entropy

MDPI

Article

Critical Comparison of MaxCal and Other Stochastic Modeling Approaches in Analysis of Gene Networks

Taylor Firman [1], Jonathan Huihui [2], Austin R. Clark [1] and Kingshuk Ghosh [1,2,*]

[1] Molecular and Cellular Biophysics, University of Denver, Denver, CO 80208, USA; taylor.firman@childrenscolorado.org (T.F.); Austin.Clark@du.edu (A.R.C.)
[2] Department of Physics and Astronomy, University of Denver, Denver, CO 80208, USA; Jonathan.Huihui@du.edu
* Correspondence: kghosh@du.edu

Abstract: Learning the underlying details of a gene network with feedback is critical in designing new synthetic circuits. Yet, quantitative characterization of these circuits remains limited. This is due to the fact that experiments can only measure partial information from which the details of the circuit must be inferred. One potentially useful avenue is to harness hidden information from single-cell stochastic gene expression time trajectories measured for long periods of time—recorded at frequent intervals—over multiple cells. This raises the feasibility vs. accuracy dilemma while deciding between different models of mining these stochastic trajectories. We demonstrate that inference based on the Maximum Caliber (MaxCal) principle is the method of choice by critically evaluating its computational efficiency and accuracy against two other typical modeling approaches: (i) a detailed model (DM) with explicit consideration of multiple molecules including protein-promoter interaction, and (ii) a coarse-grain model (CGM) using Hill type functions to model feedback. MaxCal provides a reasonably accurate model while being significantly more computationally efficient than DM and CGM. Furthermore, MaxCal requires minimal assumptions since it is a top-down approach and allows systematic model improvement by including constraints of higher order, in contrast to traditional bottom-up approaches that require more parameters or ad hoc assumptions. Thus, based on efficiency, accuracy, and ability to build minimal models, we propose MaxCal as a superior alternative to traditional approaches (DM, CGM) when inferring underlying details of gene circuits with feedback from limited data.

Keywords: gene network; inference; Maximum Caliber

Citation: Firman, T.; Huihui, J.; Clark, A.R.; Ghosh, K. Critical Comparison of MaxCal and Other Stochastic Modeling Approaches in Analysis of Gene Networks. *Entropy* **2021**, *23*, 357. https://doi.org/10.3390/e23030357

Academic Editors: Alexandre Ferreira Ramos and Gábor Balázsi

Received: 5 February 2021
Accepted: 10 March 2021
Published: 17 March 2021

1. Introduction

A common goal in synthetic biology of single cells is to design and introduce small circuits to optimize certain behaviors [1–10]. However, this requires quantitative knowledge about the details of these circuits and their subsequent optimization. Circuit characteristics are often studied by single-cell measurements using fluorescent markers on few proteins [11]. These traditional measurements—with the exception of a few emerging technologies that simultaneously monitor transcription and translation [12,13]—cannot visualize many other underlying species that are involved in dictating gene expression dynamics. For example, nucleic acids, protein-complexes, protein-nucleic acid complexes remain invisible. So, how do we extract information from such limited data? One possible solution is to track the stochastic gene expression trajectories of the fluorescently tagged proteins and use the noisy time series data to learn about the underlying invisible network [14–19]. This data mining scheme requires building stochastic models that are typically bottom-up and can be divided in two major categories. The first approach builds detailed models (DM) based on a pre-specified reaction network that include tagged proteins and other molecules such as promoter, protein-promoter complexes, etc. [20,21]. These detailed reaction schemes are usually simulated using Gillespie algorithm [22,23] to

generate noisy trajectories and analyze circuit properties. However, the same DMs can be used for inference as well. The detailed reactions—written as chemical master Equation (CME) [24] and solved using Finite State Projection method [25,26]—can yield probabilities of numbers of molecules as a function of rate parameters. Consequently, parameter inferences can be done by trying to match the measured noisy trajectory. The second approach builds coarse-grain models (CGM) using a smaller set of reactions where effective reaction rates are implemented by assuming an ad hoc mechanism that accounts for invisible species. CGMs can be made arbitrarily complex with additional fit parameters but we consider only the minimal CGMs due to their computational efficiency and an alternate to DM. Although CGMs are typically used to describe average protein numbers [27,28], CMEs can be constructed for CGMs [29–32] and solved via FSP to describe probability distributions. Both models (CGM and DM) have merits and demerits over each other. DM's are by definition more detailed and are expected to be accurate but require invoking multiple molecular species that are not always visible. Consequently, DM's involve multiple parameters and a much bigger phase space. Thus, accuracy comes at the expense of computational cost. Minimal CGM's on the other hand minimize computational burden by using effective rates based on pre-specified mechanisms. As a result, these models can be less accurate but efficient. The accuracy vs. efficiency dilemma raises two critical questions: how costly is a DM and how inaccurate is a CGM? Are CGM's accurate enough or do we need to build a third class of models that are still feasible and yet more accurate than CGM?

We have recently introduced the principle of Maximum Caliber (MaxCal) to model and infer underlying details of gene networks using stochastic gene expression data [32–36]. The adoption of MaxCal-based inference for small gene networks provides a third approach that resolves the accuracy vs. efficiency dilemma. Maximum Caliber (MaxCal) is analogous to the Maximum Entropy (MaxEnt) principle but applied to the distribution of paths/trajectories instead of states [37,38]. Similar to MaxEnt, MaxCal maximizes path entropy (caliber) subject to constraints. In the application to gene networks, MaxCal starts with a minimal set of constraints on proteins whose numbers are followed by fluorescent tags. This makes MaxCal directly amenable to model measured time series data of gene expression while not invoking additional mechanisms or auxiliary species [32]. Consequently, MaxCal avoids ad hoc assumptions on mechanisms inherent in CGM while bypassing the challenge of increased phase space inherent in DM. In addition, the top-down nature of MaxCal allows systematic model building via the imposition of constraints of a higher order on top of the minimal constraints it starts with—allowing the user to improve accuracy as needed. This is in contrast to bottom-up approaches of DM and CGM, which work on the premise that a given model must be pre-specified based on pre-imposed assumptions. Since each mechanism/model can be very different from each other, it is difficult to systematically add higher order terms in a given model to increase model complexity. Instead, a new model must be generated every time and the process of testing must start all over.

In this work we show how MaxCal provides a model that is minimal and more accurate than CGM while also being as efficient (or more) as CGM. We establish this with two basic genetic circuits: Single Gene Auto-activation (SGAA) [34] and Toggle Switch (TS) where two genes mutually repress each other [1,21,33]. We first generate synthetic data to be used as input by simulating the respective reaction networks with details of protein and nucleic acid dynamics. We subject three different models (DM, CGM and MaxCal) to infer parameters from this synthetic data for which actual model parameters are known. This serves as the benchmark for accuracy between the three models of DM, CGM, and MaxCal. By noting the computation time needed, we can assess accuracy and efficiency of all three models. The quantitive comparison shows that for small circuits used by synthetic biologists, MaxCal is efficient and reasonably accurate.

2. Materials and Methods

In this section, we describe the methods involved in generating synthetic data of stochastic gene expression using known parameters that serve as a benchmark to test the

accuracy of different models. We also describe different methods of analysis to infer model parameters from this stochastic data.

2.1. Single Gene Auto-Activation (SGAA) Circuit

We first demonstrate MaxCal's performance on a single gene auto activation circuit (SGAA). This circuit consists of a single protein that enhances its own production and has been studied extensively in different biological contexts [30,31,39–44].

2.1.1. Generating Synthetic Data

We first describe the model introduced by Elston et al. [20] to generate synthetic data mimicking experiment.

$$\begin{aligned} &\alpha \xrightarrow{g} \alpha + a \;;\; a \xrightarrow{d} \oslash \\ &a \xrightarrow{p} a + A \;;\; A \xrightarrow{r} \oslash \\ &A + A \underset{b_d}{\overset{f_d}{\rightleftarrows}} A_2 \\ &\alpha + A_2 \underset{b_p}{\overset{f_p}{\rightleftarrows}} \alpha^* \;;\; \alpha^* \xrightarrow{g^*} \alpha^* + a \end{aligned} \tag{1}$$

where a is mRNA transcribed from gene α with rate g, and gets degraded with rate d. Protein A is translated from a with basal rate p and degraded with rate r. Two A molecules complex to form a dimer A_2 with rate f_d, and dissociate back to monomers with rate b_d. The dimer A_2 can bind and unbind to the promoter with rate f_p and b_p, respectively. In the bound form, the promoter gets converted to the activated state α^* producing mRNA at an enhanced rate g^* much greater than g. Thus, A promotes self-activation. Synthetic input trajectories were generated using a Gillespie algorithm to simulate Equation (1) with $d = 0.2\,\mathrm{s}^{-1}$, $p = 0.02\,\mathrm{s}^{-1}$, $f_d = 5.0 \times 10^{-3}\,\mathrm{s}^{-1}$, $b_d = 50.0\,\mathrm{s}^{-1}$, $f_p = 6.0 \times 10^{-3}\,\mathrm{s}^{-1}$, $b_p = 3.0 \times 10^{-5}\,\mathrm{s}^{-1}$, $g = 0.05\,\mathrm{s}^{-1}$, $g^* = 0.5\,\mathrm{s}^{-1}$, $r = 1.0 \times 10^{-3}\,\mathrm{s}^{-1}$. These parameters and the reaction scheme (1) were chosen for three primary reasons: First, dimers are typical stoichiometry for regulatory proteins binding to the promoter site [20]. Next, the parameters produce noisy switch like trajectory and were used earlier to test MaxCal's inferential power [35]. Third, the rate parameter values and resulting switching times (few hours) are typical in realistic circuits [5]. Finally, to further mimic experiment, ten replicates of one hundred trajectories—each trajectory seven days long—were considered.

2.1.2. The Detailed Model for Inference

Next, we describe the three models (DM, CGM, and MaxCal) used to infer the underlying details of the circuit from the synthetic data generated. We start with the DM model first. The detailed model (DM) of SGAA is built by considering the promoter in its basal state (α), activated state (α^*), and the protein A. The reaction scheme is given by

$$\begin{aligned} &\alpha \xrightarrow{g_1} \alpha + A \;;\; A \xrightarrow{r_1} \oslash \\ &\alpha + A \underset{b_1}{\overset{f_1}{\rightleftarrows}} \alpha^* \;;\; \alpha^* \xrightarrow{g_1^*} \alpha^* + A \end{aligned} \tag{2}$$

The effective protein production rate in the non-activated and activated forms are g_1 and g_1^*, respectively. Upon defining the reaction network by Equation (2), the system can be described by a chemical master Equation (CME) to compute probability $P(N_A, t)$ of observing N_A number of A proteins at time t. The CME is given by

$$\begin{aligned} \tfrac{dP(N_A,\alpha,t)}{dt} &= (g_1\alpha + g_1^*(1-\alpha))P(N_A-1,\alpha,t) \\ &- (g_1\alpha + g_1^*(1-\alpha))P(N_A,\alpha,t) \\ &+ r_1(N_A+1)P(N_A+1,\alpha,t) - r_1 N_A P(N_A,\alpha,t) \\ &+ (N_A+1)f_1(1-\alpha)P(N_A+1,1-\alpha,t) - N_A f_1 \alpha P(N_A,\alpha,t) \\ &+ b_1\alpha P(N_A-1,1-\alpha,t) - b_1(1-\alpha)P(N_A,\alpha,t) \end{aligned} \tag{3}$$

where, $\alpha = 1$ and 0 denote the basal and the activated state (α^*) of the promoter, respectively. The CME can be expressed in the matrix form as

$$\frac{dP_i(t)}{dt} = \sum_{ij} W_{ij} P_j(t) \tag{4}$$

where, states (i) are given by different values of N_A and α and their time dependent probability is $P_i(t)$. The transition matrix W is a function of rate parameters $g_1, r_1, f_1, b_1, g_1^*$ and can be constructed from Equation (3). Using Finite State Projection (FSP) [25], the state space can be bounded and the time evolution of probability state vectors is given by

$$P_i(t_2) = [exp(W(t_2 - t_1))]_{ij} P_j(t_1) \tag{5}$$

where, t_2, t_1 are final and initial time points, respectively. Thus, Equation (5) yields the probability of being in some state i at time t_2 from an initial condition at time t_1. These conditional probabilities are used to compute the likelihood of producing the synthetic trajectory as a function of the rate parameters (see Equation (15)). The likelihood is then maximized to determine the optimum values of the parameters, thus inferring details of the model. It is important to note that DM ignores mRNA and dimer dynamics and hence is less detailed than the model (Equation (1)) created to generate the synthetic data. The choice of this reduced model is motivated to lower computational cost while constructing a model that is sufficiently fine-grained compared to CGM and MaxCal, outlined next.

2.1.3. CGM Model for Inference

The coarse-grain model (CGM) uses an approach where mass-action (MA) models are combined with CME [29–32]. Specifically, there is only one reaction describing the time evolution of protein number A given by Equation (6). This is akin to a mass-action approach where coupling with other species to capture feedback is modeled in an indirect manner by assuming an ad hoc functional form for the effective rate X. Specifically, X is given by a Hill type function invoking a cooperative model with $n = 2$ defined in Equation (6). The parameter g_2 is the basal rate in the absence of any positive feedback and the second term in X monotonically increases with A approaching an asymptotic value of g_2^*, successfully capturing auto-activation. In this model, the rate in the activated state is $g_2 + g_2^*$ and K is a parameter. Protein A is degraded with rate r_2.

$$\begin{aligned} &\alpha \xrightarrow{X} \alpha + A \;;\; A \xrightarrow{r_2} \oslash \\ &X = g_2 + g_2^* \tfrac{A^n}{A^n + K} \end{aligned} \tag{6}$$

While traditional MA models use similar functional forms to describe the time evolution of average protein number, adopting Equation (6) within Chemical Master Equation (CME) framework allows stochastic modeling. Thus, the combined approach of CME and MA (CME + MA) computes the probability ($P(N_A; t)$) of protein number (N_A) at a given time t. Specifically, the time evolution equation of $P(N_A; t)$ corresponding to the reaction scheme in Equation (6) is given by

$$\begin{aligned} \tfrac{dP(N_A,t)}{dt} &= W_1^{cgm}(N_A - 1)P(N_A - 1, t) + W_2^{cgm}(N_A + 1)P(N_A + 1, t) \\ &- [W_1^{cgm}(N_A) + W_2^{cgm}(N_A)]P(N_A, t) \end{aligned} \tag{7}$$

where $W_1^{cgm}(N) = g_2 + g_2^* \frac{N^n}{N^n + K}$ and $W_2^{cgm}(N) = r_2 N$. Using the matrix formulation, similar to DM, the likelihood of input trajectories can be calculated as a function of g_2, g_2^*, K, r_2. Maximization of the likelihood infers model parameters g_2, g_2^*, K, and r_2.

2.1.4. MaxCal Model for Inference

MaxCal modeling of SGAA and other circuits have been described in our earlier work [32,34,35]. Here we give a brief outline of the method for SGAA. Maximum Caliber maximizes the path entropy subject to constraints. We define two random variables, ℓ_α and ℓ_A, to define microscopic trajectories between a time interval t and $t + \Delta t$. The probabilities of these paths are denoted as P_{ℓ_α,ℓ_A}. Variable ℓ_α tracks the production of proteins—in the time interval between t and $t + \Delta t$—ranging between 0 and M (predefined maximum) while ℓ_A is the number of proteins from the previous step that have not undergone degradation, i.e., $0 < \ell_A < N_A$. The first term in C (in Equation (8)) is the path entropy and the remaining three terms denote three constraints on average production, average degradation, and a correlation between protein production and the proteins present. The constraints are imposed by Lagrange multipliers h_α, h_A and K_A, respectively.

$$\begin{aligned} C = -\sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} P_{\ell_\alpha,\ell_A}\log P_{\ell_\alpha,\ell_A} + h_\alpha \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_\alpha P_{\ell_\alpha,\ell_A} + \\ h_A \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_A P_{\ell_\alpha,\ell_A} + K_A \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A} \ell_\alpha \ell_A P_{\ell_\alpha,\ell_A} \end{aligned} \tag{8}$$

Maximizing the caliber subject to the three constraints yield the probability of these micro trajectories as

$$\begin{aligned} P_{\ell_\alpha,\ell_A} &= Q^{-1}\binom{N_A}{\ell_A}\exp(h_\alpha\ell_\alpha + h_A\ell_A + K_A\ell_\alpha\ell_A) \\ Q &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A}\binom{N_A}{\ell_A}\exp(h_\alpha\ell_\alpha + h_A\ell_A + K_A\ell_\alpha\ell_A) \end{aligned} \tag{9}$$

2.2. Two-Gene Toggle Switch (TS) Circuit

Next we consider a two gene mutually repressing circuit called the Toggle Switch (TS), first designed by Collins and colleagues [1].

2.2.1. Generating Synthetic Data

Synthetic data mimicking experimental time trace was created by using the reaction scheme described in reference [21], detailed below:

$$\begin{aligned} &\alpha \xrightarrow{g_3} \alpha + a \;;\; a \xrightarrow{d_3} \oslash \\ &a \xrightarrow{p_3} a + A \;;\; A \xrightarrow{r_3} \oslash \\ &\alpha + A \underset{b_3}{\overset{f_3}{\rightleftarrows}} \alpha^* \;;\; \alpha^* \xrightarrow{g_3} \alpha^* + a \;;\; \alpha^* \xrightarrow{g_3^*} \alpha^* + b \\ &\alpha \xrightarrow{g_3} \alpha + b \;;\; b \xrightarrow{d_3} \oslash \\ &b \xrightarrow{p_3} b + B \;;\; B \xrightarrow{r_3} \oslash \\ &\alpha + B \underset{b_3}{\overset{f_3}{\rightleftarrows}} \alpha' \;;\; \alpha' \xrightarrow{g_3^*} \alpha' + a \;;\; \alpha' \xrightarrow{g_3} \alpha' + b \end{aligned} \tag{10}$$

where mRNA a and b are transcribed from gene α (different loci) at rate g_3, that in turn produce proteins A and B, respectively, at rate p_3. mRNAs and proteins are degraded at rate d_3 and r_3, respectively. Mutual repression is modeled by binding of either protein (A or B) to the promoter site of α at rate f_3, altering the promoter state to different expression levels α^* (for A) and α' (for B). In the state α^*, mRNAs of B are produced at a rate g_3^* much less than g_3, while state α' produces mRNAs of A at that same slower rate g_3^*, implementing negative feedback. Unbinding of proteins convert the inactivated promoter states (α^*, α') back to the basal state α. For simplicity and as proof of concept, we assume a symmetric model in A and B. The rate values ($d_3 = 0.5\ \mathrm{s}^{-1}$, $p_3 = 0.02\ \mathrm{s}^{-1}$, $f_3 = 3.5 \times 10^{-6}\ \mathrm{s}^{-1}$, $b_3 = 2.0 \times 10^{-5}\ \mathrm{s}^{-1}$, $g_3 = 0.5\ \mathrm{s}^{-1}$, $g_3^* = 2.5 \times 10^{-3}\ \mathrm{s}^{-1}$, $r_3 = 1.0 \times 10^{-3}\ \mathrm{s}^{-1}$) were chosen following our earlier

work [35]. These values create biologically relevant switching time scales [5] between two states (state 1: low A, high B; state 2: high A, low B) while maintaining typical synthesis and degradation rates [27]. Next, the Gillespie algorithm [22] is used to simulate ten replicates of one hundred noisy trajectories (each seven days long) of protein levels that exhibit bistability and serve as our synthetic data to benchmark models.

2.2.2. Detailed Model

To infer network details, we first consider the detailed model (DM) built on a reaction network similar to Equation (10) but neglecting mRNAs. The exact reaction scheme used in DM for TS is described as

$$\begin{aligned}
&\alpha \xrightarrow{g_4} \alpha + A \;;\; A \xrightarrow{r_4} \oslash \\
&\alpha + A \underset{b_4}{\overset{f_4}{\rightleftarrows}} \alpha^* \;;\; \alpha^* \xrightarrow{g_4} \alpha^* + A \;;\; \alpha^* \xrightarrow{g_4^*} \alpha^* + B \\
&\alpha \xrightarrow{g_4} \alpha + B \;;\; B \xrightarrow{r_4} \oslash \\
&\alpha + B \underset{b_4}{\overset{f_4}{\rightleftarrows}} \alpha' \;;\; \alpha' \xrightarrow{g_4^*} \alpha' + A \;;\; \alpha' \xrightarrow{g_4} \alpha' + B
\end{aligned} \tag{11}$$

where the symbols above and below arrows indicate respective reaction rates that will be inferred from the synthetic trajectory data generated from Equation (10). As before, CME arising from this reaction network architecture can be used to compute the probability $P(N_A(t+\Delta t), N_B(t+\Delta t); N_A(t), N_B(t))$ of observing $N_A(t+\Delta t)$ and $N_B(t+\Delta t)$ number of protein molecules at time $t+\Delta t$ given there were $N_A(t)$ and $N_B(t)$ number of A and B protein molecules at time t. These probabilities are used to compute the likelihood of a given trajectory (see Equation (17)), which can then be maximized to infer rates for this system.

2.2.3. Coarse Grain Model

The coarse grain model (CGM) for TS does not explicitly model protein-promoter complexation unlike DM. Feedback is modeled by an effective production rate of A and B, given by X and Y, respectively, defined as

$$\begin{aligned}
&\alpha \xrightarrow{X} \alpha + A \;;\; A \xrightarrow{r_5} \oslash \;;\; \beta \xrightarrow{Y} \beta + B \;;\; B \xrightarrow{r_5} \oslash \\
&X = g_5 + g_5^* \tfrac{1}{B^n + K} \;;\; Y = g_5 + g_5^* \tfrac{1}{A^n + K}
\end{aligned} \tag{12}$$

The specific functional form for X ensures that large B inhibits production of A. Similarly, Y ensures that A inhibits production of B. Proteins A and B degrade with rate r_5. As before, the reaction network is used to construct the corresponding CME from which the likelihood is computed as a function of model parameters g_5^*, g_5, K, and r_5. The cooperativity parameter n was assumed to be 2, typically assumed in MA models of TS to produce bistability [27]. The inferred values of these parameters are obtained by maximizing the likelihood.

2.2.4. MaxCal

Two variables, ℓ_α and ℓ_A, were introduced to model SGAA (in Section 2.1.4) to track production of A and number of A proteins not degraded in a time interval t and $t+\Delta t$. The same variables were used for protein A in TS. Additionally, variables ℓ_β and ℓ_B were introduced to track the same but for protein B. The micro trajectories between time t and $t+\Delta t$ are labeled by stochastic variables ℓ_α, ℓ_β, ℓ_A, and ℓ_B, and the corresponding probability is denoted as $P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}$. The caliber ($C$) is defined as the path entropy (given by the first term on the right hand side of Equation (13)) along with four constraints. First, the average of ℓ_α and ℓ_β are imposed as constraints with Lagrange multiplier h_α to model protein production. Next, the average of ℓ_A and ℓ_B are constrained with Lagrange multiplier h_A capturing information about degradation. Third, correlations between production

and number of existing proteins (of the same protein type) are imposed by constraining average of $\ell_\alpha \ell_A$ and $\ell_\beta \ell_B$ by Lagrange multiplier $K_{A\alpha}$ (fourth term in the right hand side of Equation (13)). This term captures any positive feedback that may not be known about a priori (see reference [35] for more). Finally, the average of $\ell_\beta \ell_A$ and $\ell_\alpha \ell_B$ are constrained by Lagrange multiplier $K_{A\beta}$ to model cross-correlation (negative feedback) between A and B.

$$\begin{aligned} C = -\sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A}\sum_{\ell_\beta=0}^{M}\sum_{\ell_B=0}^{N_B}\Big[&P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}\log P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} \\ &+ h_\alpha(\ell_\alpha+\ell_\beta)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + h_A(\ell_A+\ell_B)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} \\ &+ K_{A\alpha}(\ell_\alpha\ell_A+\ell_\beta\ell_B)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + K_{A\beta}(\ell_\beta\ell_A+\ell_\alpha\ell_B)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}\Big], \end{aligned} \tag{13}$$

Maximizing the caliber subject to these constraints gives the path probability in terms of the Lagrange multipliers as

$$\begin{aligned} P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} &= Q^{-1}\tbinom{N_A}{\ell_A}\tbinom{N_B}{\ell_B}\exp[h_\alpha(\ell_\alpha+\ell_\beta)+h_A(\ell_A+\ell_B)+ \\ &\qquad K_{A\alpha}(\ell_\alpha\ell_A+\ell_\beta\ell_B)+K_{A\beta}(\ell_\beta\ell_A+\ell_\alpha\ell_B)]; \\ Q &= \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{N_A}\sum_{\ell_\beta=0}^{M}\sum_{\ell_B=0}^{N_B}\tbinom{N_A}{\ell_A}\tbinom{N_B}{\ell_B}\exp[h_\alpha(\ell_\alpha+\ell_\beta)+h_A(\ell_A+\ell_B)+ \\ &\qquad K_{A\alpha}(\ell_\alpha\ell_A+\ell_\beta\ell_B)+K_{A\beta}(\ell_\beta\ell_A+\ell_\alpha\ell_B)]. \end{aligned} \tag{14}$$

2.3. Calculation of Trajectory Likelihood

Parameters of DM, CGM, and MaxCal models are determined by maximizing the likelihood ($\mathscr{L}$) of the synthetic trajectory—mimicking experimental data—recording fluctuating numbers of protein with time.

2.3.1. Calculation of Trajectory Likelihood for SGAA

For SGAA, the likelihood of the trajectory is calculated as

$$\mathscr{L} = \prod_{n=1}^{\mathscr{N}} P(N_A(t+m); N_A(t=m(n-1))) = \prod_{\{i\to j\}} P_{(i\to j),m}^{\omega_{(i\to j),m}}\ , \tag{15}$$

where T is the total snapshots recorded in experiment or synthetic data, $\mathscr{N}$ is T/m rounded to the nearest integer and $\omega_{(i\to j),m}$ is the total number of transition from state i to j over m frames. States are given by number of proteins. Typically m is chosen in the same range as the average residence time (in frames) in the high and low state. Multiple frames are combined to avoid any spurious jumps in the number of proteins present in the trajectory, not allowed in the MaxCal formulation. Thus, calculating likelihood for transitions over m frames allows MaxCal to select reasonable values for the Lagrange multipliers (see reference [34] for further details). The individual transition probability of going from state i to state j in one time step in MaxCal is computed as

$$P_{i\to j} = \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i}\delta(\ell_\alpha+\ell_A-j)P_{\ell_\alpha,\ell_A}. \tag{16}$$

The transition probability $P_{(i\to j),m}$ over m frames—needed for likelihood calculation—can be obtained by arranging single time step transition in a matrix (with i, j as matrix elements) and raising the matrix to the m^{th} power. FSP formalism—originally devised by Munsky and colleagues [25]—has been used to introduce a sink state and carry out this calculation. Supplemental material in reference [36] provides a general description on how to use FSP formalism for MaxCal, including the calculation of transition probabilities for m

steps. For DM and CGM models, the calculation of $P_{(i\to j),m}$ require matrix exponentiation following Equation (5) with $t_2 - t_1 = m\Delta t$.

2.3.2. Calculation of Trajectory Likelihood for TS

For TS, the likelihood of the trajectory is calculated as

$$\begin{aligned}\mathscr{L} = \prod_{n=1}^{\mathscr{N}} P(N_A(t+m), N_B(t+m); N_A(t=m(n-1)), N_B(t=m(n-1))) \\ = \prod_{\{i\to j, k\to l\}} P_{(i\to j),(k\to l),m}^{\omega_{(i\to j),(k\to l),m}},\end{aligned} \tag{17}$$

where T is total time frames noted in data, $\mathscr{N}$ is T/m rounded to the nearest integer, $\omega_{(i\to j),(k\to l),m}$ is the total number of simultaneous transitions from i to j states (in number of protein A) and k to l (in number of protein B) over m frames, and $N_A(t)$, $N_B(t)$ denote the number of A, B proteins at time t (corresponding to frame $m(n-1)$ where n is an integer). The most likely values of h_α, h_A, $K_{A\alpha}$, $K_{A\beta}$, and M are determined by maximizing the likelihood. As before, transitions between m frames were used to avoid large spurious jumps in protein number in one step in the raw data that cannot be easily modeled in MaxCal. The transition probabilities $P(j,l,t+1;i,k,t)$ between two consecutive frames in MaxCal is defined as,

$$P(j,l,t+1;i,k,t) = P_{i\to j,k\to l} = \sum_{\ell_\alpha=0}^{M}\sum_{\ell_A=0}^{i}\sum_{\ell_\beta=0}^{M}\sum_{\ell_B=0}^{k}\delta(\ell_\alpha+\ell_A-j)\delta(\ell_\beta+\ell_B-l)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}. \tag{18}$$

with $P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}$ defined in Equation (14). The transition probability $P_{(i\to j),(k\to l),m}$ between m frames is then calculated by raising the two frame transition probability matrix to the m^{th} power, similar to the procedure described for SGAA. Similar to SGAA, calculation of the transition probability $P_{(i\to j),(k\to l),m}$ within DM and CGM formalism require matrix exponentiation in contrast to matrix multiplication needed for MaxCal.

3. Results

3.1. Comparison of Three Models for SGAA

Quantitative comparison of three SGAA models was carried out in terms of their accuracy and efficiency. Accuracy is determined by comparing inferred values of three observables (p_{eff}, p^*_{eff}, r_{eff}) against the gold standard used to generate the synthetic data (see Table 1). The true basal production rate p_{eff} and activated production rate p^*_{eff} are obtained from the parameters in reaction 1 by $p_{\text{eff}} = p(g/d)$, $p^*_{\text{eff}} = p(g^*/d)$. The degradation rate remains the same, $r_{\text{eff}} = r$. For DM, these rates can be extracted as $p_{\text{eff}} = g_1$, $p^*_{\text{eff}} = g^*_1$, $r_{\text{eff}} = r_1$, where g_1, g^*_1, and r_1 are defined in reaction 2. The effective rates for CGM (described by Equation (6)) are calculated as $p_{\text{eff}} = g_2$, $p^*_{\text{eff}} = g_2 + g^*_2$, and $r_{\text{eff}} = r_2$. The effective rates using MaxCal are calculated using

$$\begin{aligned} p_{\text{eff}} &\approx \tfrac{\langle \ell_\alpha\rangle_{N_L}}{\Delta t}, & p^*_{\text{eff}} &\approx \tfrac{\langle \ell_\alpha\rangle_{N_H}}{\Delta t}, \\ r(N) &= \frac{N - \langle \ell_A\rangle_N}{N\Delta t}, & r_{\text{eff}} &\approx \sum_N P_{eq}(N) r(N), \end{aligned} \tag{19}$$

where $\langle\cdots\rangle_i$ denotes the average of an observable for a given $N_A = i$, N_L is the peak position of the protein number distribution in the low state, N_H is the peak in the high state, and $P_{eq}(N)$ is the probability distribution of N proteins at relative equilibrium calculated using FSP [25] (see reference [35] for details). Comparing reported values in Table 1, we conclude that both DM and MaxCal models infer these underlying details reasonably well. CGM however infers a value of p^*_{eff} almost twice the true value.

Table 1. Comparison of accuracy between three models for SGAA. The first row reports the values of three known ("True") rates for effective production (p_{eff}) in the basal state, production (p^*_{eff}) in the activated state, and protein degradation (r_{eff}) used to generate the synthetic data. The inferred rates using three models, DM (second row), CGM (third row), and MaxCal (fourth row), are compared against each other and the "True" rates indicating that CGM is less accurate than DM and MaxCal. Error bars for rates were obtained by using inference on ten replicates of the input trajectory data.

Method	p_{eff} (s^{-1})	p^*_{eff} (s^{-1})	r_{eff} (s^{-1})
True	5.0×10^{-3}	50×10^{-3}	1.0×10^{-3}
DM	$(4.2 \pm 0.5) \times 10^{-3}$	$(42 \pm 5) \times 10^{-3}$	$(0.8 \pm 0.1) \times 10^{-3}$
CGM	$(4.1 \pm 0.4) \times 10^{-3}$	$(91 \pm 8.9) \times 10^{-3}$	$(1.3 \pm 0.1) \times 10^{-3}$
MaxCal	$(5.6 \pm 0.2) \times 10^{-3}$	$(43 \pm 2.1) \times 10^{-3}$	$(0.96 \pm 0.1) \times 10^{-3}$

Next, we provide a comparison of the computational efficiency between the three models by tracking the typical time needed to complete the basic unit of operation invoked in the calculation of the likelihood function. For a given model, we measured the time taken for each likelihood calculation during the entire process of inference. The averages and standard deviations of these times are noted in column 4 in Table 2. For MaxCal, the basic operation is raising the transition matrix to the m^{th} power, in contrast to the matrix exponentiation required for DM and CGM (see Equation (5)). We note that both CGM and MaxCal have identical matrix dimensions (column 3 in Table 2), constrained by the maximum number of proteins allowed ($N_{max} = 92$ reported in column 2 of Table 2). The maximum number of proteins used for FSP calculation was chosen to be significantly higher than the maximum protein number seen in the input trajectory. The total state space dimension is equal to $N_{max} + 1$ with the additional state (or "sink" state) accounting for all states with protein numbers greater than N_{max}. Despite identical matrix size, a typical step in CGM is slower than MaxCal because matrix exponentiation is slower than matrix multiplication performed m times in succession. We notice basic step calculation in DM is significantly slower than MaxCal. This is primarily due to two reasons. First, DM has almost four times larger of a matrix size—compared to MaxCal—due to explicit consideration of the promoter state (basal and activated) in combination with different protein numbers defining the state space. Next, DM requires matrix exponentiation, a much more computationally expensive operation compared to matrix multiplication. We conclude that CGM is less accurate than MaxCal and MaxCal is the method of choice over DM given its efficiency. However, for SGAA, the true power of MaxCal is not reflected since the basic step/operation is in milliseconds for DM and the cumulative time needed for inference (fifth column) is in seconds, feasible on traditional hardware. Nevertheless, the comparison conceptually shows MaxCal's ability to balance efficiency and accuracy for SGAA, in conjunction with the bigger circuit that we discuss next.

Table 2. Comparison of efficiency between three models for SGAA. Second column reports the maximum number of proteins used in FSP, third column shows the overall matrix dimension, fourth column reports the average time taken (using a CPU platform) for the basic matrix operation needed for a likelihood calculation, and fifth column reports the total time taken during the entire process of likelihood maximization to infer model parameters for the three different models (noted in the first column).

Method	Max N	Matrix Size	Unit Operation Time (ms)	Total Time (ms)
DM	92	185×185	10 ± 0.4	1263 ± 69
CGM	92	93×93	3.0 ± 0.2	813 ± 190
MaxCal	92	93×93	0.6 ± 0.5	47 ± 13

3.2. Comparison of Three Models in TS

Next, we provide a quantitative comparison of three models for TS. As before, accuracy is determined by comparing inferred values of the three effective rates, p_{eff}, p^*_{eff}, and r_{eff} against the gold standard used to generate the synthetic data (see Table 3). The true basal production rate p_{eff}, repressed production rate p^*_{eff} are obtained from the parameters in reaction (10) defined as $p_{\text{eff}} = p_3(g_3/d_3)$ and $p^*_{\text{eff}} = p_3(g^*_3/d_3)$. The degradation rate remains the same, $r_{\text{eff}} = r_3$. Within the framework of DM, these rates are extracted as $p_{\text{eff}} = g_4$, $p^*_{\text{eff}} = g^*_4$, and $r_{\text{eff}} = r_4$, where g_4, g^*_4, and r_4 are defined in reaction (11). The effective rates for CGM (described by Equation (12)) are calculated as $p_{\text{eff}} = g_5 + g^*_5/K$, $p^*_{\text{eff}} = g_5$, and $r_{\text{eff}} = r_5$. The effective rates from MaxCal are calculated as

$$\begin{aligned} p_{\text{eff}} &\approx \frac{\langle \ell_\alpha \rangle_{N_H,N_L}}{\Delta t} = \frac{\langle \ell_\beta \rangle_{N_L,N_H}}{\Delta t}, \\ p^*_{\text{eff}} &\approx \frac{\langle \ell_\alpha \rangle_{N_L,N_H}}{\Delta t} = \frac{\langle \ell_\beta \rangle_{N_H,N_L}}{\Delta t}, \\ r_A(N_A,N_B) &= \frac{N_A - \langle \ell_A \rangle_{N_A,N_B}}{N_A \Delta t}, \qquad r_B(N_A,N_B) = \frac{N_B - \langle \ell_B \rangle_{N_A,N_B}}{N_B \Delta t}, \\ r_{\text{eff}} &\approx \sum_{N_A=0}^{\infty}\sum_{N_B=0}^{\infty} P_{eq}(N_A,N_B) r_A(N_A,N_B) = \sum_{N_A=0}^{\infty}\sum_{N_B=0}^{\infty} P_{eq}(N_A,N_B) r_B(N_A,N_B), \end{aligned} \tag{20}$$

where $\langle \cdots \rangle_{i,j}$ is the average of a quantity of interest given that there are i and j number of proteins initially present of type A and B, respectively, N_H and N_L are the peaks of the protein number distribution in the basal (unrepressed) and repressed state, respectively, and $P_{eq}(i,j)$ is the relative equilibrium probability (calculated using FSP [25]) of having $N_A = i$ and $N_B = j$ proteins. The details of these definitions and results can be found in our earlier work [35]. Comparing the extracted rates against the gold standard, we find DM and MaxCal reliably infer effective rates, similar to their performance with SGAA. CGM however infers rates that differ from the "True" rates by almost an order of magnitude or even more (p^*_{eff} for example).

Table 3. Comparison of accuracy between three models for TS. The first row reports the values of three known ("True") rates for effective production (p_{eff}) in the basal state, production (p^*_{eff}) in the repressed state, and protein degradation (r_{eff}) used to generate the synthetic data. The inferred rates using the three models, DM (second row), CGM (third row), and MaxCal (fourth row), are compared against each other and the "True" rates indicate CGM is less accurate than DM and MaxCal. Error bars for rates were obtained by using inference on ten replicates of the input trajectory data.

Method	p_{eff} (s^{-1})	p^*_{eff} (s^{-1})	r_{eff} (s^{-1})
True	20×10^{-3}	0.1×10^{-3}	1.0×10^{-3}
DM	$20.7 \times 10^{-3} \pm 5.2 \times 10^{-5}$	$0.1 \times 10^{-3} \pm 2.6 \times 10^{-5}$	$1.0 \times 10^{-3} \pm 0.2 \times 10^{-3}$
CGM	$170 \times 10^{-3} \pm 30 \times 10^{-3}$	$2.3 \times 10^{-7} \pm 0.7 \times 10^{-7}$	$7.0 \times 10^{-3} \pm 2.0 \times 10^{-3}$
MaxCal	$14.6 \times 10^{-3} \pm 0.9 \times 10^{-3}$	$0.16 \times 10^{-3} \pm 0.02 \times 10^{-3}$	$0.7 \times 10^{-3} \pm 3 \times 10^{-5}$

Next, we provide a comprehensive comparison of the three models in terms of their computational efficiency (see Table 4). Similar to SGAA, DM is significantly slower than MaxCal and CGM primarily due to large matrix size. We have chosen $N_{A,max} = N_{B,max} = 59$ resulting in $59 \times 59 \times 2 \times 2 + 1 = 13{,}925$ states, considering the two states of the promoter and the sink state included in the state space of DM. The enormous dimensional explosion in DM compounded with matrix exponentiation significantly slows down the basic operation to minutes. The estimates of basic steps are obtained using a GPU platform for feasibility reasons. CPU-based computing—used in SGAA–was abandoned as it would have taken an unreasonably long time to complete DM inference. We modified the existing SciPy libraries for calculating the matrix exponential term (needed for basic operation in DM and CGM) using CuPy. To calculate matrix power needed for MaxCal, we used

existing CuPy libraries. The timing of the basic step was evaluated on two Nvidia Tesla P100 cards using CUDA version 8.0.

CGM and MaxCal are however much faster than DM. This is primarily because of significantly smaller variable space that CGM and MaxCal uses. Furthermore, we notice MaxCal is even faster compared to CGM due to MaxCal's reliance on matrix multiplication in contrast to matrix exponentiation. The differences in basic operation (column four in Table 4) translate to marked differences in the total time taken for the entire process of likelihood maximization (reported in column five in Table 4) to infer model parameters. Combining these findings we conclude for TS—similar to SGAA—MaxCal is the preferred method of inference due to its efficiency and ability to reliably infer underlying model parameters. For TS, unlike SGAA, the gain from MaxCal is readily appreciated when using standard GPU hardware with only few nodes.

The example of TS considered above illustrates the typical challenge of inferring network details using detailed models (DM) where multiple genes and species are involved. Genetic circuits larger than TS, i.e., involving more than two expressed proteins, will face even more combinatorial challenges when using models like DM to account for proteins and their promoters. This combinatorial explosion of the state space combined with the need of matrix exponentiation will render detailed models unrealistic for the purposes of inference. Although this challenge can be somewhat mitigated in CGM reducing the state space, CGMs too will face computational challenge, due to their reliance on matrix exponential that will tend to be slower for larger state space (due to multiple genes). Alternate approaches similar to MaxCal should be used where basic operations in likelihood calculation are less burdensome keeping overall computational cost manageable.

Table 4. Comparison of efficiency between three models for TS. Second column reports the maximum number of proteins used in FSP, third column shows the overall matrix dimension, and fourth column reports the average time taken (using GPU platform) for the basic matrix operation needed for a likelihood calculation and fifth column reports the average of total time taken during the entire process of likelihood maximization to infer model parameters for the three different models (noted in the first column).

Method	Max N	Matrix Size	Unit Operation Time (s)	Total Time (s)
DM	59	$13{,}925 \times 13{,}925$	223 ± 110	$106{,}878 \pm 19{,}765$
CGM	59	3482×3482	5.8 ± 0.4	$14{,}124 \pm 12{,}493$
MaxCal	59	3482×3482	0.13 ± 0.01	851 ± 60

4. Discussion

Quantitative determination of parameters in a gene network is critical to designing new circuits for synthetic biology applications. However, determining these parameters is challenging due to limited information available on few expressed proteins, much less than the actual number of species involved in such feedback networks. A powerful approach is to mine information-rich stochastic trajectories of protein numbers to infer network details. Three primary modeling schemes were considered to harness information from these noisy trajectories for two specific gene networks: Single Gene Auto Activation (SGAA) circuit and Toggle Switch (TS). The first inferential approach used a detailed modeling (DM) scheme where proteins and their interaction with corresponding genes were explicitly modeled. The second approach employed a coarse grain model (CGM) where Hill type functional forms were invoked to describe feedback, circumventing the need to explicitly model promoter and protein-promoter complexes. The third scheme used the principle of Maximum Caliber (MaxCal) which relies on the maximization of path entropy subject to constraints of protein production, degradation, and feedback. MaxCal—similar to CGM—also avoids explicit consideration of additional molecular species and provides a stochastic description of protein expression trajectories, suitable to analyze noisy gene expression data

typically measured in experiments. Using synthetic data generated from a known model, we show that DM accurately infers network details. However, DM is computationally challenging for networks with multiple proteins, even as few as two. The prohibitive computational cost of DM is due to its large state space, resulting in the exponentiation of high dimensional matrices. CGM operates on a lower dimensional state space and hence is more efficient than DM, but still suffers from computational cost for larger circuits due to matrix exponentiation. Moreover, minimal CGMs are less accurate in inferring underlying model details. MaxCal offers the much needed framework both in terms of accuracy and feasibility. In addition, MaxCal provides a systematic way to improve and select models with the same variables but including different correlations as additional constraints. However, adding more constraints will increase the number of parameters (Lagrange multipliers) over which the likelihood function needs to be maximized, slowing down the overall inference process. Nevertheless, MaxCal enjoys the computational efficiency of the basic step—to be iterated multiple times during likelihood maximization—due to relatively smaller matrix size and matrix multiplication operation.

The ability to systematically improve models by adding higher order correlation—due to MaxCal's inherent top-down nature [32,37]—is another advantage of MaxCal. In contrast, traditional approaches are bottom-up and require first imposing a reaction diagram (such as in DM) or a mechanism (functional forms in CGM). Even with small changes, different reaction networks or functional forms need to be incorporated into the model, restricting systematic model building in a perturbative manner. Furthermore, MaxCal's reliance on production and degradation variables separately allows the calculation of an effective feedback parameter useful for network characterization, design, and evolution [35,36]. These advantages strongly favor adoption of MaxCal to infer parameters in gene networks from noisy time series protein expression data. However, MaxCal's performance depends on the availability of basic information. For example, MaxCal model built on only two genes cannot produce oscillation seen in three gene repressilator circuit. MaxCal requires minimal knowledge about the existence of all three genes to correctly model repressilator [36]. Another common challenge with inference is experiments typically measure fluorescence and not protein numbers. Furthermore, fluorescence per protein is not fixed, but random. Direct application of MaxCal on raw experimental data requires experimentally determining the distribution of fluorescence per protein. In the absence of such information, we created synthetic data mimicking noisy fluorescence trajectories and demonstrated MaxCal's ability to decouple fluorescence noise from gene expression noise and infer underlying circuit details [34–36].

It is important to note, MaxCal—in spite of its relative efficiency over DM and CGM—will face increased computational challenge with circuits having multiple different genes due to increased dimensionality of matrices. With the emergence of new mass-spectrometry tools measuring multiple proteins, we are likely to face such data deluge requiring sophisticated tools of inference. Building efficient and reliable models like MaxCal and adoption of GPU platform for additional computational acceleration, as done here, are necessary steps to this direction.

Author Contributions: Conceptualization, T.F. and K.G.; methodology, T.F., J.H. and A.R.C.; software, T.F., J.H. and A.R.C.; validation, T.F. and J.H.; formal analysis, T.F., J.H. and K.G.; investigation, T.F., J.H. and K.G.; writing T.F., J.H. and K.G.; supervision, K.G.; project administration, K.G.; funding acquisition, K.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Institutes of Health grant number R15GM128162-01A1.

Data Availability Statement: The codes for SGAA and TS are available in Github with links: https://github.com/MaxCalLab/SelfPromo and https://github.com/MaxCalLab/ToggleSwitch, respectively (accessed on 16 March 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MaxCal	Maximum Caliber
DM	Detailed Model
CGM	Coarse Grain Model
SGAA	Single Gene Auto Activation
TS	Toggle Switch
CME	Chemical Master Equation
FSP	Finite State Projection

References

1. Gardner, T.; Cantor, T.C.; Collins, J. Construction of a Genetic Toggle Switch in Escherichia coli. *Nature* **2000**, *403*, 339–342. [CrossRef] [PubMed]
2. Elowitz, M.; Leibler, S. A Synthetic Oscillatory Network of Transcriptional Regulators. *Nature* **2000**, *403*, 335–338. [CrossRef]
3. Alon, U. Network Motifs: Theory and Experimental Approaches. *Nat. Rev. Genet.* **2007**, *8*, 450–461. [CrossRef] [PubMed]
4. Tsai, T.; Choi, Y.; Ma, W.; Pomerening, J.; Tang, C.; Ferrell, J. Robust, Tunable Biological Oscillations from Interlinked Positive and Negative Feedback Loops. *Science* **2008**, *321*, 126–129. [CrossRef]
5. Nevozhay, D.; Adams, R.; Itallie, E.V.; Bennett, M.; Balázsi, G. Mapping the Environmental Fitness Landscape of a Synthetic Gene Circuit. *PLOS Comput. Biol.* **2012**, *8*, e1002480. [CrossRef] [PubMed]
6. Nevozhay, D.; Adams, R.; Murphy, K.; Josic, K.; Balazsi, G. Negative auto regulation linearizes the dose–response and suppresses the heterogeneity of gene expression. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 5123–5128. [CrossRef] [PubMed]
7. Lyons, S.; Xu, W.; Medford, J.; Prasad, A. Loads Bias Genetic and Signaling Switches in Synthetic and Natural Systems. *PLoS Comput. Biol.* **2014**, *10*, e1003533. [CrossRef] [PubMed]
8. Wang, L.; Wu, F.; Flores, K.; Lai, Y.; Wang, X. Build to Understand: Synthetic Approaches to Biology. *Integr. Biol.* **2016**, *8*, 394–408. [CrossRef]
9. Mukherji, S.; van Oudenaarden, A. Synthetic Biology: Understanding Biological Design from Synthetic Circuits. *Nat. Rev. Genet.* **2009**, *10*, 859–871. [CrossRef]
10. Wu, F.; Wang, X. Applications of Synthetic Gene Networks. *Sci. Prog.* **2015**, *98*, 244–252. [CrossRef]
11. Aymoz, D.; Wosika, V.; Durandau, E.; Pelet, S. Real-time quantification of protein expression at the single-cell level via dynamic protein synthesis translocation reporters. *Nat. Commun.* **2016**, *7*, 11304. [CrossRef] [PubMed]
12. Lin, J.; Jordi, C.; Son, M.; Phan, H.; Drayman, N.; Abasiyanik, M.; Vistain, L.; Tu, H.-L.; Tay, S. Ultra-sensitive digital quantification of proteins and mRNA in single cells. *Nat. Commun.* **2019**, *10*, 3544. [CrossRef]
13. Mair, F.; Erickson, J.; Voillet, V.; Simoni, Y.; Bi, T.; Tyznik, A.; Martin, J.; Gottardo, R.; Newell, E.; Prlic, M. A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. *Cell Rep.* **2020**, *31*, 107499. [CrossRef]
14. Munsky, B.; Trinh, B.; Khammash, M. Listening to the Noise: Random Fluctuations Reveal Gene Network Parameters. *Mol. Syst. Biol.* **2009**, *5*, 318. [CrossRef] [PubMed]
15. Lillacci, G.; Khammash, M. Parameter Estimation and Model Selection in Computational Biology. *PLoS Comput. Biol.* **2010**, *6*, e1000696. [CrossRef] [PubMed]
16. Zechner, C.; Ruess, J.; Krenn, R.; Pelet, S.; Peter, M.; Lagers, J.; Koeppl, H. Moment-Based Inference Predicts Bimodality in Transient Gene Expression. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 8340–8345. [CrossRef] [PubMed]
17. Lillacci, G.; Khammash, M. A Distribution-Matching Method for Parameter Estimation and Model Selection in Computational Biology. *Int. J. Robust Nonlinear Control* **2012**, *22*, 1065–1081. [CrossRef]
18. Ruess, J.; Milias-Argeitis, A.; Lygeros, J. Designing Experiments to Understand the Variability in Biochemical Reaction Networks. *J. R. Soc. Interface* **2013**, *10*, 20130588. [CrossRef] [PubMed]
19. Lillacci, G.; Khammash, M. The Signal within the Noise: Efficient Inference of Stochastic Gene Regulation Models Using Fluorescence Histograms and Stochastic Simulations. *Bioinformatics* **2013**, *29*, 2311–2319. [CrossRef] [PubMed]
20. Kepler, T.; Elston, T. Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. *Biophys. J.* **2001**, *81*, 3116–3136. [CrossRef]
21. Lipshtat, A.; Loinger, A.; Balaban, N.; Biham, O. Genetic Toggle Switch without Cooperative Binding. *Phys. Rev. Lett.* **2006**, *96*, 188101. [CrossRef]
22. Gillespie, D. Exact Stochastic Simulation of Coupled Chemical Reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361. [CrossRef]
23. Gillespie, D.T. Stochastic simulation of chemical kinetic. *Annu. Rev. Phys. Chem.* **2007**, *58*, 35. [CrossRef] [PubMed]
24. Jong, H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* **2004**, *9*, 67. [CrossRef]
25. Munsky, B.; Khammash, M. The Finite State Projection Algorithm for the Solution of the Chemical Master Equation. *J. Chem. Phys.* **2006**, *124*, 044104. [CrossRef] [PubMed]
26. Munsky, B.; Khammash, M. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *J. Comput. Phys.* **2007**, *226*, 818. [CrossRef]

27. Phillips, R.; Kondev, J.; Theriot, J.; Garcia, H.G. *Physical Biology of The Cell*; Garland Science: New York, NY, USA, 2013.
28. Cherry, J.; Adler, F. How to make a biological switch. *J. Theor. Biol.* **2000**, *203*, 117. [CrossRef] [PubMed]
29. Zhdanov, V. Transient stochastic bistable kinetics of gene transcription during the cellular growth. *Chem. Phys. Lett.* **2006**, *424*, 394. [CrossRef]
30. Cheng, Z.; Liu, F.; Zhang, X.-P.; Wang, W. Robustness analysis of cellular memory in an autoactivating positive feedback system. *FEBS Lett.* **2008**, *582*, 3776. [CrossRef] [PubMed]
31. Frigola, D.; Casanellas, L.; Sancho, J.; Ibanes, M. Asymmetric Stochastic Switching Driven by Intrinsic Molecular Noise. *PLoS ONE* **2012**, *7*, e31407. [CrossRef]
32. Ghosh, K.; Dixit, P.; Agozzino, L.; Dill, K. The Maximum Caliber Variational Principle for Nonequilibria. *Annu. Rev. Phys. Chem.* **2020**, *71*, 213–238. [CrossRef]
33. Pressé, S.; Ghosh, K.; Dill, K. Modeling Stochastic Dynamics in Biochemical Systems with Feedback Using Maximum Caliber. *J. Phys. Chem. B* **2011**, *115*, 6202–6212. [CrossRef] [PubMed]
34. Firman, T.; Balazsi, G.; Ghosh, K. Building Predictive Models of Genetic Circuits Using the Principle of Maximum Caliber. *Biophys. J.* **2017**, *113*, 2121–2130. [CrossRef] [PubMed]
35. Firman, T.; Wedekind, S.; McMorrow, T.; Ghosh, K. Maximum Caliber Can Characterize Genetic Switches with Multiple Hidden Species. *J. Phys. Chem. B* **2018**. [CrossRef]
36. Firman, T.; Amgalan, A.; Ghosh, K. Maximum Caliber can build and infer models of oscillation in three-gene feedback network. *J. Phys. Chem. B.* **2019**, *123*, 343–355. [CrossRef]
37. Pressé, S.; Ghosh, K.; Lee, J.; Dill, K. Principle of Maximum Entropy and Maximum Caliber in Statistical Physics. *Rev. Mod. Phys.* **2013**, *85*, 1115–1141. [CrossRef]
38. Dixit, P.; Wagoner, J.; Weistuch, C.; Pressé, S.; Ghosh, K.; Dill, K. Perspective: Maximum Caliber is a General Variational Principle for Dynamical Systems. *J. Chem. Phys.* **2018**, *148*, 010901. [CrossRef] [PubMed]
39. Keller, A. Model Genetic Circuits Encoding Autoregulatory Transcription Factors. *J. Theor. Biol.* **1995**, *172*, 169–185. [CrossRef]
40. Smolen, P.; Baxter, D.; Byrne, J. Frequency, Selectivity, Multistability, and Oscillations Emerge from Models of Genetic Regulatory Systems. *Am. J. Physiol.* **1998**, *274*, C531–C542. [CrossRef]
41. Becksei, A.; Seraphin, B.; Serrano, L. Positive Feedback in Eukaryotic Gene Networks: Cell Differentiation by Graded to Binary Response Conversion. *EMBO J.* **2001**, *15*, 2528–2535.
42. Tyson, J.; Chen, K.; Novak, B. Sniffers, Buzzers, Toggles and Blinkers: Dynamics of Regulatory and Signaling Pathways in the Cell. *Curr. Opin. Cell Biol.* **2003**, *15*, 221–231. [CrossRef]
43. Bishop, L.; Qian, H. Stochastic Bistability and Bifurcation in a Mesoscopic Signaling System with Autocatalytic Kinase. *Biophys. J.* **2010**, *98*, 1–11. [CrossRef] [PubMed]
44. Faucon, P.; Pardee, K.; Kumar, R.; Li, H.; Loh, Y.-H.; Wang, X. Gene Networks of Fully Connected Triads with Complete Auto-Activation Enable Multistability and Stepwise Stochastic Transitions. *PLoS ONE* **2014**, *9*, e102873. [CrossRef] [PubMed]