PLOS ONE

RESEARCH ARTICLE

Devil in the details: Mechanistic variations impact information transfer across models of transcriptional cascades

Michael A. Rowland*, **Kevin R. Pilkiewicz**, **Michael L. Mayo**

Environmental Laboratory, U.S. Army Engineer Research and Development Center, Vicksburg, MS, United States of America

* Michael.A.Rowland@usace.army.mil

Abstract

The transcriptional network determines a cell's internal state by regulating protein expression in response to changes in the local environment. Due to the interconnected nature of this network, information encoded in the abundance of various proteins will often propagate across chains of noisy intermediate signaling events. The data-processing inequality (DPI) leads us to expect that this intracellular game of "telephone" should degrade this type of signal, with longer chains losing successively more information to noise. However, a previous modeling effort predicted that because the steps of these signaling cascades do not truly represent independent stages of data processing, the limits of the DPI could seemingly be surpassed, and the amount of transmitted information could actually *increase* with chain length. What that work did not examine was whether this regime of growing information transmission was attainable by a signaling system constrained by the mechanistic details of more complex protein-binding kinetics. Here we address this knowledge gap through the lens of information theory by examining a model that explicitly accounts for the binding of each transcription factor to DNA. We analyze this model by comparing stochastic simulations of the fully nonlinear kinetics to simulations constrained by the linear response approximations that displayed a regime of growing information. Our simulations show that even when molecular binding is considered, there remains a regime wherein the transmitted information can grow with cascade length, but ends after a critical number of links determined by the kinetic parameter values. This inflection point marks where correlations decay in response to an oversaturation of binding sites, screening informative transcription factor fluctuations from further propagation down the chain where they eventually become indistinguishable from the surrounding levels of noise.

Citation: Rowland MA, Pilkiewicz KR, Mayo ML (2021) Devil in the details: Mechanistic variations impact information transfer across models of transcriptional cascades. PLoS ONE 16(1): e0245094. https://doi.org/10.1371/journal.pone.0245094

Editor: Ramon Grima, University of Edinburgh, UNITED KINGDOM

Received: September 25, 2020

Accepted: December 22, 2020

Published: January 13, 2021

Data Availability Statement: All data is found in the manuscript.

Funding: Funding was provided by the U.S. Army's Installations and Operational Environment Basic Research program. The authors express gratitude to Dr. E. Ferguson, Technical Director of the US Army ERDC Installations and Operational Environment Program, for support of this research. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

Introduction

Studies over the past half century have made it clear that eukaryotic gene-regulatory networks are exceedingly complex. Within these networks, proteins appropriately named transcription factors (TFs) bind to regulatory elements within promoter regions of DNA to modulate the

Competing interests: The authors have declared that no competing interests exist.

transcriptional rates of genes [1]. Once TFs bind to DNA, they may recruit other elements to activate the transcription of the gene or act to block additional critical binding events needed for transcription [1, 2]. TFs are able to bind multiple sites, although with varying levels of specificity, and some genes require interactions with multiple TFs to initiate transcription [2–5]. In the *Escherichia coli* bacterium (*E. coli*), for instance, the gene regulatory network is hierarchically organized so that only a handful of "global" TFs remain unregulated by any others, with more precise regulation controlled by co-regulation with local TFs [6]. This level of complexity can generate network structures in which a gene is controlled, directly or indirectly, by many upstream TFs. The *E. coli* gene *slp*, for example, is regulated by 17 different TFs [7]. This begs an important question regarding control of this and other biologically networked systems: To what extent can gene expression be reliably influenced by fluctuations in the activity of a TF that is several regulatory links removed? In other words, to what extent can the regulatory biology effectively convey an "upstream" signaling event if the information must propagate over a noisy molecular cascade?

The activity level of a TF (e.g., its time-series abundance within the nucleus) directly influences the response of a cell to changes in the environment. Biological functions, however, are inherently noisy, in this case either from the influence of the rest of the gene regulatory network, or through physical noise, such as the impact of Brownian motion on the binding kinetics between a TF and its binding site(s) [8–12]. The ability of a system to identify a signal fluctuation from the pervasive noise, and respond to it appropriately (what we have dubbed the *fluctuation sensitivity* [13]), can be quantified as the mutual information between the input and the output signals, i.e., between the time-dependent fluctuations in the concentration of a TF and those of some directly or indirectly regulated gene product [14]. In previous work, we investigated how the information propagated across a "daisy-chain" cascade of concatenated transcriptional regulatory events varied with the length of the cascade, as well as the linearized kinetic rate constants of the regulatory interactions. We found that, under certain conditions, longer cascades could exhibit higher mutual information than shorter cascades, seemingly in violation of the data-processing inequality (DPI) [13]. No actual violation occurs, however, because the dependence of a gene's regulation on the steady-state concentrations of all upstream transcription factors ensures that the individual regulatory interactions are statistically dependent upon one another; in other words, the fluctuations in protein abundance at the beginning and ends of a cascade remain significantly correlated despite the presence of noise.

By assuming that the kinetics of transcription were sufficiently well-described by their values near a homeostatic steady state, we linearized the fully nonlinear kinetics and found that protein production should outweigh its destruction to permit growth of information across successive cascade events. What we did not previously consider was whether such a regime was truly feasible in an actual biological system. (At the very least, it could not be sustainable for infinitely long cascades due to fluctuations increasing in magnitude across the chain, which would eventually violate our assumption of small fluctuations at steady state.) In this work, we address this concern by considering a more biologically relevant model of gene regulation that takes into account the explicit binding kinetics of each TF associating and dissociating with sequences within the transcription-initiating regions of DNA. To properly capture the fully nonlinear character of this kinetic model, we simulate it *in silico* using the Stochastic Simulation Algorithm (SSA) [15]. Using these simulation data, we compute the mutual information between fluctuations in the abundance of a "source" TF and the final gene product produced by the terminus of a daisy chain composed of transcriptional-regulatory interactions that we assume rate-limits protein production. Although adequately sampling the probability mass functions underlying this mutual information turns out to be a technical challenge for longer cascades due to the increasing size of fluctuations, we ultimately find that although the fluctuation sensitivity can be enhanced by initially increasing the length of the cascade, these gains are lost as the cascade continues to grow. Applying our linearized theory to this "explicit-binding" model, we find that it grossly overestimates the quantitative value of the mutual information; it nonetheless reproduces the qualitative, nonmonotonic behavior of initial growth followed by rapid decay seen in the mutual information computed from the SSA simulations. Importantly, the theory makes it clear that the eventual quenching of the fluctuation sensitivity occurs as a result of an emergent separation in time scales between the binding kinetics and those of the actual transcription process. This almost adiabatic separation interferes with communication between the steps of the cascade, resulting in the suppression of information about TF molecule fluctuations.

Results

In our previous work, we used a generalized model for transcriptional kinetics that we linearized for small fluctuations about steady state. Our model of a regulatory cascade was simplified by assuming that the linearized rate constants of each production/destruction process were equal in value across the chain. The most trivial (and least contrived) realization of this regime would be the case where each gene is regulated by an identical mass-action rate law in which temporal changes to the concentration of its encoded protein are directly proportional to the concentration of its regulating TF:

$$\frac{d\delta R_i}{dt} = k\delta R_{i-1} - k_d\delta R_i + \eta_i \qquad [1]$$

In the above, δR_i is the time-dependent concentration fluctuation of the i^{th} TF in the signaling cascade from its steady-state mean value (in other words, the i^{th} response to the initiating signal), k is the linear rate constant for TF production, and k_d is the rate constant of TF degradation. The function η_i is a delta-correlated Brownian noise term with zero mean, which we use to approximate the stochastic fluctuations caused by all the complex cellular machinery that we neglect to model explicitly. As stated, we assume that rate constants have identical values across the chain, each noise function has an identical statistical distribution, and R_0 is understood to be the concentration of the TF that initiates the cascade. For the sake of achieving closed-form analytic results, we neglected to model the regulation of this lead TF, instead assuming that it remained at a fixed homeostatic concentration until time t, at which point it experienced an instantaneous, stochastic fluctuation (the signal) drawn from the same distribution as that characterizing the noise in each other protein concentration. We then used the metric of mutual information to study how this signal correlated with the instantaneous response of each downstream TF population in the cascade. It should be noted that this response can only be instantaneous when the discrete molecular events of transcription are coarse grained as continuum processes, which is a reasonable approximation when studying protein fluctuations across an entire cellular population.

Under the above assumptions, we ultimately found that the condition for the fluctuation sensitivity of the cascade to grow with the number of links was $k > \sqrt{2}k_d$. So long as the signal fluctuation is, on average, the same size as a typical noise fluctuation, this result is independent of the noise strength. In this regime, the steady-state concentration of each TF is magnified by a factor k/k_d relative to the concentration of its regulator, and the effect of every random concentration fluctuation is similarly magnified across succeeding generations in the cascade. The fluctuations in the source TF always travel at least one more link than any other noisy fluctuation, which means that the signal always gets magnified more than the noise. We also work in the long-time limit ($t \rightarrow \infty$), where the impact of any single noise fluctuation tends to be dampened out over time by countless other fluctuations. By contrast, the signal excitation does not occur until time t, so its impact is, by assumption, not attenuated. This ultimately enables the signal-to-noise ratio—and thus the fluctuation sensitivity—to increase with longer cascades.

Clearly, this theoretical framework sacrifices a fair amount of physical realism in exchange for tractable mathematics. Indeed, the very basis for using continuum chemical kinetics to describe the discrete regulatory processes of individual cells relies on an assumption that the dynamics of a large population of cells can effectively be treated as one giant biochemical reactor. Stochastic fluctuations in the concentration of the signal TF would also have to be taken into account in a more realistic treatment, and would surely inhibit the growth potential of the fluctuation sensitivity. Our previous work demonstrated this latter point with some simple stochastic simulations of discretized mass-action kinetics, though a statistically significant growth trend with cascade length was still observed for $k \gg k_d$. Our objective in this work is to modify those simulations to further relax the cruder assumptions of our analytic model and thereby determine whether information gains across a cascade might be expected in more biologically plausible scenarios. In addition to allowing the number of signaling proteins to fluctuate stochastically, these expanded simulations will include an explicit treatment of the nonlinear protein-binding kinetics central to transcriptional signaling and will be parametrized to describe signaling within a single cell rather than an entire cellular population.

To meet this objective, we develop and study two models in which information encoded by molecular fluctuations propagates via regulatory interactions with differing kinetic mechanisms. In our first model, protein production and destruction rates are linearly dependent upon the concentrations, and its deterministic kinetics can be expressed by the following set of differential rate laws:

$$\frac{dR_i}{dt} = kR_{i-1} - k_d R_i. \qquad [2]$$

We refer to this as the mass-action (MA) model, and note that it is essentially equivalent to **Eq** [1], except that we have expressed its kinetics in terms of the absolute concentrations in order to emphasize that they are linear by construction, and not by linearization about a steady state. We have likewise suppressed the stochastic component of these kinetics to emphasize that the fluctuations in our simulated models will be controlled by the various reaction rates, rather than being imposed, as in our original analytic model, by a simple Brownian process.

Our second model modifies the kinetic mechanisms of **Eq** [2] as a step toward biological fidelity. If R_i is once again the concentration of the i^{th} TF in the cascade, B_i is the concentration of free DNA sites that bind that TF, and $R_i{\cdot}B_i$ is the concentration of those sites that have reversibly bound a TF molecule, then the deterministic component of this second model's kinetics can be represented by the following set of differential rate laws:

$$\frac{dR_i}{dt} = q_i R_{i-1} \cdot B_{i-1} - k_+(R_i)(B_i) + k_- R_i{\cdot}B_i - k_d R_i$$

$$\frac{dR_i{\cdot}B_i}{dt} = k_+(R_i)(B_i) - k_- R_i{\cdot}B_i. \qquad [3]$$

We refer to this as the "explicit binding" (EB) model and further assume that kinetics of binding, unbinding, and protein catabolism are identical across the cascade, so that the rate constants k_+, k_-, and k_d are of identical value for all TF species. The transcriptional kinetics, which we assume rate-limits protein production, can be different for each link in the chain; however, we shall choose values of q_i that allow for a fair comparison between this model and

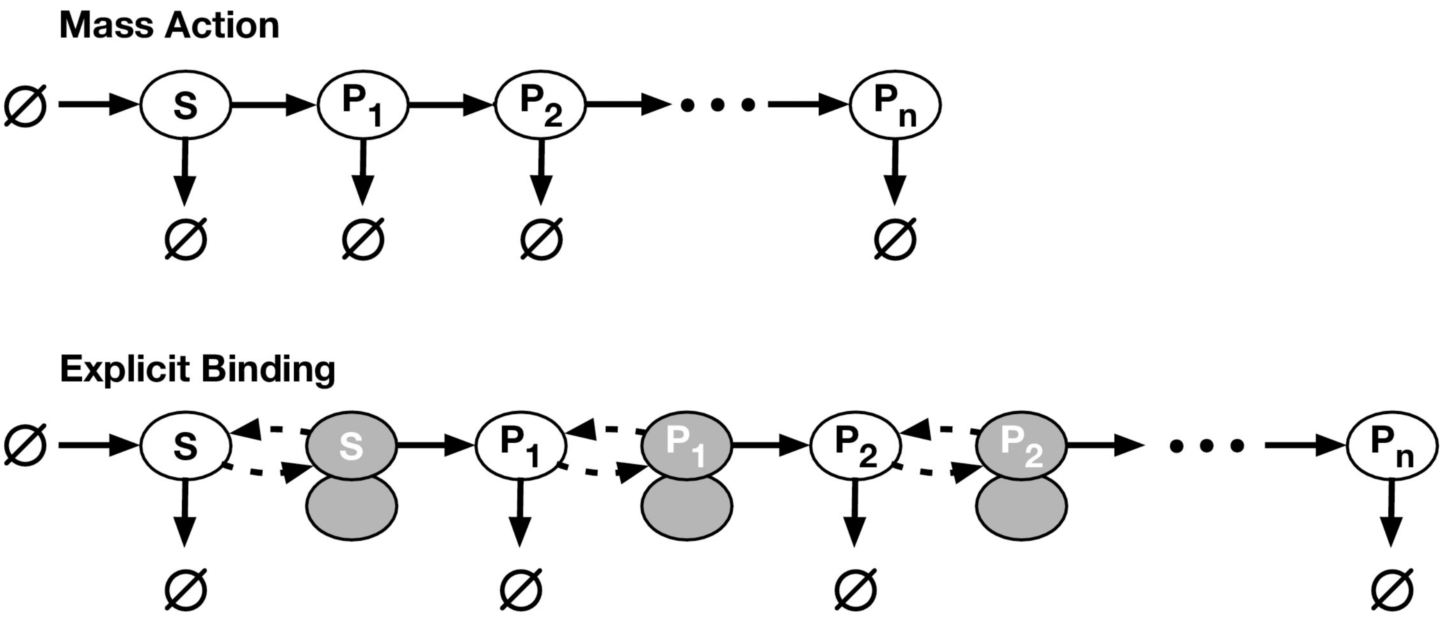

Fig 1. Diagram of the mass-action and explicit-binding models. In the mass-action model, source TF (*S*) is created and destroyed to maintain it at a steady state. *S* regulates the synthesis of gene product R_1, R_1 regulates R_2 (when present), etc. The explicit-binding model is similar to the mass-action, except a TF must first bind to a DNA binding site (bound TFs are shaded in the diagram) before it can stimulate the synthesis of its product. Note that the final product, not being a TF, does not have its own binding site.

https://doi.org/10.1371/journal.pone.0245094.g001

the simpler mass-action model in the regime where fluctuation sensitivity was predicted to grow with cascade length. Fig 1 provides a schematic of the elementary reactions that are part of a transcriptional signaling cascade described by **Eq** [2] and **Eq** [3]**.** For the sake of clarity, we once again stress that these two models will be compared through the lens of stochastic simulation. A comparison between the simulation results for each model and the predictions of our previously derived analytic model [13] will be postponed to the end of this section.

In the MA model, the average steady-state concentrations of the various species are related to one another by the following recursion:

$$\langle R_{i,0}\rangle = \frac{k}{k_d}\langle R_{i-1,0}\rangle, \quad [4]$$

wherein the angled brackets, $\langle\cdot\rangle$, denote temporal averages, and we have defined $R_{i,0}$ as the *total* concentration of the i^{th} TF protein. In the mass-action model, $R_{i,0} = R_i$, but in the explicit-binding model $R_{i,0} = R_i + R_i{\cdot}B_i$. To make a fair comparison between the two models, we want their kinetics to both fluctuate about the same set of steady-state concentrations $\{\langle R_{i,0}\rangle, \forall i\}$. If, as we have assumed, transcription initiation is rate-limiting, then we can leverage the resulting timescale separation to approximately treat the concentration $R_i{\cdot}B_i$ as if it were always at a steady state. This Briggs-Haldane quasi-steady state assumption (QSSA) amounts to an adiabatic separation of the frequent binding and unbinding kinetics and the slow, irreversible kinetics of transcription itself. Applying the QSSA to **Eq** [3], we can derive the following pair of equations:

$$q_i = \frac{k_d\langle R_i\rangle}{\langle R_{i-1,0}\rangle - \langle R_{i-1}\rangle}$$

$$\langle R_i \rangle = \frac{1}{2}[\langle R_{i,0} \rangle - B_{i,0} - K_D + \sqrt{(\langle R_{i,0} \rangle - B_{i,0} - K_D)^2 + 4K_D \langle R_{i,0} \rangle}] \qquad [5]$$

which, when combined with **Eq** [3], can be solved iteratively to select the values of the q_i required to restrict both models to identical mean steady states. Note that in the above, we have defined $B_{i,0} \equiv B_i + R_i \cdot B_i$ as the total number of binding sites for the i^{th} TF, which we assume to be fixed since these sites can neither be created nor destroyed, and we have introduced the dissociation constant $K_D \equiv k_- / k_+$. When recursively solving for the rate constants q_i, we assume that $\langle R_{0,0} \rangle$ and $B_{i,0}$ are known for all i and are inputs of the model.

In general, the average concentrations calculated from the chemical master equation do not quantitatively agree with average concentrations derived from a macroscopically valid rate equation treatment of chemical kinetics, such as given by **Eqs** [2] and [3]. This disagreement originates from the fact that the rate equation approach is valid in the thermodynamic limit wherein molecular fluctuations are, to good approximation, proportional to $\Omega^{1/2}$ (Ω being the volume of the relevant compartment or the system size); but this assumption is too restrictive for microscopic fluctuations in general. However, it can be shown using the well-known linear noise approximation [16] that for larger system size, the average calculated from the master equation obeys the macroscopic law for zero, first, and second order chemical reactions [17], which covers the region of validity of **Eqs** [2] and [3]. At mesoscopic scales, quantitative disagreement is more pronounced, but a more careful analysis of the higher orders of the system size expansion of the master equation can produce effective rate equations that are valid for any system size [17]. Despite an assumption of "small" fluctuations used to justify the linear mass-action kinetics of **Eq** [2], we found an increasing trend of mutual information for longer daisy chains simulated using the SSA for 2 to 128 source molecules [13], in qualitative agreement with predictions based on the rate equation approach.

We investigate the consequences of the kinetic mechanisms associated with the MA and EB models for transcriptional cascades of length $n = 1,\ldots,7$ through use of the stochastic simulation algorithm (SSA) [16] implemented within the KaSim v4.0 engine [18–20]. These simulations approximate the solution of the relevant chemical master equation, and, therefore, avoid many assumptions associated with the usual rate equation treatment at the cost of computational complexity. Simulations were initialized at steady state with kinetic parameters $k = 4$, $k_d = 1$, $k_+ = 0.1/4820$, and $k_- = 0.1$. The value of k was chosen to ensure that $k \gg k_d$, and $K_D = 4820$ ensures that TF proteins must typically make multiple binding attempts before a single transcription event occurs. We set $\langle R_{0,0} \rangle = 100$ and fixed $B_{i,0} = 3$ for all links in the cascade, which is both a reasonable estimate for the number of DNA promoter sites available to a TF within a cell, and a further guarantee that transcription initiation will be rate-limiting, regardless of the actual values of the rate constants q_i. (Even if the transcriptional rate constants are large, the low number of binding sites coupled with the large dissociation constant will severely limit how frequently new proteins can be produced.) Unlike in our previous theoretical approach [21], wherein the unregulated lead TF was assumed not to fluctuate in concentration until the time point of interest, we employ the more realistic assumption that R_0 obeys the following differential equation, which we also stochastically simulate:

$$\frac{dR_0}{dt} = \langle R_{0,0} \rangle - k_d R_0. \qquad [6]$$

A common issue plaguing the application of information theory metrics such as mutual information is the sensitivity of their calculated values to the choice of the bin size used in histogramming the data [22–25]. This is only an issue, however, when the underlying random variables are formally continuous, and must therefore be discretized post hoc to estimate the differential entropy. Although we are simulating the behavior of chemical kinetics models with continuous concentrations, our simulations respect the underlying discreteness of the real biochemical processes. As such, when computing the mutual information between the initial and final TF copy numbers for an n-link cascade,

$$I(R_n;R_0)=\sum_{R_n,R_0=0}^{\infty}\mathrm{p}(\mathrm{R_n},\mathrm{R_0})\log_2\left[\frac{p(R_n,R_0)}{p(R_n)p(R_0)}\right], \quad [7]$$

the probability mass functions on the right-hand side of the above expression can be interpreted as histograms with separate bins for each possible number of protein molecules. Note that the joint probability $p(R_n, R_0)$ will be a two-dimensional histogram constructed from the subset of all same-time pairs of R_n and R_0 counts.

To sufficiently sample such a large number of bins, a large number of data points is required [26]. To determine just how high a sampling density we require to obtain consistent values of the mutual information in **Eq** [7], we first simulated a pair of uncoupled transcription factors that both obey the kinetics of **Eq** [6], and whose initial, steady-state copy counts were 100 and 400, respectively. Since the stochastic fluctuations in the number of molecules of these two TFs are, by construction, independent and identically distributed (iid), their mutual information should, in principle, be exactly zero. Any finite ensemble of instantiations of this system will approach a zero mutual information only asymptotically; to determine a sufficient sampling density to approximate this condition, we simulated this system for 50 time units (the inverse unit to that of the degradation rate constant), taking snapshots of its molecular composition at regular intervals. By choosing the size of this interval differently, we were able to compile datasets containing between 10^3 and 10^6 samples, and for each dataset of a given number of samples, we performed 100 replicate simulations.

We plotted the mutual information values computed from each simulation in **Fig 2,** as a function of their sampling density. Although there is some expected variation across replicates, this pales in comparison to the variations across different sample sizes. Initially, with only 10^3 data points per simulation, we calculated around 4.5 bits transferred, which is erroneously quite large, exemplifying the ability of spurious fluctuations to bias the value of the mutual information [27]. A three orders of magnitude increase in the number of data points is required to decrease this value to 0.25 bits, far below the 1-bit threshold required to determine with precision if the signal is above or below the mean. This monotonically decreasing trend of the mutual information toward zero with increasing sampling density is well fit with a sigmoidal equation, in which a line drawn from the slope at its inflection point crosses the log-scaled sample axis (x-axis) at approximately 5.23 (**Fig 2**, red dotted line). This suggests that we need at least $10^{5.23} \approx 170{,}000$ data points to sufficiently reduce the impact of an imperfect sampling methodology on the value of the calculated mutual information.

Based on this analysis, we simulated each interacting transcriptional cascade for 50 time units, capturing snapshots of the total copy number of each molecular species every 5 x 10^{-5} time units. We then calculated the mutual information between R_0 and R_n for each of 100 replicate simulations. These replicate-averaged results are now shown in **Fig 3** as a function of cascade length for the MA (blue), the EB model (red), and a model with non-interacting/uncorrelated gene products accumulated at the same steady-state concentrations (green). We term this latter model, the non-interacting (NI) model. The error bars represent 95% confidence intervals, which we obtained by bootstrapping the results of the simulations with replacement 1000 times. Although the mutual information also increases with chain length for the explicit-binding model, this trend plateaus for chains of approximately $n \approx 3$–4 links, and

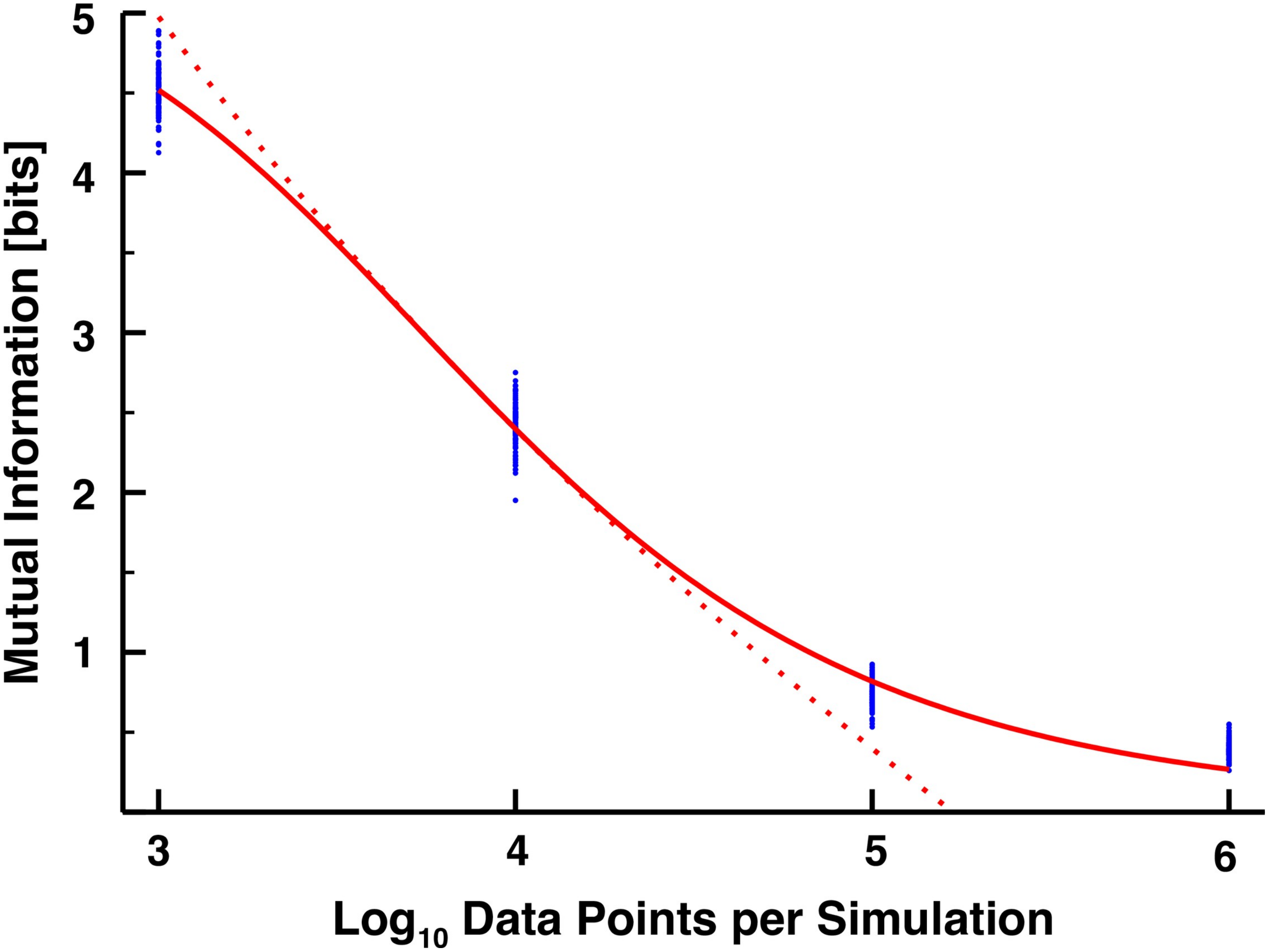

Fig 2. The mutual information (in bits) between two independent transcription factors as a function of the number of data points sampled per simulation. Different sampling densities were achieved by sampling the data more or less frequently, and 100 simulations were performed for each sampling frequency (blue dots). We fitted a sigmoidal relationship to the data ($y = y_{max}/(1+(x/K)^h)$, $y_{max} = 5.29832$, $K = 3.88923$, $h = 6.765666$, solid red curve). We then approximated the power law region of the sigmoid with the line $y = m \ln x+b$, $m = -8.9617$, $b = 5.2269$.

https://doi.org/10.1371/journal.pone.0245094.g002

for longer chains, the mutual information rapidly decays toward the value given by the non-interacting model. Although biology should generally disfavor longer chains simply due to the greater metabolic burden they place on the cell, they also appear unfavorable from a signaling perspective.

The EB model is clearly less informationally efficient than the MA model for the set of parameters chosen in Fig 3, but we now demonstrate that the EB model is less efficient for any set of parameters. To prove this, we consider only a single regulatory interaction ($n = 1$) and show that none of the three control parameters of the explicit-binding model can make it out-perform the mechanistically simpler alternative. The first parameter we consider is the ratio k/k_d, which controls the steady-state ratio of concentrations $\langle R_{1,0}\rangle/\langle R_{0,0}\rangle$, and, in Fig 4A, we plot the mutual information for both the explicit-binding and mass-action models for a single-link cascade as a function of this dimensionless parameter. We varied k/k_d from 2^{-2} to 2^2, and for

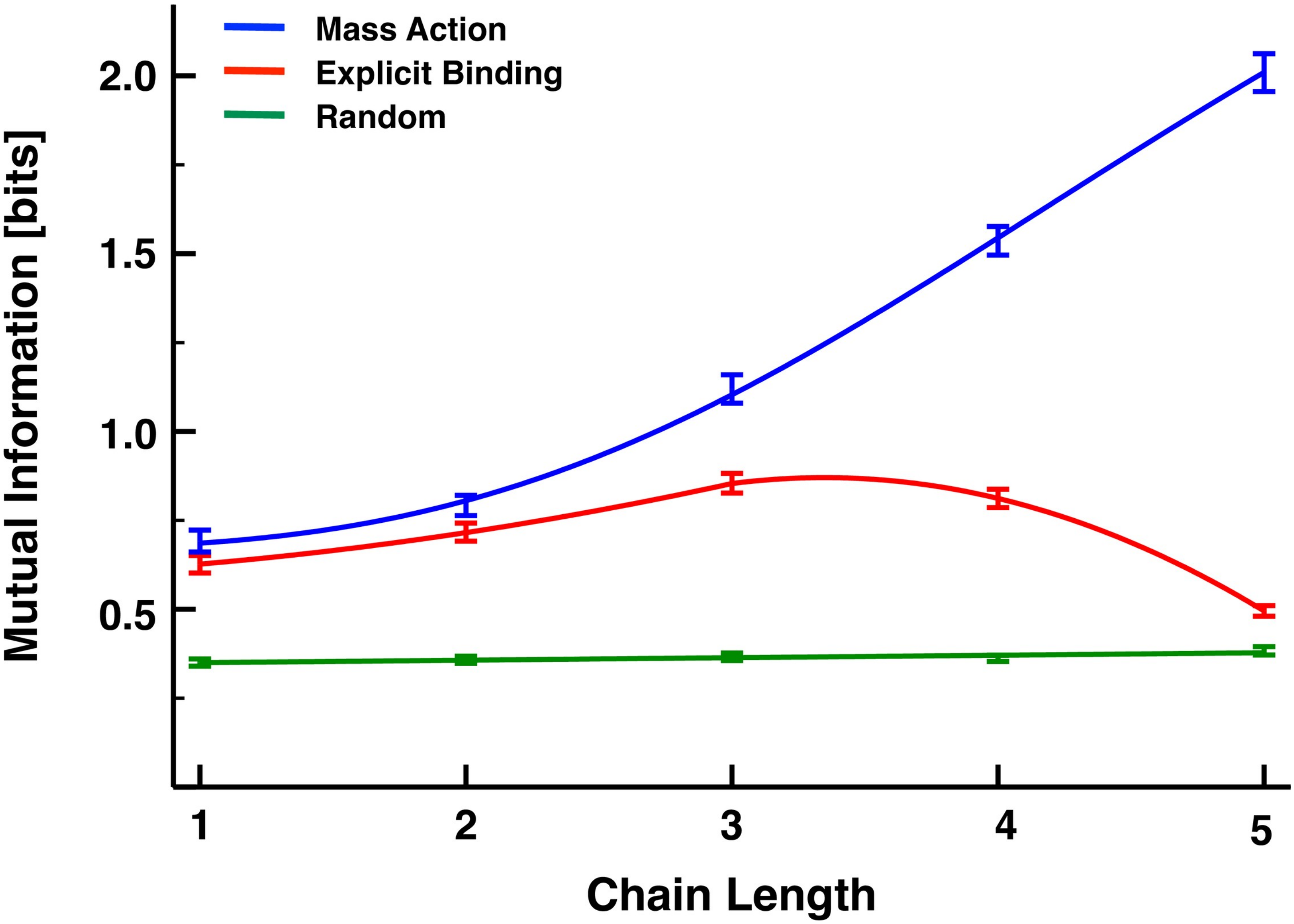

Fig 3. Mutual information as a function of chain length for the MA (blue), EB (red), and NI (green) models. This latter model, whose mutual information should be formally zero, provides a measure of the magnitude of the spurious correlations resulting from our sampling choices. The bars represent the 95% confidence intervals of the mean mutual information as measured by bootstrapping from 100 replicate simulations. The mass-action model results were fit with the sigmoid $y = y_{min}+x^h(y_{min}+y_{max})/(1+(x/K)^h)$, $y_{min} = 0.67027$, $y_{max} = 3.45238$, $K = 175.58899$, $h = 3.16516$. The explicit-binding model results were fit piecewise with the quadratic $y = ax^2+bx+c$; $a = 0.02473$, $b = 0.01435$, $c = 0.58801$ for $x \leq 3$, and $a = -0.14725$, $b = 0.91883$, $c = 0.66762$ for $x>3$.

https://doi.org/10.1371/journal.pone.0245094.g003

each value from within this interval, simulated the models exactly as before, while keeping all other parameters the same as those used to produce the curves of Fig 3. At any given value of k/k_d, we find that the two models share a similar amount of information, and in both cases this information grows roughly linearly with the logarithm of the control parameter. The best-fit line to the MA simulation data does, however, exhibit a statistically significantly steeper slope than that of the EB model (see Table 1), which suggests that there could be statistically significant differences between the fluctuation sensitivities of the two models at much larger or much smaller values of k/k_d. In the former case, this difference would favor the mass-action model even more, and in the latter case, the mutual information values would be indistinguishable from noise. (Recall from Fig 3 that for datasets with a million samples, this indistinguishability threshold fell at roughly 1/4 of a bit.)

The EB model has two control parameters that are not present in the MA or NI models: the dissociation constant K_D and the number of TF binding sites $B_{i,0}$. In Fig 4B, we fix k/k_d and

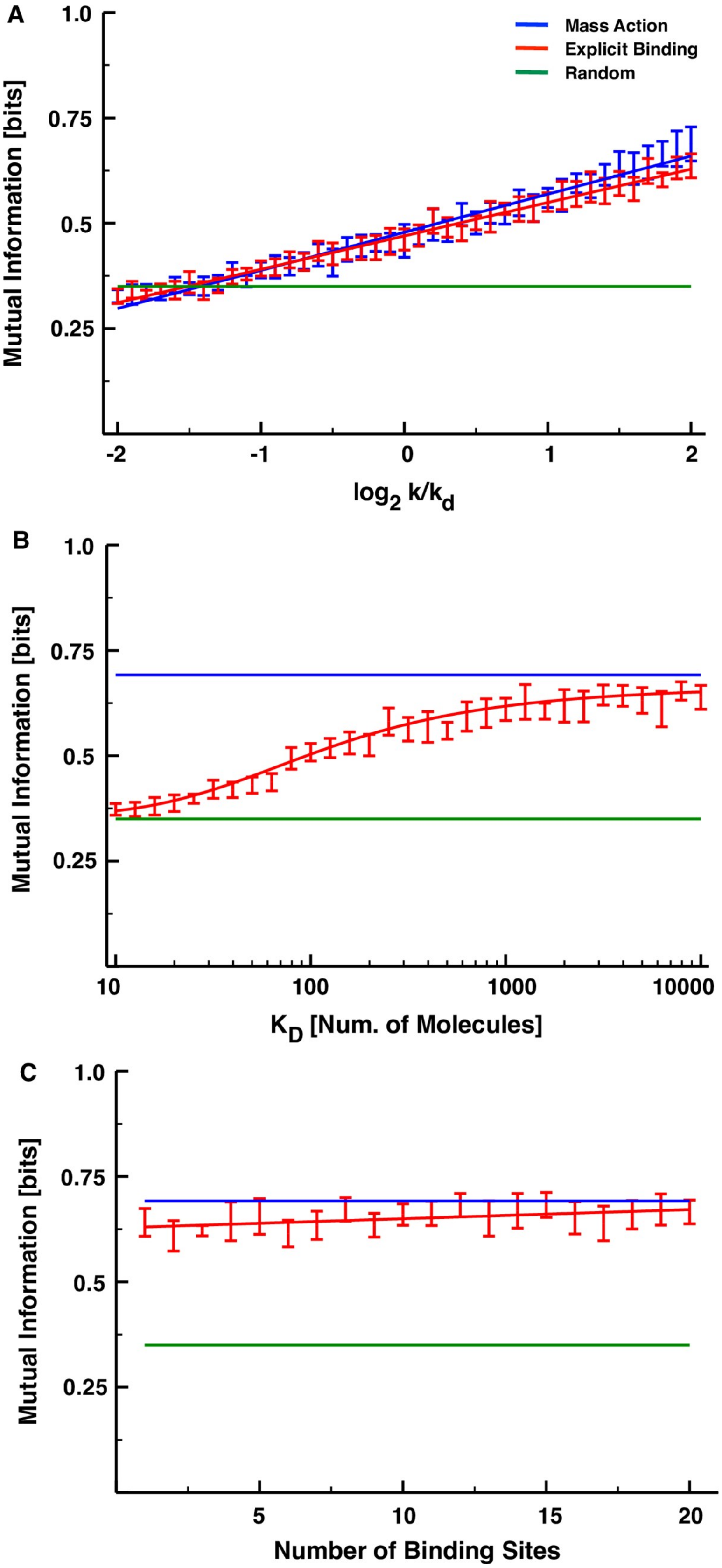
A
B
C
Mass Action
Explicit Binding
Random
Mutual Information [bits]
1.0
0.75
0.5
0.25
-2
-1
0
1
2
$\log_2 k/k_d$
10
100
1000
10000
K_D [Num. of Molecules]
5
10
15
20
Number of Binding Sites

Fig 4. Sensitivity of the mutual information of each model on the parameters. (A) Mutual information as a function of k/k_d for the MA (blue) and EB models (red). The bars represent the estimated 95% confidence intervals of the mean mutual information as determined by bootstrapping the results of 100 replicate simulations. Each model was fit to the function y = ax + b, with a = 0.090467, b = 0.478574 for the mass-action model and a = 0.079252, b = 0.469904 for the explicit-binding model. (B) Mutual information as a function of K_D for the explicit-binding model. These results were fit with the sigmoid $y = y_{min} + (y_{max}–y_{min})(x^h)/(K + x^h)$, $y_{min} = 0.356155$, $y_{max} = 0.667316$, K = 22.917036, h = 4.367485. (C) Mutual information as a function of the number of binding sites for the explicit model. These results were fit with the line y = mx + b, with m = 0.0021724, b = 0.6280985.

https://doi.org/10.1371/journal.pone.0245094.g004

$B_{i,0}$ to the values used in Fig 3, and plot the resulting mutual information as a function of K_D for a single-link cascade obeying explicit-binding kinetics. The dashed lines in the plot mark the mutual information levels for the mass-action and non-interacting models, and the sigmoid curve we use to fit the EB simulation data appears to saturate at the former for large K_D and the latter for small K_D. In other words, this means that if the binding of transcription factors to DNA is too efficient, differing concentration fluctuations cannot be discriminated by the transcriptional mechanism; but no matter how inefficient the binding becomes, the fluctuation sensitivity can never surpass the mass-action limit. In Fig 4C we repeat this exercise varying $B_{i,0}$ while keeping all other parameters fixed, and we see that there is only a very weak growth trend in the mutual information with the number of binding sites (slope = 0.0021724, p = 0.00432). If it is possible for the fluctuation sensitivity of the EB model to exceed that of the mass-action model for a sufficiently large number of binding sites, it would clearly have to be for a biologically infeasible number of them.

To better understand the trends observed in Fig 4, we analyze the EB model within the previously-developed linearized kinetics framework [14]. Starting from **Eq** [3] and applying the Briggs-Haldane QSSA for $R_i{\cdot}B_i$, we can reduce the 2n differential equations governing our n-link cascade to only n coupled equations:

$$\frac{dR_i}{dt} = \frac{q_i(B_{i,0})(R_{i-1})}{K_D + R_{i-1}} - k_d R_i. \qquad [8]$$

Taylor expanding **Eq** [8] about steady state and keeping only terms of linear order, we then get the following:

$$\frac{d\delta R_i}{dt} = \tilde{k}_i \delta R_{i-1} - k_d \delta R_i + \eta_i, \qquad [9]$$

wherein we have explicitly added the stochastic noise term and defined an effective rate constant $\tilde{k}_i$ as:

$$\tilde{k}_i \equiv \frac{q_i B_{i-1,0} K_D}{(K_D + \langle R_{i-1} \rangle)^2}. \qquad [10]$$

Table 1. Estimates of the parameters for the function $y = ax+bzx+c+dz$, fitted to the mean mutual information for increasing k/k_d values for the mass action model (z = 1) against the explicit binding (z = 0).

A	b	p-value	c	d	p-value
0.79252	0.011214	4.45e-5	0.469904	0.008671	0.00597

Significant p-values indicate that the values of b and d are relevant, meaning that the slope and intercept of the best-fit curves for the two models are significantly different.

https://doi.org/10.1371/journal.pone.0245094.t001

If we assume that $\langle R_{i,0}\rangle \approx \langle R_i\rangle$ when $\langle R_{i,0}\rangle \gg B_{i,0}$, then **Eq** [4] can be used to express **Eq** [10] in terms of the mass-action rate constant k instead of q_i:

$$\tilde{k}_i = \frac{kK_D}{K_D + \langle R_{i-1}\rangle}. \quad [11]$$

The above set of linearized rate constants can then be substituted into our previously reported, approximate formula for I_∞, which is the long-time limiting value of the mutual information transferred by an n-link signaling cascade whose kinetics consist of small concentration fluctuations about steady state (see **Eq [27]** of reference [13]):

$$I_\infty\left(n, \{\tilde{k}_i\}\right) = \frac{1}{2}\log\left\{1 + \frac{\left(\prod_{i=1}^{n} \tilde{k}_i\right)^2 / k_d^{2n}}{\sum_{m=1}^{n} \prod_{j=1}^{m-1} \left[\tilde{k}_{n-(j-1)}^2 / \left(2k_d^2\right)^{m-1}\right]}\right\}. \quad [12]$$

In **Fig 5**, we plot the mutual information of **Eq** [12] for both the explicit-binding and mass-action models as a function of cascade length, using the same parameter values from **Fig 3**.

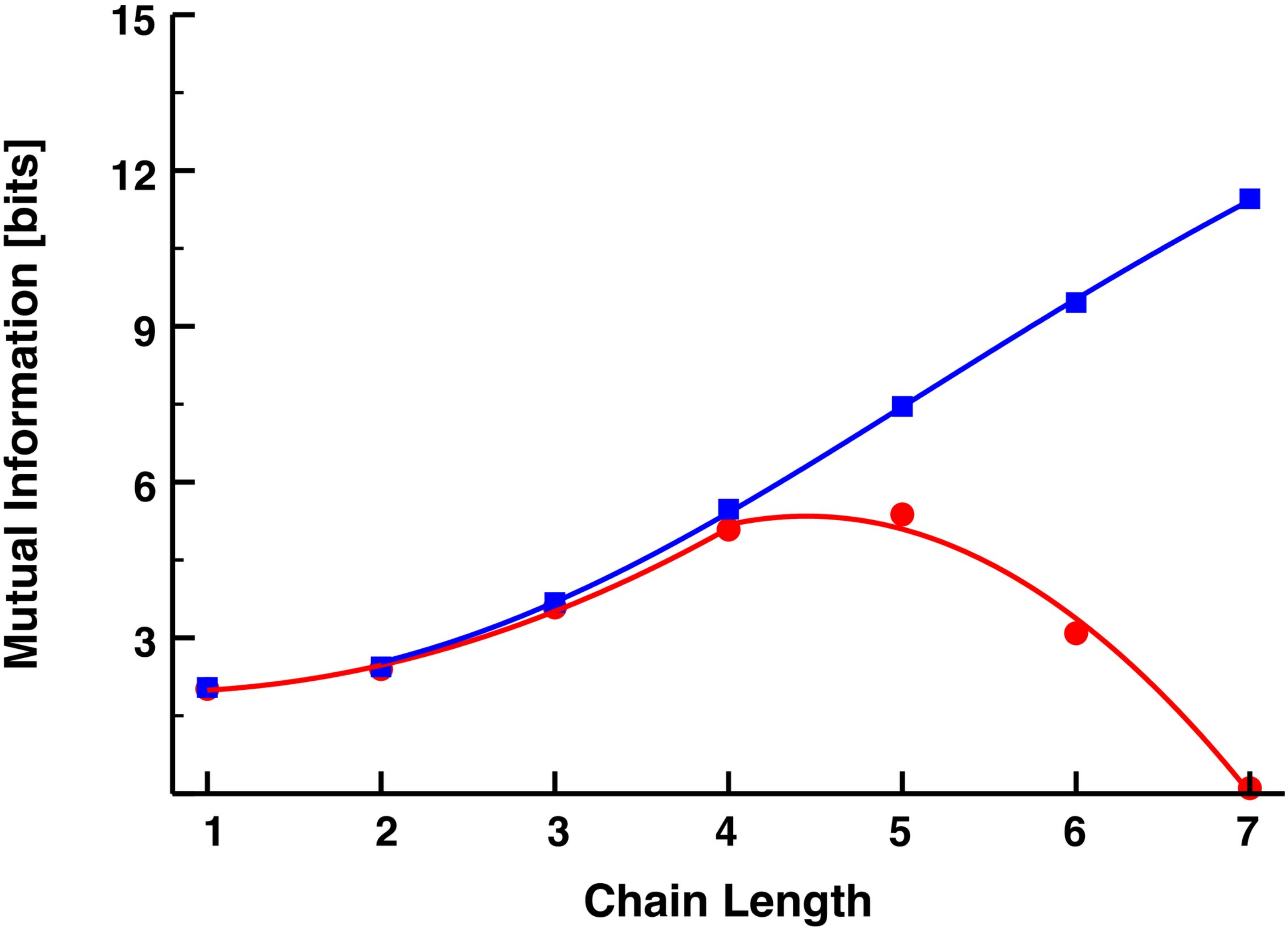

Fig 5. Theoretical predictions of the mutual information (in bits) for a signaling cascade described by the EB model kinetics (red) and MA model kinetics (blue). All parameters were chosen the same as those used in **Fig 3**, and the mutual information plots from that figure are replotted here for ease of comparison. The results from the MA simulations were fit with the sigmoidal function $y = y_{min}+x^h(y_{max}-y_{min})/(K+x^h)$, $y_{min} = 1.88496$, $y_{max} = 20.17665$, $K = 179.79606$, $h = 2.71205$. The results from the EB simulations were fit piecewise with the quadratic $y = ax^2+bx+c$; $a = 0.27913$, $b = -0.35633$, $c = 2.06838$ for $x \leq 4$; $a = -0.8188$, $b = 7.2838$, $c = -10.8557$ for $x > 4.74923792$.

https://doi.org/10.1371/journal.pone.0245094.g005

Although **Eq** [12] considerably overestimates the value of the mutual information for both models, it qualitatively captures the features of the simulation results: information associated with the EB model and its reversible binding mechanisms is bounded from above by the MA model results; and, after initially growing with the number of signaling links, it rapidly decreases towards zero. This theory also overestimates how many links it takes to reverse the growth trend in the mutual information (five rather than three), but this, along with the overall larger information values, can be attributed to our theoretical framework ignoring the noise in the signal, which would no doubt reduce the informational efficiency of the cascade.

Noting that the condition for the mutual information in **Eq** [12] to grow with n is roughly $\tilde{k}_i \gg k_d \forall i$, one can use **Eq** [11] to demonstrate that there is in fact no regime in which the EB model kinetics can achieve the monotonic growth in fluctuation sensitivity that is possible in the limit of the MA model. If on the one hand, we choose $k>k_d$, then the steady-state concentration $\langle R_{i-1} \rangle$ will grow monotonically with i, leading $\tilde{k}_i$ to invariably become smaller than k_d after some critical value of i. If, on the other hand, we choose $k<k_d$, then the sequentially decreasing values of the steady-state concentrations will eventually reduce the effective rate constant in **Eq** [11] to approximately the value of k, which is less than k_d by assumption. We can also use the above equations to account for all of the trends observed in **Fig 4** by substituting **Eq** [11] into **Eq** [12] for the case $n = 1$:

$$I_\infty\left(1, \tilde{k}_1\right) = \frac{1}{2}\log_2\left[1 + \left(\frac{kK_D}{k_d(K_D + \langle R_0 \rangle)}\right)^2\right]. \quad [13]$$

So long as the squared term inside the argument of the logarithm in **Eq** [13] is much larger than unity, the information clearly scales as $\log(k/k_d)$. As $K_D \to 0$, the information approaches zero, and as $K_D \to \infty$, the information approaches the limit of the MA model, wherein $\tilde{k}_1 = k$. Finally, the information in **Eq** [13] does not depend at all upon the number of binding sites, $B_{0,0}$.

Discussion

In this work, we set out to determine whether the ability of biological signaling cascades to sidestep the limitations of the data-processing inequality—a prediction made by a previously developed theory based upon a linearization of the fully nonlinear kinetic mechanism—was actually attainable in a model that did not rely on as many coarse approximations, and that explicitly accounted for certain aspects of the real biology of cellular transcriptional signaling. The EB model we employed, while still a gross simplification of real biology, at least required the transcription factors in charge of protein regulation to reversibly bind to DNA before being able to influence the rate of gene translation and subsequent transcription. By stochastically simulating the full nonlinear kinetics of this model, we were able to avoid making many of the approximations required to make the linearized kinetic theory algebraically tractable. Nonetheless, we still found that the information transmitted across a transcriptional signaling cascade can increase with the number of regulatory links—it just cannot grow indefinitely. After increasing for a few links, the signal abruptly becomes indistinguishable from noise after only an additional link or two. We found that our linearized theory, when applied to the EB model equations, can reproduce this phenomenology, though it grossly overestimates the absolute magnitude of the mutual information. This enabled us to justify our finding that simpler models free of reversible binding kinetics, which allow transcription factors to directly regulate protein synthesis without first binding to a DNA promoter site, provide an upper bound on the informational efficiency of the EB model, even for short cascades where both models predicted monotonic, link-by-link signal amplification. This result in particular suggests that the kinetics of protein binding serve as a sort of signal dampener that further complicates the evolutionary narrative of molecular communication in biological systems.

Due to the high number of different signaling molecules crisscrossing the cellular cytoplasm, protein binding requires a high level of specificity to be effective, and it must also be reversible, so that binding sites are not occupied longer than necessary. These constraints favor protein dissociation over association, which means that multiple cycles of binding and unbinding must typically occur before processes like transcription can successfully initiate. This results in a separation of time scales, wherein the kinetics of association and dissociation can be thought to exist at all times in a quasi-steady state with respect to the kinetics of transcription itself. When the steady-state protein concentrations grow across the length of a cascade (generally true when the rate of transcription outpaces that of protein catabolism), this quasi-steady fraction of occupied binding sites will approach saturation with each successive link. Once this saturation is reached, the number of bound proteins will, on the time scale of transcription, effectively not fluctuate. In this limit, a fluctuation in the number of free TF molecules (the signal) cannot be transmitted, since a commensurate fluctuation in the concentration of bound TF molecules cannot be induced. This essentially adiabatic regime is fundamentally why the fluctuation sensitivity of an explicit-binding cascade inevitably falls off after enough links: increasing the amount of transmitted information requires an amplification in the number of proteins, but this amplification saturates the rate of transcription, thereby rendering the kinetics insensitive to fluctuations.

In addition to limiting the length of regulatory cascades over which information can be meaningfully transmitted, a saturating rate of transcription also suppresses the absolute amount of information that can be transferred over a cascade below the single-bit threshold. Less than a single bit of information corresponds to a response of "maybe" to a "yes" or "no" question, suggesting that individual cells struggle with even a binary response to environmental changes. This low communication capacity is consistent with past investigations that have found an association between poor intracellular communication and efficient population level responses [28]. Cells typically exist as part of a large population, and adaptation to an environmental change seldom requires the participation of every single cell. Low fidelity communication within each individual cell ensures that only a fraction of the population will succeed in responding to a stimulus, and this can actually be healthier for the community as a whole by conserving resources and avoiding a population-amplified response that exceeds the scale of the triggering stimulus.

Our modeling of the effect of binding kinetics on information transmission along signaling chains is general enough to suggest a molecular role in constraining biological network structure. Gene regulatory networks, for example, may grow through an evolutionary mechanism that involves gene duplication and divergence to generate new regulatory interactions [29]. Although networks modeled statistically with this growth mechanism have some topological similarity with known gene-regulatory networks (i.e., they are "scale-free," "small world" networks [30]), they do not explicitly account for the underlying regulatory mechanisms which connect network structure with function and phenotype [31]. Our information-theoretic analyses identify a signaling "length" scale for these and possibly other molecular networks, suggesting a new mechanism of consideration in models that hope to explain the large-scale structure of molecular networks. If the structure of these networks is constrained, in part, by molecular binding events, then our theory predicts that longer chains should exhibit binding interactions that are weaker (larger K_D) than comparatively shorter chains. Experiments could test this hypothesis, for example, by comparing the value of curve-fitted rate constants for the kinetic activity of fluorescent protein reporters in shorter and longer regulatory chains of protein expression.

To determine whether such experiments are possible, we reviewed datasets from the BIOGRID database [32], which provides a number of gene regulatory networks obtained for singular and multicellular organisms. We reviewed datasets for Saccharomyces cerevisiae (baker's yeast), the Escherichia coli bacterium, Drosophila melanogaster (fruit fly), Mus musculus (house mouse), and Homo sapiens, searching them for regulatory daisy chains of 3, 4, and 5 nodes with, respectively, 2, 3, and 4 links. Specifically, we searched for regulatory daisy chains in which none of the intermediate genes exhibited interactions beyond the adjacent ones. We found only the vertebrate datasets exhibited regulatory daisy chains with up to 3 links, and no datasets we reviewed had any chains with 4 links. For example, in the mouse dataset, we identified 2699 2-link and 148 3-link daisy chains. A more thorough analysis of the functions associated with these chains is beyond the scope of our discussion, but their existence shows that cell-based expression assays could be used as a basis to test the general results from our mathematical models.

Ultimately, we have demonstrated that seemingly small mechanistic details can have a profound impact on how information flows through a system. By better understanding how the granular mechanisms of molecular signaling events impact the communication capacities of complex biological networks, we can perhaps one day use mechanistic knowledge to make predictions about network topology or vice versa. For example, the lack of long, linear cascades in the transcriptional network of the bacterium *Escherichia coli* may in fact be nature's attempt to compensate for the very limitations on information flow that we have predicted with our modeling.

Acknowledgments

The authors would like to thank Dr. Pratip Rana for analyses of several gene regulatory networks to support the possibility of future experiments. The authors express gratitude to Dr. E. Ferguson, Technical Director of the US Army ERDC Installations and Operational Environment Program, for support of this research. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

Author Contributions

Conceptualization: Michael A. Rowland, Kevin R. Pilkiewicz, Michael L. Mayo.

Formal analysis: Michael A. Rowland, Kevin R. Pilkiewicz, Michael L. Mayo.

Funding acquisition: Michael L. Mayo.

Investigation: Michael A. Rowland, Kevin R. Pilkiewicz, Michael L. Mayo.

Methodology: Michael A. Rowland, Kevin R. Pilkiewicz.

Project administration: Michael L. Mayo.

Supervision: Michael L. Mayo.

Validation: Michael A. Rowland, Kevin R. Pilkiewicz.

Visualization: Michael A. Rowland.

Writing – original draft: Michael A. Rowland.

Writing – review & editing: Michael A. Rowland, Kevin R. Pilkiewicz, Michael L. Mayo.

References

1. Latchman DS. Transcription factors: an overview. Int J Exp Pathol. 1993; 74(5):417–22. Epub 1993/10/01. PMID: 8217775; PubMed Central PMCID: PMC2002184.
2. Lemon B, Tjian R. Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 2000; 14(20):2551–69. Epub 2000/10/21. https://doi.org/10.1101/gad.831000 PMID: 11040209.
3. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ. Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol. 2004; 5(8):R56. Epub 2004/08/04. https://doi.org/10.1186/gb-2004-5-8-r56 PMID: 15287978; PubMed Central PMCID: PMC507881.
4. He X, Chen CC, Hong F, Fang F, Sinha S, Ng HH, et al. A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. PLoS One. 2009; 4 (12):e8155. Epub 2009/12/04. https://doi.org/10.1371/journal.pone.0008155 PMID: 19956545; PubMed Central PMCID: PMC2780727.
5. Bilu Y, Barkai N. The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biol. 2005; 6(12):R103. Epub 2005/12/17. https://doi.org/10.1186/gb-2005-6-12-r103 PMID: 16356266; PubMed Central PMCID: PMC1414079.
6. Martinez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol. 2003; 6(5):482–9. Epub 2003/10/24. https://doi.org/10.1016/j.mib.2003.09.002 PMID: 14572541.
7. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. Nucleic Acids Res. 2004; 32(22):6643–9. Epub 2004/12/18. https://doi.org/10.1093/nar/gkh1009 PMID: 15604458; PubMed Central PMCID: PMC545451.
8. Ahrends R, Ota A, Kovary KM, Kudo T, Park BO, Teruel MN. Controlling low rates of cell differentiation through noise and ultrahigh feedback. Science. 2014; 344(6190):1384–9. https://doi.org/10.1126/science.1252079 PMID: 24948735; PubMed Central PMCID: PMC4733388.
9. Allen LJS. Stochastic Processes with Applications to Biology. Upper Saddle River: Pearson Prentice Hall; 2003.
10. Balazsi G, van Oudenaarden A, Collins JJ. Cellular decision making and biological noise: from microbes to mammals. Cell. 2011; 144(6):910–25. https://doi.org/10.1016/j.cell.2011.01.030 PMID: 21414483; PubMed Central PMCID: PMC3068611.
11. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297(5584):1183–6. Epub 2002/08/17. https://doi.org/10.1126/science.1070919 PMID: 12183631.
12. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009; 459(7245):428–32. https://doi.org/10.1038/nature08012 PMID: 19363473; PubMed Central PMCID: PMC2858974.
13. Pilkiewicz KR, Mayo ML. Fluctuation sensitivity of a transcriptional signaling cascade. Phys Rev E. 2016; 94(3–1):032412. Epub 2016/10/16. https://doi.org/10.1103/PhysRevE.94.032412 PMID: 27739739.
14. Cover TM, Thomas JA. Elements of Information Theory. New York: Wiley; 1991.
15. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977; 81 (25):2340–61.
16. Van Kampen NG. Stochastic Processes in Physics and Chemistry: Elsevier; 1992.
17. Grima R. An effective rate equation approach to reaction kinetics in small volumes: theory and application to biochemical reactions in nonequilibrium steady-state conditions. J Chem Phys. 2010; 133 (3):035101. Epub 2010/07/24. https://doi.org/10.1063/1.3454685 PMID: 20649359.
18. Danos V, Feret J, Fontana W, Harmer R, Krivine J. Rule based modeling of biological signaling. In: Caires L, Vasconcelos VT, editors. Proceedings of CONCUR 2007. 4703 of LNCS: Spring; 2007. p. 17–41.
19. Danos V, Laneve C. Formal molecular biology. Theoretical Computer Science. 2004;325.
20. Faeder JR, Blinov ML, Hlavacek WS. Rule based modeling of biochemical networks. Complexity. 2005:22–41.
21. Mayo M, Pilkiewicz K. Multiscale Modeling of Information Conveyed by Gene-Regulatory Signaling. BICT 2015. 2016:148–51.
22. Daw CS, Finney CEA, Tracy ER. A review of symbolic analysis of experimental data. Rev Sci Instrum. 2003; 74(2):915–30. https://doi.org/10.1063/1.1531823 WOS:000180579500001.

23. Timme NM, Lapish C. A Tutorial for Information Theory in Neuroscience. eNeuro. 2018; 5(3). Epub 2018/09/14. https://doi.org/10.1523/ENEURO.0052-18.2018 PMID: 30211307; PubMed Central PMCID: PMC6131830.

24. Ross BC. Mutual Information between Discrete and Continuous Data Sets. Plos One. 2014; 9(2). ARTN e87357 https://doi.org/10.1371/journal.pone.0087357 WOS:000331711900011. PMID: 24586270

25. Camp J, Robb R. A novel binning method for improved accuracy and speed of volume image coregistration using normalized mutual information. Proc Spie. 1999; 3661:24–31. https://doi.org/10.1117/12.348572 WOS:000080862400003.

26. Arsic I, Marina N, Thiran J-P, editors. Impact of sample sizes on information theoretic measures for audio-visual signal processing. European Signal Processing Conference; 2005; Antalya, Turkey: IEEE.

27. Parmehr EG, Fraser CS, Zhang C, Leach J. An Effective Histogram Binning for Mutual Information Based Registration of Optical Imagery and 3d Lidar Data. Ieee Image Proc. 2013:1286–90. WOS:000351597601080.

28. Suderman R, Bachman JA, Smith A, Sorger PK, Deeds EJ. Fundamental trade-offs between information flow in single cells and cellular populations. Proc Natl Acad Sci U S A. 2017; 114(22):5755–60. https://doi.org/10.1073/pnas.1615660114 PMID: 28500273; PubMed Central PMCID: PMC5465904.

29. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. Nat Genet. 2004; 36 (5):492–6. https://doi.org/10.1038/ng1340 WOS:000221183000022. PMID: 15107850

30. Steinbock C, Biham O, Katzav E. Distribution of shortest path lengths in a class of node duplication network models. Physical Review E. 2017; 96(3). ARTN 032301 https://doi.org/10.1103/PhysRevE.96.032301 WOS:000408829500003. PMID: 29347025

31. Schaerli Y, Jimenez A, Duarte JM, Mihajlovic L, Renggli J, Isalan M, et al. Synthetic circuits reveal how mechanisms of gene regulatory networks constrain evolution. Mol Syst Biol. 2018; 14(9). ARTN e8102 https://doi.org/10.15252/msb.20178102 WOS:000445624900004. PMID: 30201776

32. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006; 34:D535–D9. https://doi.org/10.1093/nar/gkj109 WOS:000239307700116. PMID: 16381927