ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

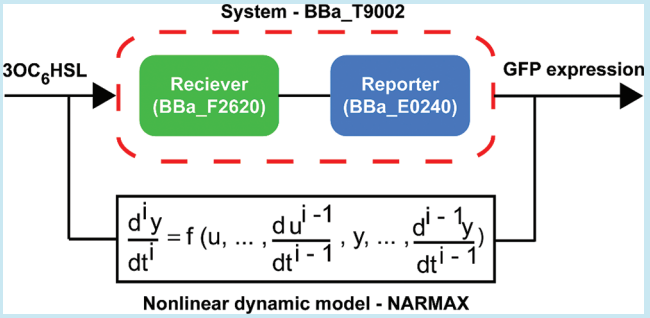
A Data-Driven Framework for Identifying Nonlinear Dynamic Models of Genetic Parts

Kirubhakaran Krishnanathan,* Sean R. Anderson, Stephen A. Billings, and Visakan Kadirkamanathan

Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K.

Supporting Information

ABSTRACT: A key challenge in synthetic biology is the development of effective methodologies for characterization of component genetic parts in a form suitable for dynamic analysis and design. In this investigation we propose the use of a nonlinear dynamic modeling framework that is popular in the field of control engineering but is novel to the field of synthetic biology: Nonlinear AutoRegressive Moving Average model with eXogenous inputs (NARMAX). The framework is applied to the identification of a genetic part *BBa_T9002* as a case study. A concise model is developed that exhibits accurate representation of the system dynamics and a structure that is compact and consistent across cell populations. A comparison is made with a biochemical model, derived from a simple enzymatic reaction scheme. The NARMAX model is shown to be comparably simple but exhibits much greater prediction accuracy on the experimental data. These results indicate that the data-driven NARMAX framework is an attractive technique for dynamic modeling of genetic parts.



KEYWORDS: *NARMAX, nonlinear system identification, biochemical models, dynamic, genetic parts*

The field of synthetic biology has progressed through a number of stages[1,2] from the early concept,[3,4] to the initial demonstrations of simple devices,[5−7] and more recently to systems composed of integrated modules that perform useful tasks.[8−11] There is now an expectation that synthetic biology will deliver solutions to global challenges, for instance, in healthcare,[12−14] food security,[15,16] and energy production.[17−19] However, there are a number of obstacles to overcome before realizing such transformative systems in practice.[20,21] In particular, challenges in characterization and design are directly linked to our ability to derive useful models of genetic parts and biosystems because the concept of model-based design is central to the engineering ethos. Therefore, methodologies for model-based design are likely to underpin the future success of top-bottom biological synthesis using off the shelf genetic parts and modules.[22,23] Here, we extend recent results in static input-output characterization[24] by developing a data-driven framework for describing the dynamic properties of genetic parts. Models derived using this framework could be specified in data sheets associated with genetic parts with the purpose of aiding in the control design and synthesis of larger systems.

The system model aids design procedures and prediction, as well as promoting high reliability and modular abstraction.[21,23] Recent advances in characterizing and describing genetic parts[24] have stimulated debate concerning the appropriate strategies for modeling, highlighting choices for the model form, such as static versus dynamic, deterministic versus stochastic, and single-cell versus population level.[21] Taking inspiration from the engineering community, we observe that in the field of control engineering, data-driven dynamic models dominate process descriptions, where techniques for modeling are derived from the domain of system identification.[25]

Data-driven system identification contrasts with modeling approaches commonly applied to biological processes. These include protein expression[26−29] and cell growth,[30] which have been modeled using simplifications of enzymatic reaction schemes, e.g., the Michaelis−Menten (MM) and Hill equations.[31] These models are simple to apply but limited, retaining only a fixed model structure that is not adjusted to the complexity of the system under investigation. At the other extreme, dynamic models used in synthetic biology have also been explored that take the form of many coupled ordinary[32−34] or stochastic[35] differential equations (ODEs and SDEs, respectively): As the system and network grows larger, the number of equations needed to describe the interlinked processes using ODEs or SDEs increases, leading to an explosion in model complexity. Models of this type are typically intractable for dynamic systems analysis and design.

In this investigation we propose the use of the Nonlinear AutoRegressive Moving Average model with eXogenous input (NARMAX) modeling framework.[36,37] We suggest that the NARMAX framework is a solution for overcoming the typical problems of models for genetic parts that are overly complex, unwieldy, and of unknown structure. The NARMAX model class provides a general nonlinear dynamic system description,

Received: January 31, 2012
Published: March 23, 2012

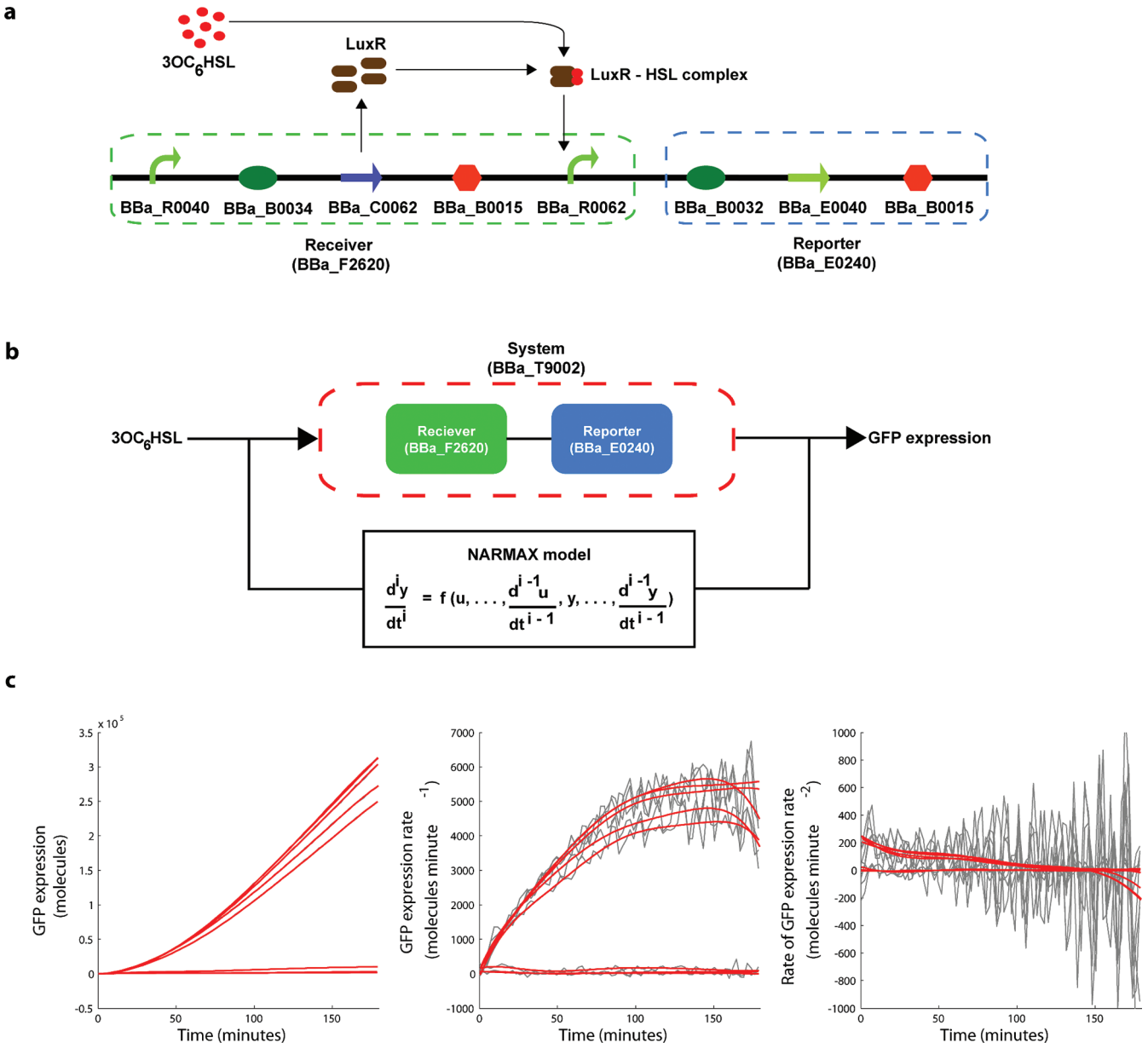


Figure 1. Genetic part *BBa_T9002*, NARMAX model representation and identification signals. (a) The *BBa_T9002* system with input and output of $3OC_6HSL$ and GFP expression, respectively. (b) NARMAX model representation of the *BBa_T9002* system. (c) The observed GFP signal and derivatives obtained from the smoothing algorithm (red) in comparison to derivatives from numerically differencing the raw GFP signal (gray).

which was developed in the field of system identification: it is a data-driven technique in terms of both parameters and structure.[38,39] In relation to the model representation issues discussed above: the NARMAX model (i) is dynamic but can represent static relationships (as can all dynamic models, for which the reverse is not true); (ii) can represent both deterministic and stochastic processes via noise models; (iii) is data-driven and hence can apply equally well to single-cell or population level representations, depending on the scale of the process observations.

The NARMAX methodology has been applied in a wide range of areas, such as biomedical engineering,[40,41] sound and vibration systems,[42,43] power generation,[44,45] and econometrics.[46] The breadth and success of these various applications demonstrates the versatility and utility of the model class. The use of the NARMAX modeling framework in the area of synthetic biology is a novel step, which we suggest will hold the following advantages:

1. Compact model descriptions: the NARMAX model can produce very compact descriptions of systems in comparison to alternatives such as the Volterra series[47] and biophysically derived ODEs/SDEs, which will be well suited to biosystem design.
2. Data-driven structure detection and parameter estimation: typical approaches to modeling biosystems require insight into the underlying biochemical processes, which often results in either oversimplified or overparameterised descriptions. The NARMAX methodology provides a framework for detecting a parsimonious set of model terms that describes the observed dynamics.
3. Integrated framework for identification, analysis, and design: there is wide supporting literature that is related

dx.doi.org/10.1021/sb300009t | *ACS Synth. Biol.* 2012, 1, 375–384

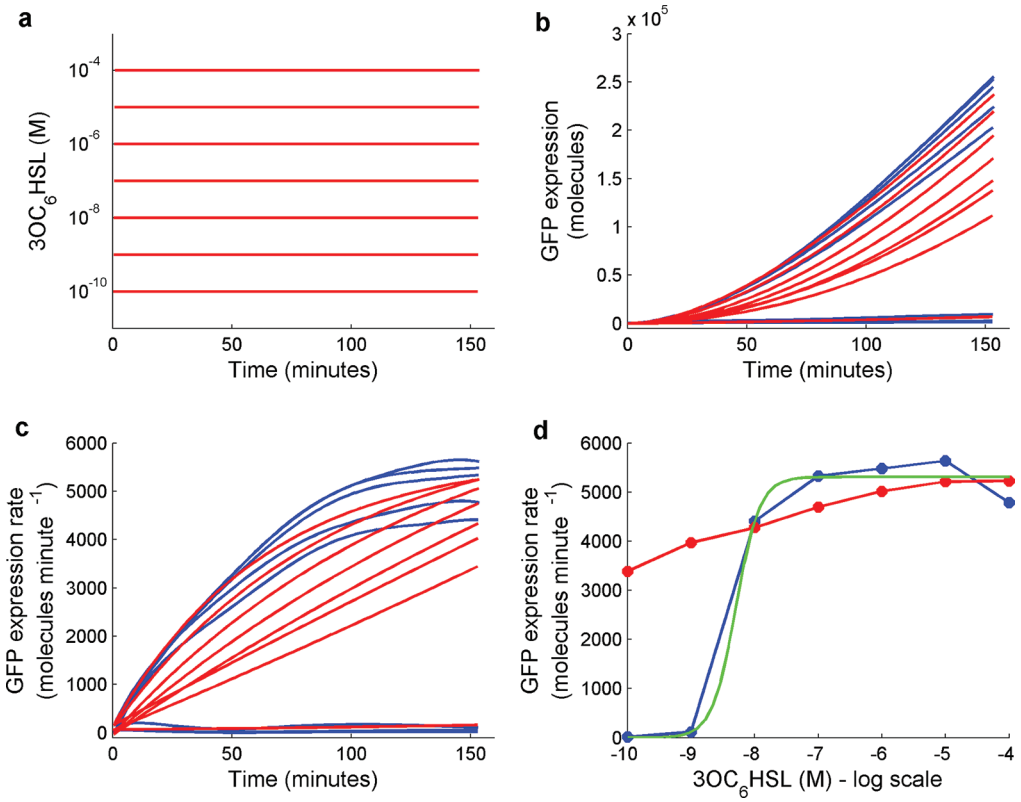


Figure 2. Biochemical model simulation for Expt 1. (a) The simulated model input signal $s(t)$. (b) Comparison of GFP expression (blue) and the model prediction $p(t)$ (red). (c) Rate of change of GFP expression (blue) and model prediction (red). (d) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), and ERS model prediction (red). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ has been omitted because of the log transformation.

not only to NARMAX model identification but also to the use of such models in dynamic systems analysis and control design.[48,49]

To demonstrate the identification approach outlined above, here we develop a continuous-time NARMAX model[50] for the genetic part *BBa_T9002* (this is a labeled identity of the genetic part in the registry of parts: http://partsregistry.org/Part:BBa_F2620). The advantage of using this system as a case study is that it has been well-studied and characterized using alternative modeling techniques[24] and the data is available online, facilitating further comparison and investigation. The genetic part *BBa_T9002* is a composite receiver-reporter quorum sensing system that is made up of *BBa_F2620* (receiver) and *BBa_E0240* (reporter). Canton et al.[24] derived a model based on the Hill equation for *BBa_F2620* indirectly using the data collected for *BBa_T9002* (where knowledge of the model for *BBa_E0240* was assumed). Here, we first take the approach of modeling the complete system using the original dynamic form of the reaction scheme from which the Hill equation is derived, which highlights certain drawbacks of that approach, and second, we identify the system using the NARMAX framework producing a compact and accurate model, useful for systems analysis.

RESULTS AND DISCUSSION

Overview. *BBa_T9002* is a quorum sensing receiver-reporter system[24] shown in Figure 1a. In the absence of tetracycline, LuxR protein is constitutively expressed and upregulated on addition of 3-ox-ohexanoyl-L-homoserine lactone ($3OC_6HSL$), in cells that have little or no TetR. The LuxR protein and $3OC_6HSL$ forms a complex that activates the LuxR regulated promoter (*BBa_R0062*) producing the receiver's (*BBa_F2620*) output, polymerases per second (PoPS). On the activation of the LuxR regulated promoter, the expression of green florescence protein GFP can be observed, which serves as the output for the *BBa_T9002* system.

Two different approaches to modeling the system were taken here: a biochemical model based on a well-known description of an enzymatic reaction scheme (ERS) and a data-driven continuous-time NARMAX model (Figure 1b). Models were identified directly in continuous-time. To estimate derivatives of the GFP expression signal used in the identificaton of each model, we used a smoothing algorithm (see Methods). This technique led to derivative estimates that were relatively noise-free compared to directly differenced signals (Figure 1c).

Identification of Dynamic and Static Biochemical Models. *Introduction.* The Michaelis–Menten equation and the Hill equation (see Methods) are well-known models used in describing enzymatic reactions. These simplified models are derived from a more complicated coupled nonlinear dynamic model (see Methods) using the assumptions that the total enzyme concentration is constant, the rate of change of enzyme–substrate complex is zero, and the substrate level is at steady state (further discussed in the Supporting Information). A drawback of the MM and Hill equations is that they do not provide a full dynamic description of the system; only the static relationship between product derivative and substrate input is captured. To address this drawback we used the full dynamic description of the enzymatic reaction

dx.doi.org/10.1021/sb300009t | *ACS Synth. Biol.* 2012, 1, 375–384

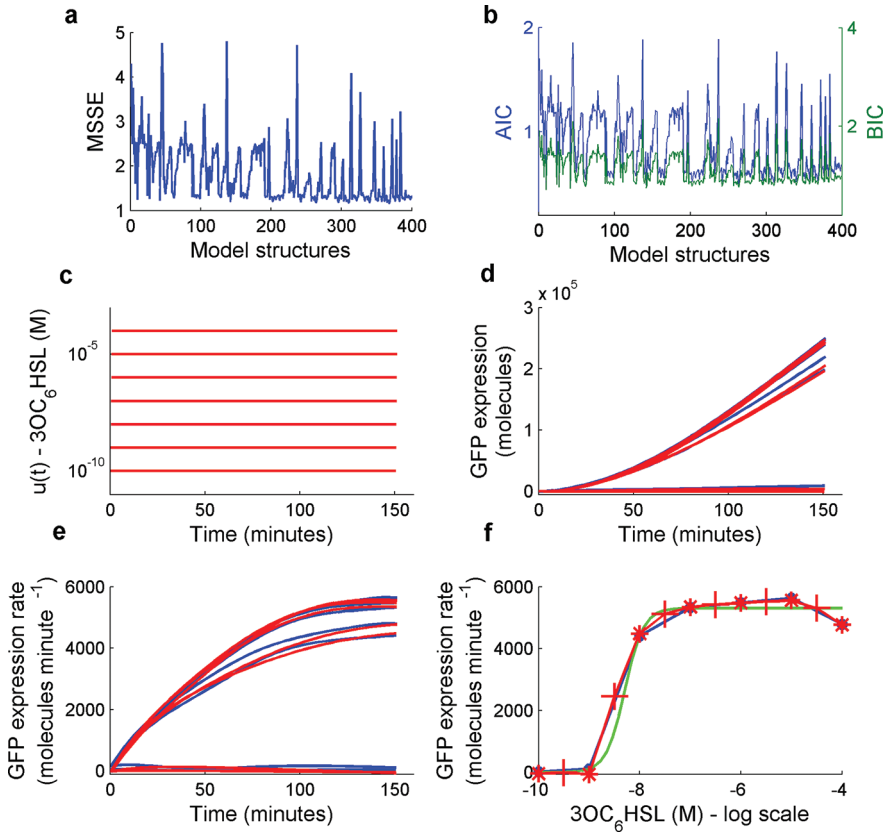


Figure 3. NARMAX structure detection and model simulation for Expt 1. (a) Mean-squared-simulation-error (MSSE) for NARMAX models with an MSSE < 5 (models are ordered by increasing complexity, i.e., number of model terms). (b) Akaike and Bayesian information criteria (AIC and BIC, respectively): the optimal model with minimum AIC and BIC is model structure 16 (models are ordered by increasing complexity, i.e., number of model terms). (c) NARMAX model input signals. (d) Observed GFP signals (blue) and NARMAX prediction (red). (e) Rate of change of GFP (blue) and NARMAX model prediction (red). (f) Rate of change of GFP expression at the 150th minute (blue), the Hill equation prediction (green), the NARMAX model prediction at observed input concentrations (red stars), and the NARMAX model prediction at interpolated input concentrations (red crosses). Note that the response corresponding to the lowest input level $3OC_6HSL = 0$ has been omitted because of the log transformation.

scheme associated with the Hill equation, which we denote here as the enzymatic reaction scheme (ERS) model (see Methods).

Enzymatic Reaction Scheme Model of BBa_T9002. The ERS model was identified from observations of the dynamic behavior of the genetic part *BBa_T*9002 (see Methods). Analyzing the simulated results in comparison to the experimental data, we noted that the estimated level of the $3OC_6HSL$ signal remained constant from the time of induction to quasi steady state (Figure 2a); it may be the case here that the concentration of $3OC_6HSL$ at the point of induction was very high in comparison to the amount used up by the cells. In addition, the observed GFP dynamics were not well described by the ERS model (Figure 2b and c): the prediction error variance of the GFP signal was 24.3%. This low prediction accuracy indicated the possibility of missing dynamics in the ERS model or that the ERS model structure was not appropriate for describing the system.

The peak values of rate of GFP expression indicated a sigmoid shape (Figure 2d), which is usually observed when cooperative binding is involved in the reaction process.[31] Therefore we used the Hill equation to fit these peak values as a function of $3OC_6HSL$. The Hill equation showed a much improved fit to the steady-state behavior compared to the ERS model prediction at the same point in time (Figure 2d).

Enzymatic Reaction Scheme Model Is Not Consistent with the Hill Equation. The Hill equation is generally considered by the synthetic community to be a useful model for describing simple switching properties in biochemical processes.[24,51,52] In this case, the Hill equation gave a much improved prediction of the steady-state GFP expression rate, compared to the ERS model (Figure 2d). However, the Hill equation failed to accurately capture the decay at high levels of $3OC_6HSL$, which was a feature across all experiments. This decay could be a result of toxicity to the cells.[53]

It is interesting to note that both the Hill equation and the ERS models are derived from the same reaction scheme (see Methods and Supporting Information). However, it is apparent from these results that the Hill equation provides a much improved model of the steady-state process. This inconsistency raises a question over the link between the Hill equation and the dynamic model on which it is predicated, and hence the interpretability of the Hill equation parameters. A possible explanation for the improvement of the Hill equation over the ERS model is that the sigmoidal form of the Hill equation is coincidentally well suited to describing the switching behavior observed in the data. Another explanation is that the optimal ERS model parameters were not obtained here due to the difficulties inherent in nonlinear parameter estimation.

In the context of biosynthesis, model inaccuracies will be problematic. If a model fails to capture the key properties of a system, then errors will be imposed on the system design. This motivates the development of alternative modeling strategies that will solve these challenges.

Identification of a Data-Driven Dynamic Model - the NARMAX Framework. *Introduction.* The NARMAX identification framework is used in a data-driven context to obtain input-output nonlinear dynamic models. For the system *BBa_T9002* we defined the input to the system as $3OC_6HSL$ concentration. The output we defined as rate of change of GFP expression. The reason for this was that it is usually preferable to model a stable system in a data-driven framework, and the initial growth in GFP is unstable, whereas its derivative is stable. Furthermore, in the data set considered here the input signal was not observed through time. Hence, we made the assumption that over the relatively short time-scale of the recording (from the time of induction to quasi steady state) the input remained constant, equivalent to the initial level of $3OC_6HSL$.

An advantage of the NARMAX framework is that the choice of model structure is data-driven. This is known as the structure detection problem, and there are a number of algorithms that can be used to automate the choices that determine model structure, e.g., dynamic order and basis function placement.[38,54,55] Structure detection is a powerful asset of the NARMAX framework because it can highlight "missing" dynamics, i.e., terms that are absent from biochemically derived models, which are required to accurately describe the system. A further advantage of the NARMAX model is that it is linear-in-the-parameters. This is a useful feature, which facilitates rapid identification and comparison of many different proposed model structures.

NARMAX Model of BBa_T9002. Typically, the superset of possible NARMAX model terms is very large, and so the structure is detected using efficient search algorithms based on, for instance, forward regression with orthogonal least-squares.[38] However, in this investigation the number of terms was relatively small (only 9 candidate terms), and so the structure was detected by an exhaustive search of all possible model structures resulting from different term combinations (a total of $2^9 = 512$ models). To detect the model with the optimal trade-off in terms of maximum accuracy and minimal complexity, we used the Akaike and Bayesian information criteria (AIC and BIC, respectively).[25] For model selection, information criteria are preferable to the use of residual-error metrics (e.g., mean-squared-simulation-error, MSSE) for model comparison. This is because ICs penalize model complexity, whereas the MSSE does not. Instead, the MSSE tends to decrease with each addition of a parameter. This point is illustrated by comparing model selection plots for Expt 1 in Figure 3a and b.

The continuous-time NARMAX model of *BBa_T9002* identified using the AIC and BIC was,

$$\ddot{y}(t) = c_1 y^2(t) + c_2 y^3(t) + c_3 \dot{y}(t) + \tilde{u}(t)$$

where $y(t)$ was the model output signal rate-of-GFP-expression, with associated parameters c_1, c_2, and c_3. The input term $\tilde{u}(t)$ was obtained from a static transformation $G(.)$ of the input signal $3OC_6HSL$, which was primarily used to describe the static switching effect in dynamics across linearly increasing levels of $3OC_6HSL$ (see Methods).

The ERS model identified above was demonstrated in simulation to be inaccurate (Figure 2), where the error variance was 24.3%. However, in contrast, the NARMAX model provided a much more accurate description of the system *BBa_T9002* (an error variance of 0.2%) while retaining a simple model structure (compare Figure 2 to Figure 3c–f). In order to test the behavior of the NARMAX model on input concentrations not used in the identification procedure, we simulated the model with additional input concentrations. The simulation results from using these intermediate inputs ($3OC_6HSL$ input concentrations: $1 \times 10^{-9.5}$, $1 \times 10^{-8.5}$, $1 \times 10^{-7.5}$, $1 \times 10^{-6.5}$, $1 \times 10^{-5.5}$, $1 \times 10^{-4.5}$ M.) demonstrated that the NARMAX model behaved as expected (Figure 3f).

In effect, we have identified a NARMAX model that seeks to describe the same relationship as part of the ERS model: the input-output dynamics between $3OC_6HSL$ and GFP expression (see eq 2 in Methods). The nonlinear dynamic terms identified using the NARMAX framework were $y^2(t)$ and $y^3(t)$. These terms have participated in greatly improved accuracy of the NARMAX description without significantly increasing the model complexity. At this stage, the biochemical features that these additional terms might represent are unknown but of interest for future work. In addition, we note that their lack of interpretability is not relevant to the utility of the model for use in design, in which context it would appear that the NARMAX model is highly preferable.

Consistent NARMAX Model Identification over a Set of Colonies. The NARMAX modeling framework was applied separately to 3 colonies, with 3 instances of the system *BBa_T9002* in each colony making a total of 9 experimental data sets: colonies 1 and 2 were used for identification of the model, and colony 3 was reserved for cross validation. The same model structure of the dynamic function was selected for all 6 experimental data sets across colonies 1 and 2. In 4 of the experimental data sets, this model structure had little or no sensitivity toward the truncation point of the data, while for the remaining 2 experimental data sets, there was some sensitivity where slight differences occurred in structure depending on the exact time at which the data was truncated. Such sensitivity is not unusual in data-driven modeling where time-domain descriptions are typically non-unique.

Parameters of the NARMAX model were fairly consistent across data sets; variation was within an order of magnitude, and consistent differences were observed between colonies (Figure 4a–c). NARMAX model simulations were similar in accuracy to those shown in Figure 3, which is indicated by the similarity in MSSE (Figure 4d). Table 1 summarizes the variability in the parameters for the dynamic NARMAX parameters.

In addition, we estimated separate static input transformation functions $G_j(.)$, for $j = 1, ..., 9$, pertaining to each data set. These were of a consistent form, with some variability over scaling (Figure 5a); a mean static transformation was also estimated using all data sets (Figure 5b). The variability in both static and dynamic parameters across data sets was likely due to heterogeneity in cell populations. This is generally caused by the total effect of intrinsic and extrinsic noise in single cells.

In order to provide a single model description for the system *BBa_T9002* across data sets, we used the mean values of dynamic parameters from Table 1 in combination with the mean static input transformation function (Figure 5b). This single model was similarly accurate in describing the dynamic

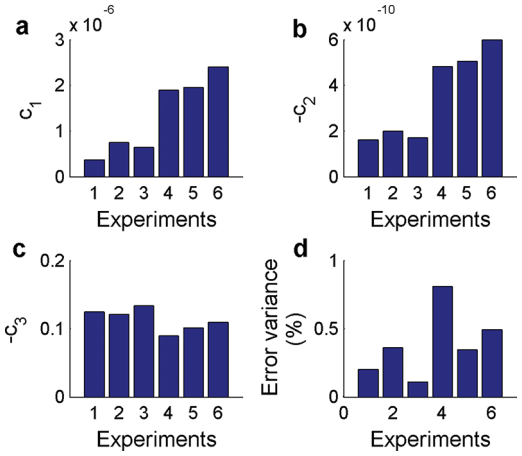


Figure 4. NARMAX model identification across colonies and experimental data sets. (a–c) Estimates of NARMAX model dynamic parameters c_1, c_2, and c_3. (d) NARMAX model mean-squared-simulation-error (MSSE) for each of 6 different experimental data sets where sets are grouped by colony: colony 1 comprises Expts 1–3; colony 2 comprises Expts 4–6.

Table 1. Mean and Variability in NARMAX Model Parameters Across Colonies

parameters	c_1	c_2	c_3
mean	1.34×10^{-6}	-3.53×10^{-10}	−0.1134
variance	7.16×10^{-13}	3.85×10^{-20}	2.61×10^{-4}

behavior of the system across both training and validation data sets (Figure 6).

Summary. In this investigation we have proposed the NARMAX modeling framework for dynamic characterization of genetic parts. This framework has particular advantages for use in top-bottom design in higher order systems: the NARMAX model is usually compact, is data-driven in both structure and parameters, and is part of a wider toolset of associated design and analysis methods. We applied the NARMAX modeling framework to the identification of a genetic part, *BBa_T9002*, for which we obtained an accurate dynamic model. In addition, we benchmarked the NARMAX model against a biochemical model, which was based on an enzymatic reaction scheme. The enzymatic reaction scheme model was shown to be inaccurate and inconsistent with its associated simplified form, the Hill equation. In contrast to the reaction scheme model, the NARMAX model provided an accurate dynamic description of the system while retaining a simple structure. On the basis of these results, the NARMAX modeling framework offers great promise for use in the characterization of synthetic biosystems.

METHODS

Experimental Data. The experimental data describing the dynamic response of the genetic part used in this investigation, *BBa_T9002*, was collected by Canton et al.[24] The data can be obtained from the BioBrick Registry of Standard Biological Parts for part *BBa_F2620* (http://partsregistry.org/Part:BBa_F2620). The experimental data consisted of GFP expression over time, observed over ∼180 min (77 time steps, sampled at intervals of ∼141 s) of *BBa_T9002* over 8 different 3-oxohexanoyl-L-homoserine lactone ($3OC_6HSL$) input concentrations: 0, 1×10^{-10}, 1×10^{-9}, 1×10^{-8}, 1×10^{-7}, 1×10^{-6}, 1×10^{-5}, 1×10^{-4} M. In this investigation we only performed systems modeling up to the point of quasi steady state behavior, which was defined as the peak of the rate of GFP expression. This truncation resulted in data records that were approximately 150 min in length (∼60 time samples; the range was 50–70 time samples). Here we analyzed data from 3 colonies of *BBa_T9002* (out of a total of 9 observed by Canton et al.[24]) There were 3 replicates for each colony, resulting in 9 expts in total analyzed here. Expts 1–3, 4–6, and 7–9 were from colonies 1, 2, and 3 respectively.

Dynamic Modeling Based on an Enzymatic Reaction Scheme. *Dynamic Representation of the Reaction Scheme.* Consider the reaction scheme

$$e + n.\, s \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} e_s \xrightarrow{k_2} e + p$$

where e denotes enzyme, s is substrate, e_s is the enzyme–substrate complex, p is product, n is the Hill coefficient, and k_1, k_{-1}, and k_2 are parameters defining the rate of reactions. In this modeling investigation we use substrate s to represent $3OC_6HSL$ and product p to represent the GFP expression. The reaction scheme can be described by the coupled dynamic nonlinear system represented using ODEs - enzymatic reaction

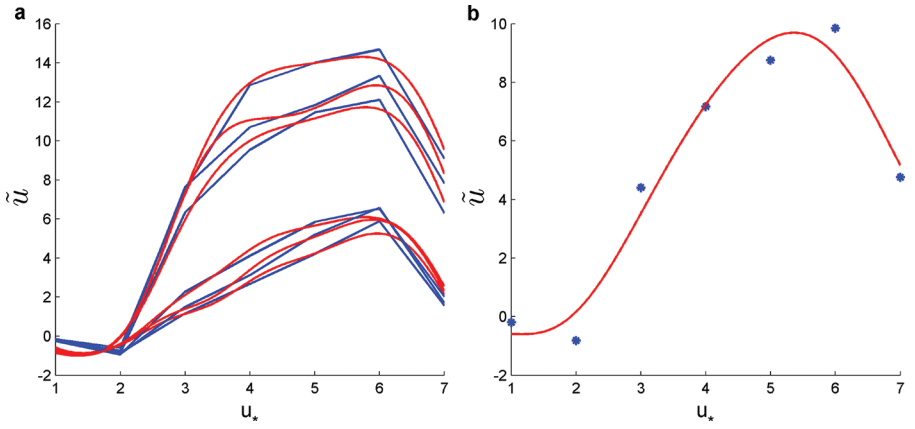


Figure 5. Static model of the input nonlinearity. The NARMAX dynamic model input $\tilde{u}(t)$ was obtained from transforming $u_*(t) = \log_{10}(gu(t))$ through a function $G(u_*(t))$, where $u(t)$ was the level of $3OC_6HSL$. (a) Separate estimates of the static function $G(.)$ (red) across 6 experimental data sets (blue) used for identification purpose. (b) Single estimate of the static function $G(.)$ (red) using 6 experimental data sets compared to the average of the experimental data curves in panel a (blue dots).

dx.doi.org/10.1021/sb300009t | *ACS Synth. Biol.* 2012, 1, 375–384

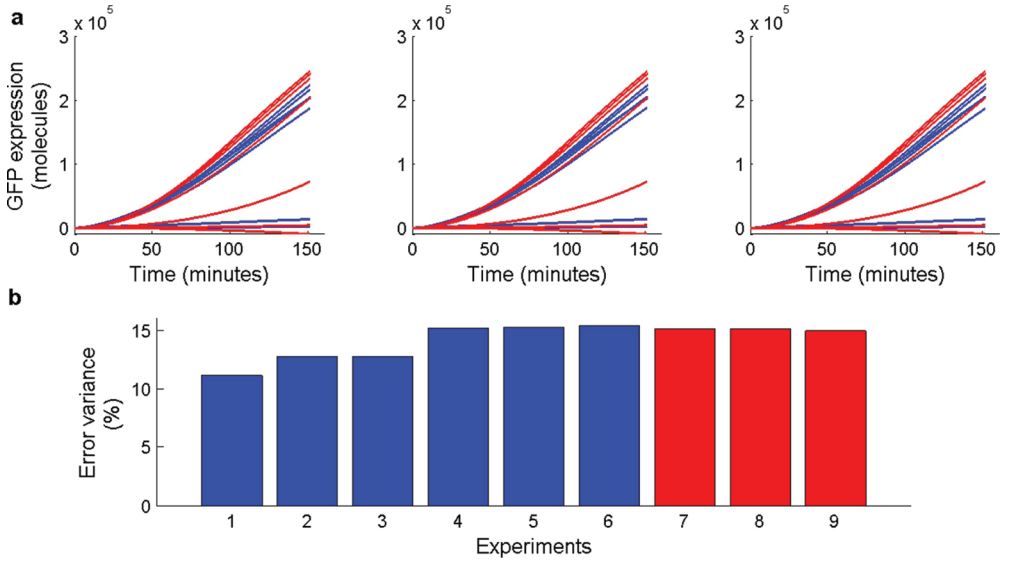


Figure 6. NARMAX model prediction on validation data. (a) A single NARMAX model with averaged parameter estimates was simulated (red) and compared to a reserved set of validation data (blue). (b) The percentage prediction error variance from the averaged NARMAX model using both estimation (blue) and validation (red) data sets.

scheme (ERS) model (derivations are provided in the Supporting Information):[31,56]

$$\dot{s}(t) = a_1 s^n(t) + a_2 \dot{p}(t) + a_3 s^n(t)\dot{p}(t) \quad (1)$$

$$\ddot{p}(t) = b_1 s^n(t) + b_2 \dot{p}(t) + b_3 s^n(t)\dot{p}(t) \quad (2)$$

where $a_1 = -nk_1e_0(t)$, $a_2 = nk_{-1}/k_2$, $a_3 = nk_1/k_2$, $b_1 = k_2k_1e_0(t)$, $b_2 = -k_1 - k_2$, $b_3 = -k_1$, and $e_0(t)$ denotes total enzyme concentration. Assuming total enzyme concentration and the rate of change of enzyme−substrate complex to be constant and equal to zero respectively, we can derive the more often-used Hill equation (also equivalent to the Michaelis−Menten equation for $n = 1$), which is a simplification of the full dynamic form described in eqs 1 and 2:[31]

$$\dot{p}(t) = \frac{V_m s_{ss}^n}{s_{ss}^n + k_p^n} \quad (3)$$

where s_{ss} is the substrate in steady-state, $k_p^n = (k_{-1} + k_2)/k_1$ and $V_m = k_2e_0(t)$.

Note that the Hill equation should provide a description of the reaction scheme that is consistent with the dynamic form presented above in eqs 1 and 2 for substrate signals that are near constant. Also note that in the case of steady-state substrate the LHS of eq 1 is by definition equal to zero, $\dot{s}(t) = 0$. Hence, any dynamic behavior in the system must be predominantly described by eq 2.

Parameter Estimation. Parameters of the ERS model were estimated using the available prior information: (i) initial $3OC_6HSL$ level, $s(t = 0)$, and (ii) GFP expression over time, $p(t)$. A separable least-squares (SLS) algorithm[57,58] was used to estimate the model parameters, where the main feature of the approach was to separate the parameters into linear and nonlinear sets, $\zeta_l = (b_1,b_2,b_3)$ and $\zeta_n = (a_1,a_2,a_3,n)$ respectively. The advantage of the SLS algorithm is that it typically converges in fewer iterations, has improved numerical conditioning and requires initialization of fewer parameters in comparison to the full nonlinear optimization problem.[57]

We define the optimization cost function for M experimental signals corresponding to different input levels of $3OC_6HSL$, with N samples per signal as

$$J = \frac{1}{MN}\|P - \Gamma(\zeta_n)\zeta_l\|_2^2 \quad (4)$$

where $P = [\mathbf{p}_1^T, ..., \mathbf{p}_M^T]^T$, $\mathbf{p}_j = [\ddot{p}_j(1), ..., \ddot{p}(N)]^T$, $\Gamma(\zeta_n) = [\gamma_1(\zeta_n)^T, ..., \gamma_M(\zeta_n)^T]^T$ and

$$\gamma_j(\zeta_n) = \begin{bmatrix} \hat{s}_j^n(1) & \dot{p}_j(1) & \hat{s}_j^n(1)\dot{p}_j(1) \\ \vdots & \vdots & \vdots \\ \hat{s}_j^n(N) & \dot{p}_j(N) & \hat{s}_j^n(N)\dot{p}_j(N) \end{bmatrix} \quad (5)$$

and $\hat{s}$ is obtained from simulation of eq 1, given the initial condition of the substrate and the nonlinear parameters ζ_n. The linear parameters ζ_l can be expressed in terms of the nonlinear parameters ζ_n using the least-squares solution

$$z_l = G(z_n)^\dagger P \quad (6)$$

where † denotes the pseudo-inverse, $\Gamma(\zeta_n)^\dagger = (\Gamma(\zeta_n)^T\Gamma(\zeta_n))^{-1}\Gamma(\zeta_n)^T$. Substituting eq 6 into eq 4 leads to the reduced optimization problem, from which the linear parameters have been eliminated:

$$\min_{\zeta_n} \frac{1}{MN}\|P - \Gamma(\zeta_n)\Gamma(\zeta_n)^\dagger P\|_2^2 \quad (7)$$

The nonlinear optimization technique used in this case was a quasi-Newton method (implemented using the Matlab function fminunc). The nonlinear parameters were initialized using a grid search, where the parameter ranges were $a_1, a_2, a_3 \in [-10, -9.9, ..., 10]$ and $n \in [0, 0.5, ..., 6]$. In implementation, the ERS model was simulated using a first order Euler approximation for computational simplicity (we verified on a subset of the data that use of higher order numerical integration methods did not alter the results).

NARMAX Modeling Framework. *NARMAX Model Representation.* In general, the NARMAX model is obtained

in a data-driven framework from samples of input-output signals. In this investigation we defined the input to the system, $u(t) \in \mathbb{R}$, as $3OC_6HSL$ concentration and the output, $y(t) \in \mathbb{R}$, as rate of change of GFP expression (due to its stable nature). For predicting GFP using the identified model, we numerically integrated the prediction from the NARMAX model. The structure of a general continuous-time NARMAX model can be defined by[50]

$$y^{(i)}(t) = F(\mathbf{z}(t)) \tag{8}$$

$$\mathbf{z}(t) = (u(t), \dot{u}(t), ..., u^{(i-1)}(t), y(t), \dot{y}(t), ..., y^{(i-1)}(t)) \tag{9}$$

where i is the differential order, $F(\mathbf{z}(t))$ is some unknown nonlinear function and $\mathbf{z}(t) \in \mathbb{R}^{2i}$ is the model input vector of system input-output derivatives. We describe the function $F(.)$ using a basis function decomposition:

$$y^{(i)}(t) = \sum_{j=1}^{L} \theta_j \phi_j(\mathbf{z}(t)) \tag{10}$$

where $\phi_j(.)$ is a basis function with associated parameter $\theta_j \in \mathbb{R}$. In this investigation we used polynomial basis functions of maximum order $q = 3$ and assumed second order system dynamics, $i = 2$.

The general form of the NARMAX model outlined above was specialized for this investigation by only considering derivatives in the output signal and no cross-product terms between input and output. This was due to the assumption that the input level of $3OC_6HSL$ was constant over the duration of each experiment, so input derivatives were zero and cross-product terms were unidentifiable (more details are given in the Supporting Information). In addition, we noted that there appeared to be a nonlinear gain variation associated with different input levels of $3OC_6HSL$, which we described using separate input gain terms k_j, for $j = 1, ..., M$, resulting in the following modification of the NARMAX model:

$$y_j^{(i)}(t) = F(y(t), \dot{y}(t), ..., y^{(i-1)}(t)) + k_j u_j(t) \tag{11}$$

for $j = 1, ..., M$ experimental signals corresponding to different constant input levels of $3OC_6HSL$.

NARMAX Model with Static Input Nonlinearity. The static nonlinear gain variation across input levels was modeled here using a function $G(.)$, which mapped the $3OC_6HSL$ input $u(t)$ to the dynamic model input $\tilde{u}(t)$, and the NARMAX model was correspondingly modified so that

$$y^{(i)}(t) = F(y(t), \dot{y}(t), ..., y^{(i-1)}(t)) + \tilde{u}(t) \tag{12}$$

where $\tilde{u}(t) = G(u_*(t))$, $u_*(t) = \log_{10}(gu(t))$ (g is a scaling parameter discussed below). The log transformation was applied to the scaled input $gu(t)$ because of the log spacing in levels of $3OC_6HSL$. The function $G(.)$ was described by the basis function decomposition

$$\tilde{u}(t) = \sum_{j=1}^{B} w_j \psi_j(u_*(t)) \tag{13}$$

where $w_j \in \mathbb{R}$ is the jth basis function parameter, B is the number of basis functions, and in this investigation we used radial basis functions, specifically the squared exponential function

$$\psi_j(u_*(t)) = \exp\left(-\frac{1}{2\sigma_j^2} \|u_*(t) - \mu_j\|_2^2\right) \tag{14}$$

where μ_j and σ_j are the respective centers and widths of the jth basis function. Basis functions were centered on the levels of the input data values $u_*(t)$ and the corresponding width parameters were heuristically tuned in the range $\sigma_j \in [1,1.5]\ \forall\ j$.

Parameter Estimation and Structure Detection. An advantage of the basis function decomposition of the NARMAX model defined in eq 10 is that it is linear-in-the-parameters, hence we can use least-squares for parameter estimation; first we define the regression equation:

$$Y^{(i)} = \Phi\theta + \varepsilon \tag{15}$$

where $Y^{(i)} = (\mathbf{y}_1^{(i)}, ..., \mathbf{y}_M^{(i)})^T$ is the model output vector of differential order i, ε is the model residual error vector, $\theta = (k_1, ..., k_M, c_1, ..., c_{L-1})^T$, is the parameter vector, and $\Phi = [U\ Y]$ is the regression matrix where

$$U = \begin{bmatrix} \mathbf{u}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{u}_M \end{bmatrix} \tag{16}$$

$$Y = \begin{bmatrix} \left(\mathbf{y}_1^{(0)}\right)^1 & \left(\mathbf{y}_1^{(1)}\right)^1 & ... & \left(\mathbf{y}_1^{(i-1)}\right)^q \\ \vdots & \vdots & & \vdots \\ \left(\mathbf{y}_M^{(0)}\right)^1 & \left(\mathbf{y}_M^{(1)}\right)^1 & ... & \left(\mathbf{y}_M^{(i-1)}\right)^q \end{bmatrix} \tag{17}$$

where $\mathbf{u}_j = (gu_{j,1}, ..., gu_{j,N})^T$, and $\mathbf{y}_j^{(i)} = (y_{j,1}^{(i)}, ..., y_{j,N}^{(i)})$ for $j = 1, ..., M$. The least-squares estimate of the parameters is

$$\hat{\theta} = \Phi^{\dagger} Y^{(i)} \tag{18}$$

where $\Phi^{\dagger} = (\Phi^T\Phi)^{-1}\Phi^T$. In order to improve the numerical conditioning of the regression matrix Φ, we rescaled the input levels using a gain g, where $g = 1 \times 10^{11}$.

The Y matrix, defined above, was setup to contain a superset of model terms composed of polynomial transformations of $y(t)$ and its derivatives. As part of the identification procedure a parsimonious model structure was detected composed of a reduced set of those terms. In this investigation the number of terms was relatively small (9 terms) and so the structure was detected by an exhaustive search of all possible term combinations ($2^9 = 512$). In order to compare models we used information criteria (IC) to obtain the optimal trade-off between model accuracy and model complexity - Akaike's and the Bayesian IC[25] (details can be found in the Supporting Information). In implementation, the NARMAX model was simulated using a first order Euler approximation.

The basis function parameters for the static nonlinear input function $G(.)$ were estimated using least-squares from the target data $\tilde{\mathbf{u}} = (k_1u_{1,0}, ..., k_Mu_{M,0})^T$, where $u_{j,0}$, for $j = 1, ..., M$, corresponded to the rescaled input levels of $3OC_6HSL$.

Signal Derivative Estimation. A problem that hampers continuous-time modeling is obtaining signal derivatives. The observed signals are typically corrupted by high frequency measurement noise: approximating the derivatives from directly

differencing the observed signal amplifies this noise. To overcome this problem, a method for derivative estimation was used here based on a Taylor-series expansion of the signal in conjunction with the Kalman smoothing technique[59] (see Supporting Information for more details).

The GFP signal, which we denote as $\tilde{y}(t)$ (where the NARMAX modeling signal $y(t) = \dot{\tilde{y}}(t)$, and the GFP derivatives can be represented in the discrete-time state-space model,

$$\mathbf{x}_{k+1} = A\mathbf{x}_k + \mathbf{w}_k \tag{19}$$

$$\tilde{y}_k = C\mathbf{x}_k + v_k \tag{20}$$

where $C = (1, 0, ..., 0)$ is the measurement matrix, $\mathbf{x}_k = (\tilde{y}_k, \dot{\tilde{y}}_k, ..., \tilde{y}_k^{(D)}) \in \mathbb{R}^{n_x}$ is the state vector at sample time k that contains the vector of GFP expression and its derivatives up to order D, $n_x = D + 1$, $\mathbf{w}_k \sim N(0,Q)$ and $\nu_k \sim N(0,R)$ are independent zero mean Gaussian white noise signals, and the state transition matrix A is described using rows that are based on the Taylor series expansion of the signal:

$$A = \begin{bmatrix} 1 & T & \frac{T^2}{2!} & \dots & \frac{T^D}{D!} \\ 0 & 1 & T & \dots & \frac{T^{D-1}}{(D-1)!} \\ \vdots & & & & \vdots \\ 0 & 0 & & \dots & 1 \end{bmatrix} \tag{21}$$

where T is the time step, which was set to a unit sample step, $T = 1$. The elements of the state noise covariance matrix Q are given by

$$q_{ij} = \frac{\sigma_w^2 T^{2D+3-(i+j)}}{(D+1-i)!(D+1-j)!(2D+3-(i+j))} \tag{22}$$

where σ_w is a tuning parameter describing the power of the state noise. In order to obtain the derivatives in the state vector, a Kalman smoother was implemented using the Rauch−Tung−Streibel recursions (further details are in the Supporting Information).[60] The GFP derivatives estimated using this algorithm were used in both the enzymatic reaction scheme modeling and the NARMAX identification.

■ ASSOCIATED CONTENT

Supporting Information

Discussion of descriptive derivations and implementation of ERS model, NARMAX model, and signal derivative estimation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: k.krishnanathan@sheffield.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge that this work was supported by the Engineering and Physical Sciences Research Council (EPSRC) UK, and a European Research Council Advanced Investigator Award. We also thank D. Endy for making the data freely available online for usage.

■ ABBREVIATIONS

NARMAX, Nonlinear AutoRegressive Moving Average model with eXogenous inputs; ERS model, enzymatic reaction scheme model

■ REFERENCES

(1) Purnick, P. E. M., and Weiss, R. (2009) The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol. 10*, 410−422.

(2) Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999) From molecular to modular cell.biology. *Nature 402*, 47.

(3) Szybalski, W. (1974) In vivo and in vitro initiation of transcription. *Adv. Exp. Med. Biol. 44*, 23−24.

(4) Szybalski, W., and Skalka, A. (1978) Nobel prizes and restriction enzymes. *Gene 4*, 181−182.

(5) Elowitz, M. B., and Leibler, S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature 403*, 335−338.

(6) Gardner, T. S., Cantor, C. R., and Collins, J. J. (2000) Construction of a genetic toggle switch in Escherichia coli. *Nature 403*, 339−342.

(7) Levskaya, A., Chevalier, A. A., Tabor, J. J., Simpson, Z. B., Lavery, L. A., Levy, M., Davidson, E. A., Scouras, A., Ellington, A. D., Marcotte, E. M., and Voigt, C. A. (2005) Synthetic biology: Engineering Escherichia coli to see light. *Nature 438*, 441−442.

(8) Ro, D. K., Paradise, E. M., Ouellet, M., Fisher, K. J., Newman, K. L., Ndungu, J. M., Ho, K. A., Eachus, R. A., Ham, T. S., Kirby, J., Chang, M., Withers, S., Shiba, Y., Sarpong, R., and Keasling, J. D. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature 440*, 940−943.

(9) Shiba, Y., Paradise, E. M., Kirby, J., Ro, D. K., and Keasling, J. D. (2007) Engineering of the pyruvate dehydrogenase bypass in Saccharomyces cerevisiae for high-level production of isoprenoids. *Metab. Eng. 9*, 160−168.

(10) Dueber, J. E., Wu, G. C., Malmirchegini, G. R., Moon, T. S., Petzold, C. J., Ullal, A. V., Prather, K. L., and Keasling, J. D. (2009) Synthetic protein scaffolds provide modular control over metabolic flux. *Nat. Biotechnol. 27*, 753−759.

(11) Kobayashi, H., Kaern, M., Araki, M., Chung, K., Gardner, T. S., Cantor, C. R., and Collins, J. J. (2004) Programmable cells: Interfacing natural and engineered gene networks. *Proc. Natl. Acad. Sci. U.S.A. 101*, 8414−8419.

(12) Khalil, A. S., and Collins, J. J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Genet. 11*, 367−379.

(13) Lu, T. K.; Collins, J. J. Engineered bacteriophage targeting gene networks as adjuvants for antibiotic therapy. *Proc. Natl. Acad. Sci.* U.S.A.2009,

(14) Weber, W., Schoenmakers, R., Keller, B., Gitzinger, M., Grau, T., Daoud-El Baba, M., Sander, P., and Fussenegger, M. (2008) A synthetic mammalian gene circuit reveals antituberculosis compounds. *Proc. Natl. Acad. Sci. U.S.A.. 105*, 9994−9998.

(15) Stocker, J., Balluch, D., Gsell, M., Harms, H., Feliciano, J., Daunert, S., Malik, K. A., and van der Meer, J. R. (2003) Development of a set of simple bacterial biosensors for quantitative and rapid measurements of arsenite and arsenate in potable Water. *Environ. Sci. Technol. 37*, 4743−4750.

(16) van der Meer, J. R., and Belkin, S. (2010) Where microbiology meets microengineering: design and applications of reporter bacteria. *Nat. Rev. Microbiol. 8*, 511−522.

(17) Alper, H., and Stephanopoulos, G. (2009) Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? *Nat. Rev. Microbiol. 7*, 715−723.

(18) Atsumi, S., Hanai, T., and Liao, J. C. (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature 451*, 86−89.

(19) Keasling, J. D., and Chou, H. (2008) Metabolic engineering delivers next-generation biofuels. *Nat. Biotechnol. 26*, 298−299.
(20) Kwok, R. (2010) Five hard truths for synthetic biology. *Nature 463*, 288−290.
(21) Arkin, A. (2008) Setting the standard in synthetic biology. *Nat. Biotechnol. 26*, 771−773.
(22) Endy, D. (2005) Foundations for engineering biology. *Nature 438*, 449−453.
(23) Andrianantoandro, E.; Basu, S.; Karig, D. K.; Weiss, R. Synthetic biology: New engineering rules for an emerging discipline. *Mol. Syst. Biol.* 2006, *2*.
(24) Canton, B., Labno, A., and Endy, D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol. 26*, 787−793.
(25) Ljung, L. (1999) *System Identification: Theory for the User*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.
(26) English, B. P., Min, W., Van Oijen, A. M., Lee, K. T., Luo, G., Sun, H., Cherayil, B. J., Kou, S. C., and Xie, X. S. (2006) Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Biol. 2*, 87−94.
(27) Gertz, J., Siggia, E. D., and Cohen, B. A. (2008) Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature 457*, 215−218.
(28) Kim, J.; White, K. S.; Winfree, E. Construction of an in vitro bistable circuit from synthetic transcriptional switches. *Mol. Syst. Biol.* 2006, *2*.
(29) Houston, J. B., and Kenworthy, K. E. (2000) In vitro-in vivo scaling of CYP kinetic data not consistent with the classical Michaelis-Menten model. *Drug Metab. Dispos. 28*, 246.
(30) Holmberg, A. (1982) On the practical identifiability of microbial growth models incorporating Michaelis-Menten type nonlinearities. *Math. Biosci. 62*, 23−43.
(31) Cornish-Bowden, A. (2004) *Fundamentals of Enzyme Kinetics*, Portland Press, Portland.
(32) Covert, M. W., Xiao, N., Chen, T. J., and Karr, J. R. (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. *Bioinformatics 24*, 2044−2050.
(33) Tian, T., and Burrage, K. (2006) Stochastic models for regulatory networks of the genetic toggle switch. *Proc. Natl. Acad. Sci. U.S.A. 103*, 8372−8377.
(34) Karlebach, G., and Shamir, R. (2008) Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol. 9*, 770−780.
(35) Wilkinson, D. J. (2009) Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat. Rev. Genet. 10*, 122−133.
(36) Leontaritis, I. J., and Billings, S. A. (1985) Input-output parametric models for non-linear systems. Part 1: Deterministic non-linear systems. *Int. J. Control 41*, 303−328.
(37) Leontaritis, I. J., and Billings, S. A. (1985) Input-output parametric models for nonlinear systems. Part 2: Stochastic nonlinear systems. *Int. J. Control 41*, 329−344.
(38) Chen, S., Billings, S. A., and Luo, W. (1989) Orthogonal least squares methods and their application to non-linear system identification. *Int. J. Control 50*, 1873−1896.
(39) Anderson, S. R., and Kadirkamanathan, V. (2007) Modelling and identification of non-linear deterministic systems in the delta-domain. *Automatica 43*, 1859−1868.
(40) Anderson, S. R., Lepora, N. F., Porrill, J., and Dean, P. (2010) Nonlinear dynamic modeling of isometric force production in primate eye muscle. *IEEE Trans. Biomed. Eng. 57*, 1554−1567.
(41) Kukreja, S. L., Galiana, H. L., and Kearney, R. E. (2003) NARMAX representation and identification of ankle dynamics. *IEEE Trans. Biomed. Eng. 50*, 70−81.
(42) Kerschen, G., Worden, K., Vakakis, A., and Golinval, J. (2006) Past, present and future of nonlinear system identification in structural dynamics. *Mech. Syst. Signal Process. 20*, 505−592.
(43) Jang, H. K., and Kim, K. J. (1994) Identification of loudspeaker nonlinearities using the NARMAX modeling technique. *J. Audio Eng. Soc. 42*, 50−60.
(44) Basso, M., Giarre, L., Groppi, S., and Zappa, G. (2005) NARX models of an industrial power plant gas turbine. *IEEE Trans. Control Syst. Technol. 13*, 599−604.
(45) Chiras, N., Evans, C., and Rees, D. (2001) Nonlinear gas turbine modeling using NARMAX structures. *IEEE Trans. Instrum. Meas. 50*, 893−898.
(46) Soofi, A. S.; Cao, L. (2002) *Modelling and Forecasting Financial Data: Techniques of Nonlinear Dynamics*, Springer, Amsterdam.
(47) Chen, S., and Billings, S. A. (1989) Representations of non-linear systems: the NARMAX model. *Int. J. Control 49*, 1013−1032.
(48) Nelles, O. (2001) *Nonlinear System Identification*, Springer, New York.
(49) Fabri, S. G.; Kadirkamanathan, V. (2001) *Functional Adaptive Control: An Intelligent Systems Approach*, Springer, New York.
(50) Coca, D., and Billings, S. A. (1999) A direct approach to identification of nonlinear differential models from discrete data. *Mech. Syst. Signal Process. 13*, 739−755.
(51) Wang, B., Kitney, R. I., Joly, N., and Buck, M. (2011) Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nat. Commun. 2*, 508.
(52) Tamsir, A., Tabor, J. J., and Voigt, C. A. (2011) Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. *Nature 469*, 212−215.
(53) Canton, B. Engineering the interface between cellular chassis and synthetic biological systems. Ph.D. thesis, Massachusetts Institute of Technology, 2008.
(54) Wei, H., Billings, S. A., and Lui, J. (2004) Term and variable selection for nonlinear models. *Int. J. Control 77*, 86−110.
(55) Baldacchino, T., Anderson, S. R., and Kadirkamanathan, V. (2012) Structure detection and parameter estimation for NARX models in a unified EM framework. *Automatica*, DOI: 10.1016/j.automatica.2012.02.021.
(56) Palsson, B. O. (1987) On the dynamics of the irreversible Michaelis-Menten reaction mechanism. *Chem Eng. Sci. 42*, 447−458.
(57) Golub, G., and Pereyra, V. (2003) Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems 19*, R1.
(58) Bruls, J., Chou, C. T., Haverkamp, B. R. J., and Verhaegen, M. (1999) Linear and non-linear system identification using separable least-squares. *Eur. J. Control 5*, 116−128.
(59) Fioretti, S., and Jetto, L. (1989) Accurate derivative estimation from noisy data: a state-space approach. *Int. J. Syst. Sci. 20*, 33−53.
(60) Kailath, T.; Sayed, A. H.; Hassibi, B. (2000) *Linear Estimation*, Prentice Hall, New York.