Biophysical Journal, Volume *113*

Supplemental Information

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

Taylor Firman, Gábor Balázsi, and Kingshuk Ghosh

Supporting Material

Application of Finite State Projection to Maximum Caliber

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

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

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

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

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

Alternate model to test MaxCal inference

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

$$\begin{gathered} \alpha \xrightarrow{g} \alpha + A \qquad\qquad A \xrightarrow{r} \varnothing \\ \alpha + A \underset{b_p}{\overset{f_p}{\rightleftarrows}} \alpha^* \qquad \alpha^* \xrightarrow{g^*} \alpha^* + A \end{gathered} \tag{S3}$$

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

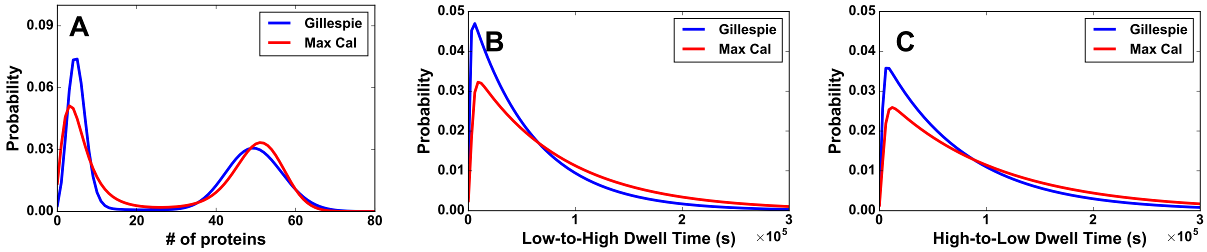

Figure S1: **Predicted distributions for alternate model agree well with the "true" distributions.** (**A**) Protein number probability distributions from synthetic input trajectories (blue) and predicted MaxCal trajectories (red). (**B**) Low state and (**C**) high state residence time probability distributions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

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

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

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

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