Biophysical Society

Building Predictive Models of Genetic Circuits Using the Principle of Maximum Caliber

Taylor Firman,[1] Gábor Balázsi,[2,3] and Kingshuk Ghosh[1,*]
[1]Department of Physics and Astronomy, Molecular and Cellular Biophysics, University of Denver, Denver, Colorado; [2]The Louis and Beatrice Laufer Center for Physical and Quantitative Biology and [3]Department of Biomedical Engineering, Stony Brook University, Stony Brook, New York

ABSTRACT Learning the underlying details of a gene network is a major challenge in cellular and synthetic biology. We address this challenge by building a chemical kinetic model that utilizes information encoded in the stochastic protein expression trajectories typically measured in experiments. The applicability of the proposed method is demonstrated in an auto-activating genetic circuit, a common motif in natural and synthetic gene networks. Our approach is based on the principle of maximum caliber (MaxCal)—a dynamical analog of the principle of maximum entropy—and builds a minimal model using only three constraints: 1) protein synthesis, 2) protein degradation, and 3) positive feedback. The MaxCal-generated model (described with four parameters) was benchmarked against synthetic data generated using a Gillespie algorithm on a known reaction network (with seven parameters). MaxCal accurately predicts underlying rate parameters of protein synthesis and degradation as well as experimental observables such as protein number and dwell-time distributions. Furthermore, MaxCal yields an effective feedback parameter that can be useful for circuit design. We also extend our methodology and demonstrate how to analyze trajectories that are not in protein numbers but in arbitrary fluorescence units, a more typical condition in experiments. This "top-down" methodology based on minimal information—in contrast to traditional "bottom-up" approaches that require ad hoc knowledge of circuit details—provides a powerful tool to accurately infer underlying details of feedback circuits that are not otherwise visible in experiments and to help guide circuit design.

INTRODUCTION

Biological function is largely dictated by gene networks that control protein expression in single cells. Understanding details of these networks and consequently building quantitative models is essential to control gene expression and ultimately regulate cellular dynamics. However, model development has been limited due to the lack of information about the complex web of interactions (including feedback regulation) that defines these networks. Typical experiments only provide partial information by measuring the expression levels of one or two proteins of interest using fluorescent tags, much less than the actual number of entities (mRNAs, promoters, nucleotides, and amino acids) involved in the process of gene expression. This problem of partial information is a key challenge for model building. Although the number of species monitored is limited, experimental read-outs contain crucial information, as they record the entire time trajectory of fluctuating protein expression levels. The stochastic nature of the trajectories is due to small copy numbers of molecules involved in these reactions (1–9). The details of noise statistics encode the details of network architecture. This provides a potentially useful avenue for inferring details of network architecture by analyzing noisy protein expression levels (10–15). Despite realizing the power of this approach (10,12,14,15), such efforts are still in their infancy. Existing models are either too simple, with limited single-cell-level predictive power, or too detailed, requiring too many unknown parameters (16). The most common stochastic approaches first define sets of reaction networks to be simulated using a Gillespie algorithm (17) or related methods and then fit different observables to determine the corresponding reaction rate parameters. A major drawback of these methods is that they are "bottom-up" and require detailed knowledge of the underlying reaction network. This is particularly challenging when networks involve feedback, a common feature in many natural networks and synthetic biology. It is currently impossible to test many of these ad hoc assumptions independently. Furthermore, these approaches can involve too many parameters that can fit the same data with multiple models, creating additional challenges for efficient parameter estimation (11). The challenge of having too many

Submitted May 30, 2017, and accepted for publication August 31, 2017.
*Correspondence: kingshuk.ghosh@du.edu
Editor: Nathalie Balaban.
https://doi.org/10.1016/j.bpj.2017.08.057

parameters is also problematic for circuit design (18–30), as it requires ways to efficiently explore parameter space to test different models, thus demanding models with the least possible number of parameters.

To circumvent these obstacles, we propose a “top-down” approach for modeling these networks. We use the principle of maximum caliber (MaxCal) to model stochastic trajectories with minimal information. We show the application of MaxCal on a simple auto-activating circuit, a common motif in many biological circuits (31). MaxCal maximizes path entropy subject to constraints, similar to maximum entropy on state space, and directly works with path trajectories. This makes MaxCal directly applicable to experimentally measured time trajectories of protein numbers. We establish the methodology on synthetic data generated using Gillespie simulations (17) of a known auto-activating circuit. These trajectory data serve as the input data—a proxy for experimental data—for MaxCal. The minimal model of MaxCal is then applied to the raw trajectory statistics in conjunction with maximum likelihood (ML) to determine representative parameters for the model. These parameters can predict other statistics of the data and quantitatively infer several underlying physical variables that are not visible otherwise. In the next section, we first describe the synthetic circuit and generation of in silico data that mimic experimental data. Next, we introduce MaxCal and its specific application to the circuit. We show how MaxCal, along with ML, can be used to infer model parameters and make predictions. Comparing these predictions against the known model allows us to benchmark the predictive capabilities of MaxCal. Finally, we discuss how the methodology can be applied when the input data are not in protein number but in arbitrary fluorescence, a common challenge in interpreting experimental data.

MATERIALS AND METHODS

Generating synthetic data for an auto-activating circuit

Considering the complexity of natural networks with many unknown or incompletely understood interactions, synthetic biologists are building mimics of frequently occurring parts of bigger networks, called network motifs (32–34). One natural network motif with important biological function that has inspired the design of many synthetic gene circuits is feedback regulation. Our previous work (35) has demonstrated the application of MaxCal on double-negative (overall positive) feedback circuits, where two genes mutually repress each other, commonly referred to as a toggle-switch circuit (36,37). Here, we consider a positive feedback circuit where a single gene auto-activates itself. As a proof of concept, we apply MaxCal to synthetic data generated in silico using a model for which the underlying parameters are known. This will serve as a proxy for experimental data and provide us with a gold standard to which we can compare when demonstrating how well MaxCal performs given stochastic trajectories. Among several models of auto-activation in different biological contexts (38–45), we adopt the one below (Eq. 1), studied by Kepler and Elston (46), to generate stochastic synthetic data that will serve to mimic experimental time traces:

$$\alpha \xrightarrow{g} \alpha + A \;;\; A \xrightarrow{r} \emptyset \;;\; A + A \underset{b_\mathrm{d}}{\overset{f_\mathrm{d}}{\rightleftarrows}} A_2 \qquad (1)$$

$$\alpha + A_2 \underset{b_\mathrm{p}}{\overset{f_\mathrm{p}}{\rightleftarrows}} \alpha^*; \quad \alpha^* \xrightarrow{g^*} \alpha^* + A.$$

In this scheme, some generic protein, A, is created from its corresponding gene, α, at a rate g, degrades at a rate r, and dimerizes into A_2 with forward and backward rates f_d and b_d, respectively. A_2 can then bind and unbind to the promoter site of α at rates f_p and b_p, respectively, sending α into or out of its activated state, α^*. In this activated state, α^* creates protein A at a much faster rate, g^*, capturing the essentials of a positive-feedback mechanism. Rates are chosen to produce switching times that are representative of experiments (31) while maintaining protein synthesis and degradation rates in the realm of typical rates (47). A Gillespie algorithm (17) was used to generate stochastic trajectories of protein (A) levels as shown in Fig. 1 A. Three major features are worth noting: 1) two clearly separated high and low states, 2) a large amount of fluctuation within each state, and 3) stochastic switching between the two states. In the next section, we first attempt to reproduce these three basic features in MaxCal using as simple a framework as possible.

MaxCal model for auto-activating circuit

Maximum caliber is a variational principle that gives a prescription for inferring dynamics by maximizing the path entropy (35,48–58), or caliber, subject to known constraints enforced via Lagrange multipliers. For the gene circuit of interest, there are three minimal constraints that must be in

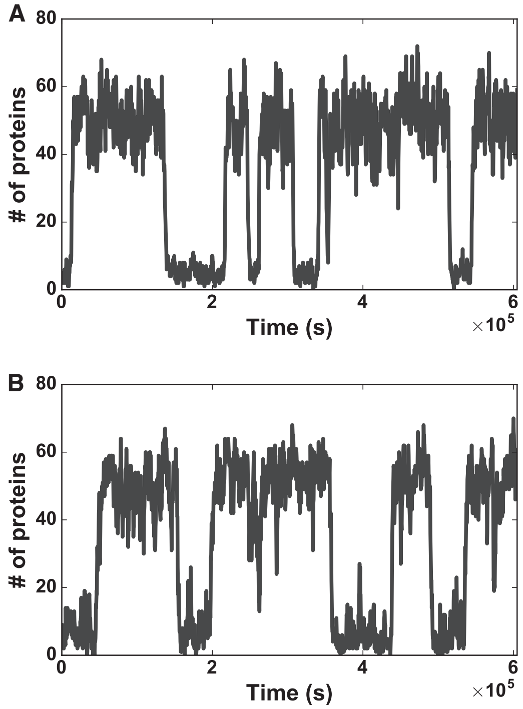

FIGURE 1 Positive feedback circuit. (A) Typical time trace of the number of proteins in a self-promotion circuit using the reaction scheme in Eq. 1 ($g = 5.0 \times 10^{-3}$ s^{-1}, $g^* = 50.0 \times 10^{-3}$ s^{-1}, $r = 1.0 \times 10^{-3}$ s^{-1}, $f_\mathrm{d} = 5.0 \times 10^{-3}$ s^{-1}, $b_\mathrm{d} = 50$ s^{-1}, $f_\mathrm{p} = 6.0 \times 10^{-3}$ s^{-1}, $b_\mathrm{p} = 3.0 \times 10^{-5}$ s^{-1}, assuming the intrinsic time unit is seconds). Data are recorded every 300 s. (B) Typical time trace of the number of proteins in the minimal model of self-promotion using MaxCal ($h_\alpha = -0.512$, $h_A = 0.585$, $K_A = 0.0298$, $M = 15$, $\Delta t = 300$ s).

place: 1) protein synthesis, 2) protein degradation, and 3) auto-activation/positive feedback. We enforce the first two by restricting the average number of proteins that are created in a discrete time interval (Δt) as well as the average number of proteins that are destroyed (35,55). To do this, we define ℓ_α as the production-state variable, which describes the number of proteins that are created in the time interval and ranges as integer values between zero and some predefined maximal value (M), i.e., $0 \le \ell_\alpha \le M$. We also define ℓ_A as the degradation-state variable, which describes the number of previously existing proteins that still exist at the end of the time interval. Clearly, ℓ_A ranges as integer values between zero and the number of proteins present at the beginning of the time interval (N_A), i.e., $0 \le \ell_A \le N_A$. The corresponding Lagrange multipliers for these two constraints are h_α and h_A, and the probability of observing a particular combination of ℓ_α and ℓ_A is defined as P_{ℓ_α,ℓ_A}. Next, we implement the constraint of positive feedback, the idea that a high number of proteins (N_A) should positively correlate with the production of A. This is done by introducing a third Lagrange multiplier, K_A, that enforces a coupling between protein production and the presence of proteins by constraining the average of $\ell_\alpha \ell_A$. This is the lowest-order term in the coupling of these two variables that must be imposed to capture the essence of feedback. Similar arguments were used to build models to describe negative feedback in toggle-switch circuitry (35). The four basic ingredients of the model, described above, yield the caliber as

$$C = -\sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} P_{\ell_\alpha,\ell_A}\log P_{\ell_\alpha,\ell_A} + h_\alpha \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_\alpha P_{\ell_\alpha,\ell_A} + h_A \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_A P_{\ell_\alpha,\ell_A} + K_A \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_\alpha \ell_A P_{\ell_\alpha,\ell_A}, \quad (2)$$

and the corresponding caliber-maximized path probabilities are

$$\begin{aligned} P_{\ell_\alpha,\ell_A} &= Q^{-1}\binom{N_A}{\ell_A}\exp(h_\alpha \ell_\alpha + h_A \ell_A + K_A \ell_\alpha \ell_A) \ ; \\ Q &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A}\binom{N_A}{\ell_A}\exp(h_\alpha \ell_\alpha + h_A \ell_A + K_A \ell_\alpha \ell_A). \end{aligned} \quad (3)$$

Using this path probability distribution, stochastic trajectories are generated using a Monte Carlo method to select a path for each time point. The system then creates and destroys the number of proteins corresponding to the ℓ_α and ℓ_A of the selected path and the time of the system advances by the predetermined Δt. A quick search of the parameter phase space (h_α, h_A, K_A, M) reveals that even with just these four parameters, bimodal behaviors like the ones seen in the self-promotion circuit of the previous section (characterized with seven parameters in Fig. 1 *A*) can be reproduced (see Fig. 1 *B*). For efficient computation, protein number probability distributions are generated using the method of finite-state projection (FSP) of Munsky and Khammash (59). This method is needed to provide a systematic way to truncate the infinite phase space of possible states, since protein number does not have an upper bound. FSP provides a rigorous self-consistent approach to ensure that the truncation error is within a pre-determined error bound (see Supporting Material for exact application).

Furthermore, the state variables ℓ_α and ℓ_A directly relate to effective protein synthesis and degradation rates analogous to g, g^*, and r in the auto-activation circuit (Eq. 1). Specifically,

$$\begin{aligned} g &= \frac{\langle \ell_\alpha \rangle_{N_A=N_L}}{\Delta t} \quad , \quad g^* = \frac{\langle \ell_\alpha \rangle_{N_A=N_H}}{\Delta t} \quad , \\ r(N) &= \frac{N - \langle \ell_A \rangle_{N_A=N}}{N\Delta t} \quad , \quad r = \sum_N P(N) r(N), \end{aligned} \quad (4)$$

where N_L and N_H are the peak values of the number of proteins in the low and high states, respectively, and $P(N)$ is the probability of having N proteins within the system (calculated via FSP).

An additional metric that could be of interest in genetic circuit design is the effective feedback metric, F. We define F as the average Pearson correlation coefficient between ℓ_α and ℓ_A:

$$F = \sum_N P(N)\frac{\langle \ell_\alpha \ell_A \rangle - \langle \ell_\alpha \rangle \langle \ell_A \rangle}{\sigma_{\ell_\alpha}\sigma_{\ell_A}}, \quad (5)$$

where σ_{ℓ_α} and σ_{ℓ_A} represent the standard deviations of ℓ_α and ℓ_A, respectively. The averages (in the numerator) and standard deviations (in the denominator) are first evaluated for a given N and then the ratio is further averaged over the protein number distribution, $P(N)$, to yield effective feedback, F. This parameter is designed to be restricted between -1 and 1 as a way to quantify the relative feedback within the system and will help to objectively compare two independent gene circuits. Although in the application presented here we expect $0 < F < 1$, we anticipate $0 > F > -1$ while describing negative-feedback circuits.

Parameter estimation via maximum likelihood

The exercise above ensures that the minimal model of MaxCal with only four parameters is capable of producing the general features of a bimodal system. Next, we proceed to benchmark the performance of the model quantitatively when given a particular stochastic trajectory to characterize. This will allow us to learn about quantitative details of the underlying network by decoding information hidden in the noisy raw trajectory. For example, we may be interested in inferring the effective synthesis/degradation rates or the degree of feedback (F), quantities that are not directly available from the raw experimental trajectory. Below, we provide the framework to quantitatively infer these specific characteristics of a network from the stochastic trajectory.

Consider an experimentally observed trajectory of sufficiently long time, T, expressed in the units of the typical timescale (Δt) used for sampling the data. In this intrinsic time unit (Δt), we have $T+1$ frames at which the protein number has been recorded. Now consider a particular transition between two subsequent frames, say t and $t+1$, in which the protein number changed from i to j. We denote the probability of this one-step (single-frame) transition as $P(j,t+1;i,t)$, which is abbreviated as $P_{i\to j}$. These one-step transition probabilities can be determined from MaxCal as

$$P_{i\to j} = \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i}\delta(\ell_\alpha + \ell_A - j)P_{\ell_\alpha,\ell_A}, \quad (6)$$

where δ is the Dirac delta function, and P_{ℓ_α,ℓ_A} are functions of the Lagrange multipliers, described by Eq. 3. The likelihood ($\mathcal{L}$) of observing the experimental trajectory given a specific set of MaxCal parameters (h_α, h_A, K_A, and M) can then be calculated as

$$\mathcal{L} = \prod_{t=0}^{T-1} P(N_{t+1}, t+1; N_t, t) = \prod_{\{i\to j\}} P_{i\to j}^{\omega_{i\to j}}, \quad (7)$$

where N_t is the number of proteins present in frame t, $\omega_{i\to j}$ is the total number of $i \to j$ one-step transitions, and the second product is over all possible transitions between different values of i and j. As outlined above (Eq. 6), $P_{i\to j}$ values are determined using MaxCal, hence the likelihood is a function of h_α, h_A, K_A, and M. Thus, we can maximize the likelihood of the trajectory to select h_α, h_A, K_A, and M.

Experiments (and our Gillespie simulations) have no upper limit on production analogous to M in MaxCal. Rare fluctuations leading to unusually large jumps in protein number (>M) in one time step will severely penalize the likelihood of parameter values that are otherwise most likely. This discontinuous jump in likelihood will erroneously eliminate the most likely set of parameters. We avoid this problem by calculating transition probabilities over multiple intervals (m frames) for a given set of MaxCal parameters. We denote the probability of a multi-step (multiple-frame) transition as $P(j, t+m; i, t)$, abbreviated as $P_{(i \to j),m}$. This slightly modifies our likelihood function, $\mathcal{L}$, as

$$\begin{aligned} \mathcal{L} &= \prod_{n=1}^{\mathcal{N}} P(N_{t+m}, t+m; N_t, t = m(n-1)) \\ &= \prod_{\{i \to j\}} P_{(i \to j),m}^{\omega_{(i \to j),m}}, \end{aligned} \tag{8}$$

where $\mathcal{N}$ is T/m rounded down to the nearest integer and $\omega_{(i \to j),m}$ is the total number of $i \to j$ transitions over m frames. An objective choice of m can be provided by using the average residence times (in frames) in the high and low states . However, our result is not sensitive to the choice of m and is robust for a range of values around the typical value.

Dealing with experimental data

Although the procedure above is applicable to synthetic data in terms of protein number, typical experimental read-outs are in arbitrary fluorescence units. Furthermore, the amount of fluorescence measured per protein is noisy and requires one to de-convolute fluorescence fluctuations from protein number fluctuations. To mimic typical experimental readouts with these challenges, we use the same synthetic data from the auto-activating circuit, but "corrupt" it to create a fluorescence trajectory in silico that is likely to be observed in an experiment. We assume the probability distribution of fluorescence intensity (I) measured per protein to be a Gaussian distribution (60–62) centered at a with a standard deviation of b, i.e., $\langle I \rangle = a$ and $\langle I^2 \rangle - \langle I \rangle^2 = b^2$. With this assumption, the fluorescence measured from N proteins would follow a probability distribution that is a convolution of N protein fluorescence distributions leading to a Gaussian distribution with mean Na and variance Nb^2. To "corrupt" simulated trajectories of protein numbers, we select a fluorescence for each time point from this distribution where the mean and variance depend on the protein number, N. Although the procedure described here assumes that the fluorescence per protein follows a Gaussian distribution, we used a similar approach for Γ distributions (63,64) as well.

With this "synthetic fluorescence trajectory" closely mimicking realistic experimental situations, we propose two strategies to infer the underlying model. In the first strategy, we assume the average fluorescence intensity per protein ($\langle I \rangle = a$) is known, possibly obtained by carrying out low-intensity photobleaching (65–77). We use a as a conversion factor to determine protein number (N) from the fluorescence intensity, f, as

$$N = \mathrm{Int}(f/a), \tag{9}$$

where the Int function yields the nearest integer with negative protein numbers being rounded to zero. Parameter estimation then proceeds in the same fashion as before when analyzing trajectories in terms of protein number over time. In this strategy, parameter estimation takes place in two steps serially: first, fluorescence to number conversion (using Eq. 9), and then MaxCal with ML (as described earlier). We call this method serial fluorescence-to-number conversion, or simply SFNC.

We propose a second strategy in which the fluorescence fluctuation is included when calculating the likelihood of a set of MaxCal parameters. In this second approach (termed parallel FNC, or PFNC), we assume that the variance—in addition to the average—in intensity fluctuation per protein is also known, i.e., both a and b are given. This can be obtained using the same photobleaching experiment mentioned above to measure the probability distribution of fluorescence per protein (65–77). With this information, we can incorporate the fluorescence distribution in the likelihood function (Eq. 8), modifying it to

$$\begin{aligned} \mathcal{L} = \prod_{n=1}^{\mathcal{N}} \Bigg(\sum_{N_t} \sum_{N_{t+m}} \Phi(N_t \,|\, f_t) P(N_{t+m}, t+m; N_t, t \\ = m(n-1)) \Phi(N_{t+m} \,|\, f_{t+m}) \Bigg) \end{aligned} \tag{10}$$

where f_t is the fluorescence at frame t, and $\Phi(N_t \,|\, f_t)$ is the conditional probability that N_t proteins are present given a fluorescence measurement of f_t. These probabilities are known and used as previously with the knowledge of the known variance and mean. The probability $P(N_{t+m}, t+m; N_t, t)$ is determined as above using MaxCal and is a function of the Lagrange multipliers and M. The new likelihood function (Eq. 10) is then maximized to determine h_α, h_A, K_A, and M.

RESULTS AND DISCUSSION

MaxCal accurately infers underlying rate parameters

Using the procedures described above, we determine h_α, h_A, K_A, and M for a given stochastic trajectory in terms of either protein number or fluorescence readout. These fully specify the minimal MaxCal model and are capable of making multiple predictions, such as of the underlying rate parameters. Effective values for the underlying production and degradation rates can be predicted using the average value of the production- and degradation-state variables, respectively (see Eq. 4). To see how well these inferred rates compare to the true values, we applied our inference method to input trajectories that are ~2000 frames long with an intrinsic sampling rate of 5 min ($\Delta t = 300$ s), equivalent to trajectories of 7 days. Furthermore, we used 100 such trajectories, equivalent to tracking protein numbers in 100 cells. These numbers were chosen to closely match typical experimental conditions (31). To quantify the variance of the effective rate estimates, we apply our method to 10 different sets of these simulations and present the average and standard deviation of the 10 sets of predicted rates. Using simulations from the reaction rates listed in Table 1, the predicted values compare well against the "true" values used to generate the synthetic data (see Table 1). The robustness of the prediction was further tested by creating synthetic data using different values of g, g^*, and r, and similar accuracies were produced. In addition, the inference scheme was applied to an alternate model of positive feedback—different from Eq. 1—to generate the synthetic data, and again, the inferred rates matched well with input values (see Supporting Material for details). However, it is important to realize that Eq. 4 is only an approximation to infer the intrinsic production and degradation rates. Thus, it is possible to have deviations between the inferred and true rates—higher than the ones reported in Table 1—whereas MaxCal captures the temporal statistics well

TABLE 1 Comparison of True Rates and Predicted Rates Using MaxCal

	True Values	Predicted Values
g (s^{-1})	5.0×10^{-3}	$5.8 \pm 0.3 \times 10^{-3}$
g^* (s^{-1})	50.0×10^{-3}	$43.0 \pm 2.2 \times 10^{-3}$
r (s^{-1})	1.0×10^{-3}	$0.95 \pm 0.05 \times 10^{-3}$
$\tau_{L \to H}$ (s)	71.5×10^3	$99.0 \pm 11.5 \times 10^3$
$\tau_{H \to L}$ (s)	86.0×10^3	$117.6 \pm 12.9 \times 10^3$
$S_{\rm I}$ (bits)	8.84	9.24 ± 0.03
$S_{\rm h}$ (bits)	9.42	9.07 ± 0.02
$S_{\rm l}$ (bits)	6.20	7.66 ± 0.04
$S_{\rm cg}$ (bits)	1.03	1.02 ± 0.01

The first column reports the "true" underlying protein synthesis and degradation rates used to create synthetic input data ($f_{\rm d} = 5.0 \times 10^{-3}$ s^{-1}, $b_{\rm d} = 50$ s^{-1}, $f_{\rm p} = 6.0 \times 10^{-3}$ s^{-1}, $b_{\rm p} = 3.0 \times 10^{-5}$ s^{-1}), average residence times in the high and low states, and corresponding path informational entropies. Synthetic input data were recorded at $\Delta t = 300$s. The second column reports the average and standard deviation of the same quantities of interest, but extracted using the MaxCal model on 10 sets of synthetic data, each consisting of 100 trajectories of 7 days.

(e.g., fluctuations in the high/low states and transitions between states).

Distributions predicted from MaxCal agree well with data

For a more detailed demonstration of how well MaxCal describes data, we further compared MaxCal-predicted distributions to that of the input data (generated from the reaction network in Eq. 1). Fig. 2 *A* shows that the protein number distribution predicted from MaxCal agrees well with the input data in that the locations and widths of the two peaks are comparable between the two approaches. Next, we compare the distribution of dwell times predicted by MaxCal to that obtained from the synthetic data. The agreement for the shape of the distribution and the average dwell times in the low and high states (see Fig. 2, *B* and *C*; Table 1) are reasonable.

The comparisons between "true" and predicted values for multiple observables show that the minimal model of MaxCal with only four parameters can make reasonable predictions for data generated with more complex models (with seven parameters). To further quantify the quality of the parameter extraction and performance of our minimal model against those of the actual model with more parameters, we compare the informational content in the "synthetic" Gillespie trajectories and trajectories generated by MaxCal using these parameters. We compute path informational entropy as (78)

$$S_I = -\sum_{i,j} P_i P_{i \to j} \log_2 \left(P_i P_{i \to j} \right), \qquad (11)$$

where P_i is the probability of having i proteins in the system and $P_{i \to j}$ is the probability of transitioning from i proteins to j proteins after a single frame. If our MaxCal model is too simple and cannot adequately capture the dynamics of the Gillespie trajectories used, its S_I will be notably different from that of the Gillespie model. We find that the MaxCal model selected by ML has only a 4.5% difference in path informational entropy compared to the "synthetic" input data from Gillespie simulations (see Table 1). This provides quantitative verification that the minimal constraints used in Eq. 2 are sufficient to describe the auto-activating circuit modeled here. The overall path entropy has contributions from three types of fluctuations: 1) within the high state, 2) within the low state, and 3) transitions between the high and low states. To further explore how MaxCal-generated path entropy captures details of these fluctuations, we compute three additional path entropies: $S_{\rm h}$, $S_{\rm l}$, and $S_{\rm cg}$. $S_{\rm h}$ and $S_{\rm l}$ are computed in the same fashion as S_I, but only consider parts of the trajectory in the high state and low state respectively (see Supporting Material for high/low state assignment). To measure $S_{\rm cg}$, the trajectory is first coarse-grained into a binary trajectory between the low state ($N_{cg} = 0$) and the high state ($N_{cg} = 1$). $S_{\rm cg}$ is then calculated in the same manner as Eq. 11. We find that MaxCal generated estimates of $S_{\rm h}$ and $S_{\rm cg}$ are in excellent agreement

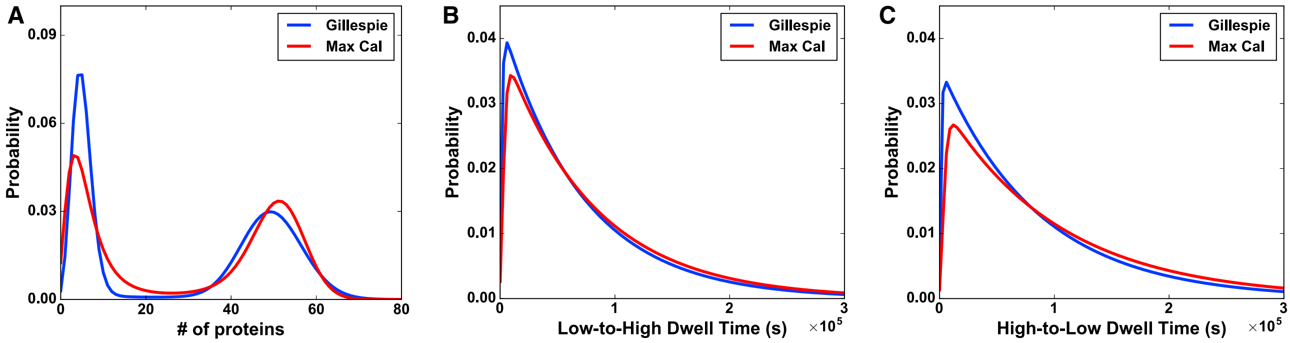

FIGURE 2 Predicted distributions agree well with the "true" distributions. (*A*) Protein-number probability distributions from synthetic input trajectories (*blue*) and predicted MaxCal trajectories (*red*). (*B*) Low-state and (*C*) high-state residence time probability distributions for synthetic input trajectories (*blue*) and predicted MaxCal trajectories (*red*). The underlying Gillespie reaction rates are the same as those used in Table 1 and the extracted MaxCal parameters used are a representative example from the 10 sets extracted to make Table 1. To see this figure in color, go online.

with the input data, whereas S_1 differs by ~24% from the input (see Table 1). The analysis above provides a quantitative measure of performance for MaxCal with a given set of constraints. These measures can be further used to determine the need for incorporating higher-order combinations of the state variables to the caliber function (Eq. 2; e.g., $\langle \ell_\alpha \ell_A^2 \rangle$, $\langle \ell_\alpha^2 \ell_A \rangle$, etc.) to develop models of higher complexity (51).

MaxCal provides an effective feedback parameter for the circuit

We also extract the effective feedback parameter, *F*, using Eq. 5. As a demonstration of its usefulness, if we compared the MaxCal parameters extracted from experimental traces with varying concentrations of inducer (31), the effective production and degradation rates might be similar, but *F* would be expected to vary with different amounts of inducer, representing the degree of coupling between the production of *A* and the concentration of *A*. To mimic the effect of varying inducer concentrations, we generated synthetic data with higher or lower promoter binding rates, f_p, to effectively increase or decrease the amount of self promotion in the system. Next, we applied our MaxCal framework to these trajectories with different levels of self-promotion. Fig. 3, *A–C*, shows that MaxCal reproduces comparable protein number distributions regardless of the degree of self-promotion. Table 2 further demonstrates that although MaxCal infers very similar production and degradation rates between the three levels of self-promotion, the effective feedback, *F*, changes accordingly.

Estimating the effective feedback parameter can be important, as it determines the onset of bimodality from unimodality as well as the relative population in the high and low states. Bimodal protein distributions and stochastic switching between the two states often dictate phenotypic variability, a characteristic of bet-hedging strategies used by microbes to evade stress such as antibiotic (31,79,80). Consequently, different strains that have evolved under different selection pressures may differentially tune their level of feedback (81). Similarly, it may be interesting to see whether strains using "resistance" or "tolerance" mechanisms to evade antibiotics (82) evolve their feedback parameters differently. Applying MaxCal on experimental trajectories of different strains evolved under different conditions to infer these feedback parameters can give us further insights into evolvability and selection. Similarly, this metric can be useful when describing circuits with negative feedback as well.

The ability to extract an effective feedback parameter is a special feature of MaxCal that provides a coarse-grained description of feedback. This is in contrast to traditional parameterization schemes that invoke auxiliary species and multiple reactions involving many parameters to describe feedback. As a result, MaxCal can provide a model with fewer parameters compared to traditional bottom-up approaches. This is true even when describing circuits with multiple species beyond the single-gene expression circuit used in this study (35,55). The success of MaxCal presented here motivates the need for future studies on synthetic data generated using more intermediate steps, such as RNA synthesis before protein synthesis. Further research must also be performed on circuits involving more species that mutually regulate each other, possibly leading to oscillatory behaviors as in the repressilator circuit of Elowitz and Liebler (83). It is also important to note that MaxCal is exactly equivalent to the master equation when describing systems without feedback, e.g., biochemical cycles where states interconvert among themselves (52,55,84).

MaxCal can be applied when dealing with noisy fluorescence trajectories

The results above illustrate the applicability of MaxCal when experimental trajectories are expressed in protein-number fluctuations. We now proceed to demonstrate the applicability of MaxCal when data are reported in noisy fluorescence trajectories instead of protein-number

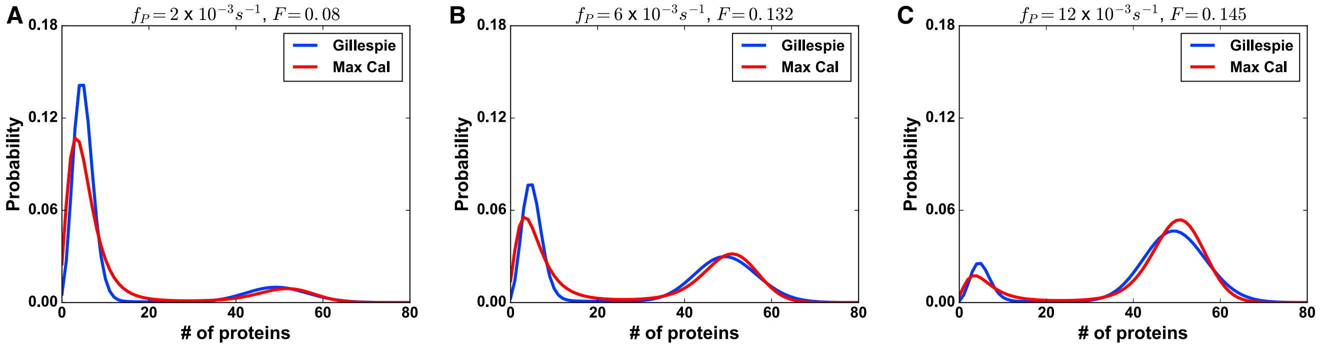

FIGURE 3 MaxCal can capture distributions under varying degrees of feedback. Protein-number probability distributions from synthetic input trajectories (*blue*) and predicted MaxCal trajectories (*red*) for different levels of self-promotion, specifically (*A*) $f_P = 2 \times 10^{-3}$ s^{-1}, (*B*) $f_P = 6 \times 10^{-3}$ s^{-1}, and (*C*) $f_P = 12 \times 10^{-3}$ s^{-1}. The underlying Gillespie reaction rates are the same as those used in Table 2 and the extracted MaxCal parameters used are a representative example from the 10 sets extracted to make Table 2. To see this figure in color, go online.

TABLE 2 Comparison of Feedback Parameter Using Different Promoter Binding Rates

f_p	Extracted g	Extracted g^*	Extracted r	F
2.0×10^{-3} s^{-1}	$4.3 \pm 0.3 \times 10^{-3}$ s^{-1}	$34.9 \pm 2.5 \times 10^{-3}$ s^{-1}	$0.81 \pm 0.06 \times 10^{-3}$ s^{-1}	0.080 ± 0.006
6.0×10^{-3} s^{-1}	$5.8 \pm 0.3 \times 10^{-3}$ s^{-1}	$43.0 \pm 2.2 \times 10^{-3}$ s^{-1}	$0.95 \pm 0.05 \times 10^{-3}$ s^{-1}	0.132 ± 0.005
12.0×10^{-3} s^{-1}	$5.7 \pm 0.4 \times 10^{-3}$ s^{-1}	$38.8 \pm 2.9 \times 10^{-3}$ s^{-1}	$0.80 \pm 0.06 \times 10^{-3}$ s^{-1}	0.145 ± 0.006

Each row reports the average and SD of the extracted production and degradation rates as well as the effective feedback, F, for different values of f_p. Similar to Fig. 2, for all three cases, $g = 5.0 \times 10^{-3}$ s^{-1}, $g^* = 50.0 \times 10^{-3}$ s^{-1}, $r = 1.0 \times 10^{-3}$ s^{-1}, $f_d = 5.0 \times 10^{-3}$ s^{-1}, $b_d = 50$ s^{-1}, $b_p = 3.0 \times 10^{-5}$ s^{-1}, and 10 sets of synthetic data, each equivalent to 100 trajectories of 7 days, were used to extract predicted values and standard deviations.

trajectories. We use both methodologies, SFNC and PFNC, as described earlier, to infer the underlying model from the noisy data. Fig. 4 and Table 3 show the performance of these strategies tested against "corrupted" synthetic data. SFNC, based on only the knowledge of average intensity per protein, performs well when the fluctuation in fluorescence per protein is sufficiently small compared to the average fluorescence (see Table 3). However, SFNC starts to deviate significantly from the "true" values when noise increases, e.g., noise is >100% of the average fluorescence. In fact, at these levels of noise, it becomes increasingly difficult to determine the unique ML function, and the corresponding values of rate parameters start to deviate largely from the true values. Considering this deficiency, SFNC should not be used for noise levels >100%. PFNC, on the other hand, does not suffer from any such issues. PFNC infers rates with reasonable accuracy even when noise is as high as 200% (see Table 3, *bottom rows*). The success of PFNC is further demonstrated by comparing "true" and predicted distributions of protein numbers and dwell times (see Fig. 4) at this level of noise. PFNC performs better than SFNC due to the incorporation of fluorescence fluctuation within its ML procedure. Although the above results were extracted from data using a Gaussian fluorescence distribution, we carried out similar exercises using a Γ distribution for the fluorescence per protein (63,64), with similarly accurate results. This highlights the need for carrying out controlled photobleaching experiments to learn about the average as well as the noise in the fluorescence per protein to faithfully infer underlying dynamics. In summary, the exercise above demonstrates broad applicability of MaxCal, even when experimental data are not in protein number but in fluorescence with high fluctuation.

CONCLUSIONS

We use the principle of maximum caliber (MaxCal)—akin to the principle of maximum entropy applied to describe path probabilities—to model protein-number fluctuations as observed in genetic circuits. We demonstrate the application of MaxCal in a positive feedback circuit, a common motif in many naturally occurring and synthetic circuits. Specifically, we consider a single-gene auto-activating circuit where a minimal model based on MaxCal was developed with three physical constraints: protein synthesis, protein degradation, and positive feedback. Through this analysis, we make four key conclusions. First, the minimal model is capable of producing the switch-like behavior of the circuit. Second, the model shows its usefulness to quantitatively infer underlying parameters. To mimic raw data from experiment, synthetic data were generated using a Gillespie algorithm with a known reaction network model to produce trajectories of fluctuating protein numbers. MaxCal correctly infers underlying rates when compared to the "known" values. Furthermore, MaxCal-predicted distributions agree well with the ones derived from the input data. Third, MaxCal provides an effective feedback parameter to characterize these circuits that can be useful for circuit

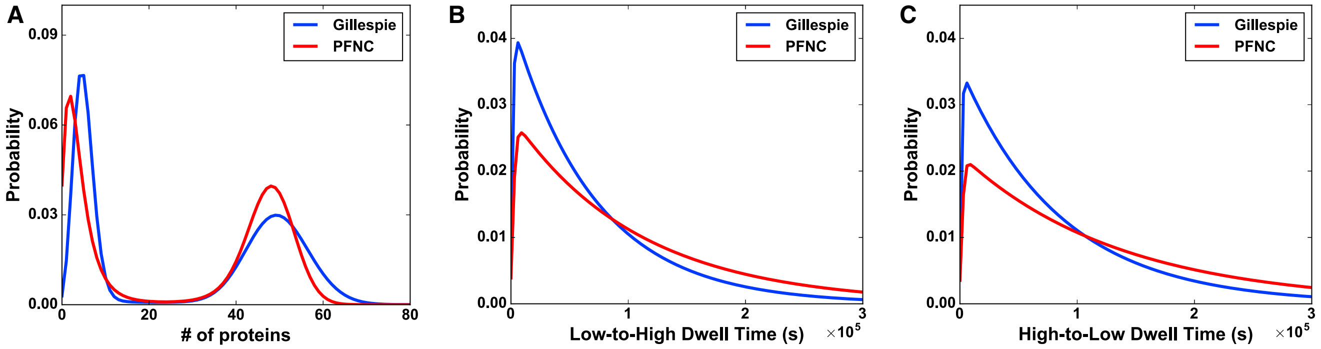

FIGURE 4 Predicted distributions from fluorescence trajectories with 200% noise using PFNC. (*A*) Protein-number probability distributions from "true" synthetic input trajectories (*blue*) and MaxCal trajectories using PFNC inference strategy (*red*). Comparisons between "true" and PFNC are also shown for (*B*) low-state and (*C*) high-state residence time probability distributions. The underlying Gillespie reaction rates are the same as those used in Table 3 and the extracted MaxCal parameters are a representative example from the 10 sets extracted to make the last row of Table 3. To see this figure in color, go online.

TABLE 3 Effective Rates from Fluorescence Trajectories

Noise	g (s^{-1})	g^* (s^{-1})	r (s^{-1})	Method
True	5.0×10^{-3}	50.0×10^{-3}	1.0×10^{-3}	
0%	$5.8 \pm 0.3 \times 10^{-3}$	$43.0 \pm 2.2 \times 10^{-3}$	$0.95 \pm 0.05 \times 10^{-3}$	MaxCal
50%	$6.1 \pm 0.3 \times 10^{-3}$	$46.2 \pm 2.6 \times 10^{-3}$	$1.03 \pm 0.06 \times 10^{-3}$	SFNC
50%	$5.3 \pm 0.3 \times 10^{-3}$	$41.5 \pm 2.1 \times 10^{-3}$	$0.93 \pm 0.05 \times 10^{-3}$	PFNC
100%	$6.8 \pm 0.4 \times 10^{-3}$	$50.5 \pm 2.7 \times 10^{-3}$	$1.08 \pm 0.05 \times 10^{-3}$	SFNC
100%	$5.7 \pm 0.3 \times 10^{-3}$	$48.0 \pm 2.5 \times 10^{-3}$	$1.09 \pm 0.06 \times 10^{-3}$	PFNC
150%	$6.1 \pm 0.3 \times 10^{-3}$	$55.2 \pm 3.5 \times 10^{-3}$	$1.26 \pm 0.08 \times 10^{-3}$	PFNC
200%	$6.6 \pm 0.3 \times 10^{-3}$	$63.3 \pm 3.9 \times 10^{-3}$	$1.47 \pm 0.08 \times 10^{-3}$	PFNC

The first row reports the "true" underlying protein synthesis and degradation rates used to create synthetic input data (same rates and conditions as in Table 1). The second row reports the average and standard deviation of MaxCal-inferred rates when trajectories are in protein number. Rows 3–8 report extracted rates for synthetically corrupted trajectories generated using different levels of noise in fluorescence per protein compared to the average (indicated in column 1) and different methods of extraction (SFNC and PFNC, as indicated in column 5).

design as well as analysis of differently evolved strains. Finally, we show how similar methods can be applied when the raw trajectory is in fluorescence rather than protein number, a typical attribute of experimental data. We demonstrate this by "corrupting" the same synthetic protein number trajectories with Gaussian fluctuation to create noisy fluorescence trajectories. In the regime of low fluorescence noise, the average fluorescence per protein can be used to convert traces back to protein number, followed by MaxCal to infer the model (SFNC). However, higher levels of noise require a more integrated approach (PFNC), where a model's likelihood is calculated by combining both MaxCal-generated transition probabilities and fluorescence fluctuation. Using fluorescence-corrupted trajectories, we show that PFNC can infer underlying rates and distributions of observables even when the relative noise is fairly high. The method presented here demonstrates the potential application of MaxCal to broader problems in gene networks involving feedback, even when data are presented in fluorescence.

SUPPORTING MATERIAL

Supporting Materials and Methods, one figure, and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)31016-0.

AUTHOR CONTRIBUTIONS

T.F., G.B., and K.G. designed research. T.F. performed research. T.F. and K.G. analyzed data, and T.F., G.B., and K.G. wrote the article.

ACKNOWLEDGMENTS

We thank Steve Pressé, Brian Munsky, Tamás Székely, Tristan Bereau, and Ken Dill for many stimulating and helpful discussions. We acknowledge the High Performance Computing facility at the University of Denver for computing assistance.

We acknowledge support from the National Science Foundation (award number 1149992), the Research Corporation for Science Advancement (as a Cottrell Scholar), and the PROF grant from the University of Denver. G.B. was supported by National Institutes of Health National Institute of General Medical Sciences (NIH-NIGMS) grant R35GM122561 and by a Laufer Center for Physical and Quantitative Biology endowment.

REFERENCES

1. Ozbudak, E. M., M. Thattai, …, A. van Oudenaarden. 2002. Regulation of noise in the expression of a single gene. *Nat. Genet.* 31:69–73.
2. Kaern, M., T. C. Elston, …, J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6:451–464.
3. Paulsson, J. 2004. Summing up the noise in gene networks. *Nature.* 427:415–418.
4. Samoilov, M., S. Plyasunov, and A. P. Arkin. 2005. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc. Natl. Acad. Sci. USA.* 102:2310–2315.
5. Sánchez, A., and J. Kondev. 2008. Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. USA.* 105:5081–5086.
6. Shahrezaei, V., and P. S. Swain. 2008. The stochastic nature of biochemical networks. *Curr. Opin. Biotechnol.* 19:369–374.
7. Elowitz, M. B., A. J. Levine, …, P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science.* 297:1183–1186.
8. Tao, Y. 2004. Intrinsic and external noise in an auto-regulatory genetic network. *J. Theor. Biol.* 229:147–156.
9. Beard, D. A., and H. Qian. 2008. Chemical Biophysics: Quantitative Analysis of Cellular Systems. University Press, Cambridge.
10. Munsky, B., B. Trinh, and M. Khammash. 2009. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* 5:318.
11. Lillacci, G., and M. Khammash. 2010. Parameter estimation and model selection in computational biology. *PLOS Comput. Biol.* 6:e1000696.
12. Zechner, C., J. Ruess, …, H. Koeppl. 2012. Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. USA.* 109:8340–8345.
13. Lillacci, G., and M. Khammash. 2012. A distribution-matching method for parameter estimation and model selection in computational biology. *Int. J. Robust Nonlinear Control.* 22:1065–1081.
14. Ruess, J., A. Milias-Argeitis, and J. Lygeros. 2013. Designing experiments to understand the variability in biochemical reaction networks. *J. R. Soc. Interface.* 10:20130588.
15. Lillacci, G., and M. Khammash. 2013. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics.* 29:2311–2319.
16. Kauffman, S. 2004. A proposal for using the ensemble approach to understand genetic regulatory networks. *J. Theor. Biol.* 230:581–590.
17. Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361.

18. Guet, C. C., M. B. Elowitz, …, S. Leibler. 2002. Combinatorial synthesis of genetic networks. *Science.* 296:1466–1470.

19. Hasty, J., M. Dolnik, …, J. J. Collins. 2002. Synthetic gene network for entraining and amplifying cellular oscillations. *Phys. Rev. Lett.* 88:148101.

20. Stricker, J., S. Cookson, …, J. Hasty. 2008. A fast, robust and tunable synthetic gene oscillator. *Nature.* 456:516–519.

21. Tsai, T. Y., Y. S. Choi, …, J. E. Ferrell, Jr. 2008. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science.* 321:126–129.

22. Gore, J., and A. van Oudenaarden. 2009. Synthetic biology: the yin and yang of nature. *Nature.* 457:271–272.

23. Mukherji, S., and A. van Oudenaarden. 2009. Synthetic biology: understanding biological design from synthetic circuits. *Nat. Rev. Genet.* 10:859–871.

24. Ellis, T., X. Wang, and J. J. Collins. 2009. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.* 27:465–471.

25. Kittisopikul, M., and G. M. Süel. 2010. Biological role of noise encoded in a genetic network motif. *Proc. Natl. Acad. Sci. USA.* 107:13300–13305.

26. Khalil, A. S., and J. J. Collins. 2010. Synthetic biology: applications come of age. *Nat. Rev. Genet.* 11:367–379.

27. Moon, T. S., C. Lou, …, C. A. Voigt. 2012. Genetic programs constructed from layered logic gates in single cells. *Nature.* 491:249–253.

28. Wu, M., R. Q. Su, …, X. Wang. 2013. Engineering of regulated stochastic cell fate determination. *Proc. Natl. Acad. Sci. USA.* 110:10610–10615.

29. Wu, F., and X. Wang. 2015. Applications of synthetic gene networks. *Sci. Prog.* 98:244–252.

30. Wu, F., R. Q. Su, …, X. Wang. 2017. Engineering of a synthetic quadrastable gene network to approach Waddington landscape and cell fate determination. *eLife.* 6:e23702.

31. Nevozhay, D., R. M. Adams, …, G. Balázsi. 2012. Mapping the environmental fitness landscape of a synthetic gene circuit. *PLOS Comput. Biol.* 8:e1002480.

32. Alon, U. 2007. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8:450–461.

33. Lyons, S. M., W. Xu, …, A. Prasad. 2014. Loads bias genetic and signaling switches in synthetic and natural systems. *PLOS Comput. Biol.* 10:e1003533.

34. Wang, L. Z., F. Wu, …, X. Wang. 2016. Build to understand: synthetic approaches to biology. *Integr. Biol.* 8:394–408.

35. Pressé, S., K. Ghosh, and K. A. Dill. 2011. Modeling stochastic dynamics in biochemical systems with feedback using maximum caliber. *J. Phys. Chem. B.* 115:6202–6212.

36. Gardner, T. S., C. R. Cantor, and J. J. Collins. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature.* 403:339–342.

37. Lipshtat, A., A. Loinger, …, O. Biham. 2006. Genetic toggle switch without cooperative binding. *Phys. Rev. Lett.* 96:188101.

38. Keller, A. D. 1995. Model genetic circuits encoding autoregulatory transcription factors. *J. Theor. Biol.* 172:169–185.

39. Smolen, P., D. A. Baxter, and J. H. Byrne. 1998. Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. *Am. J. Physiol.* 274:C531–C542.

40. Becskei, A., B. Séraphin, and L. Serrano. 2001. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* 20:2528–2535.

41. Tyson, J. J., K. C. Chen, and B. Novak. 2003. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.* 15:221–231.

42. Cheng, Z., F. Liu, …, W. Wang. 2008. Robustness analysis of cellular memory in an autoactivating positive feedback system. *FEBS Lett.* 582:3776–3782.

43. Bishop, L. M., and H. Qian. 2010. Stochastic bistability and bifurcation in a mesoscopic signaling system with autocatalytic kinase. *Biophys. J.* 98:1–11.

44. Frigola, D., L. Casanellas, …, M. Ibañes. 2012. Asymmetric stochastic switching driven by intrinsic molecular noise. *PLoS One.* 7:e31407.

45. Faucon, P. C., K. Pardee, …, X. Wang. 2014. Gene networks of fully connected triads with complete auto-activation enable multistability and stepwise stochastic transitions. *PLoS One.* 9:e102873.

46. Kepler, T. B., and T. C. Elston. 2001. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* 81:3116–3136.

47. Phillips, R., J. Kondev, and J. Theriot. 2009. Physical Biology of The Cell. Garland Science, New York, NY.

48. Ghosh, K., K. A. Dill, …, R. Phillips. 2006. Teaching the principles of statistical dynamics. *Am. J. Phys.* 74:123–133.

49. Seitaridou, E., M. M. Inamdar, …, K. Dill. 2007. Measuring flux distributions for diffusion in the small-numbers limit. *J. Phys. Chem. B.* 111:2288–2292.

50. Wu, D., K. Ghosh, …, R. Phillips. 2009. Trajectory approach to two-state kinetics of single particles on sculpted energy landscapes. *Phys. Rev. Lett.* 103:050603.

51. Otten, M., and G. Stock. 2010. Maximum caliber inference of nonequilibrium processes. *J. Chem. Phys.* 133:034119.

52. Pressé, S., K. Ghosh, …, K. A. Dill. 2010. Dynamical fluctuations in biochemical reactions and cycles. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 82:031905.

53. Ghosh, K. 2011. Stochastic dynamics of complexation reaction in the limit of small numbers. *J. Chem. Phys.* 134:195101.

54. Pressé, S., J. Peterson, …, K. Dill. 2014. Single molecule conformational memory extraction: p5ab RNA hairpin. *J. Phys. Chem. B.* 118:6597–6603.

55. Pressé, S., K. Ghosh, …, K. Dill. 2013. Principle of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* 85:1115–1141.

56. Dixit, P. D., and K. A. Dill. 2014. Inferring microscopic kinetic rates from stationary state distributions. *J. Chem. Theory Comput.* 10:3002–3005.

57. Dixit, P. D., A. Jain, …, K. A. Dill. 2015. Inferring transition rates of networks from populations in continuous-time Markov processes. *J. Chem. Theory Comput.* 11:5464–5472.

58. Wan, H., G. Zhou, and V. A. Voelz. 2016. A maximum-caliber approach to predicting perturbed folding kinetics due to mutations. *J. Chem. Theory Comput.* 12:5768–5776.

59. Munsky, B., and M. Khammash. 2006. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* 124:044104.

60. Lawrimore, J., K. S. Bloom, and E. D. Salmon. 2011. Point centromeres contain more than a single centromere-specific Cse4 (CENP-A) nucleosome. *J. Cell Biol.* 195:573–582.

61. Taniguchi, Y., P. J. Choi, …, X. S. Xie. 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science.* 329:533–538.

62. Tsekouras, K., T. C. Custer, …, S. Pressé. 2016. A novel method to accurately locate and count large numbers of steps by photobleaching. *Mol. Biol. Cell.* 27:3601–3615.

63. Friedman, N., L. Cai, and X. S. Xie. 2006. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* 97:168302.

64. McLean, P., C. Smolke, and M. Salit. 2016. Characterizing the non-normal distribution of flow cytometry measurements from transiently expressed constructs in mammalian cells. Published online June 9, 2016. 10.1101/057950.

65. Coffman, V. C., and J. Q. Wu. 2012. Counting protein molecules using quantitative fluorescence microscopy. *Trends Biochem. Sci.* 37:499–506.

66. Coffman, V. C., P. Wu, …, J. Q. Wu. 2011. CENP-A exceeds microtubule attachment sites in centromere clusters of both budding and fission yeast. *J. Cell Biol.* 195:563–572.

67. Engel, B. D., W. B. Ludington, and W. F. Marshall. 2009. Intraflagellar transport particle size scales inversely with flagellar length: revisiting the balance-point length control model. *J. Cell Biol.* 187:81–89.

68. Leake, M. C., J. H. Chandler, …, J. P. Armitage. 2006. Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature.* 443:355–358.

69. Ulbrich, M. H., and E. Y. Isacoff. 2007. Subunit counting in membrane-bound proteins. *Nat. Methods.* 4:319–321.

70. Das, S. K., M. Darshi, …, H. Bayley. 2007. Membrane protein stoichiometry determined from the step-wise photobleaching of dye-labelled subunits. *ChemBioChem.* 8:994–999.

71. Shu, D., H. Zhang, …, P. Guo. 2007. Counting of six pRNAs of ϕ29 DNA-packaging motor with customized single-molecule dual-view system. *EMBO J.* 26:527–537.

72. Delalez, N. J., G. H. Wadhams, …, J. P. Armitage. 2010. Signal-dependent turnover of the bacterial flagellar switch protein FliM. *Proc. Natl. Acad. Sci. USA.* 107:11347–11351.

73. Demuro, A., A. Penna, …, I. Parker. 2011. Subunit stoichiometry of human Orai1 and Orai3 channels in closed and open states. *Proc. Natl. Acad. Sci. USA.* 108:17832–17837.

74. Hastie, P., M. H. Ulbrich, …, L. Chen. 2013. AMPA receptor/TARP stoichiometry visualized by single-molecule subunit counting. *Proc. Natl. Acad. Sci. USA.* 110:5163–5168.

75. Arumugam, S. R., T. H. Lee, and S. J. Benkovic. 2009. Investigation of stoichiometry of T4 bacteriophage helicase loader protein (gp59). *J. Biol. Chem.* 284:29283–29289.

76. Pitchiaya, S., J. R. Androsavich, and N. G. Walter. 2012. Intracellular single molecule microscopy reveals two kinetically distinct pathways for microRNA assembly. *EMBO Rep.* 13:709–715.

77. Pitchiaya, S., V. Krishnan, …, N. G. Walter. 2013. Dissecting non-coding RNA mechanisms in cellulo by single-molecule high-resolution localization and counting. *Methods.* 63:188–199.

78. Schneidman, E., M. J. Berry, 2nd, …, W. Bialek. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature.* 440:1007–1012.

79. Balaban, N. Q., J. Merrin, …, S. Leibler. 2004. Bacterial persistence as a phenotypic switch. *Science.* 305:1622–1625.

80. Levy, S. F., N. Ziv, and M. L. Siegal. 2012. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol.* 10:e1001325.

81. González, C., J. C. Ray, …, G. Balázsi. 2015. Stress-response balance drives the evolution of a network module and its host genome. *Mol. Syst. Biol.* 11:827.

82. Brauner, A., O. Fridman, …, N. Q. Balaban. 2016. Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nat. Rev. Microbiol.* 14:320–330.

83. Elowitz, M. B., and S. Leibler. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature.* 403:335–338.

84. Ge, H., S. Pressé, …, K. A. Dill. 2012. Markov processes follow from the principle of maximum caliber. *J. Chem. Phys.* 136:064108.

Biophysical Journal, Volume *113*

Supplemental Information

Building Predictive Models of Genetic Circuits Using the Principle of Maximum Caliber

Taylor Firman, Gábor Balázsi, and Kingshuk Ghosh

Supporting Material

Application of Finite State Projection to Maximum Caliber

Since protein number theoretically has no upper limit, our self-promotion gene network would be considered an open system, a problematic condition for analytically calculating protein number distributions. However, Finite State Projection (FSP) circumvents this problem by truncating the infinite phase space of protein number down to some relatively high, *finite* maximum. The probabilities of any protein numbers higher than this maximum are combined into one collective state, or 'sink', and the probability of being in this sink provides a measurement of how much error has accumulated in the distribution due to the truncation. As such, this rigorous technique can provide analytical probability distributions within objective levels of error. For a full explanation of the technique, see the original work of Munsky and Khammash [J. Chem. Phys. 124:044104 (2006)]. However, one slight modification must be made for application to MaxCal. Within section II of Munsky et al, the chemical master equation for every possible reaction in the finite reaction space can be rewritten as

$$\dot{\mathbf{P}}\left(\mathbf{X};t\right) = \mathbf{A}\cdot\mathbf{P}\left(\mathbf{X};t\right), \tag{S1}$$

where $\mathbf{X}$ is a column vector representing the different states of the system (in our case, the number of proteins present), $\mathbf{P}\left(\mathbf{X};t\right)$ is a column vector containing the probabilities of the different states at time t, and $\mathbf{A}$ is the state reaction matrix where each element of the matrix is a combination of the reaction propensities going from one state (corresponding to the column) to another (corresponding to the row). In a reaction network designed for a Gillespie simulation, these propensities would simply be the reaction rate multiplied by the stoichiometry of the reactants. For our MaxCal system, these propensities would simply be the probability of transitioning from one protein level to another (defined by equations 3 and 6 of the main text) and time would be renormalized into units of Δt. From there, we can calculate transition probabilities over multiple frames (m) to within an acceptable error using the exponential matrix of $\mathbf{A}$,

$$\mathbf{P}\left(\mathbf{X};m\Delta t\right) = \exp\left(\mathbf{A}m\right)\mathbf{P}\left(\mathbf{X};0\right). \tag{S2}$$

To find the effective equilibrium distribution for the number of proteins in the system, we can set the time as a number large enough to ensure the system is at relative equilibrium, e.g. 100 times the average dwell time, and perform the same matrix exponentiation.

Alternate model to test MaxCal inference

To further test the accuracy of MaxCal, the inference method described in the main text was applied to an alternate model of self-promotion that has monomers binding to the promoter site rather than dimers. The reaction scheme is represented as

$$\begin{array}{cc} \alpha \xrightarrow{g} \alpha + A & A \xrightarrow{r} \oslash \\ \alpha + A \underset{b_p}{\overset{f_p}{\rightleftarrows}} \alpha^* & \alpha^* \xrightarrow{g^*} \alpha^* + A \end{array} \tag{S3}$$

where some generic protein A is created from its corresponding gene α at a rate of g, degrades at a rate of r, and binds to the promoter site, α, with forward and backward rates of f_p and b_p respectively. This sends α into or out of its activated state α^*, which creates protein A at a much faster rate g^*. This again captures the essentials of a positive feedback mechanism, but represents a different level of non-linearity and cooperativity in Hill-type models. This circuit is motivated by the earlier work of Lipshtat, Loinger, Balaban, and Biham [Phys. Rev. Lett. 96:188101 (2006)] demonstrating that bimodality in toggle switch circuits can be obtained without cooperative binding. Similarly, we also notice the above model can produce bimodality for this positive feedback circuit. Using reaction rates similar to those utilized for the model in the main text, the inferred rates and distributions are displayed in Table S1 and Figure S1 respectively. These results demonstrate that an acceptable level of accuracy can be generated using MaxCal, regardless of the exact molecular underpinnings of the circuit being considered.

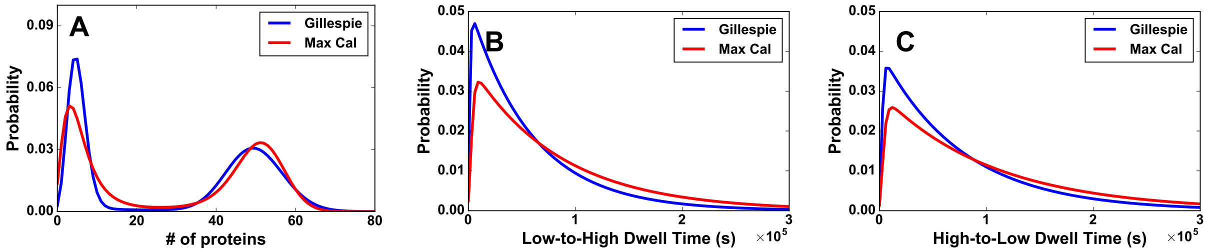

Figure S1: **Predicted distributions for alternate model agree well with the "true" distributions.** (**A**) Protein number probability distributions from synthetic input trajectories (blue) and predicted MaxCal trajectories (red). (**B**) Low state and (**C**) high state residence time probability distributions for synthetic input trajectories (blue) and predicted MaxCal trajectories (red). Underlying Gillespie reaction rates are the same as those used in Table S1 and the extracted MaxCal parameters used are a representative example from the ten sets extracted to make Table S1.

	True Values	Predicted Values
g (s^{-1})	5.0×10^{-3}	$6.2 \pm 0.1 \times 10^{-3}$
g^* (s^{-1})	50.0×10^{-3}	$45.8 \pm 1.4 \times 10^{-3}$
r (s^{-1})	1.0×10^{-3}	$1.01 \pm 0.03 \times 10^{-3}$
$\tau_{L \to H}$ (s)	59.0×10^{3}	$85.2 \pm 3.0 \times 10^{3}$
$\tau_{H \to L}$ (s)	78.7×10^{3}	$105.5 \pm 5.5 \times 10^{3}$
S_I (bits)	8.86	9.23 ± 0.03
S_h (bits)	9.38	9.02 ± 0.02
S_l (bits)	6.25	7.66 ± 0.02
S_{cg} (bits)	1.02	1.01 ± 0.01

Table S1: **Comparison of true rates and predicted rates using MaxCal on alternate self-promotion model.** The first column reports “true” underlying protein synthesis and degradation rates used to create synthetic input data ($f_p = 3.56 \times 10^{-6}$ s^{-1}, $b_p = 1.65 \times 10^{-5}$ s^{-1}), average residence times in the high and low states, and corresponding path informational entropies. Synthetic input data was recorded at $\Delta t = 300$s. The second column reports the average and standard deviation of the same quantities of interest, but extracted using the MaxCal model on ten sets of synthetic data, each consisting of 100 trajectories of 7 days.

High and low state assignment for S_h, S_l, S_{cg}, and dwell times

To assign parts of a trajectory to the low and high state, the locations of the low and high state peaks are used as thresholds ($N = 5$ and $N = 50$ in the case of the Gillespie distribution (blue) of Figure 2A in the main text). Once the protein level is less (greater) than or equal to the lower (upper) threshold, the system is considered to be in the low (high) state. It then remains in that state until it reaches the opposite threshold. This is done to reduce the amount of false positive state switches associated with a single high/low threshold ($N = 25$ in the case of the Gillespie distribution (blue) of Figure 2A in the main text).