less coefficients v. The additional advantage of this simplified ODE model is that (3) follows the so-called law of mass-action, a popular general empirical law of molecular dynamics (Hornos et al., 2005).

Given the two models $\mathcal{M}_1$ and $\mathcal{M}_2$, a natural question arises as to whether either one of them may meaningfully describe the data in Table 1. In order to answer it, we need the assessment of the accuracy of the model predictions associated with $\mathcal{M}_1$ and $\mathcal{M}_2$, which we discuss next.

2.3. Model validation and discrimination

To describe the relationship between the solution of ODE models (2) and (3), and the actual mRNA expressions for the genes of interest summarized in Table 1, we follow Rempala et al. (2007) and employ the additive data model as follows.

Denote the time-course vectors of mRNA expression for each of the TFs by $y_{ij}^L(\mathbf{t})$, $i=\{A, M, P\}$, $j=1, \ldots, r_i$, with $\mathbf{t}=(t_1, \ldots, t_K)'$. Here $K = 3$ with the time-points (in hourly units) $t_1 = 1$, $t_2 = 6$, $t_3 = 24$; A, M, and P stand for amylase, Mist1, and PSP respectively, with the corresponding number of experiments (see Table 1), $r_A = 3$, $r_M = 7$, and $r_P = 4$. Let $\mathbf{v}_i$, $i = \{A, M, P\}$ be the vector of parameters of the ODE models for each gene (i.e., for $\mathcal{M}_1$, $\mathbf{v}_A = (v_1, v_2, v_3)'$, $\mathbf{v}_M = v_4$, and $\mathbf{v}_P = (v_5, v_6, v_7)'$ and for $\mathcal{M}_2$, $\mathbf{v}_A = (v_1, v_2)'$, $\mathbf{v}_M = v_3$, and $\mathbf{v}_P = (v_4, v_5)'$). If $y_i^M(\mathbf{t}|\mathbf{v}_i)$ is the solution of the ODE model (either under $\mathcal{M}_1$ or $\mathcal{M}_2$) corresponding to the i-th TF, then $y_{ij}^L(\mathbf{t})$ is assumed to be related to $y_i^M(\mathbf{t}|\mathbf{v}_i)$ as follows

$$y_{ij}^L(\mathbf{t}) = y_i^M(\mathbf{t}|\mathbf{v}_i) + b_i(\mathbf{t}) + \epsilon_{ij}(\mathbf{t}), \qquad i=\{A, M, P\},\ j=1, \ldots, r_i, \tag{4}$$

where $b_i(\mathbf{t})$ ($\equiv \mathbf{b}_i$) is the vector of parameters accounting for the model bias, and $\epsilon_{ij}(\mathbf{t})$ is the vector of random errors. Since, after the initial measurement, we have three time points for each mRNA, both models $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same number of nine bias parameters. Since $\mathcal{M}_1$ and $\mathcal{M}_2$ are parametrized by seven and five ODE coefficients respectively, for $\mathcal{M}_1$ the model (4) has a total of 16 parameters while for $\mathcal{M}_2$ it has 14.

For the purpose of model validation, we need to provide the credibility bounds in (4) and this can be done, for instance, by considering the joint posterior distribution of the model parameters. The Bayesian framework is particularly appropriate in our case due to the limited number of observations available and hence potential problems with the identifiability of the standard maximum likelihood estimators (see, e.g., Wilkinson, 2009). As in Rempala et al. (2007) and Bayarri et al. (2007), we assume the form of the likelihood function for $y_{ij}^L(\mathbf{t})$ as follows:

$$p(y_{ij}^L(\mathbf{t})|\mathbf{v}_i, \mathbf{b}_i) \sim \mathrm{MVN}(\mu_i,\ diag(\delta_i)), \qquad i=\{A, M, P\},\ j=1, \ldots, r_i, \tag{5}$$

where $\mathrm{MVN}(\mu_i,\ diag(\delta))$ is a multivariate normal distribution with a mean vector $\mu_i = (y_i^M(t_1|\mathbf{v}_i) + b_i(t_1), \ldots, y_i^M(t_K|\mathbf{v}_i) + b_i(t_K))'$, and the variance–covariance matrix with diagonal elements δ_i, where $\delta_i = [\exp\{-t_k/5\}]_k$, $k=1, \ldots, K(K=3)$, and off-diagonal elements zero due to the assumed independence of each culture well (see Section 2.1). The variance terms for the longitudinal levels of mRNA are taken as decreasing exponentially with time—the assumption that seems consistent with the data pattern observed in Figure 2. The form of the likelihood (5) may be justified by the fact that in many kinetic biochemical systems (including those considered here) the fluctuations around ODE models are approximately Gaussian (see, e.g., Ge and Qian, 2009; Andersson and Britton, 2000, Chapter 5).

To complete the statistical model setup, we need to define flexible prior distributions for the ODE parameters $\mathbf{v} = (\mathbf{v}_A, \mathbf{v}_M, \mathbf{v}_P)'$ as well as the bias terms $\mathbf{b} = (\mathbf{b}_A, \mathbf{b}_M, \mathbf{b}_P)'$. Due to the parameter values support and the covariance structure considerations, the convenience choices here are a multivariate log-normal prior for $\mathbf{v}$ and a multivariate normal one for $\mathbf{b}$. Whereas the appropriateness of these particular distributions may be debated, their form does not seem to particularly effect the results of our analysis presented below. Define

$$p(\mathbf{v}) \sim \mathrm{Log-normal}(\nu, \Lambda), \tag{6}$$

where ν is a mean vector and Λ is a variance–covariance matrix, empirically estimated from the data, as well as

$$p(\mathbf{b}) \sim \mathrm{MVN}(\mathbf{0},\ \Delta), \tag{7}$$

where the variance–covariance matrix Δ is a block diagonal matrix with the same diagonal blocks $\mathbf{C} = [\exp\{-(t_m - t_n)^2\}]_{mn}$, $m, n = 1, \ldots, K$. Thus, in our case ($K = 3$), Δ is a 9×9 matrix. The exponential, block-diagonal form of Δ takes into account the possible dependancies among biases between and within

the trajectory points of the three mRNAs. From the forms of the likelihood function of (5) and the priors (6) and (7), we see that the posterior distribution is proportional to

$$p(\mathbf{v}, \mathbf{b}|y_A^L(\mathbf{t}), y_M^L(\mathbf{t}), y_P^L(\mathbf{t})) \propto \left[\prod_{i=\{A,M,P\}} \prod_{j=1}^{r_i} p(y_{ij}^L(\mathbf{t})|\mathbf{v}_i, \mathbf{b}_i)\right] p(\mathbf{v})p(\mathbf{b}). \tag{8}$$

To obtain this posterior distribution, the usual Markov Chain Monte Carlo (MCMC) method based on the Gibbs sampler as in Bayarri et al. (2007) was used, but other alternatives could have been also entertained (see, for instance, Zhan et al., 2011, or Jones, 2011). Regardless of the detail of the particular MCMC technique, consider now $\mathbf{v}_{il}$ and $\mathbf{b}_{il}$, $i = \{A, M, P\}$, $l = 1, \ldots, n$ as the collection of posterior samples obtained from the converged chain. Using these posterior samples, we can approximate the posterior means of the ODE parameter $\mathbf{v}_i$, $i = \{A, M, P\}$, (cf, Eq. (A.1) in Appendix A), which in turn allows us to approximate the mean of the pure model prediction $y_i^M(\mathbf{t}|\hat{\mathbf{v}}_i)$ and the bias-corrected prediction (cf, Eq. (A.3) in Appendix A) for the i-th mRNA ($i = \{A, M, P\}$). The model prediction $(1 - \alpha)\%$ credibility bounds for the i-th mRNA ($i = \{A, M, P\}$) may be obtained by calculating $\alpha/2$ and $(1 - \alpha/2)$ quantiles of the empirical distribution based on $y_i^M(\mathbf{t}|\mathbf{v}_{il}) + \mathbf{b}_{il}$, $l = 1, \ldots, n$. Further details and relevant formulas are provided in Appendix A.

2.4. Results

The results of the analysis based on (8) are summarized in Table 2 and illustrated in Figure 3, which shows, for each mRNA, the pure model and the bias-corrected predictions, credibility envelopes for both $\mathcal{M}_1$ and $\mathcal{M}_2$ ODE models given by (2) and (3), as well as the actual data. From the plots in Figure 3, it appears that both $\mathcal{M}_1$ and $\mathcal{M}_2$ have validity for the mRNA expression data from Table 1, as for all three mRNAs in respective models, their 95% credibility envelopes include both the pure and the bias-corrected model predictions (cf., Rempala et al., 2007; Bayarri et al., 2007). Moreover, the width of the credibility envelopes in both models seems similar for each of the respective genes. However, when comparing the differences between the pure and the bias-corrected model predictions, it is clear that those under $\mathcal{M}_2$ are considerably smaller, especially for amylase and PSP mRNAs. This seems to indicate that, albeit the uncertainties of the prediction in both models are similar (which is also seen more directly by comparing the variances of the model predictions computed via the formula (A.5) in Appendix A), the simplified model $\mathcal{M}_2$, despite its sparser parametrization, has smaller bias than $\mathcal{M}_1$. Plots in Figures 4 and 5 further support this conclusion. In Figure 4, variances of each gene for both models across times are seen as not very different. However, in Figure 5, $\mathcal{M}_2$ appears to have small biases relative to $\mathcal{M}_1$, with the similar overall longitudinal bias pattern for both models.

Denote for brevity the observed data by $\mathbf{y} = (y_A^L(\mathbf{t}), y_M^L(\mathbf{t}), y_P^L(\mathbf{t}))$ and let θ_i be the parameter vector $(\mathbf{v}, \mathbf{b})_i$ for the model $\mathcal{M}_i$ ($i = 1,2$), with the respective prior $p(\theta_i|\mathcal{M}_i)$. In order to quantitatively discriminate between the two models, we again utilize the results of the MCMC analysis based on Equation (8), in the spirit of the Fisher discriminant principle alluded to in the introduction. To this end, we consider the Bayes factor (BF) criterion (see, e.g., Hastie et al., 2001, Chapter 7; for a possible alternative based on Bayesian information (see, e.g., Jones, 2011, and Skanda and Lebiedz, 2010)

$$BF = \frac{p(\mathbf{y}|\mathcal{M}_1)}{p(\mathbf{y}|\mathcal{M}_2)} = \frac{\int p(\mathbf{y}|\theta_1, \mathcal{M}_1)p(\theta_1|\mathcal{M}_1)d\theta_1}{\int p(\mathbf{y}|\theta_2, \mathcal{M}_2)p(\theta_2|\mathcal{M}_2)d\theta_2} \tag{9}$$

where both the numerator and the denominator integrals are now approximated by the numerically averaging $p(\mathbf{y}|\theta_i, \mathcal{M}_i)p(\theta_i|\mathcal{M}_i)$ across the posterior sample obtained from the MCMC algorithm under respective models. The numerical value $BF \approx 0.0001$ indicates that under the equal prior preference for both $\mathcal{M}_1$ and $\mathcal{M}_2$ the posterior probability of $\mathcal{M}_2$ is considerably larger (10,000-fold) than that of $\mathcal{M}_1$ and thus the simplified, mass action-based ODE system (2) is strongly preferred over the originally proposed Hill-function-based ODE system (3).

2.5. Additional results with synthetic data

In order to assess the scalability of the method to more realistic biological networks, in addition to (2) and (3), a larger synthetic system was also considered in the model discrimination analysis. This synthetic

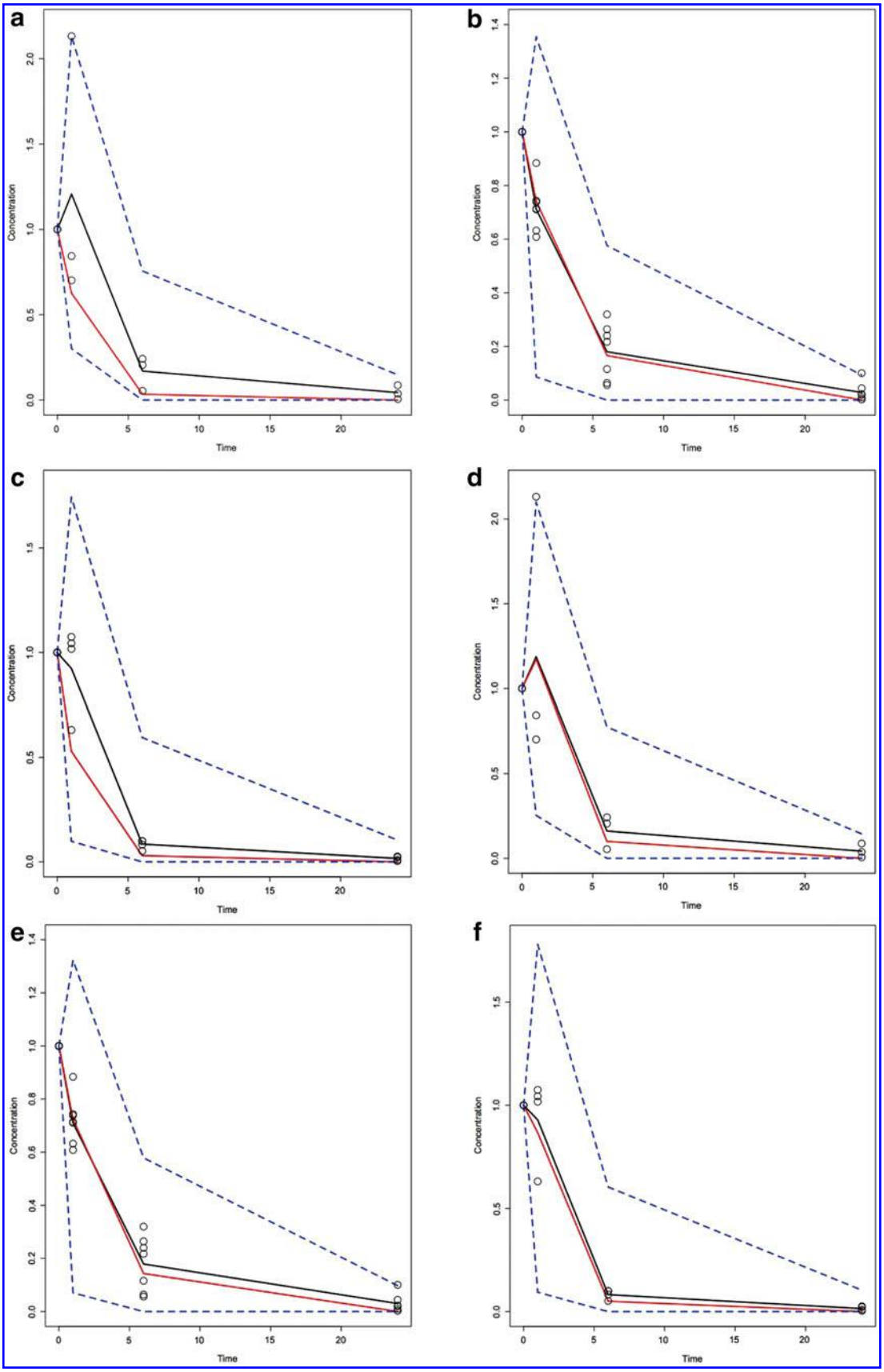

FIG. 3. The 95% posterior credibility bounds for $\mathcal{M}_1$ and $\mathcal{M}_2$. Actual data (circle) versus the pure model prediction $y_i^M(\mathbf{t}, \hat{\mathbf{v}}_i)$ (red line) and the bias-corrected model prediction $\hat{y}_i^L(\mathbf{t})$ (black line) along with the 95% credibility bounds (blue dotted line): (**a**) amylase for $\mathcal{M}_1$; (**b**) Mist1 for $\mathcal{M}_1$; (**c**) PSP for $\mathcal{M}_1$; (**d**) amylase for $\mathcal{M}_2$; (**e**) Mist1 for $\mathcal{M}_2$; and (**f**) PSP for $\mathcal{M}_2$.

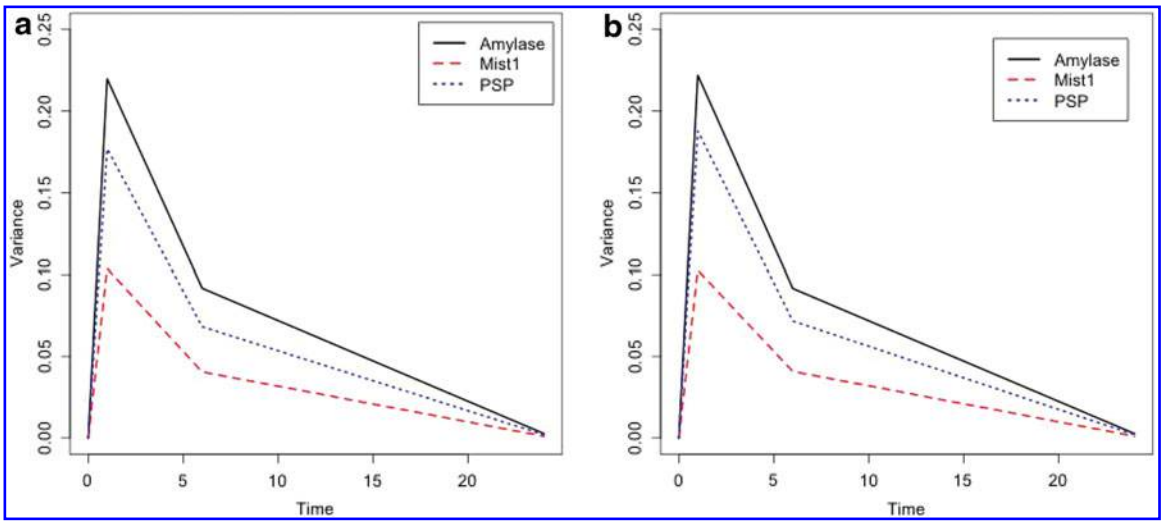

FIG. 4. Variance of the model prediction for mRNA data: (**a**) $\mathcal{M}_1$ and (**b**) $\mathcal{M}_2$.

system model is described by an ODE set with 12 species and 24 parameters (see Appendix B) and comes from the simplified version of the ''Model1'' of the most recent DREAM6 Estimation of Model Parameters Challenge. The synthetic data was generated from this ODE model under the assumptions (4). Both the correct ($\mathcal{M}_1$) and slightly altered ($\mathcal{M}_2$) ODE systems were compared as above, in an effort to discriminate between the two alternatives. The complete descriptions of $\mathcal{M}_1$ and $\mathcal{M}_2$ are provided in (B.1) and (B.2), in Appendix B. Both sets of ODE are seen to have the same number of parameters, with the difference introduced in the growth models of the species Y_9 and Y_{11} in both sets. The details on data generation are provided in Appendix B. The generated data and fitted models are presented in Figure 6, which shows little difference in terms of the model fit. In order to analyze the fit in more detail, the same MCMC method was applied to obtain 95% credibility bounds for models $\mathcal{M}_1$ and $\mathcal{M}_2$. The results are plotted in Figures 7 and 8 and include both the pure and the bias-corrected model predictions. From the plots it appears that both models have similar patterns of predictions for all species, with no obvious differences between the pure and the bias-correct versions. Some prediction bias is visible in both, for instance, at time-point 1 for species Y_{12} in $\mathcal{M}_1$ and for species Y_9, Y_{10}, and Y_{11} in $\mathcal{M}_2$. Overall, it seems hard to discriminate between two models on the basis of Figures 7 and 8 alone, and the BF criterion has to be invoked. The large BF value in (9) (BF = 2348.4) allows us to identify the model $\mathcal{M}_1$ as having considerably larger posterior

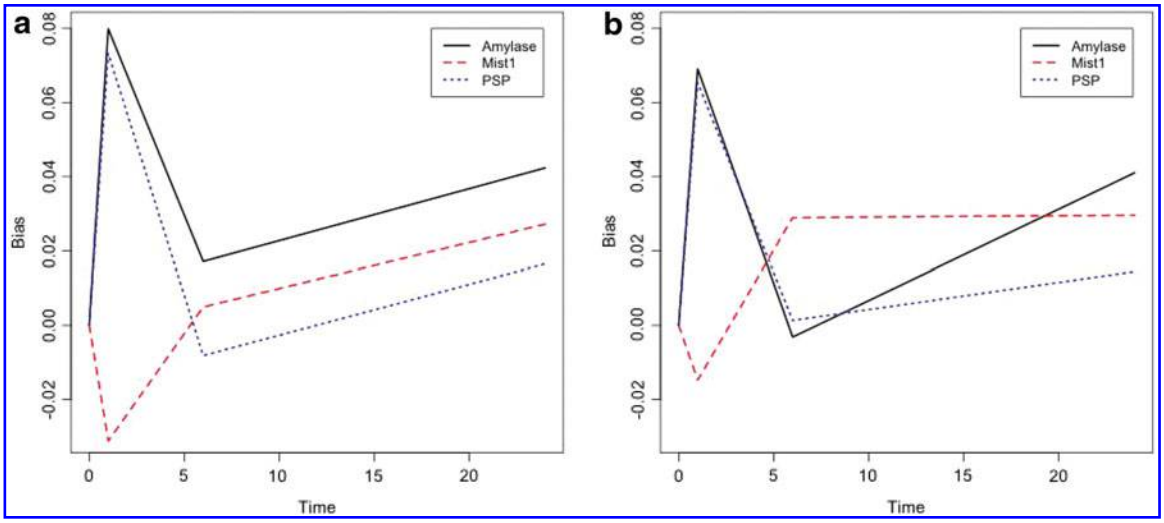

FIG. 5. Bias of the model prediction for mRNA data: (**a**) $\mathcal{M}_1$ (**b**) $\mathcal{M}_2$.

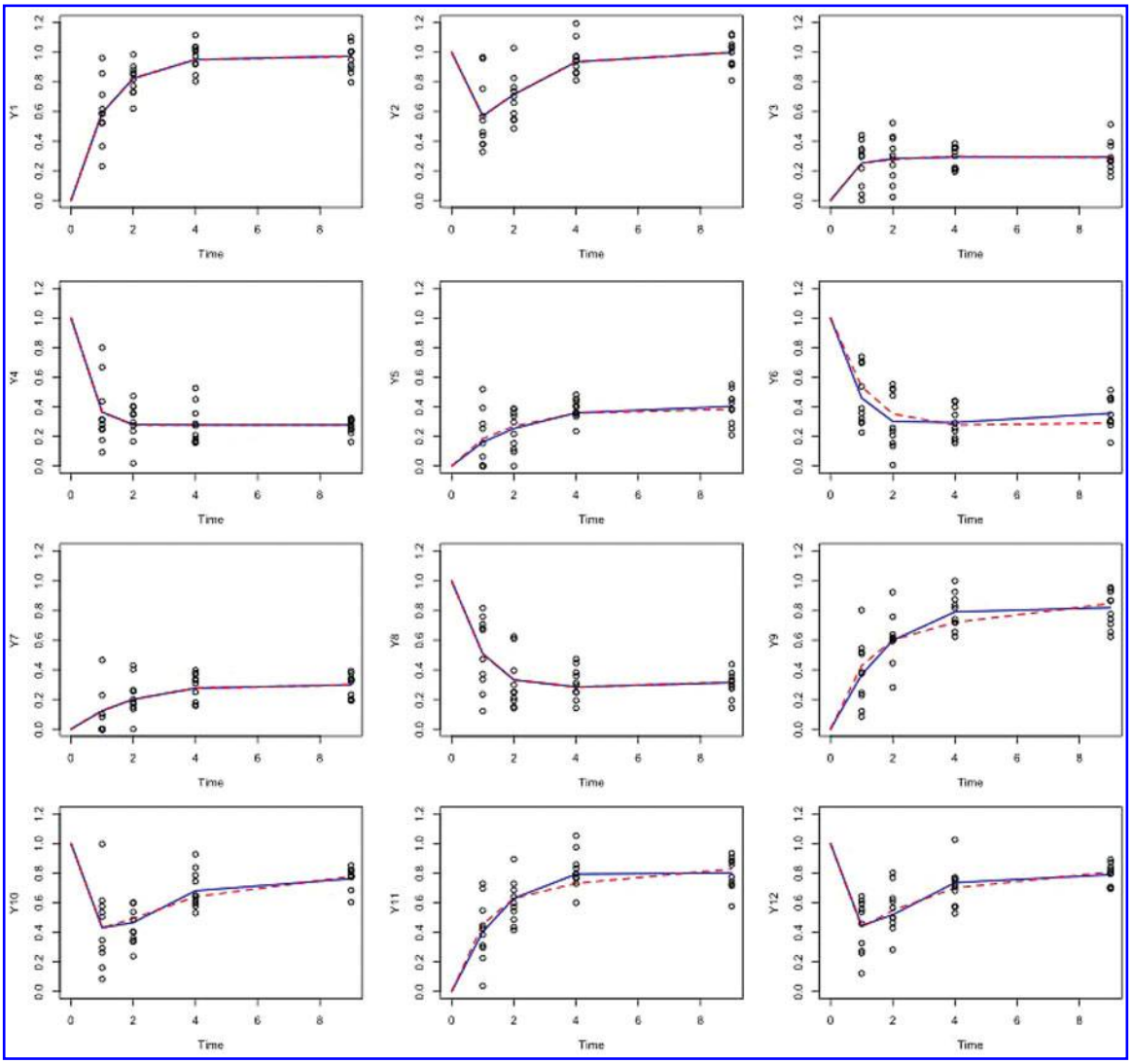

FIG. 6. Synthetic time-course data and fitted lines for $\mathcal{M}_1$ and $\mathcal{M}_2$. Circles are synthetic data points; a blue solid line represents the correct model $\mathcal{M}_1$ fitted to the data using the least-square method, and a red dotted line represents the fitted partially incorrect model $\mathcal{M}_2$.

probability of being correct than the partially altered model $_2$ of (B.2) and, consequently, to select the correct one.

3. DISCUSSION

Selecting the model that best describes a given biological system is increasingly relevant in the modern era of systems biology and ever-improving molecular data collection methods. In the context of discriminating between competing kinetic ODE-based models of interactions among several biomolecular species, we presented here a possible solution to the problem with a new, conceptually simple and flexibly applicable data-based discrimination rule. The rule is based on analyzing the appropriate Bayes factor and draws from the fields of Bayesian analysis, computer model validation, and modern MCMC analysis. Specifically, the proposed rule is based on the pairwise comparisons of the model-specific posterior distributions induced by the data against those predicted by the corresponding ODE models and is similar in spirit to a longitudinal version of the celebrated Fisher's discriminant analysis. Due to its fully Bayesian implementation (as opposed to an approximation, like, e.g., the Bayesian information criterion), the method

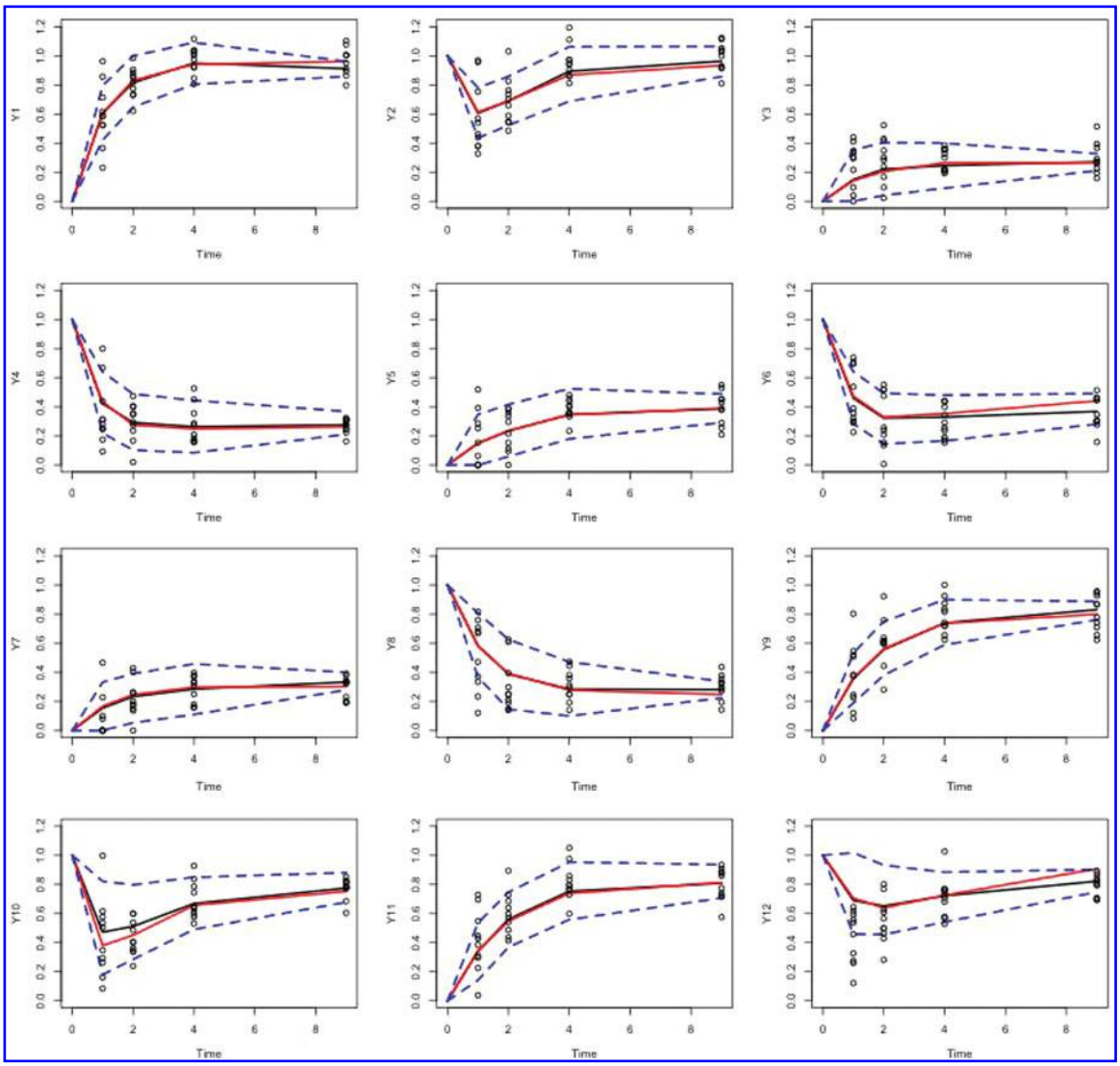

FIG. 7. The 95% posterior credibility bounds for $\mathcal{M}_1$. Simulated data (circle) versus the pure model prediction $y_i^M(\mathbf{t}, \hat{\mathbf{v}}_i)$ (red line) and the bias-corrected model prediction $\hat{y}_i^L(\mathbf{t})$ (black line) along with the 95% credibility bounds (blue dotted line).

allows us to take properly into account the competing models' complexity, and hence, automatically penalizes models that attempt to over-fit the data. In two particular examples considered in this article—de-differentiation process of a mammalian salivary gland and a large synthetic model of nonlinear dynamics from the DREAM6 competition—the proposed method worked well, despite the small size of the longitudinal data set, multiple noise sources, a relatively high experimental variation between technical replicates, and parametric similarity of the competing ODE models. In spite of these challenges, the Bayesian criterion in both examples clearly identified the models with less bias and with biochemically more realistic dynamics as the more likely ones.

Whereas the immediate goal of the proposed method is to discriminate between the two competing, predefined models, one could possibly envision the use of a similar criterion sequentially in order to build up, in a stepwise process, the most likely posterior model among a set of dynamically proposed competitors. The potential candidates for comparisons could be generated based on the current model and either the available biological knowledge or a random walk across the model space. However, such model-building considerations, being computationally challenging, would with no doubt require further refinements of the currently proposed discrimination method.

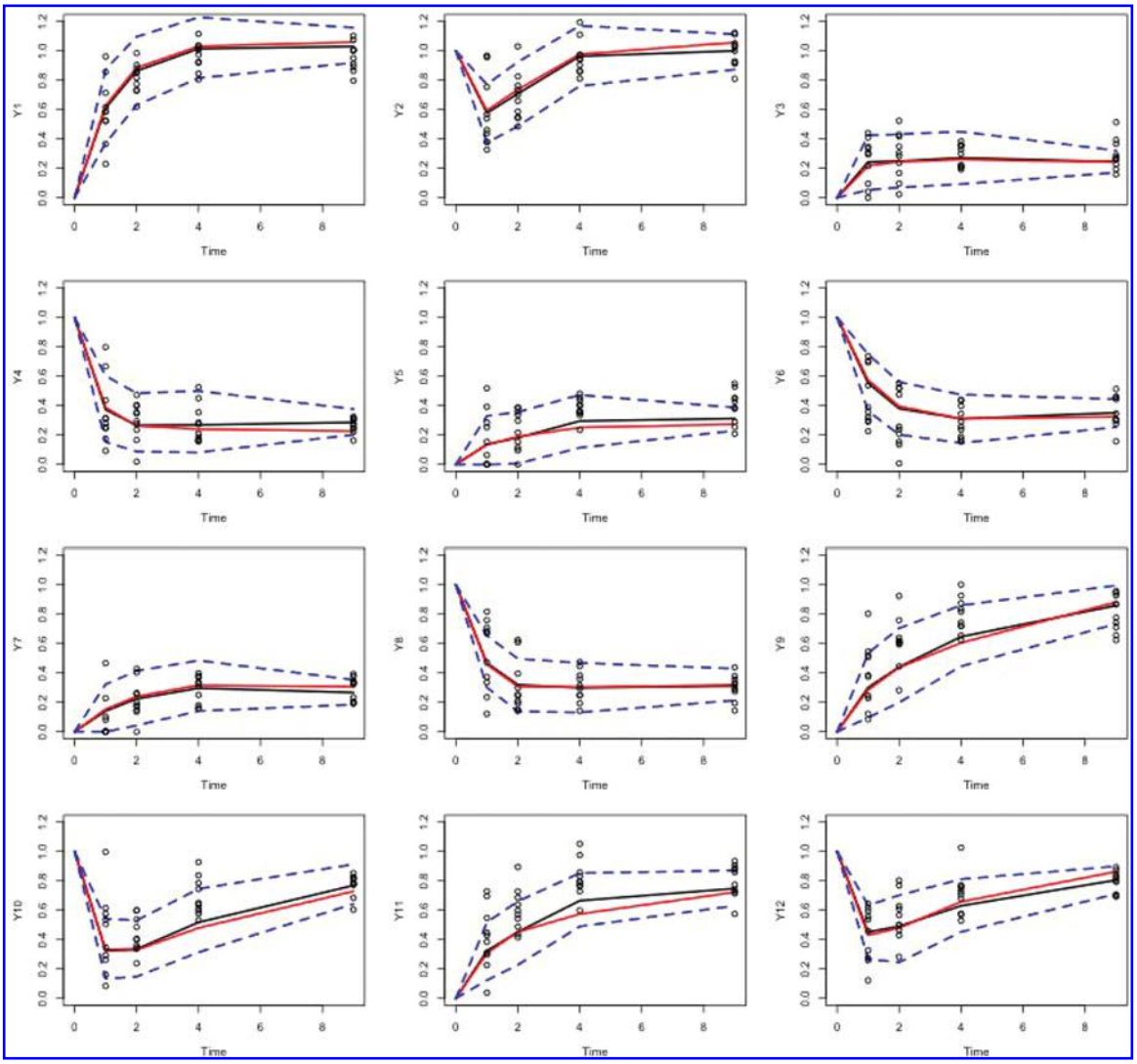

FIG. 8. The 95% posterior credibility bounds for $\mathcal{M}_2$. Simulated data (circle) versus the pure model prediction $y_i^M(\mathbf{t}, \hat{\mathbf{v}}_i)$ (red line) and the bias-corrected model prediction $\hat{y}_i^L(\mathbf{t})$ (black line) along with the 95% credibility bounds (blue dotted line).

The empirically demonstrated relative robustness and sensitivity of the proposed selection criterion indicates that it might be appropriate for use with other molecular models, particularly in sparse-data scenarios. While in our example we used a specific parametric form of the likelihood function underlying the data, both the parametric assumptions as well as the MCMC algorithms applied could be easily modified to accommodate the specifics of other systems. In our settings, the Gaussian likelihood model seemed justified by the approximation results from biochemical stochastic dynamics of aggregated cellular systems. However, when considering, for instance, data from next-generation (high-throughput)

TABLE 2. PARAMETER ESTIMATES (POSTERIOR MEANS) AND THEIR STANDARD DEVIATIONS UNDER $\mathcal{M}1$ AND $\mathcal{M}2$

Model	*Mean & SD*	v_1	v_2	v_3	v_4	v_5	v_6	v_7
$\mathcal{M}_1$	$\hat{v}_i$	21.4868	4.0750	1.6180	0.2990	20.2618	2.1828	4.5004
	$SD(\hat{v}_i)$	16.0030	2.8402	1.5623	0.0557	16.6145	2.3336	2.1067
$\mathcal{M}_2$	$\hat{v}_i$	1.7622	0.9648	0.3237	1.4291	1.2157		
	$SD(\hat{v}_i)$	2.0204	0.7260	0.0526	1.4561	0.7214		

sequencing, the use of Poisson likelihood might be more appropriate. In general, with suitable adjustments due to the nature of various experimental data, the methodology presented here should be broadly applicable to a wide variety of biological settings and experimental designs.

4. APPENDIX A

Throughout this section, we use the terminology developed within the computer validation framework as described, for instance, in Bayarri et al. (2007). For further discussions and other recent alternatives, see, e.g., Zhan et al. (2011). Let $\mathbf{v}_{il}$ and $\mathbf{b}_{il}$, $l=\{A, M, P\}$, $l=1, \ldots, n$ be n posterior samples obtained by the converged MCMC algorithm as discussed in the article. The quantities of interest may be approximately estimated from these samples as follows. The estimated posterior means for $\mathbf{v}_i$ and $\mathbf{b}_i$ may be approximated by, respectively,

$$\hat{\mathbf{v}}_i = \frac{1}{n}\sum_{l=1}^{n} \mathbf{v}_{il}, \quad i=\{A, M, P\}, \tag{A.1}$$

and

$$\hat{\mathbf{b}}_i = \frac{1}{n}\sum_{l=1}^{n} \mathbf{b}_{il}, \quad i=\{A, M, P\}. \tag{A.2}$$

Similarly, the bias-corrected model prediction for $y_i^L(\mathbf{t})$ may be approximated by

$$\hat{y}_i^L(\mathbf{t}) = \frac{1}{n}\sum_{l=1}^{n} \left(y_i^M(\mathbf{t}, \mathbf{v}_{il}) + \mathbf{b}_{il}\right), \quad i=\{A, M, P\}, \tag{A.3}$$

where $y_i^M(\mathbf{t}, \mathbf{v}_{il})$ is the vector of mean pure model predictions. The bias of the pure model prediction may be approximated by

$$\hat{\mathbf{b}}_i^{\hat{\mathbf{v}}_i} = \hat{y}_i^L(\mathbf{t}) - y_i^M(\mathbf{t}, \hat{\mathbf{v}}_i), \quad i=\{A, M, P\}. \tag{A.4}$$

Finally, the variance of the bias-corrected predictor (A.3) may be approximately calculated as

$$V\{\hat{y}_i^L(\mathbf{t})\} = \frac{1}{n}\sum_{l=1}^{n} \left(y_i^M(\mathbf{t}, \mathbf{v}_{il}) + \mathbf{b}_{il} - \hat{y}_i^L(\mathbf{t})\right)^2, \quad i=\{A, M, P\}. \tag{A.5}$$

5. APPENDIX B

For the synthetic data analysis of Section 2.5, the correct ODE model $\mathcal{M}_1$ is given by the following ODE system (''Model1'' in DREAM6).

$$\begin{aligned}
\frac{dY_1(t)}{dt} &= v_1 - v_2 Y_1(t),\\
\frac{dY_2(t)}{dt} &= v_3 Y_1(t) - v_4 Y_2(t),\\
\frac{dY_3(t)}{dt} &= v_5 \frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_{12}(t)} - v_6 Y_3(t),\\
\frac{dY_4(t)}{dt} &= v_7 Y_3(t) - v_8 Y_4(t),\\
\frac{dY_5(t)}{dt} &= v_9 \frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_4(t)} - v_{10} Y_5(t),\\
\frac{dY_6(t)}{dt} &= v_{11} Y_5(t) - v_{12} Y_6(t),\\
\frac{dY_7(t)}{dt} &= v_{13} \frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_{10}(t)} - v_{14} Y_7(t),
\end{aligned}$$

$$\begin{aligned}
\frac{dY_8(t)}{dt} &= v_{15}Y_7(t) - v_{16}Y_8(t),\\
\frac{dY_9(t)}{dt} &= v_{17}\frac{1}{1+Y_8(t)} - v_{18}Y_9(t),\\
\frac{dY_{10}(t)}{dt} &= v_{19}Y_9(t) - v_{20}Y_{10}(t),\\
\frac{dY_{11}(t)}{dt} &= v_{21}\frac{1}{1+Y_8(t)} - v_{22}Y_{11}(t),\\
\frac{dY_{12}(t)}{dt} &= v_{23}Y_{11}(t) - v_{24}Y_{12}(t).
\end{aligned} \tag{B.1}$$

A competing model $\mathcal{M}_2$ is defined by the following set of partially altered ODE, where the alterations (in the growth rates equations for Y_9 and Y_{11}) were made so as to preserve the overall model dynamics.

$$\begin{aligned}
\frac{dY_1(t)}{dt} &= v_1 - v_2Y_1(t),\\
\frac{dY_2(t)}{dt} &= v_3Y_1(t) - v_4Y_2(t),\\
\frac{dY_3(t)}{dt} &= v_5\frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_{12}(t)} - v_6Y_3(t),\\
\frac{dY_4(t)}{dt} &= v_7Y_3(t) - v_8Y_4(t),\\
\frac{dY_5(t)}{dt} &= v_9\frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_4(t)} - v_{10}Y_5(t),\\
\frac{dY_6(t)}{dt} &= v_{11}Y_5(t) - v_{12}Y_6(t),\\
\frac{dY_7(t)}{dt} &= v_{13}\frac{Y_2(t)}{1+Y_2(t)}\frac{1}{1+Y_{10}(t)} - v_{14}Y_7(t),\\
\frac{dY_8(t)}{dt} &= v_{15}Y_7(t) - v_{16}Y_8(t),\\
\frac{dY_9(t)}{dt} &= v_{17}Y_6(t) - v_{18}Y_9(t),\\
\frac{dY_{10}(t)}{dt} &= v_{19}Y_9(t) - v_{20}Y_{10}(t),\\
\frac{dY_{11}(t)}{dt} &= v_{21}Y_6(t) - v_{22}Y_{11}(t),\\
\frac{dY_{12}(t)}{dt} &= v_{23}Y_{11}(t) - v_{24}Y_{12}(t).
\end{aligned} \tag{B.2}$$

In order to generate a synthetic time-course dataset for 12 species ($Y_1(\mathbf{t}), \ldots, Y_{12}(\mathbf{t})$), we used the correct set of ODE models of (B.1), which is a simplified version of the Model1 from DREAM6. The parameters of both ODE models, as well as the initial conditions, followed the setting of Model1 (i.e., $\mathbf{v} = (v_1, \ldots, v_{24})' = (1, \ldots, 1)'$ and $Y_1(0) = Y_3(0) = Y_5(0) = Y_7(0) = Y_9(0) = Y_{11}(0) = 0$ and $Y_2(0) = {}_2(0) = Y_4(0) = Y_6(0) = Y_8(0) = Y_{10}(0) = Y_{12}(0) = 1$). In the experiment described in the current article, we considered 4 time-points and 10 experiments for each species (which is consistent with typical DREAM6 scenarios). The time-course vector for each species is therefore $y_{ij}^L(\mathbf{t}), i = 1, \ldots, 12$, $j = 1, \ldots, 10$, with $\mathbf{t} = (t_1, \ldots, t_4)'$, where $t_1 = 1$, $t_2 = 2$, $t_3 = 4$, and $t_4 = 9$. The bias $b_i(\mathrm{t})$ was generated from the multivariate normal distribution with the mean vector 0 and the variance–covariance matrix $\mathrm{C} = [\exp\{-((t_m - t_n)/10)^2\}]_{mn}$, $m, n = 1, \ldots, 4$, and the random error ϵ_{ij} $(\mathbf{t})$ was randomly sampled from the multivariate normal distribution with the mean vector 0 and the diagonal variance–covariance matrix with diagonal element δ_i, where $\delta_i = [\exp\{-t_k/3\}]_k$, $k = 1, \ldots, 4$. The data for 10 time-course experiments we generated from (B.1) and used in the analysis under the parametric noise (4) model. Both the generated data and the least-squares fitted ODE model trajectories are presented in Figure 6.

ACKNOWLEDGMENTS

The research of all the authors was partially funded by the U.S. National Institutes of Health under grant R01DE019243 to D.S.D. and G.A.R. The authors are grateful to the reviewers for their helpful comments.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Andersson, H., and Britton, T. 2000. *Stochastic Epidemic Models and Their Statistical Analysis* (Lecture Notes in Statistics, Vol. 151). Springer, New York.

Baker, C.L., Loros, J.J., and Dunlap, J.C. 2011. The circadian clock of neurospora crassa. *FEMS Microbiol. Rev*. 36, 95–110.

Bayarri, M., Berger, J., Paulo, R., and Sacks, J. 2007. A framework for validation of computer models. *Technometrics* 49, 138–154.

Brown, E., Choe, Y., Luithardt, H., and Czeisler, C. 2000. A statistical model of the human core-temperature circadian rhythm. *American Journal of Physiology-Endocrinology and Metabolism* 279, E669–E683.

Carrillo, N., Ceccarelli, E., and Roveri, O. 2010. Usefulness of kinetic enzyme parameters in biotechnological practice. *Biotechnol. Genet. Eng. Rev.* 27, 367–82.

Chang, J., Yuan, Y.-J., Yang, D.-Y., and Liu, X.-D. 2011. Pore network model and simulation of transport process for grain drying. *Chemical Engineering & Technology* 34, 1049–1056.

El Samad, H., Khammash, M., Petzold, L., and Gillespie, D. 2005. Stochastic modeling of gene regulatory networks. *Int. J. Rob. Nonlin. Control* 15, 691–711.

Fisher, R. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179–188.

Ge, H., and Qian, H. 2009. Thermodynamic limit of a nonequilibrium steady state: Maxwell-type construction for a bistable biochemical system. *Phys. Rev. Lett.* 103, 148103.

Gin, E., Crampin, E.J., Brown, D.A., et al. 2007. A mathematical model of fluid secretion from a parotid acinar cell. *Journal of Theoretical Biology* 248, 64–80.

Gorr, S.-U., Venkatesh, S.G., and Darling, D.S. 2005. Parotid secretory granules: crossroads of secretory pathways and protein storage. *J. Dent. Res.* 84, 500–509.

Grundmann, O., Mitchell, G.C., and Limesand, K.H. 2009. Sensitivity of salivary glands to radiation: from animal models to therapies. *J. Dent. Res.* 88, 894–903.

Harunaga, J., Hsu, J.C., and Yamada, K.M. 2011. Dynamics of salivary gland morphogenesis. *J. Dent. Res.* 90, 1070–1077.

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.

Hecker, M., Lambeck, S., Toepfer, S., et al. 2009. Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* 96, 86–103.

Hornos, J., Schultz, D., Innocentini, G., et al. 2005. Self-regulating gene: An exact solution. *Physical Review E* 72, 051907.

Jolma, I.W., Laerum, O.D., Lillo, C., and Ruoff, P. 2010. Circadian oscillators in eukaryotes. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 533–549.

Jones, R.H. 2011. Bayesian information criterion for longitudinal and clustered data. *Stat. Med.* 30, 3050–3056.

Liao, F., Zhu, X., Wang, Y., and Zuo, Y. 2005. The comparison of the estimation of enzyme kinetic parameters by fitting reaction curve to the integrated michaelis-menten rate equations of different predictor variables. *J. Biochem. Biophys. Methods* 62, 13–24.

Likhoshvai, V., and Ratushny, A. 2007. Generalized hill function method for modeling molecular processes. *J. Bioinform. Comput. Biol.* 5, 521–531.

Luo, X., Shin, D.M., Wang, X., et al. 2005. Aberrant localization of intracellular organelles, ca2+ signaling, and exocytosis in mist1 null mice. *J. Biol. Chem.* 280, 12668–12675.

Morton, V.L., Dykeman, E.C., Stonehouse, N.J., et al. 2010. The impact of viral RNA on assembly pathway selection. J. Mol. Biol. 401, 298–308.

Neff, K.L., Offord, C.P., Caride, A.J., et al. 2011. Validation of fractal-like kinetic models by time-resolved binding kinetics of dansylamide and carbonic anhydrase in crowded media. *Biophys. J.* 100, 2495–2503.

Novak, M., and Tyson, J. 2004. A model for restriction point control of the mammalian cell cycle. *Journal of Theoretical Biology* 230, 563–579.

Rempala, G., Ramos, K., Kalbfleisch, T., and Teneng, I. 2007. Validation of a mathematical model of gene transcription in aggregated cellular systems: application to l1 retrotransposition. *Journal of Comput. Biol.* 14, 339–349.

Skanda, D., and Lebiedz, D. 2010. An optimal experimental design approach to model discrimination in dynamic biochemical systems. *Bioinformatics* 26, 939–945.

Soneji, S., Huang, S., Loose, M., et al. 2007. Inference, validation, and dynamic modeling of transcription networks in multipotent hematopoietic cells. *Hematopoietic Stem Cells VI* 1106, 30–40.

Sorribas, A., Hernández-Bermejo, B., Vilaprinyo, E., and Alves, R. 2007. Cooperativity and saturation in biochemical networks: a saturable formalism using Taylor series approximations. *Biotechnol. Bioeng* 97, 1259–77.

Steimer, J.-L., Dahl, S.G., De Alwis, D.P., et al. 2010. Modelling the genesis and treatment of cancer: the potential role of physiologically based pharmacodynamics. *Eur. J. Cancer* 46, 21–32.

Sveiczer, A., Csikasz-Nagy, A., Gyorffy, B., et al. 2000. Modeling the fission yeast cell cycle: Quantized cycle times in wee1(-) cdc25 delta mutant cells. *Proceedings of the National Academy of Sciences of the United States of America* 97, 7865–7870.

Tyson, J., Chen, K., and Novak, B. 2001. Network dynamics and cell physiology. *Nature Reviews Molecular Cell Biology* 2, 908–916.

Venkatesh, S., Kim, J., Li, J., et al. 2012. Differentiation model of salivary gland network. Submitted manuscript, Dept of Biostatistics, Georgia Regents University.

Wang, G. 2007. Estimation of the proliferation and maturation functions in a physiologically structured model of thymocyte development. *Journal of Mathematical Biology* 54, 761–786.

Wilkinson, D.J. 2009. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet.* 10, 122–133.

Zhan, Z., Fu, Y., Yang, R.-J., and Peng, Y. 2011. An enhanced Bayesian-based model validation method for dynamic systems. *Journal of Mechanical Design* 133, 041005.

Address correspondence to:
Grzegorz A. Rempala
Department of Biostatistics
Georgia Regents University
Augusta, GA 30912

E-mail: grzes68@me.com