Title: Listening to the noise: Random Fluctuations Reveal Gene Network Parameters

Author(s): Brian Munsky, Z# 226675 (CCS3 and CNLS) Mustafa Khammash (UCSB)

Intended for: Submission to Science
Hi,

I would like to request an LA-UR for a paper I am submitting to Science. This is not an initial submission and not a final draft. I would very much appreciate it if this paper is **not** made publicly available. Below is the relevant information and an attached LAUR request form. Please let me know what else I need to include.

Thanks,
Brian

Authors (First at LANL second at UCSB):
Brian Munsky (CCS3 and CNLS)
Mustafa Khammash (UCSB)

My time code: x9pm

Title: Listening to the noise: Random Fluctuations Reveal Gene Network Parameters

Abstract: The cellular environment is abuzz with noise. The origin of this noise is attributed to the inherent random motion of reacting molecules that take part in gene expression and post expression interactions. In this noisy environment, clonal populations of cells exhibit cell-to-cell variability that frequently manifests as significant phenotypic differences within the cellular population. The stochastic fluctuations in cellular constituents induced by noise can be measured and their statistics quantified. We show that these random fluctuations carry within them valuable information about the underlying genetic network. Far from being a nuisance, the ever-present cellular noise acts as a rich source of excitation that, when processed through a gene network, carries its distinctive fingerprint that encodes a wealth of information about that network. We demonstrate that in some cases the analysis of these random fluctuations enables the full identification of network parameters, including those that may otherwise be difficult to measure. This establishes a potentially powerful approach for the identification of gene networks and offers a new window into the workings of these networks.

Paper Type: Full paper/Report

Journal Name: Science
Listening to the noise: Random Fluctuations Reveal Gene Network Parameters

Brian Munsky\textsuperscript{1,2}, and Mustafa Khammash\textsuperscript{2}\ast

Abstract

The cellular environment is abuzz with noise. The origin of this noise is attributed to the inherent random motion of reacting molecules that take part in gene expression and post expression interactions. In this noisy environment, clonal populations of cells exhibit cell-to-cell variability that frequently manifests as significant phenotypic differences within the cellular population. The stochastic fluctuations in cellular constituents induced by noise can be measured and their statistics quantified. We show that these random fluctuations carry within them valuable information about the underlying genetic network. Far from being a nuisance, the ever-present cellular noise acts as a rich source of excitation that, when processed through a gene network, carries its distinctive fingerprint that encodes a wealth of information about that network. We demonstrate that in some cases the analysis of these random fluctuations enables the full identification of network parameters, including those that may otherwise be difficult to measure. This establishes a potentially powerful approach for the identification of gene networks and offers a new window into the workings of these networks.

Noise and its impact on cellular networks has attracted much interest. It has been shown to be a key source of cellular variability \cite{1} and can cause a multitude of effects such as stochastic switching \cite{2} stochastic resonance and stochastic focusing \cite{3}. Several key factors conspire to create the random fluctuations (noise) that are inherent in cellular networks and contribute to their significance. One derives from the very nature of chemical reactions at the molecular level where random thermal motion of reactant species translates into randomness in the sequence and timing of chemical reactions. This in turn leads to randomness in the number of molecular species. While randomness can often be inconsequential at the macroscopic scale, additional factors contribute to its significance at the cellular level. One of these has

\textsuperscript{1}CCS-3 and the Center for NonLinear Studies; Los Alamos National Lab; Los Alamos, NM 87545
\textsuperscript{2}Center for Control, Dynamical Systems and Computations; University of California; Santa Barbara, CA 93106
\ast to whom correspondence should be addressed. Email: khammash@engr.ucsb.edu
to do with the integer nature of reactants at the molecular level and reflects the fact that these reactants are present in whole numbers e.g. genes, RNAs, and proteins. Random fluctuations in the number of reactants would not be so important if it weren’t for the fact that these reactants are often present in very small quantities, e.g. a few copies of a certain gene, tens of copies of its RNA products, and so on, thereby making the variation relative to the mean quite large. When one further takes into consideration that these fluctuating species can control events downstream, e.g transcription factors that control gene expression, their fluctuations may take on added significance as they are amplified through the network.

A main goal of computational modeling, whether deterministic or stochastic, is to obtain deeper understanding of biological phenomena and its impact on cell function and to guide experimental design. Computational sensitivity/robustness analyses can reduce models to their critical components and thereby illuminate the most important aspects of a system. The process of constructing such models requires that key parameters be measured or inferred in such a manner as to be consistent with experimental observations. In the context of model identification, “noise” often seems like bad news—it corrupts real data and its effects must be overcome through noise filtering. Seen from a different perspective, however, noise may be a great bonus. Just as a white noise input helps to identify the parameters of a dynamical system [4], so too may the characterization of noise transmission help to identify and understand natural mechanisms. For example, in [5] the authors examine a transcription/translation process with two different parameter sets. They find that two different parameter sets may have the same mean protein expression and ODE descriptions, but it is only by considering the noise in the system that one can distinguish between the parameter sets. Similar observations on the effects of parameters on noise propagation have been made in many studies including [6, 7]. While the authors of these studies correctly observe these differences, researchers are only now beginning to explore the full portent of such differences. In one example, the authors of [8] show that the dynamics of statistical cumulants can enable the identification of reaction rate parameters for some gene regulatory networks. The authors observe that if one can control some aspect of the system, such as a gene transcription rate, then one can identify parameters from observed output responses. In this paper, we examine the possibility of identifying system parameters and mechanisms directly from noise measurements without additional control, and we demonstrate that the analysis of these measurements enables the full identification of network parameters. We illustrate our approach on two stochastic models of gene regulatory networks.

**Gene Expression Model.** We adopt the stochastic gene expression model in [6] where the genetic system is assumed to be characterized by the number of mRNA molecules and the number of protein molecules, which are modeled as the integer random variables $R$ and $P$, respectively. Transcription, translation, and degradation events act to change the state of the system by altering the number of mRNA and protein
molecules. For mRNA, transcription and degradation of mRNA are modeled as random events that occur according to exponentially distributed waiting times that depend on the transcription rate $k_r$ and degradation rate $\gamma_r$. Thus, given a state of $r$ mRNA molecules, the probability that a single mRNA molecule is degraded within the next small time increment $dt$ is given by $r \cdot (\gamma_r \cdot dt)$. In a similar way, translation and degradation of proteins are modeled by random events with exponentially distributed waiting times dictated by the translation rate $k_p$ and the protein degradation rate $\gamma_p$. The resulting stochastic model is thus represented by a continuous-time discrete-state Markov process. The probability of finding the system in a given state ($R(t) = r$, $P(t) = p$) is fully characterized by the system's chemical master equation, from which the evolution of moments $E[R(t)], E[P(t)], E[R^2(t)], E[P^2(t)], \ldots$ can be described (see supplementary material).

Our first finding is that all model parameters are identifiable from transient moment measurements. Noise induced fluctuations lead to cell-to-cell variability in the population of mRNA and proteins. When measured using single-cell observations via flow cytometry, flow activation cell sorting (FACS) or a related approach, the statistics of these fluctuations offer a surprisingly rich source of information about system parameters. We assume that one is able to determine first and second-order moments, or equivalently means, variances, and covariances of cellular proteins and mRNA. At a given time, $t$, each such measurement yields a vector of five quantities: $v(t) = (E[R(t)], E[P(t)], E[R^2], E[P^2], E[R|P])$. Our first result is that transient measurements of these moments contains sufficient information to allow us to determine the system parameters from as few as two such measurements. More precisely, given the measurements $v(t_0)$
and \( v(t_1) \) at two distinct time instants \( t_0 \) and \( t_1 \) where \( t_0 < t_1 \), there often exists a unique set of parameters \( k_r, k_p, \gamma_r, \gamma_p \) that could have produced these measurements. Furthermore, an implicit expression for these parameters has been derived and is presented in the supplemental material. We illustrate this for transcription only: Suppose the mRNA mean and variance are known at two times \( t_0 < t_1 < \infty \), and let these be represented by \((\mu_0, \sigma_0)\) and \((\mu_1, \sigma_1)\) respectively. Then the transcription parameters are fully identifiable, and

\[
\gamma_r = -\frac{1}{2} \log \left( \frac{\sigma_0^2 - \mu_1}{\sigma_1^2 - \mu_0} \right), \quad k_r = \gamma_r \frac{\mu_1 - \exp(-\gamma_r \tau) \mu_0}{1 - \exp(-\gamma_r \tau)},
\]

where \( \tau := t_1 - t_0 \).

Thus, first and second order statistics, measured at two distinct times, contain sufficient information for full identifiability of the model parameters.

Although parameters are identifiable from transient moment measurements, we propose that it is impossible to identify all parameters from stationary moments. Means, variances, and other statistics of cellular constituents are commonly measured in the steady-state after all the transients have died away. This presents a lost opportunity to peek into the cell's inner workings and to recover the network parameters—by the time the network has reached stationarity, its parameters are concealed. More succinctly, stationary moments are insufficient for full parameter identifiability. Figure 1(D and E) illustrate this result by plotting the protein and mRNA distributions at two different times and two different parameters sets. At 1000s the two parameter sets produce very different distributions for the protein levels, but at 5000s the two parameter sets produce distributions that are nearly indistinguishable. The supplemental materials provide a precise mathematical proof that indeed no matter the order of moments used, stationary moments of any order \( v_{\infty} = \lim_{t \to \infty} \left( \mathbb{E}[R(t)], \mathbb{E}[P(t)], \mathbb{E}[R^2(t)], \mathbb{E}[R(t)P(t)], \ldots, \mathbb{E}[R^N(t)P^N(t)] \right) \) are insufficient to uniquely identify the model parameters \( k_r, k_p, \gamma_r, \gamma_p \). In fact, there always exists an infinite set of parameters that yields the statistics \( v_{\infty} \).

The results reported above establish the principle that transient moments carry within them information that allows one to identify all model parameters. Achieving such identification using measurements at only two time instants assumes exact measurement of the statistics and that statistics of all the state variables are available. If one has either incomplete or inaccurate measurements of the statistics, it is often possible to identify model parameters by compensating for the loss of information through additional time measurements of the statistics. While inaccuracies in the moments due to measurement error can be reduced by using larger cell populations, additional time measurements of the same cell population can also achieve a reduction in identification error. Suppose \( v_j := v(t_j) \) has been measured at equally separated points in
Such measurements may have errors yielding $\hat{v}_j$. The collection of measurements $\hat{v}_j$ may be used to extract the best parameter set, in a least square sense. This is illustrated in Fig. 2 where

The plotted variable, $\Gamma$, is the median percent identification error per average percent measurement error. All data are separated by a time step period of 300s. All initial conditions are unknown and are simultaneously identified (not shown). (A) The identification is done with all five elements of the first two moments: $v = (E[R], E[P], E[R^2], E[P^2], E[RP])$. (B) Solid and dashed lines correspond to the identification with protein mean and second moment only ($E[P], E[P^2]$). Dotted lines correspond to identification with means of protein and mRNA only ($E[R], E[P]$). (C) The effectiveness of parameter identification from various noisy measurement sets taken at twenty time points separated by 1 min each.

The effect of measurement error on the accuracy of identified model parameters is shown. In general, utilizing more measurements contributes to improved numerical conditioning. Likewise, one can also include additional time measurements in order to compensate for incomplete measurability of the statistical moments. Suppose for example that it is difficult or impractical to measure mRNA statistics, and that one can only measure the cell to cell variability of proteins. In this case, with additional time measurements, the parameters and the unknown dynamics of the mRNAs can be resolved from the protein data alone. Of course with limited data availability and greater number of unknowns, one should expect that measurements noise will be a larger concern. This is indeed the case; Fig. 2B shows that the identification with protein data alone is more sensitive to measurement noise than the estimation with all five moments. It should also be noted that while all parameters can be identified simply from the mean values of mRNA and protein, this identification is much more sensitive to measurement noise (See Figs. 2A-C). Furthermore, with data on the
protein mean alone, it is impossible to simultaneously identify all four parameters and the unknown initial conditions.

**Identification of more complex networks.** The authors of [9] engineered a genetic toggle switch to be used as a sensor of environmental influences, such as radiation or external chemical signals. This switch is a construction of two genes, each of whose protein, λCI or LacI, inhibits the production of the other (see Fig. 3A). With exposure to Ultraviolet light (UV) or mitomycin C (MMC), the SOS pathway results in RecA coproteases, which increase the degradation rate of λCI. As a result, different amounts of UV or MMC change the trade-off between λCI and LacI molecules. The output of the mechanism is GFP, which is assumed to be expressed at the same level as LacI. Depending upon environmental conditions, the system exhibits a bias toward one phenotype or another. Various models have been proposed to describe this system, including a deterministic model [9] as well as a stochastic version [10]. This study considers a stochastic model similar to that in [10] and aims to identify kinetic parameters from the experimental data of [9] (see reproduced histograms in Fig. 3(B-D)). The model consists of four basic reactions:

\[
\begin{align*}
R_1 & \quad \delta_0 \\
R_2 & \quad u \rightarrow \delta \\
R_3 & \quad v \rightarrow \delta \\
R_4 & \quad v \rightarrow \delta 
\end{align*}
\]

and the rates of these reactions, \( w(u, v, \Lambda) = [w_1(u, v, \Lambda), \ldots, w_4(u, v, \Lambda)] \) depend upon the populations of the proteins \( u \) and \( v \) as well as a set of unknown parameters, \( \Lambda = [(k_{12}, k_{21}, k_{22}, \delta_u, \delta_v)] \), according to:

\[
\begin{align*}
w_1 &= k_{11} + \frac{k_{12}}{1 + k_{13}v^3}; \\
w_2 &= \delta_uu; \\
w_3 &= k_{21} + \frac{k_{22}}{1 + k_{23}u^3}; \\
w_4 &= \delta_vv,
\end{align*}
\]

where \( u \) and \( v \) are used to refer to λCI and LacI, respectively. In the model, the λCI degradation parameter, \( \delta_u \), has been allowed to vary as an unknown function of the applied UV or MMC. Fig. 3(B-D) show distributions of LacI at 3 different UV radiation levels: 0, 6, and 12 and \( J/m^2 \). Each of these levels is assumed to correspond to a different value of \( \delta_u \). The remaining parameters are the same for all three figures.

The object of this study is to identify the parameters \( \Lambda \) from experimental data in Fig. 3(B-D) taken from [9]. These distributions are far more complicated than those in the previous transcription/translation example and cannot adequately be captured except with a large number of high order moments. Furthermore, since the propensity functions, \( w_1 \) and \( w_3 \) are nonlinear functions of the populations of LacI and λCI, there are no closed expressions for these moments. Instead, we must solve the chemical master equation (CME) [13]–an infinite dimensional system of linear ordinary differential equations, 

\[
\dot{P}(t, \Lambda) = A(\Lambda) \cdot P(t, \Lambda),
\]

which describes the evolution of probabilities for every possible combination of LacI and λCI populations and the parameter set \( \Lambda \). The CME has no known solution for this system, but our recently developed
finite state projection approach [11] allows us to approximate its solution within any desired degree of accuracy. That is for any \( \varepsilon > 0 \) we can systematically find a finite projection system \( \mathbf{P}^{FSP}(t, \Lambda) = \mathbf{A}_J(\Lambda) \cdot \mathbf{P}^{FSP}(t, \Lambda) \) such that:

\[
\left\| \begin{bmatrix} \mathbf{P}_J(t, \Lambda) \\ \mathbf{P}_J'(t, \Lambda) \end{bmatrix} - \begin{bmatrix} \mathbf{P}^{FSP}(t, \Lambda) \\ 0 \end{bmatrix} \right\|_1 \leq \varepsilon, \text{ and } \mathbf{P}^{FSP}(0, \Lambda) = \mathbf{P}_J(0, \Lambda),
\]

where the index vector \( J \) denotes the set of states included in the projection, \( \mathbf{P}_J \) is the corresponding probabilities of those states, and \( \mathbf{A}_J \) is the corresponding principle submatrix of \( \mathbf{A} \) [11, 12]. With the FSP solution approach in hand, the identification procedure is relatively simple—we find the parameter arguments, \( \Lambda^* \), that minimizes the one norm difference between the measured distribution \( \mathbf{P}(t) \) and the numerical solution of that distribution:

\[
\Lambda^* := \text{argmin}_\Lambda \left\| \mathbf{P}(t) - \mathbf{P}(t, \Lambda) \right\|_1.
\]

An initial guess is made for the parameters, and then this set is updated iteratively using both gradient and non-gradient based searches until the analytical distribution matches the experimental distribution as closely as possible.

Fig. 3 shows that the identified model matches the experimental measured distributions from [9] fairly well. In addition to matching the effect of UV radiation on the switch, the same parameters provide an equally good match for the effect of MMC concentration (see Fig. 3E-G), even though this data has not been used in the identification procedure. It is natural at this point to ask if the identified parameters are unique. In fact they are not. As can be seen in Fig. 3 B-D equally good model fits can be achieved with more than one parameter set (see also Parameter sets 2 and 3 of Table 1 in supplemental material). Although the identification has drastically narrowed the space of possible parameter sets, it has not yet reduced that space to a unique point. This lack of uniqueness suggests that the data is not sufficiently rich for complete identifiability. Indeed the data used in this identification includes only information on the marginal distribution of LacI. The question arises whether adding information about the distribution of λcI would yield better identifiability. This is suggested by the observation that the two parameter sets which achieve an equally good match of the LacI distributions produce different distributions of λcI (see Fig. 4). It is then natural to use both the λcI and LacI distributions in the identification procedure. Since λcI is not available, we have used a single parameter set (Parameter Set 1 in Table 1) to numerically simulate these distributions. The identification procedure was then conducted on this simulated data.

Using the simulated data as just described, we found that using the joint distributions of LacI and λcI at 0 and 50 seconds yielded a unique parameter set and was therefore sufficient for full parameter
Figure 3: (A) Basic schematic of the toggle model comprised of two inhibitors: λcI inhibits the production of Lacl and vice-versa. In the model, the synthesis rates of λcI and Lacl are non-linear functions of their counterparts. Environmental influences (UV radiation and MMC) increase the degradation rate of λcI and affect the tradeoff between the two regulators. Toggle switch Lacl distributions in various conditions of DNA damage. (B-G) In B-D the histograms are experimentally measured data from [9], and the blue and red dashed lines correspond to model fits with parameters sets 1 and 2 from Table 1 under three different levels of UV radiation. The degradation parameter $\delta u$ is the only parameter that changes between the six figures. The identification of the remaining 7 parameters is conducted using only B-D.

Identifiability. As for the case of gene transcription and translation, transient measurements are crucial in the identification of the toggle parameters. For example, when the identification was carried out with measurements at 0 and 1000 seconds, the identified parameters were no longer unique, as some of the dynamics had already decayed by the time of the second measurement.

We have thus demonstrated that stochastic fluctuations in cellular species induced by noise carry valuable identifying information about the underlying network. The statistical information contained in these fluctuations when properly extracted and processed is sufficiently rich as to enable the unique identification of all the network parameters. In principle, this can be accomplished when accurate distributions taken at two distinct time points are available. More time points are needed if the distributions are noisy, but the idea remains the same. Such distributions are increasingly available as single cell measurement technology attains wider use. Hence, the proposed exploitation of cellular noise establishes a potentially powerful approach for the identification of gene networks and offers a new window into the workings of these networks.
Figure 4: Comparison of the probability distributions for LaCl and λcI with the same two parameter sets as in Fig. 3 and no UV radiation. (A) Although the marginal distributions for LaCl are nearly indistinguishable in Fig. 3, there are noticeable differences in the marginal distribution of λcI and the joint distributions in B.

References


1 Supplementary materials.

Implicit expression for transcription and translation.

In this supplemental section, we derive explicit expressions for the evolution of the first two moments in the simple gene transcription and translation process. For this derivation, let $R$ denote the population of mRNA molecules, and let $P$ denote the population of proteins in the system. As above, these populations change through four reactions:

\[
\begin{align*}
\emptyset &\rightarrow R \\
R &\rightarrow \emptyset \\
R &\rightarrow R + P \\
P &\rightarrow \emptyset
\end{align*}
\]

for which the propensity functions (or stochastic reaction rates) are

\[
\begin{align*}
w_1 &= k_r + k_{21}P, \\
w_2 &= \gamma_r R, \\
w_3 &= k_p R, \quad \text{and} \\
w_4 &= \gamma_p P.
\end{align*}
\]

Here the term $k_{21}$ corresponds to a feedback effect that the protein is assumed to have on the transcription process. In positive feedback, $k_{21} > 0$, the protein increases transcription; in negative feedback, $k_{21} < 0$, the protein inhibits transcription. For the results in the main text, this feedback term has been set to zero.

The master equation [13] for this system can be written:

\[
\begin{align*}
\dot{P}_{i,j}(t) &= -(k_r + k_{21}j + \gamma_r i + k_p i + \gamma_p j)P_{i,j}(t) \\
&\quad + (k + k_{21}j)P_{i-1,j}(t) \\
&\quad + \gamma(i + 1)P_{i+1,j}(t) \\
&\quad + k_p i P_{i,j-1}(t) \\
&\quad + \gamma_p (j + 1)P_{i,j+1}(t),
\end{align*}
\]

where $P_{i,j}(t)$ is the probability that $(R, P) = (i, j)$ at the time $t$, conditioned on some initial probability distribution $P(t_0)$. In this expression, the first negative term corresponds to the probability of transitions that begin at the state $(R, P) = (i, j)$ and leave to another state, and the remaining positive terms corre-
spond to the reactions that begin at some other state \((R, P) \neq (i, j)\) and transition into the state \((i, j)\).

The mean populations of mRNA and protein molecules can be written as:

\[
v_1(t) = E\{R\} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i P_{i,j}(t)
\]

\[
v_2(t) = E\{P\} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} j P_{i,j}(t).
\]

The derivatives of these mean values are found simply by substituting (1) into (2):

\[
\dot{v}_1(t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i \dot{P}_{i,j}(t) = k_r + k_{21} v_2 - \gamma_r v_1,
\]

and

\[
\dot{v}_2(t) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} j \dot{P}_{i,j}(t) = k_p v_1 - \gamma_p v_2.
\]

Similarly, expressions for the second uncentered moments can be written:

\[
v_3 = E\{RR\} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i^2 P_{i,j},
\]

\[
v_4 = E\{PP\} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} j^2 P_{i,j},
\]

\[
v_5 = E\{RP\} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} ij P_{i,j},
\]

and evolve according to the set of ordinary differential equations:

\[
\dot{v}_3 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i^2 \dot{P}_{i,j}(t)
\]

\[= k_r + (2k_r + \gamma_r)v_1 - 2\gamma_r v_3 + k_{21} v_2 + 2k_{21} v_5,\]

\[
\dot{v}_4 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} j^2 \dot{P}_{i,j}
\]

\[= k_p v_1 + \gamma_p v_2 - 2\gamma_p v_4 + 2k_p v_5,\]

\[
\dot{v}_5 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} ij \dot{P}_{i,j}
\]

\[= k_p v_3 + k_r v_2 + k_{21} v_4 - (\gamma_r + \gamma_p) v_5.\]
Altogether the various components of the first two moments,

\[ v(t) := \begin{bmatrix} E\{R\} & E\{RR\} & E\{P\} & E\{PP\} & E\{RP\} \end{bmatrix}^T, \]

evolve according to the linear time invariant ODE:

\[
\dot{v} = \begin{bmatrix}
-\gamma_r & k_{21} & 0 & 0 & 0 \\
-k_p & -\gamma_p & 0 & 0 & 0 \\
\gamma_r + 2k_r & k_{21} & -2\gamma_r & 0 & 2k_{21} \\
k_p & \gamma_p & 0 & -2\gamma_p & 2k_p \\
0 & k_r & k_p & k_{21} & -\gamma_r - \gamma_p
\end{bmatrix} v + \begin{bmatrix} k_r \\
k_r \\
k_r \\
0 \\
0
\end{bmatrix}
\]

= \mathbf{A}v + \mathbf{b} \quad (4)

These expressions now fully characterize the dynamics of the first two moments of mRNA and protein molecules. With these expressions one can now begin to identify the various parameters: \([k_r, \gamma_r, k_p, \gamma_p, k_{21}]\) from properly chosen experimental data sets.

**Non-Identifiability from Stationary Distributions**

In this supplemental section, we show conclusively that the parameters of the transcription/translation model cannot be identified from invariant distributions alone. Suppose that the moments of the probability distribution described in (4) has an invariant distribution:

\[ v_\infty = \lim_{t\to\infty} [v_1, v_2, v_3, v_4, v_5]^T. \]

These steady state moments must satisfy the expression:

\[ \mathbf{A}v_\infty - \mathbf{b} = 0, \quad (5) \]

which can be rewritten in terms of the unknown parameters as:

\[ \Psi_\infty \Lambda = \lim_{t\to\infty} \Psi(t) \Lambda = 0, \]
where

\[
\Psi(t) = \begin{bmatrix}
1 & -v_1 & 0 & 0 & v_2 \\
1+2v_1 & v_1 - 2v_3 & 0 & 0 & v_2 + 2v_5 \\
0 & 0 & v_3 & -v_2 & 0 \\
0 & 0 & v_1 + 2v_5 & v_2 - 2v_4 & 0 \\
v_2 & -v_5 & v_3 & -v_5 & v_4
\end{bmatrix}
\]

In Eqn. (5) there are two possible cases: (1) the rank of the matrix is full and we are left with the trivial solution \( \Lambda = 0 \), or (2) the matrix has a null-space spanned by \( \{\phi_1, \ldots, \phi_p\} \) and there are an infinite number of parameter sets that will result in the same invariant distribution:

\[
\Lambda = \sum_{i=1}^{p} \alpha_i \phi_i, \text{ for any } [\alpha_1, \ldots, \alpha_p] \in \mathbb{R}^p.
\]

So long as the parameters enter linearly into the propensity functions \( w(x) = \sum_{i=1}^{M} c_i f(x) \), then one can extend this argument for any finite number of \( n \) moments of the stationary distribution. This tells us that the steady state distribution cannot provide enough information to uniquely identify the set of system parameters. Additional information is needed. For example, if the rank of the null space is one, then the knowledge of any one parameter from the set \( \Lambda \) can provide an additional linearly independent equation, and can enable the unique determination of the parameters. If the rank of the null space is \( p \), then at least \( p \) additional, linearly independent, pieces of information will be required.

**Implicit Expressions for the Identification of Transcription and Translation Parameters from Transient Data**

In this supplemental section, we show how one can obtain an implicit analytical expression for transcription and translation parameters in the absence of feedback \( (k_{12} = 0) \). For this we define the following variables:

\[
\begin{bmatrix}
z_1(t) \\
z_2(t) \\
z_3(t) \\
z_4(t)
\end{bmatrix} =
\begin{bmatrix}
\mu_r \\
\sigma_{rr} - \mu_r \\
\mu_p \\
\sigma_{rp}
\end{bmatrix}
= 
\begin{bmatrix}
v_1 \\
v_3 - v_1^2 - v_1 \\
v_2 \\
v_5 - v_1v_2
\end{bmatrix}.
\]

These can be shown to evolve according to the linear ODE:

\[
\frac{dz(t)}{dt} =
\begin{bmatrix}
-\gamma_r & 0 & 0 & 0 \\
0 & -2\gamma_r & 0 & 0 \\
k_p & 0 & -\gamma_p & 0 \\
k_p & k_p & 0 & -(\gamma_r + \gamma_p)
\end{bmatrix}
\begin{bmatrix}
z_1(t) \\
z_2(t) \\
z_3(t) \\
z_4(t)
\end{bmatrix} +
\begin{bmatrix}
k_r \\
0 \\
0 \\
0
\end{bmatrix}
\]
The first two equations yield \( k_r \) and \( \gamma_r \) as discussed above. With these, one can solve for the \( z_1(t) \) and \( z_2(t) \):

\[
z_1(t) = e^{-\gamma_r(t-t_1)} z_1(t_1) + \frac{k_r}{\gamma_c} \left( 1 - e^{-\gamma_r(t-t_1)} \right),
\]

\[
z_2(t) = e^{-2\gamma_r(t-t_1)} z_2(t_1),
\]

and plug these expressions into the third and fourth equations. This gives the following expressions for the solution:

\[
\begin{bmatrix}
  z_3(t_2) \\
  z_4(t_2)
\end{bmatrix} =
\begin{bmatrix}
  e^{-\gamma_p(t_2-t_1)} z_3(t_1) + k_p \int_{t_1}^{t_2} e^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau \\
  e^{-(\gamma_r+\gamma_p)(t_2-t_1)} z_4(t_1) + k_p \int_{t_1}^{t_2} e^{-(\gamma_r+\gamma_p)(t_2-\tau)} (z_1(\tau) + z_2(\tau)) d\tau.
\end{bmatrix}
\]

One can combine many of the known quantities to gather a simpler expression

\[
\begin{bmatrix}
  z_3(t_2) \\
  z_4(t_2)
\end{bmatrix} =
\begin{bmatrix}
  e^{-\gamma_p(t_2-t_1)} z_3(t_1) + k_p \int_{t_1}^{t_2} e^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau \\
  e^{-\gamma_p(t_2-t_1)} C_1 + k_p \int_{t_1}^{t_2} e^{-\gamma_p(t_2-\tau)} C_2(\tau) d\tau.
\end{bmatrix}
\]

where

\[
C_1 = e^{-\gamma_r(t_2-t_1)} z_4(t_1), \quad \text{and}
\]

\[
C_2(\tau) = e^{-\gamma_r(t_2-\tau)} (z_1(\tau) + z_2(\tau))
\]

are known expressions. Solving the first expression in terms of \( k_p \) and substituting that expression into the second yields the implicit expression for \( \gamma_p \):

\[
z_4(t_2) = e^{-\gamma_p(t_2-t_1)} C_1 + \frac{z_3(t_2) - e^{-\gamma_p(t_2-t_1)} z_3(t_1)}{\int_{t_1}^{t_2} e^{-\gamma_p(t_2-\tau)} z_1(\tau) d\tau} \int_{t_1}^{t_2} e^{-\gamma_p(t_2-\tau)} C_2(\tau) d\tau.
\]

An explicit expression for \( \gamma_p \) does not appear to be immediately obvious. However, by substituting in the known expressions for \( C_1, C_2(\tau), \) and \( z_3(t_1) \), one can easily plot the the left hand side of this expression as a function of \( \gamma_p \). For example, consider the system with the parameter set:

\[
\Lambda =
\begin{bmatrix}
  k_r \\
  \gamma_r \\
  k_p \\
  \gamma_p
\end{bmatrix} =
\begin{bmatrix}
  0.05 \\
  0.005 \\
  0.05 \\
  0.001
\end{bmatrix}.
\]
Figure 5: Implicit determination of the protein degradation rate. The red curve corresponds to the right hand side of Eqn. 7 versus $\gamma_p$. The horizontal dashed line corresponds to the measured value of $z_4(t_2)$. From the figure, one can correctly determine that $\gamma_p = 0.001$ (vertical dashed line).

and the initial condition at $t_1 = 0$ of

$$z(t_1) = \begin{bmatrix} z_1(t_1) \\ z_2(t_1) \\ z_3(t_1) \\ z_4(t_1) \end{bmatrix} = \begin{bmatrix} 1 \\ 0.2 \\ 10 \\ 5 \end{bmatrix}$$

The corresponding response at $t_2 = 100s$ is

$$z(t_2) = \begin{bmatrix} z_1(t_2) \\ z_2(t_2) \\ z_3(t_2) \\ z_4(t_2) \end{bmatrix} = \begin{bmatrix} 4.541 \\ 0.07358 \\ 23.07 \\ 14.82 \end{bmatrix}$$

Assuming that the quantities $z(t_1)$ and $z(t_2)$ are known exactly, then it is relatively easy to identify the first two parameters $k_r$ and $k_p$ and substitute these into the expression (7). This expression can then be plotted as a function of the unknown $\gamma_p$ as shown in Figure 5. The value of $\gamma_p$ is the value at which the expression crosses the measured value for $z_4(t_2)$, which can be found using a simple line search. From the figure it is obvious that this intersection does indeed correspond to the correct value of $\gamma_p = 0.001$. Once $k_r$, $\gamma_r$, and $\gamma_p$ are all known, it is simple to solve for $k_p$ using (6).
Supplemental Material on the Identification of the Toggle Switch

In the main text, we attempted to identify the parameters from the distribution of LacI at a single point in time. We found that there are many possible parameter sets that will succeed in matching this distribution but which have very different distributions for LacI. As a numerical experiment, we have used Parameters Set I to generate the full joint distribution of LacI and LacI, which we then use as the target distribution in the identification procedure.

Using the full distribution at a time of 1000s allowed for a reasonably close identification of the ten parameters (see Parameter set 3 in Table 1), but the identification is still not unique. In this case, closer examination reveals that there is much stronger correlation between the parameters. In particular, the values for all of the production and degradation parameters \( k_{12}, k_{21}, k_{22}, \delta_{u}, \delta_{u}(0), \delta_{u}(6), \delta_{u}(12) \) are all a constant factor of 1.19 removed from the correct parameters. Thus, we have uniquely established the ratio between all of these parameters but not their exact numbers. This suggests that the identification is very close, and may be complete with a little additional information.

Closer examination of the model with Parameter Set 1 reveals that at least some of the transient modes have died out on a time scale less than the chosen 1000s. This can be seen readily by comparing the distributions computed with Parameter Sets I and 3 at different times. In Fig. 6 the marginal distribution of LacI at different times as computed from Set 1 (solid blue line) and Set 3 (dashed red line). For very short times of 1 or 10s (top two rows), these distributions are distinguishable from one another. However, after a short transient time of 100s, the two distributions are indistinguishable (compare bottom three rows). In essence, conducting the identification at 1000s is effectively the same as identifying the system after it has already reached some lower dimensional manifold. As was the case above, we discover that it is impossible

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Set 1</th>
<th>Set 2: ( f_u ) at 10^5s</th>
<th>Set 3: ( f_{u,v} ) at t=10^5s</th>
<th>Set 4: ( f_{u,v} ) at t=50s</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k_{11} )</td>
<td>0.0099</td>
<td>0.0089 (0.90)</td>
<td>0.011 (1.13)</td>
<td>0.0098 (0.996)</td>
</tr>
<tr>
<td>( k_{12} )</td>
<td>40.3</td>
<td>27.5 (0.68)</td>
<td>48.1 (1.19)</td>
<td>40.2 (0.998)</td>
</tr>
<tr>
<td>( k_{21} )</td>
<td>1.35</td>
<td>2.49 (1.84)</td>
<td>1.62 (1.19)</td>
<td>1.35 (0.998)</td>
</tr>
<tr>
<td>( k_{22} )</td>
<td>10.1</td>
<td>18.6 (1.84)</td>
<td>12.1 (1.19)</td>
<td>10.1 (0.998)</td>
</tr>
<tr>
<td>( \delta_u )</td>
<td>0.26</td>
<td>0.47 (1.84)</td>
<td>0.31 (1.19)</td>
<td>0.26 (0.998)</td>
</tr>
<tr>
<td>( \delta_{u}(0) )</td>
<td>0.53</td>
<td>0.40 (0.75)</td>
<td>0.64 (1.19)</td>
<td>0.53 (0.998)</td>
</tr>
<tr>
<td>( \delta_{u}(6) )</td>
<td>0.77</td>
<td>0.57 (0.74)</td>
<td>0.91 (1.19)</td>
<td>0.77 (0.998)</td>
</tr>
<tr>
<td>( \delta_{u}(12) )</td>
<td>1.19</td>
<td>1.46 (1.23)</td>
<td>1.42 (1.19)</td>
<td>1.19 (0.998)</td>
</tr>
<tr>
<td>( k_{13} )</td>
<td>0.0025</td>
<td>0.0040 (1.61)</td>
<td>0.0025 (1.0)</td>
<td>0.0025 (1.00)</td>
</tr>
<tr>
<td>( k_{23} )</td>
<td>0.0084</td>
<td>0.025 (2.93)</td>
<td>0.009 (1.10)</td>
<td>0.0084 (0.998)</td>
</tr>
</tbody>
</table>

Table 1: Four parameter sets that yield the same distribution for LacI for the stochastic toggle model. Parameter Set 1 has been identified directly from the experimental data using a time of \( t = 1000s \). Parameter set 2 is identified from the distribution of LacI at \( t = 1000s \). Parameter set 3 is identified from the full joint LacI and LacI distribution at \( t = 1000s \). Parameter set 4 is identified from the full distribution at \( t = 50s \). For parameter sets 2-4, the values in the parenthesis denote the ratios of the identified parameters to the "true" parameters of Set 1.
Figure 6: Comparison of the marginal distribution of λcI at different UV levels of {0, 6, 12} J/m² at different points in time {1, 10, 100, 10³, 10⁴} s for two different sets of parameters. Parameter Set 1 corresponds to the solid blue lines, and Set 3 corresponds to the dashed red line. Note that the two sets produce results that match very well after a short transient time.

To uniquely identify all parameters from insufficiently rich dynamics. However, if instead we attempt the identification at a shorter time step of 50s before these dynamics have fully decayed, then the identification becomes possible, and we are able to uniquely identify every parameter each within an error of 0.5% (See parameter set 4 in Table 1).