Published in final edited form as:

J Theor Biol. 2012 May 7; 300: 232–241. doi:10.1016/j.jtbi.2012.01.037.

Identifying a static nonlinear structure in a biological system using noisy, sparse data

Joshua R. Porter^a, John S. Burg^{b,1}, Peter J. Espenshade^b, and Pablo A. Iglesias^{a,*}

Joshua R. Porter: josh.porter@jhu.edu; John S. Burg: jsburg@stanford.edu; Peter J. Espenshade: peter.espenshade@jhmi.edu

^aDepartment of Electrical & Computer Engineering, Johns Hopkins University, 105 Barton Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA

^bDepartment of Cell Biology, Johns Hopkins University School of Medicine, Physiology 107B, 725 N. Wolfe Street, Baltimore, MD 21205, USA

Abstract

When part of a biological system cannot be investigated directly by experimentation, we face the problem of structure identification: how can we construct a model for an unknown part of a mostly-known system using measurements gathered from its input and output? This problem is especially difficult to solve when the measurements available are noisy and sparse, i.e. widely and unevenly spaced in time, as is common when measuring biological quantities at the cellular level. Here we present a procedure to identify a static nonlinearity embedded between two dynamical systems using noisy, sparse measurements. To reduce the level of error caused by measurement noise, we introduce the concept of weighted-sum predictability. If we make the input and output subsystems weighted-sum predictable and normalize the measurements to their weighted sum, we achieve better noise reduction than through normalizing to a loading control. We then interpolate the normalized measurements to obtain continuous input and output signals, with which we solve directly for the input-output characteristics of the unknown static nonlinearity. We demonstrate the effectiveness of this structure identification procedure by applying it to identify a model for ergosterol sensing by the proteins Sre1 and Scp1 in fission yeast. Simulations with this model produced outputs consistent with experimental observations. The techniques introduced here will provide researchers with a new tool by which biological systems can be identified and characterized.

Keywords

Signaling cascade; System identification; Structure identification; Modeling

© 2012 Elsevier Ltd. All rights reserved.

*Corresponding author. Tel.: +1 410 5166026; fax: +1 410 5165566, pi@jhu.edu (Pablo A. Iglesias).

¹Present address: Department of Molecular & Cellular Physiology, Stanford University School of Medicine, Beckman Center B177, 279 Campus Drive, Stanford, CA 94305, USA.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Models of biological systems are commonly constructed by investigating the parts of a system experimentally, then synthesizing knowledge of the parts into a model of the whole system (Leloup and Goldbeter, 2003; Chen et al., 2004; El-Samad et al., 2005). This technique fails, however, when the limitations of existing experimental methods prevent direct investigation of part of a system. In that case, a researcher might resort to an alternative approach: given partial knowledge of a system and experimental measurements from the system input and output, can a phenomenological model (structure and parameters) for the unknown part be identified? If successful, this structure identification could lead to a model-driven understanding of that part that precedes its experimental understanding and even informs the design of future experiments.

This problem of structure identification has been well studied (Sjöberg et al., 1995), and structure identification techniques developed for man-made systems have been successfully applied to many biological systems. Notable recent examples include cellular systems whose structures have been identified by frequency response analysis (Ang et al., 2011; Mettetal et al., 2008). In other biological systems, however, structure identification is hampered by the experimental data available. Biological measurements are often noisier than measurements from electrical or mechanical systems because of measurement techniques that are not as well developed. Furthermore, because of technical limitations and the labor involved in collecting biological data, such data are often sparse: the inputs and outputs are often sampled only a few times during an experiment, and only a few experiments may be available to identify a model for the unknown part. Standard temporal filtering techniques cannot be used to attenuate noise in a sparsely sampled signal, making the combination of noise and sparsity especially difficult to work with.

In this context, we focus on one class of systems: those in which a static nonlinearity connects two nonlinear dynamical systems (Figure 1). These systems have been used to model a number of physiological and neurological processes (Kearney and Hunter, 1983; Emerson et al., 1992; Su et al., 2007, 2010). They are especially well suited to model a step in a biological signaling cascade, in which a change in a dynamical system upstream is transduced nonlinearly into a response in the dynamical system downstream. Here, we address the case in which the two dynamical systems are known and the static nonlinearity is to be identified. Our goal is not simply to construct a “working” model for the nonlinearity but to gain insight into the biological processes that create it. Accordingly, we want to probe the response of the nonlinearity over a wide range of inputs, making frequency response analysis an inappropriate approach since it requires linearizing the system about an operating point. The data available to identify the nonlinearity are noisy, sparse measurements gathered from the two known dynamical systems during several experiments.

To effectively perform structure identification with these limited measurements, we propose a procedure based on a combination of experimental and modeling techniques. In section 2, we formally describe this procedure: the type of system and measurement data under consideration, the requirements and steps of the procedure, and the reason why it works. We demonstrate the usefulness of this procedure in section 3 by applying it to a real biological problem with noisy, sparse data. Finally, in section 4, we evaluate the procedure by showing that the model constructed in section 3, when given experimental inputs, produces simulated outputs consistent with experimental outputs.

A preliminary version of this work was presented at the 2011 Conference on Information Sciences and Systems (Porter et al., 2011).

2. Method Overview

We consider the biological system in Figure 1, defined by the following set of ordinary differential equations:

$$N_1: \begin{cases} \dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1, \mathbf{u}_1); \mathbf{x}_1(t_0) = \mathbf{x}_{1,0} \\ y_1 = g_1(\mathbf{x}_1) \\ \mathbf{m}_1 = [m_{1,1} \dots m_{1,p_1}]^T = \mathbf{h}_1(\mathbf{x}_1) \end{cases} \quad (1)$$

$$u_2 = f_{\text{NL}}(y_1) \quad (2)$$

$$N_2: \begin{cases} \dot{\mathbf{x}}_2 = \mathbf{f}_2(\mathbf{x}_2, u_2); \mathbf{x}_2(t_0) = \mathbf{x}_{2,0} \\ \mathbf{y}_2 = \mathbf{g}_2(\mathbf{x}_2) \\ \mathbf{m}_2 = [m_{2,1} \dots m_{2,p_2}]^T = \mathbf{h}_2(\mathbf{x}_2) \end{cases} \quad (3)$$

This system consists of a known “input subsystem” N_1 and a known “output subsystem” N_2 connected by an unknown static nonlinearity f_{NL} . N_1 has physiological inputs $\mathbf{u}_1$, which may include signals entering the system from upstream in a signaling cascade or experimental perturbations to N_1 . For our purposes, N_1 has a single output y_1 , the input to f_{NL} . The output of f_{NL} , u_2 , serves as the input to N_2 , whose physiological outputs $\mathbf{y}_2$ can be passed downstream in a cascade. Each subsystem N_i has a set of p_i measurable quantities $\mathbf{m}_i$, which are limited by the experimental methods available and need not directly correspond to the physiological quantities of interest. We note that the structure of this system resembles that of cascade-type systems such as Wiener, Hammerstein, and linear-nonlinear-linear (LNL) systems (Korenberg and Hunter, 1986; Hunter and Korenberg, 1986). Numerous examples of biological models taking this form can be found in the literature (Kearney and Hunter, 1983; Emerson et al., 1992; Su et al., 2007, 2010).

We want to construct a model for f_{NL} using samples from $\mathbf{m}_1$ and $\mathbf{m}_2$ measured during experiments $k = 1, \dots, q$ at sampling times t_ℓ , $\ell = 1, \dots, s_k$. For clarity, the notation in Table 1 will be used throughout this paper to refer to specific measurements. The biological signals $\mathbf{m}_1$ and $\mathbf{m}_2$ are noisy and sparsely sampled, creating a structure identification challenge distinct from that of previously studied cascade-type systems.

2.1. Requirements

For this procedure, we require that the experiments used to gather data, the methods of measuring $\mathbf{m}_1$ and $\mathbf{m}_2$, and the subsystems N_1 and N_2 meet several conditions:

- Req. 1. *f_{NL} domain sampling*: The set of experiments used to identify f_{NL} must adequately sample its domain – that is, they must include experimental treatments $\mathbf{u}_1(t)$ that generate signals $y_1(t)$ covering the physiologically relevant range of inputs to f_{NL} .
- Req. 2. *Co-measurement*: $\mathbf{m}_1$ must consist of at least two measurable quantities, and at each sampling time, these quantities must be measured together by the same instrument. The same must be true for $\mathbf{m}_2$, although the instrument need not be the same as that measuring $\mathbf{m}_1$.
- Req. 3. *Weighted-sum predictability*: A weighted sum of all measurable quantities of $\mathbf{m}_1$ must be a known function of time independent of the individual quantities. The same must be true for $\mathbf{m}_2$. Mathematically, for $i = \{1, 2\}$, there must exist a known constant weighting vector $\mathbf{c}_i > 0$ and a known function of time $K_i(t)$ such that

$$\mathbf{c}_i^T \mathbf{m}_i(t) = K_i(t), \quad \forall t \geq 0. \quad (4)$$

One way to satisfy this condition is when N_i describes the dynamics of a substance X that is converted between several forms, each of which is measured by $\mathbf{m}_i$. The rate of input of X into N_i , $v_{\text{in}}^X(t)$, must be known independently of $\mathbf{m}_i$ for the duration of each experiment, and the rate constant for removal of X from N_i , k_{rem}^X , must be the same for all forms of X. Let $\mathbf{c}_i^T \mathbf{m}_i(t) = K_i(t)$ be the total amount of X in N_i at time t . Then

$$\frac{dK_i(t)}{dt} = v_{\text{in}}^X(t) - k_{\text{rem}}^X K_i(t) \quad (5)$$

and $K_i(t)$ can be found for the experiment given an initial condition $K_i(0)$. In this case, by the definition of $K_i(t)$, the weighting vector $\mathbf{c}_i$ is chosen to represent the amount of X in each of its forms. For example, if $m_{i,1}$ is molecules of X monomer and $m_{i,2}$ is molecules of X dimer, then $\mathbf{c}_i = [1 \ 2]^T$ is a logical choice.

Of the requirements listed here, this one may be the most restrictive, but several common types of biological systems satisfy it or can be modified slightly to satisfy it. For example, a metabolic pathway in which metabolites are serially converted from one form to another can satisfy this requirement in the way described above, as can a protein that takes multiple measurable forms. Section 3 of this paper presents examples of biological systems that satisfy this requirement.

Req. 4. *Computing continuous $\mathbf{m}_1 \rightarrow y_1$* : From the known structure of N_1 , we must be able to derive a system N_1' such that given a vector of continuous measurement signals $\mathbf{m}_1(t)$, we can use N_1' to compute the continuous signal $y_1(t)$ being output by N_1 concurrent with $\mathbf{m}_1(t)$.

Req. 5. *Computing continuous $\mathbf{m}_2 \rightarrow u_2$* : From the known structure of N_2 , we must be able to derive a system N_2^{-1} such that given a vector of continuous measurement signals $\mathbf{m}_2(t)$, we can use N_2^{-1} to compute the continuous signal $u_2(t)$ that generated $\mathbf{m}_2(t)$ via N_2 .

2.2. Procedure

With these requirements satisfied, we identify a model for f_{NL} as follows:

Step 1. *Normalize the measurements*: For each experiment at each sampling time t_ℓ , we normalize the samples of $\mathbf{m}_1$ and $\mathbf{m}_2$ to their weighted sums described by req. 3. We calculate $\bar{m}_{i,j}$, the j th normalized measurement of $\mathbf{m}_i$, as follows:

$$\bar{m}_{i,j}(t_\ell) = \frac{m_{i,j}(t_\ell)}{\mathbf{c}_i^T \mathbf{m}_i(t_\ell)} K_i(t_\ell) \quad (6)$$

By construction, $\bar{\mathbf{m}}_i$ has the same units as $\mathbf{m}_i$.

Step 2. *Interpolate the normalized measurements*: For each experiment, we generate continuous signals $\bar{\mathbf{m}}_1(t)$ and $\bar{\mathbf{m}}_2(t)$ over the duration of the experiment by interpolating the normalized measurements $\bar{\mathbf{m}}_1(t_\ell)$ and $\bar{\mathbf{m}}_2(t_\ell)$ with piecewise cubic Hermite interpolating polynomials, described in section 2.3.2.

Step 3. *Solve for f_{NL}* : For each experiment, we use the system N'_1 specified by req. 4 to compute $y_1(t)$ from $\mathbf{m}_1(t)$. Similarly, we use the system N_2^{-1} specified by req. 5 to compute $u_2(t)$ from $\mathbf{m}_2(t)$. Thus, we estimate the input and output of f_{NL} over the duration of each experiment.

Step 4. *Inspect the result*: For each experiment, we plot $u_2(t)$ versus $y_1(t)$ over the duration of the experiment, i.e. the experiment's putative trajectory through (y_1, u_2) -space. We overlay these trajectories from several experiments to produce an input-output plot of f_{NL} , which we can then characterize further by inspection and curve-fitting.

2.3. Justification

2.3.1. Normalizing—Normalizing the samples of $\mathbf{m}_1$ and $\mathbf{m}_2$ to their weighted sums incorporates the property of weighted-sum predictability (req. 3) into the data used to identify f_{NL} . Here we show that in most cases, weighted-sum normalizing reduces the percent error in the samples due to measurement noise. We also compare this strategy with that of normalizing to a loading control, a common technique in biological experimentation.

For this analysis we consider two types of measurement noise: systemic measurement noise, which equally affects all measurable quantities of $\mathbf{m}_i$, and component measurement noise, which affects each measurable quantity $m_{i,j}$ independently of the others. Because of req. 2, differences in the loading of biological samples in the instrument measuring $\mathbf{m}_i$ lead to systemic measurement noise. Component measurement noise describes other sources of random error.

We model both types of noise as normally distributed random variables that multiply the measurements. Let $e_{i,s} \sim \mathcal{N}(0, \sigma_{i,s}^2)$ be the systemic measurement noise affecting $\mathbf{m}_i$, and for $j = 1, \dots, p_i$, let $e_{i,j} \sim \mathcal{N}(0, \sigma_{i,j}^2)$ be the component measurement noise affecting $m_{i,j}$. Thus, $\sigma_{i,s}$ and $\sigma_{i,j}$ are the levels of systemic and component measurement noise, respectively. All $e_{i,j}$ are independent of each other and of $e_{i,s}$. Then, for $j = 1, \dots, p_i$, we measure samples of the random variable $\hat{m}_{i,j}$, the noisy estimate of the true measurable quantity $m_{i,j}$:

$$\hat{m}_{i,j} = m_{i,j}(1 + e_{i,s})(1 + e_{i,j}) \quad (7)$$

The percent error in $\hat{m}_{i,j}$ from $m_{i,j}$ is a random variable $e_{i,j}^M$:

$$e_{i,j}^M = \frac{\hat{m}_{i,j} - m_{i,j}}{m_{i,j}} = e_{i,s} + e_{i,j} + e_{i,s} e_{i,j} \quad (8)$$

To see how weighted-sum normalizing affects the percent error due to measurement noise, we normalize the noisy measurements $\hat{m}_{i,j}$ as described in section 2.2, obtaining the random variable $\widehat{\widehat{m}}_{i,j}$:

$$\widehat{\widehat{m}}_{i,j} = \frac{\hat{m}_{i,j}}{\mathbf{c}_i^T \widehat{\mathbf{m}}_i} K_i(t) = \frac{m_{i,j}(1 + e_{i,j})K_i(t)}{\sum_{z=1}^{p_i} c_{i,z} m_{i,z}(1 + e_{i,z})} \quad (9)$$

We note that weighted-sum normalizing makes the effect of the systemic measurement noise disappear completely. The percent error in $\widehat{m}_{i,j}$ from $\bar{m}_{i,j}$ is a random variable $\bar{e}_{i,j}^M$:

$$\bar{e}_{i,j}^M = \frac{\widehat{m}_{i,j} - \bar{m}_{i,j}}{\bar{m}_{i,j}} = \frac{\sum_{z=1, z \neq j}^{p_i} c_{i,z} m_{i,z} (e_{i,j} - e_{i,z})}{\sum_{z=1}^{p_i} c_{i,z} m_{i,z} (1 + e_{i,z})} \quad (10)$$

To normalize m_i to a loading control, we find a substance that is not included in m_i but can be measured concurrent with m_i by the same instrument. The measured quantity of this substance, the “loading control,” must remain at a constant level $m_{i,L}$ for the duration of each experiment. Here we assume that the loading control occurs naturally in the system; if it must be manually added to each sample, that introduces additional error. The loading control is subject to the same systemic measurement noise as m_i along with its own

component measurement noise $e_{i,L} \sim \mathcal{N}(0, \sigma_{i,C}^2)$, which is independent of all other noise sources. Thus, when we quantify the loading control, we sample the random variable $\hat{m}_{i,L}$, the noisy estimate of $m_{i,L}$:

$$\hat{m}_{i,L} = m_{i,L} (1 + e_{i,S}) (1 + e_{i,L}) \quad (11)$$

We normalize m_i to the loading control by dividing each measurement $\hat{m}_{i,j}$ by our loading control measurement $\hat{m}_{i,L}$, obtaining the random variable $\widehat{\tilde{m}}_{i,j}$:

$$\widehat{\tilde{m}}_{i,j} = \frac{\hat{m}_{i,j}}{\hat{m}_{i,L}} = \frac{m_{i,j} (1 + e_{i,j})}{m_{i,L} (1 + e_{i,L})} \quad (12)$$

Like weighted-sum normalizing, this technique completely eliminates the effect of the systemic measurement noise. The percent error in $\widehat{\tilde{m}}_{i,j}$ from $\tilde{m}_{i,j}$ is a random variable $\tilde{e}_{i,j}^M$:

$$\tilde{e}_{i,j}^M = \frac{\widehat{\tilde{m}}_{i,j} - \tilde{m}_{i,j}}{\tilde{m}_{i,j}} = \frac{e_{i,j} - e_{i,L}}{1 + e_{i,L}} \quad (13)$$

Now we can assess the effect of normalizing by comparing the percent errors without normalizing ($e_{i,j}^M$), after weighted-sum normalizing ($\bar{e}_{i,j}^M$), and after normalizing to a loading control ($\tilde{e}_{i,j}^M$). For an error metric, we consider the expected values of the percent errors squared averaged over the measurable quantities of m_i :

$$\text{error} = \frac{1}{p_i} \sum_{j=1}^{p_i} E \left[(e_{i,j}^M)^2 \right] \quad (14)$$

and likewise for $\bar{e}_{i,j}^M$ and $\tilde{e}_{i,j}^M$. This is a fair way to assess the effect of weighted-sum normalizing, which may improve the percent error in one measurable quantity while worsening the error in another.

Figure 2 shows this error metric evaluated at several levels of component measurement noise ($\sigma_{i,C}$) and systemic measurement noise ($\sigma_{i,S}$) for the three normalizing conditions. Figure 2a compares the errors when there are two measurable quantities ($p_i = 2$), and Figure

2b does the same for three measurable quantities ($p_i = 3$). In both cases we let $K_i(t) = 1$, i.e.

we picked weighted measurement values $c_{i,z}m_{i,z}$ such that $\sum_{z=1}^{p_i} c_{i,z}m_{i,z} = 1$. To ensure numerical stability in computing the expected values, we evaluated the probability density functions for $e_{i,j}$ and $e_{i,L}$ only over the range $[-3\sigma_{i,C}, 3\sigma_{i,C}]$, which makes up 99.7% of the probability space for these random variables.

We can see from Figure 2 that weighted-sum normalizing consistently yields a lower average expected percent measurement error than normalizing to a loading control. In most cases weighted-sum normalizing also leads to lower error than not normalizing at all, particularly at high levels of systemic measurement noise. The exception to this is when component measurement noise is high, systemic measurement noise is low, and one weighted measurement $c_{i,j}m_{i,j}$ dominates the rest.

2.3.2. Interpolation—Interpolating the normalized samples of m_1 and m_2 generates continuous “synthetic data” that can be used with continuous-time models for N_1' and N_2^{-1} (reqs. 4 and 5) to identify f_{NL} . Since the samples are sparse in time, an exact interpolation of all samples is reasonable. If done well, this can contribute to accurate structure identification; if not, it can yield misleading results.

For this procedure, we chose the piecewise cubic Hermite interpolating polynomial (PCHIP) constructed by the `pchip` function in MATLAB (MathWorks, Natick, MA) (Moler, 2004) as our interpolation method. This polynomial was designed to represent physical reality for applications in science and engineering (Fritsch and Carlson, 1980), and it has two properties that make it especially appropriate for this context. First, the PCHIP is monotone on each interval between two consecutive measurements, i.e. it does not “overshoot” the measurements, which would introduce information not justified by the data. Additionally, the PCHIP is continuous to its first derivative, which is reasonable provided that the signal being measured is expected to be continuous.

3. Application

We used this structure identification procedure to investigate a link in the Sre1 signaling pathway of the fission yeast *S. pombe* (Porter et al., 2010). This pathway (Figure 3a) mediates the yeast’s adaptation to low-oxygen conditions (Hughes et al., 2005). Briefly, the proteins Sre1 and Scp1 sense oxygen indirectly by sensing the cell membrane component ergosterol (Porter et al., 2010), the synthesis of which requires oxygen. When Sre1-Scp1 detect a drop in ergosterol, Sre1 is activated to increase transcription of genes that help the yeast to survive in a low-oxygen environment (Todd et al., 2006). We sought to model the relationship between ergosterol concentration and the rate coefficient for Sre1 activation in order to gain insight into the biological mechanism for ergosterol sensing by Sre1-Scp1.

3.1. Defining the systems

Ergosterol sensing by Sre1-Scp1 can be thought of as a static nonlinearity connecting two dynamical systems (Figure 3b). The input subsystem N_1 is the metabolic pathway that synthesizes ergosterol in the endoplasmic reticulum (ER) of a yeast cell (Espenshade and Hughes, 2007). We simplify this complex pathway by considering only the chemical species found in the greatest abundance. Our model of N_1 begins with squalene (S), which is converted to lanosterol (L), 24-methylene lanosterol (M), and finally ergosterol in the ER (E_{ER}). Ergosterol is subsequently transported to the plasma membrane (E_{PM}) (Zinser et al., 1991; Baumann et al., 2005). The state variables of N_1 are concentrations of these five chemical species.

The physiological inputs to this process (u_1) include concentrations of oxygen, enzymes involved in the pathway, and pathway inhibitors. Our model of N_1 does not explicitly describe how each of these inputs affects the pathway, as this knowledge is not required to identify f_{NL} . The output y_1 is the concentration of ergosterol in the ER (E_{ER}), the pool sensed by Sre1-Scp1. The measurable quantities m_1 are the amounts of each chemical species found in a fixed-size population of cells, measured simultaneously by gas chromatography to satisfy req. 2 (Hughes et al., 2007). We note that this technique cannot discriminate between ergosterol in the ER and in the PM. Measuring these separate pools of ergosterol is possible (Radhakrishnan et al., 2008) but would violate req. 2, increase experimental complexity, and most likely add measurement noise.

Using the notation in Table 2, we write model equations for N_1 in the form of (1):

$$\dot{x}_1 = \begin{bmatrix} \dot{S} & \dot{L} & \dot{M} & \dot{E}_{ER} & \dot{E}_{PM} \end{bmatrix}^T = \begin{bmatrix} k_{ps} - f_{SL}(S, L, u_1) - \beta S \\ f_{SL}(S, L, u_1) - f_{LM}(L, M, u_1) - \beta L \\ f_{LM}(L, M, u_1) - f_{ME}(M, E_{ER}, u_1) - \beta M \\ f_{ME}(M, E_{ER}, u_1) - k_{TS}(E_{ER} - pE_{PM}) - \beta E_{ER} \\ k_{TS} \frac{A_{ER}}{A_{PM}} (E_{ER} - pE_{PM}) - \beta E_{PM} \end{bmatrix} \quad (15)$$

$$x_1(t_0) = \begin{bmatrix} S_0 & L_0 & M_0 & E_{ER,0} & E_{PM,0} \end{bmatrix}^T \quad (16)$$

$$y_1 = E_{ER} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix} x_1 \quad (17)$$

$$m_1 = \begin{bmatrix} S^{amt} \\ L^{amt} \\ M^{amt} \\ E_{total}^{amt} \end{bmatrix} = \begin{bmatrix} A_{ER} & 0 & 0 & 0 & 0 \\ 0 & A_{ER} & 0 & 0 & 0 \\ 0 & 0 & A_{ER} & 0 & 0 \\ 0 & 0 & 0 & A_{ER} & A_{PM} \end{bmatrix} x_1 \quad (18)$$

The output subsystem N_2 is the process in which Sre1 is activated in response to a decrease in ER ergosterol. Sre1 (X) and Scp1 (W), which are located in the ER membrane, bind to form a complex (Y) (Hughes et al., 2005). When this complex senses a decrease in ER ergosterol, Sre1 is activated (X_A), and Scp1 is recycled to its unbound state (Espenshade and Hughes, 2007). Active Sre1, in turn, promotes further production of its inactive precursor (Todd et al., 2006), and unbound inactive Sre1 and active Sre1 are degraded (Hughes et al., 2009; Lee et al., 2011). The state variables of N_2 are concentrations of these four chemical species.

For our purposes, the input u_2 to this process is f_A , the rate coefficient for Sre1 activation, which we take to be a function of ergosterol in the ER. The physiological output y_2 is the concentration of active Sre1, which signals for the yeast's hypoxic response (Todd et al., 2006). The measurable quantities m_2 are the amounts of inactive ($X + Y$) and active (X_A) Sre1 in a fixed-size population of cells, which we measure simultaneously by quantitative Western blotting, satisfying req. 2 (Hughes et al., 2007).

Using the notation in Table 2, we write model equations for N_2 in the form of (3):

$$\dot{\mathbf{x}}_2 = \begin{bmatrix} \dot{X} \\ \dot{W} \\ \dot{Y} \\ \dot{X}_A \end{bmatrix} = \begin{bmatrix} k_{pX} + \frac{k_{pXX} X_A^h}{T_{pXX} + X_A^h} - k_1 X W + k_2 Y - (\beta + k_{dX}) X \\ k_{pW} - k_1 X W + k_2 Y + f_A Y - \beta W \\ k_1 X W - k_2 Y - f_A Y - \beta Y \\ f_A \frac{A_{ER}}{V_{nuc}} Y - (\beta + k_{dXA}) X_A \end{bmatrix} \quad (19)$$

$$\mathbf{x}_2(t_0) = \begin{bmatrix} X_0 & W_0 & Y_0 & X_{A,0} \end{bmatrix}^T \quad (20)$$

$$\mathbf{y}_2 = X_A = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_2 \quad (21)$$

$$\mathbf{m}_2 = \begin{bmatrix} X_A^{amt} + Y_A^{amt} \\ X_A^{amt} \end{bmatrix} = \begin{bmatrix} A_{ER} & 0 & A_{ER} & 0 \\ 0 & 0 & 0 & V_{nuc} \end{bmatrix} \mathbf{x}_2 \quad (22)$$

Connecting N_1 and N_2 is f_{NL} , the unknown ergosterol sensing function. Its input is $y_1 = E_{ER}$, and its output is $u_2 = f_A$; thus, it represents the relationship between the concentration of ergosterol in the ER and the rate coefficient for Sre1 activation. We know that ergosterol is co-localized with the Sre1-Scp1 complex in the ER membrane and that Sre1-Scp1 senses ergosterol, likely via Scp1's sterol sensing domain (Porter et al., 2010). Consequently, we assume that the sensing does not rely on intracellular transport, which is slow, and does involve molecules binding, which is fast relative to other processes in the system. Thus, we can reasonably model ergosterol sensing with a static function.

3.2. Satisfying requirements

Req. 1 is a condition on the experiments, and we have shown that the measurement methods for $\mathbf{m}_1$ and $\mathbf{m}_2$ satisfy req. 2. We now show that the subsystems N_1 and N_2 (with some modification) satisfy the remaining three requirements.

3.2.1. Requirement 3—To show that N_1 satisfies req. 3, we choose a weighting vector

$\mathbf{c}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}^T$ since one molecule of squalene makes one molecule of lanosterol, which makes one molecule of 24-methylene lanosterol, and so forth. With this weighting vector, we find $K_1(t)$:

$$K_1(t) = \mathbf{c}_1^T \mathbf{m}_1(t) = A_{ER} \begin{bmatrix} 1 & 1 & 1 & 1 & \frac{A_{PM}}{A_{ER}} \end{bmatrix} \mathbf{x}_1(t) \quad (23)$$

Since our measurements come from a fixed-size population of cells, we assume that A_{ER} and A_{PM} are constant. Thus, we differentiate (23) to obtain

$$\dot{K}_1(t) = A_{ER} \begin{bmatrix} 1 & 1 & 1 & 1 & \frac{A_{PM}}{A_{ER}} \end{bmatrix} \dot{\mathbf{x}}_1(t) = k_{pS} A_{ER} - \beta K_1(t). \quad (24)$$

Given an initial condition $K_1(0)$ and values of k_{pS} and β for the duration of the experiment, we can find $K_1(t)$ for the duration of the experiment, and req. 3 is satisfied. We specifically consider two simple cases: First, suppose that the experimental treatment does not change the rate of squalene production (k_{pS}). Then, if β is constant and N_1 begins the experiment in steady state, it remains in steady state, i.e. $K_1(t) = k_{pS} A_{ER} / \beta$ for $t > 0$. Second, suppose that

the experimental treatment completely blocks squalene production, i.e. $k_{pS} = 0$ for $t > 0$. In that case, if β is constant and N_1 begins the experiment in steady state, $K_1(t) = (k_{pS} A_{ER} / \beta) e^{-\beta t}$ for $t > 0$.

The subsystem N_2 , as described by (19–22), does not satisfy req. 3. To make N_2 satisfy this requirement and enable the error-reducing benefits of weighted-sum normalizing, we made several genetic modifications to the yeast. We mutated the promoter for the *sre1⁺* gene to prevent active Sre1 from increasing the production of its precursor, which effectively set $k_{pXX} = 0$. We also deleted two genes involved in degrading unbound inactive Sre1 and active Sre1, setting $k_{dX} = 0$ and $k_{dXA} = 0$. These mutations lead to the following simplified model for N_2 dynamics:

$$\dot{x}_2 = \begin{bmatrix} \dot{X} \\ \dot{W} \\ \dot{Y} \\ \dot{X}_A \end{bmatrix} = \begin{bmatrix} k_{pX} - k_1 X W + k_2 Y - \beta X \\ k_{pW} - k_1 X W + k_2 Y + f_A Y - \beta W \\ k_1 X W - k_2 Y - f_A Y - \beta Y \\ f_A \frac{A_{ER}}{V_{nuc}} Y - \beta X_A \end{bmatrix} \quad (25)$$

With these changes, the rates of Sre1 production and degradation are no longer dependent on the state variables of N_2 , and the modified subsystem N_2 described by (25) and (20–22)

satisfies req. 3. We choose a weighting vector $c_2 \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ since one molecule of inactive Sre1 yields one molecule of active Sre1. With this weighting vector, we find $K_2(t)$:

$$K_2(t) = c_2^T m_2(t) = \begin{bmatrix} A_{ER} & 0 & A_{ER} & V_{nuc} \end{bmatrix} x_2(t) \quad (26)$$

As before, we differentiate (26) to obtain

$$\dot{K}_2(t) = \begin{bmatrix} A_{ER} & 0 & A_{ER} & V_{nuc} \end{bmatrix} \dot{x}_2(t) = k_{pX} A_{ER} - \beta K_2(t). \quad (27)$$

Given an initial condition $K_2(0)$ and values of k_{pX} and β for the duration of the experiment, we can find $K_2(t)$ for the duration of the experiment, and req. 3 is satisfied. Since our experimental treatments do not target N_2 directly, we specifically consider the simple case in which both k_{pX} and β are constant. Then, if N_2 begins the experiment in steady state, it remains in steady state, i.e. $K_2(t) = k_{pX} A_{ER} / \beta$ for $t > 0$.

3.2.2. Requirement 4—To show that N_1 satisfies req. 4, we first observe from (18):

$$m_{1,4} = E_{total}^{amt} = A_{ER} E_{ER} + A_{PM} E_{PM} \quad (28)$$

$$y_1 = E_{ER} = \frac{m_{1,4}}{A_{ER}} - \frac{A_{PM}}{A_{ER}} E_{PM} \quad (29)$$

Substituting (29) into the last row of (15) gives

$$\dot{E}_{PM} = - \left(k_{TS} \left(1 + \frac{A_{ER}}{A_{PM}} p \right) + \beta \right) E_{PM} + \frac{k_{TS}}{A_{PM}} m_{1,4} \quad (30)$$

Thus, (29) and (30) define the system N_1' with which we can compute $y_1(t)$ from $m_1(t)$, satisfying req. 4.

3.2.3. Requirement 5—To show that the modified subsystem N_2 satisfies req. 5, we consider the total amount of Sre1, $X_{\text{total}}^{\text{amt}} = X^{\text{amt}} + Y^{\text{amt}} + X_A^{\text{amt}}$. We note that $X_{\text{total}}^{\text{amt}} = K_2(t)$ and recall from section 3.2.1 that if k_{pX} and β are constant and N_2 begins an experiment in steady state, then $K_2(t)$ is constant. Similarly, we consider the total amount of Scp1, $W_{\text{total}}^{\text{amt}} = W^{\text{amt}} + Y^{\text{amt}}$. Then

$$\dot{W}_{\text{total}}^{\text{amt}} = \dot{W}^{\text{amt}} + \dot{Y}^{\text{amt}} = A_{\text{ER}} \dot{W} + A_{\text{ER}} \dot{Y} = k_{\text{pW}} A_{\text{ER}} - \beta W_{\text{total}}^{\text{amt}}. \quad (31)$$

If k_{pW} and β are constant and N_2 begins an experiment in steady state, then $W_{\text{total}}^{\text{amt}}$ is constant.

Now, since $Y^{\text{amt}} = A_{\text{ER}} Y$, we find

$$\dot{Y}^{\text{amt}} = A_{\text{ER}} \dot{Y} = \frac{k_1}{A_{\text{ER}}} X^{\text{amt}} W^{\text{amt}} - k_2 Y^{\text{amt}} - f_A Y^{\text{amt}} - \beta Y^{\text{amt}}. \quad (32)$$

Substituting $X^{\text{amt}} = X_{\text{total}}^{\text{amt}} - Y^{\text{amt}} - X_A^{\text{amt}}$, $W^{\text{amt}} = W_{\text{total}}^{\text{amt}} - Y^{\text{amt}}$, $f_A Y^{\text{amt}} = X_A^{\text{amt}} + \beta X_A^{\text{amt}}$, and $X_A^{\text{amt}} = m_{2,2}$ yields

$$\dot{Y}^{\text{amt}} = \frac{k_1}{A_{\text{ER}}} (X_{\text{total}}^{\text{amt}} - Y^{\text{amt}} - m_{2,2})(W_{\text{total}}^{\text{amt}} - Y^{\text{amt}}) - k_2 Y^{\text{amt}} - (\dot{m}_{2,2} + \beta m_{2,2}) - \beta Y^{\text{amt}}. \quad (33)$$

Finally, we rearrange the fourth line of (25):

$$u_2 = f_A = (\dot{m}_{2,2} + \beta m_{2,2}) / Y^{\text{amt}} \quad (34)$$

Thus, (33) and (34) define the system N_2^{-1} with which we can compute $u_2(t)$ from $m_2(t)$, satisfying req. 5.

3.3. Results

In several experiments (described in Appendix A), we subjected *S. pombe* cells, genetically modified as previously described, to treatments that impaired the cells' ability to synthesize ergosterol (Porter et al., 2010). These treatments included low levels of oxygen, which is required for ergosterol synthesis, as well as drugs that inhibit particular steps in ergosterol synthesis. After applying a treatment to a population of cells at $t = 0$, we grew the cells for six hours and measured amounts of ergosterol and its precursors (m_1) and inactive and active Sre1 (m_2) as described previously. We took samples up to nine times over the course of each experiment, concentrating about half of the samples during the first hour of the experiment to best observe the changes occurring immediately after treatment. The regimen of treatments included some that lowered ergosterol to as little as 10% of its mean starting value, satisfying requirement 1.

3.3.1. Step 1: normalize the measurements—Figure 4 shows samples of m_1 (a) and m_2 (b) from one experiment before and after normalizing to their weighted sums. The drug applied in this experiment blocks the conversion of lanosterol to 24-methylene lanosterol;

thus, it should not change the total amount of ergosterol plus its precursors. Normalizing the samples of m_1 enforces this condition.

The error reduction effect on m_2 appears to be stronger. The total amount of Sre1 should be constant throughout the experiment, and normalizing the m_2 samples enforces this. As a result, the amount of active Sre1 increases during the experiment, consistent with the decrease in ergosterol. Collectively, these measurements are more in line with physical expectations after they have been normalized to their weighted sums, consistent with the claim that this normalizing method reduces measurement error.

3.3.2. Steps 2–3: interpolate the normalized measurements and solve for f_{NL} —

From the data in Figure 4a, Figure 5a shows the interpolated signal $\bar{m}_{1,4}(t)$ (left) and the signal $y_1(t)$ computed via N_1' (right). Similarly, from the data in Figure 4b, Figure 5b shows the interpolated signal $\bar{m}_{2,2}(t)$ (left) and the signal $u_2(t)$ computed via N_2^{-1} (right). We note that $u_2(t)$ contains a slight “bounce” between two of the original sampling times, an artifact of the interpolation process.

3.3.3. Step 4: inspect the result—For several experiments, we used the computed signals $y_1(t)$ and $u_2(t)$ to plot the experiments’ putative trajectories through (y_1, u_2) -space (Figure 6a, upper panel). Based on these estimates of the input-output behavior of f_{NL} , we sought a function $u_2(y_1)$ to describe it. Having seen that interpolation introduces artifacts into the data, we tried to minimize this effect by considering only the points of the (y_1, u_2) trajectories at the original sampling times of each experiment (Figure 6a, lower panel). To fit a function to these points, we minimized the sum of squared differences between the logarithm of the function and that of the points.

We first tried to fit an exponential function $u_2 = ae^{by_1}$ to the (y_1, u_2) points, which fit with $R^2 = 0.66$ (not shown). To try to improve this, we looked for a better function and came up with $u_2 = a/(1 + by_1^n)$. This function was suggested by the structure of the SREBP pathway, the mammalian homolog of the yeast Sre1 pathway. Such a function would arise if Sre1-Scp1 had “on” and “off” states with the transition between them cooperatively mediated by ergosterol (Porter et al., 2010). This function fit the data points with $R^2 = 0.75$. While the noise in the data makes it difficult to be certain about what is going on, the fit of this function supports the hypothesis that Sre1-Scp1 senses ergosterol by the mechanism described here.

Additionally, we tried adding a constant term to the function to represent a basal rate of Sre1 activation. The function $u_2 = c + a/(1 + by_1^n)$ fit the data points with $R^2 = 0.78$. Because of the relatively small improvement, we judged this to be an over-fit and decided to keep the function with three parameters.

3.4. Comparison of normalizing methods in practice

To show the usefulness of weighted-sum normalizing in practice, we repeated the structure identification procedure with the same data but normalized the samples of m_1 to loading controls instead of to their weighted sum. To control for the number of cells measured, we divided each sample of m_1 by the density of the cells sampled at the same time. To control for variability in the instrument measuring m_1 , we added a reference amount of cholesterol to each sample and divided each measurement of m_1 by the measurement of cholesterol at that time point. We note that since this method involved both measuring a new quantity (cell density) with a different instrument and adding a new substance to each sample, the error should exceed that found by our calculation in section 2.3.1.

Figure 6b shows the (y_1, u_2) trajectories calculated with the $\mathbf{m}_1$ samples normalized to these loading controls. For comparison, Figure 6c shows the (y_1, u_2) trajectories calculated without normalizing the $\mathbf{m}_1$ samples. For consistency, the samples of $\mathbf{m}_2$ were normalized to their weighted sum for all three figures. To compare these results, we calculated Spearman's rank correlation coefficient (ρ) for the sampled points along the (y_1, u_2) trajectories. Normalizing $\mathbf{m}_1$ to its weighted sum yielded a higher correlation ($\rho = -0.72$) than normalizing to loading controls ($\rho = -0.52$) or not normalizing ($\rho = -0.53$). Similarly, normalizing $\mathbf{m}_1$ to its weighted sum resulted in a better fit of the function $u_2 = a/(1 + by_1^n)$ to the sampled points as measured by R^2 . Qualitatively, the scatter among the points in Figure 6b and Figure 6c is high enough to make it difficult to pick a function to fit. While weighted-sum normalizing is far from perfect at reducing measurement error, it is a clear improvement over these alternatives.

4. Evaluation

To evaluate the model of f_{NL} obtained by this structure identification procedure, we performed simulations and compared them to experimental data. For several experiments, we used the normalized, interpolated measurements $\widehat{\mathbf{m}}_1(t)$ as the inputs to the system of N'_1 , the identified f_{NL} , and N_2 . We compared the simulation outputs with the normalized measurements $\widehat{\mathbf{m}}_2$ and found them to be qualitatively consistent. The quantified measurement noise increased from its steady-state estimate, which we expect is due to experimental variability.

4.1. Error metric for normalized data

To compare simulation results with normalized measurements, we first defined a suitable error metric. Simply measuring the percent error between them is misleading because a given level of component measurement noise $\sigma_{2,C}$ affects the percent error in $\widehat{\mathbf{m}}_2$ differently depending on the value of $\mathbf{m}_2$ (Figure 2). To avoid this distortion, we used the following method to estimate the level of component measurement noise in our measurements.

We begin with the measurement noise model described in section 2.3.1, observing from (9) that for any j , $\widehat{m}_{i,j}$ is a ratio of normally distributed random variables, and making a substitution from (4):

$$\widehat{m}_{i,j} = \frac{m_{i,j}(1 + e_{i,j})K_i(t)}{\sum_{z=1}^{p_i} c_{i,z}m_{i,z}(1 + e_{i,z})} = \frac{m_{i,j}K_i(t) + m_{i,j}K_i(t)e_{i,j}}{K_i(t) + \sum_{z=1}^{p_i} c_{i,z}m_{i,z}e_{i,z}} \quad (35)$$

Now we consider: if $Z = (\mu_N + Z_N)/(\mu_D + Z_D)$, where $Z_N \sim \mathcal{N}(0, \sigma_N^2)$, $Z_D \sim \mathcal{N}(0, \sigma_D^2)$, and $\rho = \text{cov}(Z_N, Z_D)/(\sigma_N\sigma_D)$, then the random variable

$$T = \frac{\mu_D Z - \mu_N}{\sqrt{\sigma_D^2 Z^2 - 2\rho\sigma_D\sigma_N Z + \sigma_N^2}} \quad (36)$$

is distributed $\mathcal{N}(0, 1)$, provided that $\mu_D + Z_D$ is unlikely to assume a negative value (Geary, 1930). This is known as the Geary-Hinkley transformation (Hayya et al., 1975).

Since $e_{i,z} \sim \mathcal{N}(0, \sigma_{i,c}^2)$ for all z , $\widehat{m}_{i,j}$ as shown in (35) fits into this framework with $\mu_N = m_{i,j}K_i(t)$ and $\mu_D = K_i(t)$. To find σ_N , σ_D , and ρ , we note that

$$\begin{bmatrix} Z_N \\ Z_D \end{bmatrix} = \begin{bmatrix} m_{i,j}K_i(t)e_{i,j} \\ \sum_{z=1}^{p_i} c_{i,z}m_{i,z}e_{i,z} \end{bmatrix} = \begin{bmatrix} 0 & \dots & m_{i,j}K_i(t) & \dots & 0 \\ c_{i,1}m_{i,1} & \dots & c_{i,j}m_{i,j} & \dots & c_{i,p_i}m_{i,p_i} \end{bmatrix} \begin{bmatrix} e_{i,1} \\ \vdots \\ e_{i,j} \\ \vdots \\ e_{i,p_i} \end{bmatrix}. \quad (37)$$

Since $\begin{bmatrix} Z_N \\ Z_D \end{bmatrix}$ is a linear transformation of the normal random vector $\begin{bmatrix} e_{i,1} & \dots & e_{i,p_i} \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \sigma_{i,C}^2 \mathbf{I})$, we can compute the joint probability distribution of Z_N and Z_D (Stark and Woods, 2001):

$$\begin{bmatrix} Z_N \\ Z_D \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{i,C}^2 m_{i,j}^2 K_i^2(t) & \sigma_{i,C}^2 m_{i,j}^2 c_{i,j} K_i(t) \\ \sigma_{i,C}^2 m_{i,j}^2 c_{i,j} K_i(t) & \sigma_{i,C}^2 \sum_{z=1}^{p_i} (c_{i,z} m_{i,z})^2 \end{bmatrix}\right) \quad (38)$$

Thus, $\sigma_N = \sigma_{i,C} m_{i,j} K_i(t)$, $\sigma_D = \sigma_{i,C} \sqrt{\sum_{z=1}^{p_i} (c_{i,z} m_{i,z})^2}$, and $\rho = m_{i,j} c_{i,j} / \sqrt{\sum_{z=1}^{p_i} (c_{i,z} m_{i,z})^2}$. Substituting these into (36) and noting that the ratio Z is $\widehat{m}_{i,j}$ yields

$$T_{i,j} = \frac{\widehat{m}_{i,j} K_i(t) - m_{i,j} K_i(t)}{\sigma_{i,C} \sqrt{\widehat{m}_{i,j}^2 \sum_{z=1}^{p_i} (c_{i,z} m_{i,z})^2 - 2 \widehat{m}_{i,j} m_{i,j}^2 c_{i,j} K_i(t) + m_{i,j}^2 K_i^2(t)}} \sim \mathcal{N}(0, 1). \quad (39)$$

Therefore, we can define

$$T'_{i,j} = \frac{\widehat{m}_{i,j} K_i(t) - m_{i,j} K_i(t)}{\sigma_{i,C} \sqrt{\widehat{m}_{i,j}^2 \sum_{z=1}^{p_i} (c_{i,z} m_{i,z})^2 - 2 \widehat{m}_{i,j} m_{i,j}^2 c_{i,j} K_i(t) + m_{i,j}^2 K_i^2(t)}} \sim \mathcal{N}(0, \sigma_{i,C}^2). \quad (40)$$

The Geary-Hinkley transformation is valid as long as the denominator of (9) is positive, which is satisfied in practice as long as at least one measurement from $\mathbf{m}_2$ is positive and the rest are nonnegative.

This suggests a simple way of estimating $\sigma_{2,C}$: we assume that the simulation outputs are the true measurable quantities $\mathbf{m}_2$ and that the normalized measurements are the noisy normalized measurable quantities $\widehat{\mathbf{m}}_2$. For the j th measurable output quantity, we compute $T'_{2,j}$ for each sample in each experiment by (40), and we compute the sample standard deviation $s_{2,C}$ of the resulting values of $T'_{2,j}$.

4.2. Predicted component measurement noise

To assess whether our simulations could produce plausible results, we estimated the level of component measurement noise in the $\mathbf{m}_1$ and $\mathbf{m}_2$ measurement processes by examining the

samples at $t = 0$ in the five experiments used to identify f_{NL} (Figure 6) and seven additional experiments. Since these are samples of the pre-treatment steady state of the system, their variation provides a good estimate of the level of component measurement noise. From the normalized samples $\widehat{\mathbf{m}}_1(0)$, we computed $T'_{1,4}$ since $m_{1,4}$ (ergosterol) is the input to N'_1 . We estimated the true starting measurement vector $\mathbf{m}_1(0)$ as that which gave the samples of $T'_{1,j}$, $j = 1, \dots, 4$, sample means closest to zero, subject to the normalizing constraint (4). The standard deviation of our samples of $T'_{1,4}$ was $s_{1,C} = 0.11$. Similarly, since $m_{2,2}$ (active Sre1) is the input to N_2^{-1} , we computed $T'_{2,2}$ from the normalized samples $\widehat{\mathbf{m}}_2(0)$ and found the sample standard deviation to be $s_{2,C} = 0.26$.

4.3. Simulation results and assessment

For the five “identification” experiments used to create the model of f_{NL} and seven additional “validation” experiments, we performed simulations to assess how well the system containing f_{NL} could reproduce observed outputs when given observed inputs. We simulated the system of N'_1, f_{NL} , and N_2 with the input $\widehat{\mathbf{m}}_1(t)$ and compared the simulated output $\mathbf{m}_2(t)$ with the normalized measurements $\widehat{\mathbf{m}}_2(t_\ell)$ (Figure 7). The simulations produced results qualitatively consistent with the measurements.

To quantitatively estimate the noise level, we computed $T'_{2,2}$ using (40) for each sample measured at $t_\ell > 0$ in these twelve experiments. Assuming zero mean, the sample standard deviation of $T'_{2,2}$ was $s_{2,C} = 0.43$ for the identification experiments ($n = 40$) and $s_{2,C} = 0.38$ for the validation experiments ($n = 52$). The similarity between these estimates of the noise level suggests that the model of f_{NL} did not simply result from over-fitting the data but rather describes the underlying biological process. While the noise level estimates are higher than that made at steady state, we expect that this is largely because of experimental variability in cells undergoing treatment. We note that the measurements differed greatly even between identical experiments (Figure 7, compare left and right columns).

5. Discussion and Conclusion

Biology stands to gain much from quantitative approaches common in engineering and other sciences, but such approaches are hampered by difficulties in obtaining a sufficient number of accurate measurements. Consequently, there is a great need for creative ways of extracting biological knowledge from noisy, sparse data. When this includes both the structure and the parameters of part of a biological system, a structure identification method such as that presented here is an appropriate tool to attack the problem. Our approach effectively identified a static nonlinearity embedded in a biological signaling cascade using indirect measurements from the systems connected to the input and output of the nonlinearity. We reduced the influence of measurement noise through weighted-sum normalizing, then filled in the gaps in the sparse data by interpolating. The resulting description of the nonlinearity (Figure 6a), while still corrupted by noise, was coherent enough that we could fit a function to it and discriminate between alternative best-fit functions. This revealed an aspect of the signaling cascade that would have been difficult to understand directly by experimentation.

Among techniques to lower measurement noise in biological data, weighted-sum normalizing fills an important niche. In general, one can reduce measurement noise by experimental techniques or by post-processing. Most post-processing methods are not applicable when the data are sparse, as they rely on temporal filtering. Among experimental

methods for reducing measurement noise, the simplest is to average multiple samples of each measurement. If we model noise as in section 2.3.1, it can be shown that the average expected value of percent error squared (Figure 2) decreases by a factor of n when n independent samples of a measurement are averaged. The downside is that additional samples can be costly and laborious. For cases in which this extra data is not available, we found weighted-sum normalizing to be superior both in theory and in practice to the more common technique of normalizing to loading controls.

The procedure presented here, while not universally applicable to biological systems, can be employed with several common classes of systems as discussed in section 2.1. Perhaps the most tightly limiting requirements are the general form of the system (Figure 1), including the need to approximate the unknown connecting part as a static function, and the need for the input and output subsystems to be weighted-sum predictable. The latter can sometimes be achieved artificially by mutating parts of a subsystem, as demonstrated in section 3.2.1. If the requirements can be met, our procedure has the advantage that it makes no a priori assumption about the form of the unknown static function, leaving that judgment to the person inspecting the results (Figure 6a).

This procedure advances our ability to identify unknown parts within biological systems using realistic, less-than-ideal measurements. When the measurements from a system are noisy and sparse and the known system components are sensitive to noisy signals, structure identification by any means encounters a fundamental barrier raised by limited information. This method pushes back that barrier by incorporating additional information from known system properties (weighted-sum normalizing) and making well-founded assumptions based on experimental observations (interpolation), enabling structure identification in systems where it otherwise would have been impossible.

Acknowledgments

We thank the members of the Iglesias and Espenshade labs for much helpful input during the course of this work. This research was funded by NIH grant HL-077588. JRP gratefully acknowledges the support of an Abel Wolman, a Wysocki, and a Bowles Fellowship. PJE is an Established Investigator of the American Heart Association.

References

- Ang, J.; Ingalls, B.; McMillen, D. Probing the input-output behavior of biochemical and genetic systems: System identification methods from control theory. In: Johnson, ML.; Brand, L., editors. Computer Methods, Part C. Vol. 487 of Methods in Enzymology. Vol. 10. Academic Press; 2011. p. 279-317.
- Baumann NA, Sullivan DP, Ohvo-Rekilä H, Simonot C, Pottekat A, Klaassen Z, Beh CT, Menon AK. Transport of newly synthesized sterol to the sterol-enriched plasma membrane occurs via nonvesicular equilibration. *Biochemistry*. Apr; 2005 44 (15):5816–5826. [PubMed: 15823040]
- Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell*. Aug; 2004 15 (8):3841–3862. [PubMed: 15169868]
- El-Samad H, Kurata H, Doyle JC, Gross CA, Khammash M. Surviving heat shock: control strategies for robustness and performance. *Proc Natl Acad Sci U S A*. Feb; 2005 102 (8):2736–2741. [PubMed: 15668395]
- Emerson RC, Korenberg MJ, Citron MC. Identification of complex-cell intensive nonlinearities in a cascade model of cat visual cortex. *Biol Cybern*. Feb; 1992 66 (4):291–300. [PubMed: 1550879]
- Espenshade PJ, Hughes AL. Regulation of sterol synthesis in eukaryotes. *Annu Rev Genet*. 2007; 41:401–427. [PubMed: 17666007]
- Fritsch FN, Carlson RE. Monotone piecewise cubic interpolation. *SIAM J Numer Anal*. Apr; 1980 17 (2):238–246.

- Geary RC. The frequency distribution of the quotient of two normal variates. *J Royal Stat Soc.* 1930; 93 (3):442–446.
- Hayya J, Armstrong D, Gressis N. A note on the ratio of two normally distributed variables. *Manage Sci.* Jul; 1975 21 (11):1338–1341.
- Hughes AL, Lee CYS, Bien CM, Espenshade PJ. 4-methyl sterols regulate fission yeast SREBP-Scap under low oxygen and cell stress. *J Biol Chem.* Jun; 2007 282 (33):24388–24396. [PubMed: 17595166]
- Hughes AL, Todd BL, Espenshade PJ. SREBP pathway responds to sterols and functions as an oxygen sensor in fission yeast. *Cell.* Mar; 2005 120 (6):831–842. [PubMed: 15797383]
- Hughes BT, Nwosu CC, Espenshade PJ. Degradation of SREBP precursor requires the ERAD components Ubc7 and Hrd1 in fission yeast. *J Biol Chem.* Jul; 2009 284 (31):20512–20521. [PubMed: 19520858]
- Hunter IW, Korenberg MJ. The identification of nonlinear biological systems: Wiener and Hammerstein cascade models. *Biol Cybern.* Nov; 1986 55 (2):135–144. [PubMed: 3801534]
- Kearney RE, Hunter IW. System identification of human triceps surae stretch reflex dynamics. *Exp Brain Res.* Jun; 1983 51 (1):117–127. [PubMed: 6884461]
- Korenberg MJ, Hunter IW. The identification of nonlinear biological systems: LNL cascade models. *Biol Cybern.* Nov; 1986 55 (2):125–134. [PubMed: 3801533]
- Lee CYS, Yeh TL, Hughes BT, Espenshade PJ. Regulation of the Sre1 hypoxic transcription factor by oxygen-dependent control of DNA binding. *Mol Cell.* Oct; 2011 44 (2):225–234. [PubMed: 22017871]
- Leloup JC, Goldbeter A. Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci U S A.* Jun; 2003 100 (12):7051–7056. [PubMed: 12775757]
- Mettetal JT, Muzzey D, Gómez-Urbe C, van Oudenaarden A. The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. *Science.* Jan; 2008 319 (5862):482–484. [PubMed: 18218902]
- Moler, CB. Numerical Computing with Matlab. Society for Industrial and Applied Mathematics; Philadelphia, PA: Jan. 2004
- Porter JR, Burg JS, Espenshade PJ, Iglesias PA. Ergosterol regulates sterol regulatory element binding protein (SREBP) cleavage in fission yeast. *J Biol Chem.* Dec; 2010 285 (52):41051–41061. [PubMed: 20959444]
- Porter, JR.; Iglesias, PA.; Burg, JS.; Espenshade, PJ. Overcoming data limitations to identify a static nonlinearity in a biological signaling cascade. 45th Annual Conference on Information Sciences and Systems; Mar. 2011
- Radhakrishnan A, Goldstein JL, McDonald JG, Brown MS. Switch-like control of SREBP-2 transport triggered by small changes in ER cholesterol: a delicate balance. *Cell Metab.* Dec; 2008 8 (6):512–521. [PubMed: 19041766]
- Sjöberg J, Zhang Q, Ljung L, Benveniste A, Delyon B, Glorennec PY, Hjalmarsson H, Juditsky A. Nonlinear black-box modeling in system identification: a unified overview. *Automatica.* Dec; 1995 31 (12):1691–1724.
- Stark, H.; Woods, JW. Probability and Random Processes with Applications to Signal Processing. 3. Prentice Hall; Upper Saddle River, NJ: Aug. 2001
- Su SW, Huang S, Wang L, Celler BG, Savkin AV, Guo Y, Cheng TM. Optimizing heart rate regulation for safe exercise. *Ann Biomed Eng.* Mar; 2010 38 (3):758–768. [PubMed: 19953322]
- Su SW, Wang L, Celler BG, Savkin AV. Oxygen uptake estimation in humans during exercise using a Hammerstein model. *Ann Biomed Eng.* Nov; 2007 35 (11):1898–1906. [PubMed: 17687652]
- Todd BL, Stewart EV, Burg JS, Hughes AL, Espenshade PJ. Sterol regulatory element binding protein is a principal regulator of anaerobic gene expression in fission yeast. *Mol Cell Biol.* Apr; 2006 26 (7):2817–2831. [PubMed: 16537923]
- Zinser E, Sperka-Gottlieb CD, Fasch EV, Kohlwein SD, Paltauf F, Daum G. Phospholipid synthesis and lipid composition of subcellular membranes in the unicellular eukaryote *Saccharomyces cerevisiae*. *J Bacteriol.* Mar; 1991 173 (6):2026–2034. [PubMed: 2002005]

Appendix A. Experimental Conditions

For the experiments used to identify the model, we grew cultures of haploid *S. pombe* cells, genetically modified as described in section 3.2.1, to log phase at 30 °C in yeast extract plus supplements. At time = 0, we treated cells by adding drugs or depleting oxygen (Table A.1); further details are available in Porter et al. (2010). For each treatment, we measured the cell density at each sample time and computed the doubling time by fitting an exponential function to the cell densities at times ≥ 1 h (all $R^2 \geq 0.98$). In all cases, the cells continued to grow exponentially during treatment with some variation in doubling time from that of untreated cells (3.0 h).

Table A.1

Treatments and cell doubling times for experiments used to identify the model of ergosterol sensing.

Treatment	Step(s) inhibited	Doubling time
Drug 1 (terbinafine)	$S \rightarrow L$	5.3 h
Drug 2 (25-thialanosterol)	$L \rightarrow M$	3.6 h
Drug 3 (itraconazole)	$M \rightarrow E_{ER}$	2.7 h
0.5% oxygen	$S \rightarrow L, M \rightarrow E_{ER}$	2.4 h
0.2% oxygen	$S \rightarrow L, M \rightarrow E_{ER}$	3.3 h

Highlights

- We present a method to identify an unknown component within a biological system.
- The measurements available for this are noisy and sparse in time.
- We reduce error in the measurements by normalizing them to their weighted sum.
- We then interpolate the measurements and solve for the unknown.
- This method worked when we applied it to a real biological system.

Figure 1.

Type of biological system of interest. Arrows represent the physiological path of signaling. Lines without arrows represent our ability to measure quantities in the system experimentally.

Figure 2.

(a) Average expected value of percent error squared without normalizing, after weighted-sum normalizing, and after normalizing to a loading control for two measurable quantities ($p_i = 2$). The weighted measurement $c_{i,1}m_{i,1}$ is varied from 0 to 1, and $c_{i,2}m_{i,2} = 1 - c_{i,1}m_{i,1}$. (b) The same for three measurable quantities ($p_i = 3$). $c_{i,1}m_{i,1}$ is varied from 0 to 1, $c_{i,2}m_{i,2}$ is varied from 0 to $c_{i,1}m_{i,1}$, and $c_{i,3}m_{i,3} = 1 - c_{i,1}m_{i,1} - c_{i,2}m_{i,2}$.

Figure 3.

(a) The Sre1 signaling pathway in *S. pombe*. Symbols are defined in the text. (b) Ergosterol sensing in this pathway can be described in the form of Figure 1.

Figure 4.

Measurements obtained while treating *S. pombe* cells with a drug that blocks the conversion of lanosterol to 24-methylene lanosterol.

Figure 5.

Normalized measurements $\bar{m}_{1,4}$ (a) and $\bar{m}_{2,2}$ (b) from the data shown in Figure 4 were interpolated to generate continuous signals. Signals $y_1(t)$ (a) and $u_2(t)$ (b) were computed via N_1' and N_2^{-1} , respectively.

Figure 6.

(top) Putative trajectories of five experiments through (y_1, u_2) -space, computed with the m_1 data normalized three different ways. The irregularity in the trajectories is a consequence of measurement noise. (bottom) A curve $u_2 = a/(1 + by_1^n)$ was fitted to the data sampled at the original sampling times of each experiment (points), with R^2 shown. Spearman's rank correlation coefficient (ρ) was computed for these points as a nonparametric measure of their monotonicity.

Figure 7.

Simulated signals $m_{2,2}(t)$ (curves) compared with experimental measurements $\widehat{m}_{2,2}(t_\ell)$ (points) for the five “identification” experiments used to identify f_{NL} and seven additional “validation” experiments. Measured points $\widehat{m}_{2,2}(t_\ell)$ that lie within a shaded region have values of $T'_{2,2}$ less than the given threshold.

Table 1

Notation for measurements used to identify a biological system.

Vector of measurements from subsystem N_i at time t	$\mathbf{m}_i(t)$
j th measurement from subsystem N_i at time t	$m_{i,j}(t)$

Table 2Notation for the models of ergosterol biosynthesis (N_1) and Sre1 dynamics (N_2).

k_{pS}	rate of squalene production	$7.9 \times 10^{-4} \mu\text{m}^{-2} \text{h}^{-1}$ ^a
$f_{(1)(2)}$	rate of conversion of chemical species (1) to species (2)	b
k_{TS}	rate coefficient for ergosterol transport along the ER-PM concentration gradient	2.0h^{-1} ^{c,d}
p	accessibility coefficient of ergosterol in the PM	0.06 ^d
k_{pX}	constant rate of Sre1 production	$0.391 \text{ molecule } \mu\text{m}^{-2} \text{h}^{-1}$ ^d
k_{pXX}	maximum variable rate of Sre1 production	b
T_{pXX}	level of active Sre1 yielding half-maximum variable Sre1 production	b
h	Hill coefficient for variable Sre1 production	b
k_{pW}	constant rate of Scp1 production	$0.078 \text{ molecule } \mu\text{m}^{-2} \text{h}^{-1}$ ^d
k_1	rate coefficient for Sre1-Scp1 binding	$500 \mu\text{m}^2 \text{ molecule}^{-1} \text{h}^{-1}$ ^d
k_2	rate coefficient for Sre1-Scp1 unbinding	0.2h^{-1} ^d
k_{dX}	rate coefficient for degradation of unbound inactive Sre1	b
k_{dXA}	rate coefficient for degradation of active Sre1	b
β	coefficient for dilution of a chemical species due to exponential cell growth	0.23h^{-1} ^d
A_{ER}	surface area of the ER	$290 \mu\text{m}^2$ ^d
A_{PM}	surface area of the PM	$130 \mu\text{m}^2$ ^d
V_{nuc}	volume of the nucleus	b

^a Chosen so that $K_1(0) = 1$; see section 3.2.1.^b Value not used in final calculations.^c Baumann et al. (2005).^d Porter et al. (2010).