Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

Biophysical Journal
Article

Biophysical Society

MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division

Andrew Torres,[3] Spencer Cockerell,[3] Michael Phillips,[3] Gábor Balázsi,[2] and Kingshuk Ghosh[1,3,*]
[1]Molecular and Cellular Biophysics, University of Denver, Denver, Colorado; [2]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, New York; and [3]Department of Physics and Astronomy, University of Denver, Denver, Colorado

ABSTRACT Gene expression is inherently noisy due to small numbers of proteins and nucleic acids inside a cell. Likewise, cell division is stochastic, particularly when tracking at the level of a single cell. The two can be coupled when gene expression affects the rate of cell division. Single-cell time-lapse experiments can measure both fluctuations by simultaneously recording protein levels inside a cell and its stochastic division. These information-rich noisy trajectory data sets can be harnessed to learn about the underlying molecular and cellular details that are often not known a priori. A critical question is: How can we infer a model given data where fluctuations at two levels—gene expression and cell division—are intricately convoluted? We show the principle of maximum caliber (MaxCal)—integrated within a Bayesian framework—can be used to infer several cellular and molecular details (division rates, protein production, and degradation rates) from these coupled stochastic trajectories (CSTs). We demonstrate this proof of concept using synthetic data generated from a known model. An additional challenge in data analysis is that trajectories are often not in protein numbers, but in noisy fluorescence that depends on protein number in a probabilistic manner. We again show that MaxCal can infer important molecular and cellular rates even when data are in fluorescence, another example of CST with three confounding factors—gene expression noise, cell division noise, and fluorescence distortion—all coupled. Our approach will provide guidance to build models in synthetic biology experiments as well as general biological systems where examples of CSTs are abundant.

SIGNIFICANCE Gene expression and cell division dynamics are both stochastic and are experimentally measurable. Analyzing fluctuating time series data of gene expression alone has proven to be a useful avenue to infer underlying details—not available otherwise—of gene networks. The principle of maximum caliber (MaxCal), similar to maximum entropy (MaxEnt) principle but applied to trajectories, has been particularly successful for this purpose. In this work, we further extend MaxCal formalism to infer quantitative models from stochastic trajectories where gene expression and cell division noise are coupled. Using synthetic data we show that MaxCal models can predict underlying details of gene network and protein-dependent division rates from these coupled stochastic trajectories. We also show MaxCal's applicability even when data are in fluorescence and not in protein number, typical in experiments. Success of MaxCal may further motivate collection and analysis of fluctuating time series data to build quantitative models in other branches of biophysics where fluctuations at different levels are coupled.

INTRODUCTION

Synthetic biologists are building new circuits to manipulate cell behavior and ultimately reprogram organismal phenotypes for practical applications (1–17). With the advent of single-cell technologies it is possible to monitor time evolution of multiple cells with gene expression inside individual cells recorded at different time points (18). However, gene expression dynamics is stochastic due to the small number of molecules inside a cell (19–27). The ability to follow a single cell also affords recording of stochastic cellular division events when a mother cell gives rise to two daughter cells. Consequently, experimentally recorded time trajectories encode both the gene expression and cell division noise simultaneously. Sometimes the two stochastic processes, operating at different levels, can even be coupled. Consider a specific

Submitted January 24, 2023, and accepted for publication May 18, 2023.

*Correspondence: kghosh@du.edu

Andrew Torres and Spencer Cockerell contributed equally to this work.

Editor: Jianhua Xing.

https://doi.org/10.1016/j.bpj.2023.05.017

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

protein whose stochastic expression level dictates fitness such as cellular division rate. In a synthetic system studied by Balazsi and coworkers, cell division has been shown to be slower when a particular protein—conferring antibiotic resistance—is expressed in high number (28) in the absence of any stress. Similar decrease or nonmonotonic dependence of cell division rate on the levels of various proteins has been observed in natural genetic systems as well (29–31). When cells are exposed to drugs or other environmental stress, protective protein levels can enhance the division rate (28,32). In such a scenario, fluctuations in protein level and cell division dynamics are coupled and the combined stochastic trajectories can have intriguing topology (see Fig. 1).

Besides the type of coupling mentioned above, there is also a natural coupling between gene expression and cell division. Cells that duplicate faster grow faster in size before dividing. The growth in size dilutes protein content, effectively reducing protein concentration. Thus, cell division and protein abundance are necessarily coupled, either by protein dilution or by direct effect of protein level on fitness/growth rate.

Details of genetic networks give rise to the specific features of noisy trajectories. Conversely, noisy trajectories can hold valuable clues about underlying details of a biological network. This realization has led to detailed studies of stochastic trajectories (33–39) to determine underlying parameters of biological systems that are not known otherwise. Inferring meaningful parameters and/or observables from complex stochastic trajectories is a common problem in other areas of biology as well (40–42). We now ask: Does the idea of inference from noisy time series data work even when we have fluctuating trajectories where protein number and cell division are coupled (similar to Fig. 1)? Can we faithfully infer details of the gene network as well as cellular division and its dependence on the protein number from these stochastic trajectories? In this paper we address this specific problem of "coupled stochastic trajectories" (CSTs) at different scales using the principle of maximum caliber (MaxCal).

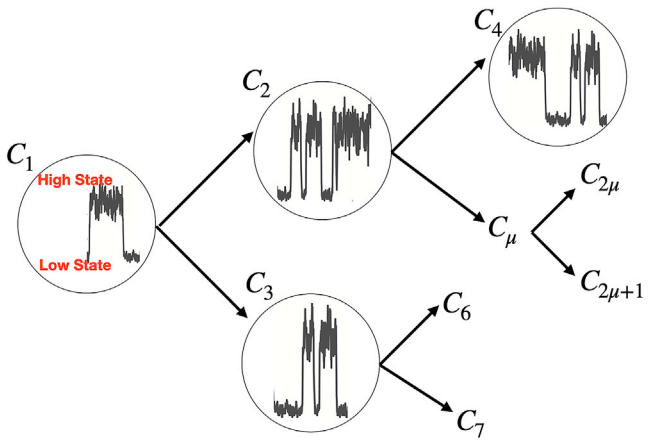

FIGURE 1 Example of cell tree lineage starting from a single mother cell. Each cell (*circle*) has its own stochastic gene expression, e.g., switching between high and low states. Cell division is also stochastic, leading to coupled stochastic trajectories (CSTs). To see this figure in color, go online.

MaxCal is similar in principle to maximum entropy (MaxEnt) but applied to trajectories (43–45). MaxCal maximizes the path entropy subject to constraints and yields trajectory probabilities. Different sets of constraints define different classes of models. MaxCal can build minimal models starting with a few basic constraints, an attractive feature when many details of the underlying networks are not known a priori. For example, a gene network with feedback can involve many underlying molecular actors (such as RNA, protein-RNA complex, protein-protein complex) that are not directly seen in experiments and have unknown interactions. Traditional modeling approaches typically assume the existence of many of these unseen actors and draw reaction networks between them to capture feedback. Stochastic models are then built with these networks with rates of the reactions as parameters (46,47). Corresponding chemical master equations (48) can be solved using a finite state projection method (49,50). This is a bottom-up approach. Another commonly used approach relies on mass-action type formalisms embedded in a chemical master equation where ad hoc nonlinear functional forms are assumed for reaction rates (45,51–53)). The nonlinear functional forms emulate feedback without invoking intermediate species. MaxCal, in contrast, is a top-down approach that avoids assuming ad hoc species or functional forms. It uses only basic system information in the form of constraints. Subsequently, path entropy is maximized subject to the specific set of constraints (model) to predict trajectory probabilities. These trajectory probabilities are used to compute the likelihood of a model given trajectory data. Thus, MaxCal is particularly well suited for inferring a model and relevant parameters from the noisy trajectory data of gene circuits (39,45,54–56). MaxCal also offers a more efficient inference tool compared to traditional bottom-up models (57). This is due to its ability to build minimal models—without invoking auxiliary species/molecular actors—reducing both the state space and numbers of parameters (see (57) for details).

However, MaxCal-based inference has not been applied to CSTs, where different stochastic processes at different levels—such as genetic and cellular—give rise to trajectories that are interwoven with each other. In this paper we apply MaxCal to these higher-dimensional problems with intricate trajectory topology (similar to Fig. 1). We generate synthetic noisy time series data coupling gene expression and cellular division using a known model (protein production, degradation rates, cell division rates, etc.). We then integrate MaxCal within Bayesian formalism to select appropriate models and extract key information

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

from those data. These details are described below in the materials and methods. The parameters of the known model serve as benchmarks and compare well against MaxCal's inferred parameters, indicating MaxCal's ability to detect molecular and cellular details from a CST. MaxCal can also predict several distributional quantities. Finally, we demonstrate MaxCal's ability to infer even when data are presented in noisy fluorescence and not in protein number, typical in experiments. These findings—detailed in the results and discussion—show MaxCal's promise to bridge scales describing stochastic processes at different levels and analyze CSTs that are frequent in biology.

MATERIALS AND METHODS

Generating synthetic data coupling gene expression and cell division

We consider a synthetic circuit in which a gene auto-activates and behaves in a switch-like manner, toggling between the high (H) and low (L) protein number states. Cell division is dependent on the protein state; for example, it is slow when protein number is in the H state and fast when in the L state. Auto-activation in biological circuits is well studied (53,58–64). We adopt the scheme proposed by Kepler and Elston (46) to model the gene network as

$$\begin{aligned} &\alpha \xrightarrow{g} \alpha + A; \quad A \xrightarrow{r} \emptyset; \quad A + A \underset{b_d}{\overset{f_d}{\rightleftarrows}} A_2 \\ &\alpha + A_2 \underset{b_p}{\overset{f_p}{\rightleftarrows}} \alpha^*; \quad \alpha^* \xrightarrow{g^*} \alpha^* + A \\ &C_\mu \xrightarrow{d_A} C_{2\mu} + C_{2\mu+1}. \end{aligned} \tag{1}$$

The gene α produces protein A at a basal rate g. The protein A degrades at a rate of r, and dimerizes to A_2 with forward and backward rates of f_d and b_d, respectively. The dimer can bind and unbind the promoter site α at a rate of f_p and b_p, respectively. The dimer-bound state is an activated state α^*, producing protein A at an enhanced rate g^* (greater than g), modeling the positive feedback. Finally, cell division is modeled by the stochastic reaction where mother cell C_μ divides in two daughter cells $(C_{2\mu}, C_{2\mu+1})$ with rate d_A. Upon division, mother cells on average distribute half of their proteins to each of their daughter cells following a binomial distribution. The division rate depends on this protein number N_A, capturing the coupling between gene expression and cell division. This dependence $d_A(N_A)$ is assumed to be sigmoidal, with slow division rate $d_H = 1 \times 10^{-5}\mathrm{s}^{-1}$ (~24 h) at high protein numbers ($N_A > 25$), and fast division rate $d_L = 3 \times 10^{-5}\mathrm{s}^{-1}$ at low protein numbers ($N_A < 25$). Specifically, we assume the following functional form to simulate protein number-dependent cell division

$$d_A(N_A) = d_L + \frac{d_H - d_L}{1 + \exp\left(\beta\,(25 - N_A)\right)} \tag{2}$$

The sharp-sigmoidal dependence is realized by setting $\beta = 1$. The specific choice of $N = 25$ as the threshold for the sigmoidal rate dependence is motivated by the equilibrium protein number distribution that has two well-separated peaks for $N > 25$ and $N < 25$. This distribution arises from the specific choice of gene network parameters used in this article (see caption of Table 1). Different thresholds can be chosen when using different parameters. We also generated a smoother sigmoidal with more gradual dependence on protein number using $\beta = 0.1$. We considered the function above with $\beta = 1$ and $\beta = 0.1$ as two representative examples among many other possibilities (28–30).

MaxCal model

MaxCal maximizes path entropy or *caliber* subject to constraints enforced by Lagrange multipliers (43,45,54,65–71). For the gene circuit we impose three minimal constraints: 1) protein synthesis, 2) protein degradation, and 3) auto-activation/positive feedback. We do this by introducing averages of several variables over a time interval, Δt. We use a production state variable ℓ_α to describe the number of proteins synthesized within the time frame, with a maximum allowed value of M. Thus, we have $0 \le \ell_\alpha \le M$. Next, the degradation state variable ℓ_A is defined as the number of proteins that remain at the end of the time frame, satisfying $0 \le \ell_A \le N_A$, where N_A is the number of proteins at the beginning of the time step. The constraint of positive feedback is modeled by coupling ℓ_α and ℓ_A and imposing the constraint by the corresponding Lagrange multiplier K_A. This is the first order of coupling between the two variables (see (39) for more). The production and degradation constraints are imposed by Lagrange multipliers h_α and h_A, respectively. Next, cell division is modeled by introducing another state variable $\ell_c = 0, 1$, with $\ell_c = 1$ describing division. Finally, the coupling between cell division and protein number is modeled by constraining the average of $\ell_c \ell_A$. These last two constraints are introduced by Lagrange multipliers h_c and K_{Ac}, respectively. Additional second-order coupling between ℓ_c and ℓ_α is ignored since cell division is not expected to directly depend on protein synthesis. Ignoring this term also keeps the model minimal and lowers computational cost. However, the formalism is general enough to add such constraints as needed. The model can also be generalized by coupling variables at different time points if the underlying dynamics has memory. The probability of a path in the time interval Δt is thus described by $P_{\ell_\alpha,\ell_A,\ell_c}$, with corresponding path entropy $-P_{\ell_\alpha,\ell_A,\ell_c} \log P_{\ell_\alpha,\ell_A,\ell_c}$. These five basic constraints combined with the path entropy yield the caliber as

$$\begin{aligned} \mathcal{C} = &- \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} P_{\ell_\alpha,\ell_A,\ell_c} \log P_{\ell_\alpha,\ell_A,\ell_c} + h_\alpha \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} \ell_\alpha P_{\ell_\alpha,\ell_A,\ell_c} \\ &+ h_A \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} \ell_A P_{\ell_\alpha,\ell_A,\ell_c} + K_A \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} \ell_\alpha \ell_A P_{\ell_\alpha,\ell_A,\ell_c} \\ &+ h_c \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} \ell_c P_{\ell_\alpha,\ell_A,\ell_c} + K_{Ac} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_c=0}^{1} \ell_A \ell_c P_{\ell_\alpha,\ell_A,\ell_c} \end{aligned} \tag{3}$$

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

and the corresponding caliber-maximized path probabilities are

$$\begin{aligned} P_{\ell_\alpha,\ell_A,\ell_c} &= Q^{-1}\binom{N_A}{\ell_A}\exp\left(h_\alpha\ell_\alpha + h_A\ell_A + K_A\ell_\alpha\ell_A + h_c\ell_c + K_{Ac}\ell_A\ell_c\right) \\ Q &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A}\sum_{\ell_c=0}^{1}\binom{N_A}{\ell_A}\exp\left(h_\alpha\ell_\alpha + h_A\ell_A + K_A\ell_\alpha\ell_A + h_c\ell_c + K_{Ac}\ell_A\ell_c\right). \end{aligned} \tag{4}$$

The state variables ℓ_α, ℓ_A, and ℓ_c directly relate to protein synthesis and degradation rates g,g^*, and r in the auto-activation circuit as well as division rates in the low and high states as in Eq. 2. Specifically, these relations are

$$\begin{aligned} d_L &= \frac{\langle\ell_c\rangle_{N_A=N_L}}{\Delta t}; d_H = \frac{\langle\ell_c\rangle_{N_A=N_H}}{\Delta t}; g = \frac{\langle\ell_\alpha\rangle_{N_A=N_L}}{\Delta t}; \\ g^* &= \frac{\langle\ell_\alpha\rangle_{N_A=N_H}}{\Delta t}; r = \frac{N_H - \langle\ell_A\rangle_{N_A=N_H}}{N_H\Delta t} \end{aligned} \tag{5}$$

where N_L and N_H are the number of A proteins found in the low and the high state, respectively. The basal production rate (g) is extracted from the average of ℓ_α when protein number is low (N_L), whereas the activated production rate (g^*) is extracted from ℓ_α when protein number is high (N_H) (also see (39) for these definitions used to model only the gene circuit). The degradation rate (r) is the fractional decrease in protein number, related to average ℓ_A; here we use only the high state (N_H) because it contains the most information on degradation. Similarly, for sharp-sigmoidal dependence of division on protein number, the low and high state division rates (d_L, d_H) are extracted from the average of ℓ_c in the low and high states, respectively, due to abundance of data in these states.

From the path probabilities, Eq. 4, we obtain transition probabilities for each observable change occurring in a single time step (Δt). Inspecting a single-cell labeled μ, which is not dividing, the abundance of protein A changes from i to j according to the joint probability

$$\begin{aligned} \mathcal{P}(i\to j\ \&\ \mu\to\mu) &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i} \\ &\times\sum_{\ell_c=0}^{1}\delta(\ell_\alpha+\ell_A-j)\,\delta(\ell_c)\,P_{\ell_\alpha,\ell_A,\ell_c} \equiv \mathcal{Y}_{ji} \end{aligned} \tag{6}$$

Note that $P_{\ell_\alpha,\ell_A,\ell_c}$ and the normalization factor Q are functions of the initial protein count, $N_A = i$. The shorthand matrix notation $\mathcal{Y}_{ji}$ is introduced for brevity, particularly for the likelihood (see the following section). Also, Eq. 6 is actually independent of the specific cell label μ; our notation indicates that the cell in question is retaining its identity, i.e., not dividing, across this time step. The cell label is included here, and in the following, for clarity and to emphasize connection to data (experimental or synthetic).

When a cell μ does divide, it results in two daughter cells labeled $2\mu; 2\mu+1$. The initial protein count i in mother cell μ can change to j in daughter labeled 2μ and k in daughter labeled $2\mu+1$. Typical experiments will have a sampling time and may not be able to detect protein numbers immediately after cell division. In the time window following cell division, each daughter cell will stochastically evolve their gene network. This compound transition is broken in three contributions: 1) cell μ divides when it has i proteins, 2) given that a division has taken place, daughter 2μ has j proteins at the end of the time frame where data collection happens, and 3) similarly, given division, daughter $2\mu+1$ has k proteins at the same time frame. We write the probability for this process as

$$\begin{aligned} \mathcal{P}(\mu, i\to 2\mu, j\,;\,2\mu+1, k) &= \mathcal{P}(\mu, i\to 2\mu; 2\mu+1) \\ &\times \mathcal{P}(i\to j|\mu\to 2\mu; 2\mu+1) \\ &\times \mathcal{P}(i\to k|\mu\to 2\mu; 2\mu+1) \end{aligned} \tag{7}$$

where the first term on the right hand side is the probability for division of mother cell μ alone and is given by

$$\begin{aligned} \mathcal{P}(\mu, i\to 2\mu; 2\mu+1) &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i} \\ &\times\sum_{\ell_c=0}^{1}\delta(\ell_c - 1)\quad P_{\ell_\alpha,\ell_A,\ell_c} \equiv \mathcal{X}_i \end{aligned} \tag{8}$$

and remaining terms involve the conditional probability for changing protein count i to j given that division has taken place, which reads

$$\begin{aligned} \mathcal{P}(i\to j|\mu\to 2\mu; 2\mu+1) &= \mathcal{X}_i^{-1} \quad \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i} \\ &\times\sum_{\ell_c=0}^{1}\delta(\ell_\alpha+\ell_A-j)\quad\delta(\ell_c-1)\quad P_{\ell_\alpha,\ell_A,\ell_c} \equiv \mathcal{Y}'_{ji}. \end{aligned} \tag{9}$$

With the above shorthand, we may write the r.h.s. of Eq. 7 as $\mathcal{X}_i\mathcal{Y}'_{ji}\mathcal{Y}'_{ki}$.

The stationary distribution $P(N)$, giving the probability of finding N proteins in a single cell at any time, is a practical quantity and can be used for comparison with data or for prediction. It can be calculated from the effective protein transition matrix $\mathcal{A}$, allowing for transitions with and without division,

$$\mathcal{A}_{ji} = \mathcal{Y}_{ji} + \mathcal{X}_i\mathcal{Y}'_{ji}. \tag{10}$$

Watching a single cell for a long time corresponds to applying $\mathcal{A}$ many times, i.e., raising it to a large power u. Then, $P(N)$ may be interpreted as the vector resulting from multiplication by an (arbitrary) initial vector $P_0(N)$,

$$P(N) = \sum_{i=0}^{N_{max}}(\mathcal{A}^u)_{Ni}P_0(i), \tag{11}$$

where we introduce a cutoff in maximum protein number, N_{max}. For the system considered here, we used a value of $N_{max} = 80$ throughout, which captures all protein fluctuations from stochastic simulations we ran. A more rigorous approach would involve finite state projection proposed by Munsky and Khammash (49) to determine errors with assumed values of N_{max} as was done in our earlier work (39). For simplicity and based on our earlier work, we assumed a fixed cutoff here. Note, $P(N)$

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

can also be calculated as the eigenvector of $\mathcal{A}$ corresponding to eigenvalue 1.

Parameter estimation via likelihood

Equations above outline a MaxCal model capable of describing a system of dividing cells, each having independent auto-activation circuits. Below we describe how to choose a MaxCal model given data, i.e., analyzing a CST within a Bayesian inference framework. The first step is to compute the likelihood of a model given trajectory data similar to Fig. 1.

Consider an experimental trajectory of a cell division lineage, starting with one cell, of sufficiently long time $\mathcal{T}$ in the units of timescale Δt. This results in $\mathcal{T}+1$ snapshots of CST data which contain information on the number of cells in the system and the protein number associated with each cell. Each snapshot is given a frame label, $t \in [0, \mathcal{T}]$.

Suppose a particular CST trajectory tree (similar to Fig. 1) gives rise to λ unique cell identities observed over all $\mathcal{T}+1$ snapshots. Each unique cell's identifying number, $\mu \in [1, \lambda]$, is observed from its birth frame, denoted $\tau_\mu \in [0, \mathcal{T}-1]$, until its division event frame, $d_\mu \in [0, \mathcal{T}-1]$. Shortly after frame d_μ begins, cell μ loses its identity: it divides into two new cells with their own identifiers, 2μ and $2\mu+1$. These new cells are first observed in the next frame: $\tau_{2\mu} = \tau_{2\mu+1} = d_\mu + 1$. The likelihood of observing protein numbers of cell μ, given a full trajectory $N_{\mu,t}$ from its birth frame τ_μ to division frame d_μ, and its daughters' protein numbers $N_{2\mu,d_\mu+1}$ and $N_{2\mu+1,d_\mu+1}$ at the frame immediately after division, depends on the set of MaxCal parameters $(h_\alpha, h_A, K_A, h_c, K_{Ac})$ via

$$\begin{aligned}\mathcal{L}_\mu^{\text{div}} &= \left(\prod_{t=\tau_\mu}^{d_\mu-1} \mathcal{P}\left(\mu, N_{\mu,t} \to \mu, N_{\mu,t+1}\right)\right) \quad \mathcal{P}\left(\mu, N_{\mu,d_\mu} \to 2\mu, N_{2\mu,d_\mu+1} \quad ; \quad 2\mu+1, N_{2\mu+1,d_\mu+1}\right) \\ &= \left(\prod_{t=\tau_\mu}^{d_\mu-1} \mathcal{Y}_{N_{\mu,t+1}\ N_{\mu,t}}\right) \mathcal{X}_{N_{\mu,d_\mu}} \mathcal{Y}'_{N_{2\mu,d_\mu+1}\ N_{2\mu,d_\mu}} \mathcal{Y}'_{N_{2\mu+1,d_\mu+1}\ N_{2\mu+1,d_\mu}}.\end{aligned} \tag{12}$$

In the above expression, the first set of terms within parentheses accounts for all transitions in protein numbers when the cell has not divided. Assuming these are independent, we write that contribution as a product of transitions in protein number for all steps before division. The last three terms describe cell division alongside transitions in protein number for the two daughter cells, capturing the coupling between cell division and protein number observed in data. The conditional probability $\mathcal{Y}'$ of protein number change given a cell has divided is written in Eq. 9. This choice does not impose any specific mechanism of protein partitioning from the mother to the daughter cell, rather it assumes the same gene expression circuit operates at the moment of division. However, if additional information such as partitioning statistics of proteins from mother to daughter cell is known, it can be used. For example, a possible partitioning statistics can follow a binomial distribution given by

$$\mathcal{Y}^b_{ji} = \binom{i}{j} \Big/ 2^i \tag{13}$$

where i is the number of proteins in the mother cell, j and $i-j$ are the number of proteins inherited by the two daughter cells. The last observed generation of cells have not divided, so their trajectories will give simpler likelihoods,

$$\begin{aligned}\mathcal{L}_\mu^{\text{no div}} &= \prod_{t=\tau_\mu}^{\mathcal{T}-1} \mathcal{P}\left(\mu, N_{\mu,t} \to \mu, N_{\mu,t+1}\right) \\ &= \prod_{t=\tau_\mu}^{\mathcal{T}-1} \mathcal{Y}_{N_{\mu,t+1}\ N_{\mu,t}}.\end{aligned} \tag{14}$$

The likelihood of observing the whole CST tree with λ unique cell labels is the product of all individual likelihoods,

$$\mathcal{L}_{CST} = \prod_{\mu=1}^{\lambda} \mathcal{L}_\mu, \tag{15}$$

using the division or no division cases, Eqs. 12 or 14, as appropriate for each cell. Note that the total cell population C_{tot} at the end of the experiment, will be less than the number of labels, i.e., $C_{tot} < \lambda$, because each label encapsulates the cell's generation. For example, if Fig. 1 up to C_2, C_6, C_7 is taken as a full CST tree, the ending population number is $C_{tot} = 3$ with a unique cell identity set written as $\{1, 2, 3, 6, 7\}$, giving $\lambda = 5$ unique cell identity labels.

These synthetic CSTs and also real life CSTs have no limit on how many proteins a single cell can create in a single step transition while the MaxCal framework presented so far only accounts for a maximum limit of M created proteins. This discrepancy is known to penalize the most likely model parameters erroneously (39). This can be avoided by considering transitions across m subsequent frames as shown in our earlier works (39,55). Hence, we sample the original trajectory in units of $m\Delta t$. This new sampling redefines the birth (τ_μ) and death (d_μ) frames for each cell (μ). With this new definition of frames, the modified likelihood can be written as

$$\begin{aligned}\mathcal{L}^{\text{div}}_{\mu,m} = &\left(\prod_{t=\tau_\mu}^{d_\mu-1} \mathcal{Y}^m_{N_{\mu,t+1}\ N_{\mu,t}}\right)\Bigg[\sum_{n=1}^{m}\sum_{j=0}^{N_{max}} \left(\mathcal{Y}^{n-1}\right)_{jN_{\mu,d_\mu}} \\ &\times \mathcal{X}_j \left(\mathcal{Y}^{m-n}\mathcal{Y}'\right)_{N_{2\mu,d_\mu+1}\ j} \left(\mathcal{Y}^{m-n}\mathcal{Y}'\right)_{N_{2\mu+1,d_\mu+1}\ j}\Bigg]\end{aligned} \tag{16}$$

The first term in Eq. 16 describes protein number transitions without division, as in Eq. 12 but across m steps of the original sampling time (Δt). The second term, in square brackets, captures protein number transitions alongside division, allowing for division to happen at any particular frame within the last m-frame interval. $\mathcal{Y}'$ is given by Eq. 9 without assuming any additional mechanism of protein partitioning from mother to daughter cells. However, a specific form of partitioning proteins from mother to daughter cells can be used if a mechanism is known such as binomial splitting given by Eq. 13.

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

In similarity to Eq. 14, we write the likelihood for the last generation of cells (those that do not divide) in steps of m frames as

$$\mathcal{L}_{\mu,m}^{\text{no div}} = \left(\prod_{t=\tau_\mu}^{\mathcal{T}-1} \mathcal{Y}^m_{N_{\mu,t+1}\ N_{\mu,t}} \right). \tag{17}$$

where $\mathcal{T}$ is now the total number of frames in the new sampling (with $m\Delta t$).

The parameter m must be chosen carefully such that, for the vast majority of the given CST tree trajectory, just one or zero division events take place in any step of m frames. Any cases with more divisions within m frames are excluded in the trajectory products in Eqs. 16 and 17. The overall framework allows for the possibility of having more than one division event in m frames, by adding more complexity to this second term in Eq. 16. However, we keep the above formalism for simplicity and chose m to minimize cases where two or more cell division events may take place in the time step $m\Delta t$.

The likelihood for the full CST tree using m-frame intervals is obtained from the product as in Eq. 15, using Eqs. 16 or 17 for each individual cell's likelihood, for cells that divide or do not divide, respectively.

The CST likelihood function ($\mathcal{L}_{CST}$) allows us to carry out Bayesian inference given the trajectory data. The likelihood is a function of MaxCal parameters $(h_\alpha, h_A, K_A, h_c, K_{Ac})$ and can be thought of as a Bayes theorem probability function with no known prior distribution. With this formalism, we can consider the probability associated with each point $(h_\alpha, h_A, K_A, h_c, K_{Ac})$ given a particular CST data set as,

$$P(h_\alpha, h_A, K_A, h_c, K_{Ac}|CST) = \frac{P(h_\alpha, h_A, K_A, h_c, K_{Ac})\mathcal{L}_{CST}(h_\alpha, h_A, K_A, h_c, K_{Ac})}{P(CST)} \tag{18}$$

Without any knowledge about the prior, the posterior distribution of MaxCal parameters given CST data are simply proportional to the likelihood of observing data given a model (parameters). Thus, the most likely model (parameters) can be determined by maximizing the likelihood. However, Bayesian formalism gives the entire probability distribution of parameters, which can be used to obtain averages $(\langle h_\alpha\rangle, \langle h_A\rangle, \langle K_A\rangle, \langle h_c\rangle, \langle K_{Ac}\rangle)$ and standard deviations $(\sigma_{h_\alpha}, \sigma_{h_A}, \sigma_{K_A}, \sigma_{h_c}, \sigma_{K_{Ac}})$ of the parameters. The average of any quantity F is calculated using

$$\langle F\rangle = \left[\int_{\{\}} F(h_\alpha, h_A, K_A, h_c, K_{Ac})\mathcal{L}_{CST}(h_\alpha, h_A, K_A, h_c, K_{Ac}) \right] \left[\int_{\{\}} \mathcal{L}_{CST}(h_\alpha, h_A, K_A, h_c, K_{Ac}) \right]^{-1} \tag{19}$$

where $\{\}$ in the integration denotes all values of Lagrange multipliers. Standard deviation of quantity F is calculated using $\sigma_F = \sqrt{\langle F^2\rangle - \langle F\rangle^2}$. The posterior distributions are calculated using Metropolis-Hastings algorithm (72) by drawing one million samples.

Dealing with experimental data

The synthetic data used to find the likelihoods in the procedure above assume CST data to be in the form of protein number for each cell. In a typical experimental setting, data are instead recorded in arbitrary fluorescence units. To mimic experimental CST data, we distort the synthetic data by converting protein number to fluorescence using a probability distribution. Protein to fluorescence conversion is not one fixed number, rather it is noisy due to several uncertainties. We assume fluorescence per protein is Gaussian (41,73,74) with average a and variance b^2. The fluorescence from N proteins will be a convolution of N Gaussian distributions, each with mean a and standard deviation b. The resulting distribution will be Gaussian with mean Na and variance Nb^2. Drawing from this distribution, we convert the synthetic data in protein number trajectory ($N_{\mu,t}$) to noisy fluorescence trajectory ($f_{\mu,t}$). We vary levels of corruption in the conversion to fluorescence by varying b/a. Specifically, we use $a = 100$ with $b = 30$ and $b = 50$ yielding a fluorescence to protein conversion noise level (b/a) of 30% and 50%, typical in experiments (see (55) for more details). When inferring from this trajectory we assume the fluorescence to protein number distribution is known, i.e., a and b are known. Additional but separate photobleaching experiments can be done to determine a and b (75–87). Incorporating this distribution into Eq. 16, we obtain a new likelihood for dividing cell μ with transitions across m frames as

$$\mathcal{L}_{\mu,m}^{\text{fluor,div}} = \left(\prod_{t=\tau_\mu}^{d_\mu-1} \sum_{i,j=0}^{N_{max}} \mathcal{Y}^m_{ji} \ \ \Phi\left(j|f_{\mu,t+1}\right) \ \ \Phi\left(i|f_{\mu,t}\right) \right) \times \left[\sum_{n=1}^{m} \sum_{i,j,k,\ell=0}^{N_{max}} \left(\mathcal{Y}^{n-1}\right)_{ji} \mathcal{X}_j \left(\mathcal{Y}^{m-n}\mathcal{Y}'\right)_{kj} \left(\mathcal{Y}^{m-n}\mathcal{Y}'\right)_{\ell j} \Phi \left(i|f_{\mu,d_\mu}\right) \Phi\left(k|f_{2\mu,d_\mu+1}\right) \Phi\left(\ell|f_{2\mu+1,d_\mu+1}\right) \right] \tag{20}$$

For the last generation of cells, which do not divide, we have

$$\mathcal{L}_{\mu,m}^{\text{fluor,no div}} = \left(\prod_{t=\tau_\mu}^{\mathcal{T}-1} \sum_{i,j=0}^{N_{max}} \mathcal{Y}^m_{ji} \Phi\left(j|f_{\mu,t+1}\right) \Phi\left(i|f_{\mu,t}\right) \right). \tag{21}$$

The likelihood for the full CST tree is obtained as before, using Eq. 15 with dividing and nondividing cells referring to Eqs. 20 and 21, respectively.

Assuming the Gaussian distribution for fluorescence from protein count, $p(f|N) = (2Nb^2\pi)^{-1/2} \exp\left(-(f-Na)^2/2Nb^2\right)$, we obtain the distribution of protein count given fluorescence using Bayes' theorem,

$$\Phi(N|f) = \frac{p(f|N)\ P(N)}{\sum_N p(f|N)\ P(N)}. \tag{22}$$

Here, we also invoke the stationary distribution $P(N)$, as in Eq. 11.

RESULTS AND DISCUSSION

MaxCal can infer underlying rate parameters for sharp-sigmoidal protein number-dependent division rate

Using the procedures described above, we determine optimal Lagrange multipliers h_α, h_A, K_A, h_c, K_{Ac} in two ways. First, we maximize the likelihood (Eq. 15) for a given stochastic trajectory. In the second method, we find the

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

TABLE 1 Inferred effective rates from protein number trajectories using MaxCal is compared against the true rates

	g (s^{-1})	g^* (s^{-1})	r (s^{-1})	d_L (s^{-1})	d_H (s^{-1})
True rates	5.0×10^{-3}	50.0×10^{-3}	1.0×10^{-3}	3.0×10^{-5}	1.0×10^{-5}
Predicted rates (maximum likelihood)	4.66	50.0	1.013	3.0	0.8
Predicted rates (using posterior)	4.66 ± 0.05	50.0 ± 0.1	1.013 ± 0.004	3.0 ± 0.2	0.8 ± 0.2

The first row reports true underlying protein synthesis and degradation rates used to create synthetic input data (with $f_d = 5.0\times10^{-3}s^{-1}$, $b_d = 50s^{-1}$, $f_p = 6.0\times10^{-3}s^{-1}$, $b_p = 3.0\times10^{-5}s^{-1}$) assuming the intrinsic timescale is in seconds, as well as the cell division time in the high and low states ($\beta = 1$ was used). Synthetic input data were recorded at $\Delta t = 300$s. The second row reports the same quantities of interest, but extracted using the minimal MaxCal model and maximum likelihood with corresponding Lagrange multipliers $h_\alpha = -0.583$, $h_A = 0.266$, $K_A = 0.0376$, $h_c = -4.66$, $K_{Ac} = -0.0396$. The third row reports average rates and their standard deviation obtained from the posterior distribution of the Lagrange multipliers. The average and standard deviation of the Lagrange multipliers are $h_\alpha = -0.583\pm0.005$, $h_A = 0.26\pm0.01$, $K_A = 0.0376\pm0.0005$, $h_c = -4.66\pm0.08$, $K_{Ac} = -0.040\pm0.007$. Inference was carried out by sampling data with $\Delta t = 300$s, $m = 6$, $M = 16$.

average and standard deviation from the full posterior distribution of the Lagrange multipliers (Eq. 19) for the same trajectory. The stochastic trajectory in terms of protein number continued for a total of 200,000 s starting from a single cell with zero proteins expressed and gave rise to $\lambda = 415$ unique cell identities (following $\beta = 1$ cell division dynamics) at the end of the allotted time. Data on protein number (N_A) for each cell as well as number of cells were recorded after every $\Delta t = 300$ s time step. These numbers were chosen to match reasonable experimental conditions (6).

Effective values for the underlying protein production, protein degradation, and cell division rate parameters were determined for both the maximum likelihood and average likelihood MaxCal models by using (Eq. 5). The standard deviations of the average rate parameters were also calculated using the posterior distributions (provided in the supporting material). These results are given in Table 1 and compared against the known values to provide a quantitative estimate of MaxCal's ability to infer from CST.

MaxCal inferred underlying cell division rates (d_H, d_L), effective protein production rate (g^*) in the activated state, and degradation rate (r) match well with the "true" rates used to create the synthetic CST data. Only the true rate value of the low state protein creation rate g is not within the standard deviation from the rates inferred by the MaxCal model. This discrepancy could be because MaxCal is not an exact mapping of the underlying model, in fact it is a low-dimensional description of the detailed model used to generate the data. As a result, inferred rates from (Eq. 5) are only approximations of the underlying model and deviations are expected. Nevertheless, the inferred average value of g is within 7% of the true value. Posterior distributions of the inferred rates and their comparison to the true values can be found in the supporting material (see Fig. S1). As a further check, we compared MaxCal-predicted protein number distributions with that of the synthetic data (see Fig. 2). Protein number distribution for the synthetic data was gathered by following the gene network dynamics for a single cell using the autoactivation scheme used in Eq. 1 without the cell division dynamics (rest of the parameters were same as the ones reported in the caption of Table 1).

The number of skipped frames $m = 6$ was chosen to assure sufficient cell division events have been recorded while also mimicking experimentally realistic sampling times akin to that of observing an actively dividing cell colony every 30 min. The inference was further carried out for an upper bound $m = 12$ corresponding to a maximum observational time interval of 1 h. The results for $m = 12$ case (see Fig. S2 and Table S1) are in good agreement with true values, showing the robustness of the inference scheme irrespective of the choice of m values.

The inference above was done with $\mathcal{Y}'$ given by Eq. 9, which does not assume specific details about how proteins from mother cells are distributed to the daughter cells. As a proof of concept we also performed additional calculation assuming proteins from mother cell are partitioned to the daughter cells following a binomial distribution ($\mathcal{Y}' = \mathcal{Y}^b$ given by Eq. 13). The assumed mechanism is consistent with the creation of the synthetic data. Not surprisingly, inferred rates agree well with true rates (see Table S3).

An additional analysis was done by generating CSTs where cell division rates in the high and low states are identical ($d_H = d_L = 3.0\times10^{-5}s^{-1}$). We first carried out

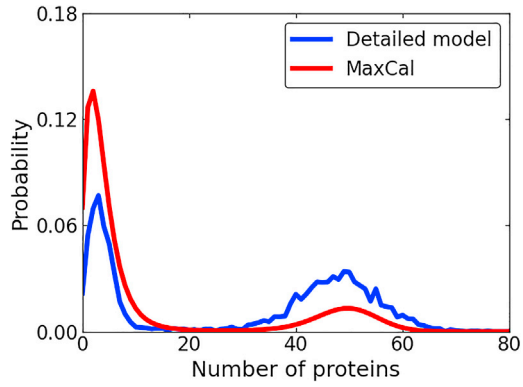

FIGURE 2 MaxCal-predicted protein number distributions (*red*) agree well with the detailed model (*blue*) generated distribution for $\beta = 1$. Underlying reaction rates used for the detailed model are reported in generating synthetic data coupling gene expression and cell division section and legend in Table 1. Lagrange multipliers for MaxCal model are the ones inferred using the maximum likelihood optimization reported in the legend to Table 1. To see this figure in color, go online.

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

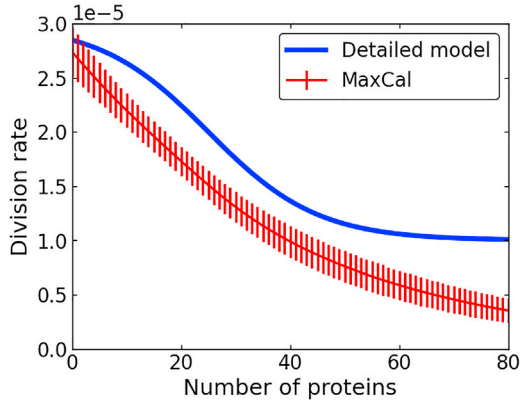

FIGURE 3 Protein number-dependent division rate $d(N)$ predicted by MaxCal (*red graph*) is compared against the true division rate (*blue*) generated using $\beta = 0.1$ in Eq. 2. MaxCal-predicted rates and error bars were calculated from the rate distributions (shown in the supporting material) obtained from the posterior distribution of the Lagrange multipliers. To see this figure in color, go online.

inference using $\mathcal{Y}'$ given by Eq. 9. Inferred rates for the gene network and cell division rates again agree reasonably well with the true values used to generate the data. Cell division rates in the high and low protein states are comparable but not equal. Next we carried out the inference using $\mathcal{Y}' = \mathcal{Y}^b$. The inferred rates are again in good agreement with the true rates for gene network. Moreover, cell division rates in the high and low states are closer and are within the standard deviation. The average of the Lagrange multiplier K_{Ac} with its standard deviation encloses zero, implying that cell division and protein number are not coupled in this control case. Results for this control study can be found in the supporting material (see Fig. S3 and Tables S2 and S4).

MaxCal can infer smooth-sigmoidal protein number-dependent division rate

Cellular division can have both sharp-sigmoidal and smooth-sigmoidal (gradual) dependence on gene expression (30). We generated synthetic CST data using $\beta = 0.1$ in Eq. 2 to model a less sensitive (smooth-sigmoidal) protein number-dependent cellular division, in contrast to $\beta = 1$ used to describe sharp-sigmoidal dependence discussed above. We inferred the optimal Lagrange multiplier values by maximizing the likelihood of the synthetic trajectory data. The inferred values of the Lagrange multipliers predict division rate ($d(N)$) as a function of protein number (N). This is in contrast to the $\beta = 1$ data set where we predicted only two relevant division rates, d_H and d_L. These two rates (d_H and d_L) represent almost uniform division rate in all of high and all of low states, respectively. However, for $\beta = 0.1$ with smooth dependence of division rate on the protein number, it is useful to predict the entire spectrum of rate values. Inferring complete protein number-dependent division rates (fitness) is useful to learn evolution of these circuits inside an organism under different stressors (88). Fig. 3 shows comparison between MaxCal-predicted protein number-dependent division rates (*red*) against the true rates (*blue*) used to generate the synthetic data. Error bars are calculated from the posterior distribution of the Lagrange multipliers. MaxCal-predicted rates show maximum deviation from the true rate values at high protein numbers. It is important to realize proposed Hamiltonian is only an approximation of the mechanism used to generate the synthetic data. The observed discrepancy between the predicted and true rates is a reflection of this approximation. Moreover, the region of maximum discrepancy has insufficient statistics due to inherently small probability of protein numbers in this regime. On the other hand, MaxCal-predicted rates agree well with true values where protein number distribution is significant. As an example, when protein number probability is greater than 0.05 the deviation between the predicted and the true division rate is at most 11%.

As before, predicted values of the parameters of the gene network are in reasonable agreement with the true values (see Table 2). Posterior distributions of these rates are provided in Fig. S4. We also notice MaxCal-predicted protein number distribution agrees well with the true distribution (see Fig. S5). As with the sharp-sigmoid ($\beta = 1$) and control cases, we carried out additional inference using $\mathcal{Y}^b$ to better model the synthetic data. As expected, inferred parameters of the gene network

TABLE 2 Inferred effective rates from protein number trajectories using MaxCal is compared against the true rates for a CST generated with $\beta = 0.1$ in Eq. 2

	g (s^{-1})	g^* (s^{-1})	r (s^{-1})
True rates	5.0×10^{-3}	50.0×10^{-3}	1.0×10^{-3}
Predicted rates (maximum likelihood)	4.77	50.0	1.013
Predicted rates (posterior)	4.77 ± 0.04	49.9 ± 0.1	1.014 ± 0.004

The first row reports true underlying protein synthesis in the basal (g) and activated state (g^*) and degradation rate (r) used to create the synthetic input data (with $f_d = 5.0 \times 10^{-3} s^{-1}$, $b_d = 50 s^{-1}$, $f_p = 6.0 \times 10^{-3} s^{-1}$, $b_p = 3.0 \times 10^{-5} s^{-1}$, $\beta = 0.1$) assuming the intrinsic timescale is in seconds. Synthetic input data were recorded at $\Delta t = 300$s. The second row reports the same quantities of interest, but extracted using the maximum likelihood MaxCal model with corresponding Lagrange multipliers ($h_\alpha = -0.573$, $h_A = 0.276$, $K_A = 0.0369$, $h_c = -4.80$, $K_{Ac} = -0.0371$). The third row reports average rates and their standard deviation obtained from the posterior distribution of the Lagrange multipliers. The average and standard deviation of the Lagrange multipliers are ($h_\alpha = -0.573 \pm 0.004$, $h_A = 0.28 \pm 0.01$, $K_A = 0.0369 \pm 0.0004$, $h_c = -4.80 \pm 0.08$, $K_{Ac} = -0.038 \pm 0.006$). Inference was carried out with $m = 6$ and $M = 16$.

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

TABLE 3 Inferred effective rates from synthetic fluorescence trajectory generated with sharp-sigmoidal protein number-dependent division rate ($\beta = 1$) are compared against the true rates

	g (s^{-1})	g^* (s^{-1})	r (s^{-1})	d_L (s^{-1})	d_H (s^{-1})
True rates	5.0×10^{-3}	50.0×10^{-3}	1.0×10^{-3}	3.0×10^{-5}	1.0×10^{-5}
0% (max likelihood)	4.66	50.0	1.013	3.0	0.8
$b/a = 30\%$ (maximum likelihood)	4.69	50.6	1.051	3.0	0.9
$b/a = 50\%$ (maximum likelihood)	4.35	51.34	1.089	3.0	0.9
$b/a = 30\%$ (using posterior)	4.6 $\pm$0.1	50.7 $\pm$0.1	1.050 $\pm$0.005	3.0 $\pm$0.2	0.8 $\pm$0.2
$b/a = 50\%$ rates (using posterior)	4.35 $\pm$0.05	51.37 $\pm$0.09	1.089 $\pm$0.006	3.0 $\pm$0.2	0.9 $\pm$0.2

The first row reports true underlying protein synthesis, degradation, and cell division rates used to create the synthetic input data (same rates and conditions as Table 1). The second row reports MaxCal inferred rates when trajectories are in protein number (same as values reported in the second row of Table 1). Rows 3 to 4 report inferred rates using maximum likelihood optimization on synthetically corrupted trajectories generated using different values of b/a (indicated in column 1). Rows 5 and 6 report the average and standard deviation of effective rates (in the same fluorescence trajectories) using the full posterior distribution of the Lagrange multipliers.

and protein number-dependent cell division rates (with error) agree reasonably well with the true rates (see Table S5 and Fig. S6).

MaxCal can infer models from noisy fluorescence trajectories

The above section describes the utility of MaxCal when CST data are expressed in protein number. However, as argued before, experimental readouts are often in fluorescence instead of protein numbers. Fluorescence per protein can be a Gaussian distribution with a given mean (a) and variance (b^2). Using the procedure described in the materials and methods, we created two synthetic trajectories in fluorescence where $b/a = 0.3$ and $b/a = 0.5$. The underlying protein number trajectory was identical to the one used earlier with parameters noted in Table 1. For the protein number-dependent cell division we first considered the sigmoidal dependence with $\beta = 1$. We inferred underlying parameters with MaxCal using two procedures, maximum likelihood optimization (rows 3 and 4 in Table 3) and computing averages using the full posterior distributions (rows 5 and 6 in Table 3). MaxCal inferred rates match reasonably well with true rates even when data are significantly distorted due to noisy fluorescence as high as $b/a = 0.5$. As expected, the discrepancy between the inferred and true rates increases with increased level of corruption (greater b/a). MaxCal-predicted protein number distribution (*blue*) agrees well with the actual protein number distribution (*red*) for $b/a = 0.3$, but deviates for higher degree of corruption with $b/a = 0.5$ (see Fig. 4).

Next, we challenged MaxCal to predict smooth-sigmoidal protein number-dependent division rate ($\beta = 0.1$) when the observed trajectory is reported in fluorescence. We notice reasonable agreement between the true rate dependence (*blue*) and MaxCal-predicted division rates (*red*) inferred from synthetic trajectories generated with $b/a = 0.3$ (left panel Fig. 5) and $b/a = 0.5$ (right panel Fig. 5). Error bars are again calculated using the posterior distribution. Inferred parameters of the gene networks also agree will with the true values (see Table 4). MaxCal-predicted protein number distribution matches with the true distribution when $b/a = 0.3$ but starts to deviate for $b/a = 0.5$ (see Fig. S7). These overall agreements demonstrate that MaxCal can reasonably infer underlying details of gene networks as well as protein number-dependent cell division

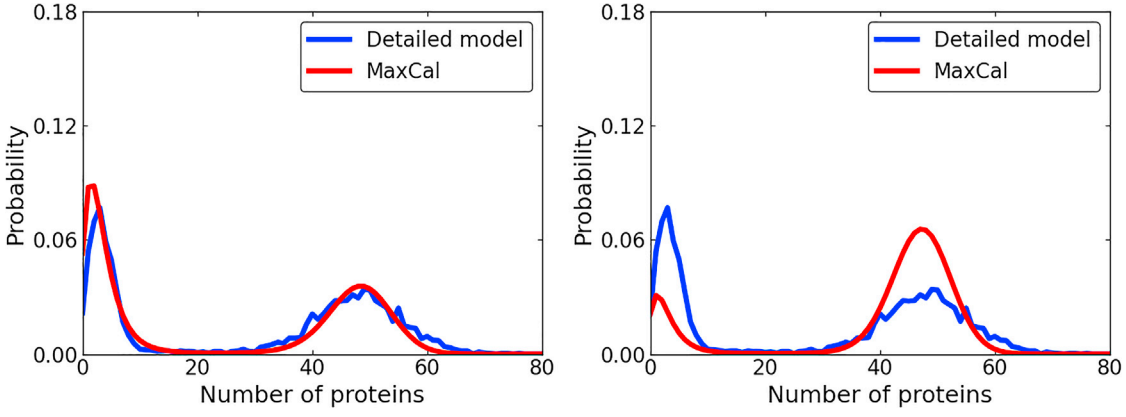

FIGURE 4 MaxCal-predicted protein number distributions (in *red*) generated from maximum likelihood Lagrange multipliers inferred from fluorescence trajectories with $b/a = 30\%$ (*left*) and 50% (*right*) are compared against the true distribution (*blue*). Synthetic trajectories were generated using $\beta = 1$. To see this figure in color, go online.

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

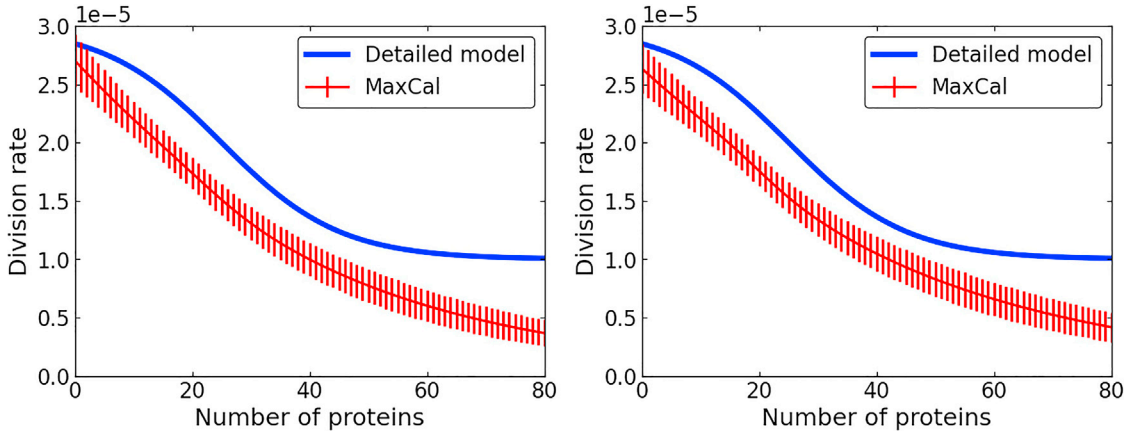

FIGURE 5 Protein number-dependent division rate $d(N)$ predicted by MaxCal (*red*) compares well against the true division rate (*blue*) when $\beta = 0.1$. Two synthetic fluorescence trajectories were generated with $b/a = 0.3$ (*left*) and $b/a = 0.5$ (*right*) and analyzed to infer the underlying rate dependence using MaxCal. To see this figure in color, go online.

rates even when cell division, gene expression, and protein to fluorescence conversion noise are all coupled. From our previous analysis of protein number trajectories, we noticed inference using $\mathcal{Y}^r$ performs reasonably well and only a slight improvement is gained by using $\mathcal{Y}^b$. Thus, we omitted additional calculations using $\mathcal{Y}^b$ for fluorescence trajectories, although such calculations can be performed if desired.

CONCLUSION

We developed a formalism to describe the CSTs of cell division and gene expression using the principle of MaxCal. The underlying gene network is a single-gene auto-activating switch that slows down cell division when in the activated state. The minimal MaxCal model maximizes path entropy subject to five constraints. Three of the five constraints capture details of gene network by constraining average protein production, degradation rates, and feedback/auto-activation. Two additional constraints impose average cell division rate and its coupling to protein level. The corresponding five Lagrange multipliers are determined from synthetic trajectory data within a Bayesian formalism. Inferred Lagrange multipliers can be further used to determine five underlying rates of the systems that are not directly accessible from time trajectory data. These are effective rates of: protein production in the basal and activated states, protein degradation, and cell division rates in the high and low states (of gene expression for sharp-sigmoidal case). Using synthetic trajectories generated with known rates, we show that MaxCal can infer these otherwise unknown rates reasonably well. MaxCal can also infer details of the network and protein number-dependent cell division rates when cellular division depends on protein number in a gradual manner, instead of a sharp-sigmoidal dependence. Bayesian formalism allows predicting errors and full posterior distribution of the inferred rates. Finally, we show MaxCal inference scheme can be used to provide reasonable estimates of rates even when data are in fluorescence and not in protein numbers, typical in experiments. The success of MaxCal-based inference presented here will motivate future studies of other complex genetic circuits where cell division depends on protein number and the underlying protein number distribution is not bimodal, for example in case of oscillatory circuits such as the repressilator (3, 56). Beyond gene network, MaxCal can also be used to model other complex biological systems where processes at different scales are coupled, and time dependent CSTs are experimentally measured but underlying details are unknown.

TABLE 4 Parameters of the genetic network inferred from the synthetic fluorescence trajectory generated with slowly varying protein number-dependent division rate ($\beta = 0.1$) compared against the true values (first row)

	$g\ (s^{-1})$	$g^*\ (s^{-1})$	$r\ (s^{-1})$
True rates	5.0×10^{-3}	50.0×10^{-3}	1.0×10^{-3}
0% (maximum likelihood)	4.77	50.0	1.013
$b/a = 30\%$ (maximum likelihood)	4.81	50.56	1.055
$b/a = 50\%$ (maximum likelihood)	4.46	51.39	1.097
$b/a = 30\%$ (using posterior)	4.81 ± 0.05	50.57 ± 0.09	1.055 ± 0.004
$b/a = 50\%$ (using posterior)	4.46 ± 0.04	51.39 ± 0.08	1.097 ± 0.004

Second, third, and fourth rows report inferred values for different levels of b/a using maximum likelihood approach, while the last two rows report average rates and their standard deviations using the full posterior distribution.

Please cite this article in press as: Torres et al., MaxCal can infer models from coupled stochastic trajectories of gene expression and cell division, Biophysical Journal (2023), https://doi.org/10.1016/j.bpj.2023.05.017

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2023.05.017.

AUTHOR CONTRIBUTIONS

G.B. and K.G. designed the research. A.T., S.C., and M.P. performed the research. A.T. and S.C. contributed equally to the manuscript. All authors analyzed results and participated in writing.

ACKNOWLEDGMENTS

K.G. acknowledges support from the National Institutes of Health award number R15GM128162 and R01GM13890. G.B. was supported by the National Institutes of Health, NIGMS MIRA Program (R35 GM122561) and by the Laufer Center for Physical and Quantitative Biology. M.P. is supported by the National Science Foundation MPS-Ascend Postdoctoral Research Fellowship under grant no. DMR-2213103.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Brophy, J., K. Magallon, …, J. Dinneny. 2022. Synthetic gene circuits as a means of reprogramming plant roots. *Science.* 377:747–751.
2. Gardner, T., C. Cantor, and J. Collins. 2000. Construction of a genetic toggle switch in Escherichia coli. *Nature.* 403:339–342.
3. Elowitz, M. B., and S. Leibler. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature.* 403:335–338.
4. Alon, U. 2007. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8:450–461.
5. Tsai, T. Y. C., Y. S. Choi, …, J. E. Ferrell, Jr. 2008. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science.* 321:126–129.
6. Nevozhay, D., R. M. Adams, …, G. Balázsi. 2012. Mapping the environmental fitness landscape of a synthetic gene circuit. *PLoS Comput. Biol.* 8, e1002480.
7. Nevozhay, D., R. M. Adams, …, G. Balázsi. 2009. Negative auto regulation linearizes the dose-response and suppresses the heterogeneity of gene expression. *Proc. Natl. Acad. Sci. USA.* 106:5123–5128.
8. Lyons, S. M., W. Xu, …, A. Prasad. 2014. Loads bias genetic and signaling switches in synthetic and natural systems. *PLoS Comput. Biol.* 10, e1003533.
9. Wang, L. Z., F. Wu, …, X. Wang. 2016. Build to understand: synthetic approaches to biology. *Integr. Biol.* 8:394–408.
10. Mukherji, S., and A. van Oudenaarden. 2009. Synthetic biology: understanding biological design from synthetic circuits. *Nat. Rev. Genet.* 10:859–871.
11. Wu, F., and X. Wang. 2015. Applications of synthetic gene networks. *Sci. Prog.* 98:244–252.
12. Gómez Tejeda Zañudo, J., M. T. Guinn, …, R. Albert. 2019. Towards control of cellular decision-making networks in the epithelial-to-mesenchymal transition. *Phys. Biol.* 16, 031002.
13. Charlebois, D. A., G. Balázsi, and M. Kærn. 2014. Coherent feedforward transcriptional regulatory motifs enhance drug resistance. *Phys. Rev.* 89, 052708.
14. Khalil, A. S., and J. J. Collins. 2010. Synthetic biology: applications come of age. *Nat. Rev. Genet.* 11:367–379.
15. Abil, Z., X. Xiong, and H. Zhao. 2015. Synthetic biology for therapeutic applications. *Mol. Pharm.* 12:322–331.
16. Isaacs, F. J., and L. You. 2009. A brave new synthetic world. *Genome Biol.* 10:302.
17. Zhou, Z., Y. Liu, …, N. Hao. 2023. Engineering longevity-design of a synthetic gene oscillator to slow cellular aging. *Science.* 380:376–381.
18. Aymoz, D., V. Wosika, …, S. Pelet. 2016. Real-time quantification of protein expression at the single-cell level via dynamic protein synthesis translocation reporters. *Nat. Commun.* 7, 11304.
19. Ozbudak, E. M., M. Thattai, …, A. van Oudenaarden. 2002. Regulation of noise in the expression of a single gene. *Nat. Genet.* 31:69–73.
20. Kaern, M., T. C. Elston, …, J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6:451–464.
21. Paulsson, J. 2004. Summing up the noise in gene networks. *Nature.* 427:415–418.
22. 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.
23. Sanchez, A., and J. Kondev. 2008. Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. USA.* 105:5081–5086.
24. Shahrezaei, V., and P. S. Swain. 2008. The stochastic nature of biochemical networks. *Curr. Opin. Biotechnol.* 19:369–374.
25. Elowitz, M. B., A. J. Levine, …, P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science.* 297:1183–1186.
26. Tao, Y. 2004. Intrinsic and external noise in an auto-regulatory genetic network. *J. Theor. Biol.* 229:147–156.
27. Beard, D. A., and H. Qian. 2008. Chemical Biophysics: Quantitative Analysis of Cellular Systems. University Press, Cambridge.
28. Nevozhay, D., R. M. Adams, …, G. Balázsi. 2012. Mapping the environmental fitness landscape of a synthetic gene circuit. *PLoS Comput. Biol.* 8, e1002480.
29. Dekel, E., and U. Alon. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature.* 436:588–592.
30. Keren, L., J. Hausser, …, E. Segal. 2016. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell.* 166:1282–1294.e18.
31. Scott, M., C. W. Gunderson, …, T. Hwa. 2010. Interdependence of cell growth and gene expression: origins and consequences. *Science.* 330:1099–1102.
32. Farquhar, K. S., D. A. Charlebois, …, G. Balázsi. 2019. Role of network-mediated stochasticity in mammalian drug resistance. *Nat. Commun.* 10:2766.
33. Munsky, B., B. Trinh, and M. Khammash. 2009. Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.* 5:318.
34. Lillacci, G., and M. Khammash. 2010. Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.* 6, e1000696.
35. Zechner, C., J. Ruess, …, H. Koeppl. 2012. Moment-based inference predicts bimodality in transient gene expression. *Proc. Natl. Acad. Sci. USA.* 109:8340–8345.
36. 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.
37. 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.
38. 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.

39. Firman, T., G. Balázsi, and K. Ghosh. 2017. Building predictive models of genetic circuits using the principle of maximum caliber. *Biophys. J.* 113:2121–2130.

40. Lipinski-Kruszka, J., J. Stewart-Ornstein, …, H. El-Samad. 2015. Using dynamic noise propagation to infer causal regulatory relationships in biochemical networks. *ACS Synth. Biol.* 4:258–264.

41. 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.

42. Sgouralis, I., and S. Pressé. 2017. An introduction to infinite HMMs for single-molecule data analysis. *Biophys. J.* 112:2021–2029.

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

44. Dixit, P. D., J. Wagoner, …, K. A. Dill. 2018. Perspective: maximum caliber is a general variational principle for dynamical systems. *J. Chem. Phys.* 148, 010901.

45. Ghosh, K., P. D. Dixit, …, K. A. Dill. 2020. The maximum caliber variational principle for nonequilibria. *Annu. Rev. Phys. Chem.* 71:213–238.

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

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

48. Jong, H. 2004. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9:67.

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

50. Munsky, B., and M. Khammash. 2007. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *J. Comput. Phys.* 226:818–835.

51. Zhdanov, V. P. 2006. Transient stochastic bistable kinetics of gene transcription during the cellular growth. *Chem. Phys. Lett.* 424:394–398.

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

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

54. 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.

55. Firman, T., S. Wedekind, …, K. Ghosh. 2018. Maximum caliber can characterize genetic switches with multiple hidden species. *J. Phys. Chem. B.* 122:5666–5677. https://doi.org/10.1021/acs.jpcb.7b12251.

56. Firman, T., A. Amgalan, and K. Ghosh. 2019. Maximum Caliber can build and infer models of oscillation in three-gene feedback network. *J. Phys. Chem. B.* 123:343–355.

57. Firman, T., J. Huihui, …, K. Ghosh. 2021. Critical comparison of MaxCal and other stochastic modeling approaches in analysis of gene networks. *Entropy.* 23:357.

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

59. 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.

60. Becksei, A., B. Seraphin, and L. Serrano. 2001. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* 15:2528–2535.

61. 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.

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

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

64. 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.

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

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

67. 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.

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

69. Pressé, S., K. Ghosh, …, K. A. Dill. 2010. Dynamical fluctuations in biochemical reactions and cycles. *Phys. Rev. E.* 82, 031905.

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

71. Pressé, S., J. Peterson, …, K. Dill. 2014. Single molecule conformational memory extraction: P5ab RNA hairpin. *J. Phys. Chem. B.* 118:6597–6603.

72. McKay, D. 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge.

73. 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.

74. 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.

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

76. 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.

77. 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.

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

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

80. 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.

81. Shu, D., H. Zhang, …, P. Guo. 2007. Counting of six pRNAs of Phi29 DNA-packaging motor with customized single-molecule dual-view system. *EMBO J.* 26:527–537.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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.

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