Multiscale Modeling of Information Conveyed by Gene-Regulatory Signaling

Michael Mayo[*]
Environmental Laboratory, US Army Engineer
Research and Development Center
Vicksburg, MS 39180
Michael.L.Mayo@usace.army.mil

Kevin Pilkiewicz
Environmental Laboratory, US Army Engineer
Research and Development Center
Vicksburg, MS 39180
Kevin.R.Pilkiewicz@usace.army.mil

ABSTRACT

Cells leverage signaling molecules to carry information about the cellular state to receptors that regulate protein synthesis in order to suit the cell's dynamically evolving needs. This regulation remains efficient and robust, despite that substantial stochasticity pervades the sub-cellular environment. In electronic and wireless signaling systems, the mutual information quantifies the extent to which information in a signal can be received across a communications channel. Applying this same metric to gene-regulatory interactions can better clarify how these biological signaling systems mitigate environmental noise. In this paper we study the information-transmission characteristics of a single gene-regulatory interaction by employing an exactly solvable master equation model for the production and degradation of individual proteins. This molecular-scale description is then coupled to a mass-action kinetics model of dynamic protein concentrations in a macroscopic sample of cells, enabling parameter values to be obtained by experiments performed using cell-based assays. We find that the mutual information depends monotonically on two parameters: one which characterizes stochastic variations in the concentration of signaling molecules, and the other the ratio of kinetic production to degradation rates of the regulated protein.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; G.3 [**Probability and Statistics**]: Stochastic processes; H.1.1 [**Information systems**]: Models and Principles— *Systems and Information Theory*

Keywords

Mutual information, Gene-regulatory networks, Master equation, Poisson distribution

1. INTRODUCTION

---
[*]Author to whom correspondence may be addressed.

Living cells leverage imperfect biological "hardware" to decode information regarding the cellular state that is conveyed through molecular signaling events. The cellular concentrations of signaling molecules–such as transcription factor proteins, which are synthesized from genes and regulated by other transcription factors–are subject to great variability due to both intrinsic and extrinsic noise [9], thereby contributing uncertainty to the individual cellular states of a clonal population. Understanding how biology overcomes this uncertainty to extract useful information is a topic of intense investigation [11].

One metric that captures information-transmission characteristics of a signal-response system is the mutual information. If s is a number of transcription factor proteins that encode a "source" state, and if r is the number of "response" proteins that sense and decode this state, then the mutual information, $I(r; s)$, is given by the following standard expression [2]:

$$I(r; s) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} p(r, s) \log_2 \frac{p(r, s)}{p(r)\, p(s)}. \qquad (1)$$

Eq. 1 can be interpreted as the information, measured in bits, conveyed to the response about the source state. Here, $I(r, s) = 0$ if s and r are uncorrelated: $p(r, s) = p(r)p(s)$. It is maximized when s uniquely determines r: $p(r|s) = 1$.

A challenge to evaluating Eq. 1 is estimating the joint probability, $p(r, s)$, which links the source to the response. Below we report a simple mathematical model of protein production stimulated by a transcriptional-regulatory interaction, and use it to understand characteristics of the information transmission via Eq. 1. This model requires only two independent parameters, and the joint probability between source and receptor molecules can be derived analytically in the long-time limit. We will further show that, despite describing a molecular-scale process, our model can be parameterized using data from macroscopic experiments.

2. MATHEMATICAL MODEL

We consider the relationship between transcription factor and response-protein concentrations within a single cell, wherein the response protein abundance is determined by a single, stimulatory transcriptional interaction. For simplicity, we ignore any diffusion-limited behavior on the response-protein yield, and treat transformative processes as chemical "reactions" by using the formalism of reaction-limited chemical

kinetics.

In our model, the source and response states are represented by the number of accumulated molecules s and r in a single cell, which we treat as random variables drawn, respectively, from the distributions $p(s)$ and $p(r)$. For simplicity we take $p(s)$ as the bounded uniform distribution:

$$p(s) = \begin{cases} \frac{1}{2\lambda+1} & \langle s \rangle - \lambda \leq s \leq \langle s \rangle + \lambda \\ 0 & \text{otherwise} \end{cases}, \qquad (2)$$

wherein $\langle s \rangle$ is the expected number of transcriptional activators per cell, and 2λ is the width of the distribution. Despite its simplicity, Eq. 2 exhibits qualitative features similar to those of a Gaussian distribution with mean $\langle s \rangle$ and variance $\sigma^2 = \lambda(\lambda+1)/3$.

2.1 Master equation

We propose a master equation to model the conditional probability distribution, $p(r|s)$, for each of the infinite but enumerable states indexed by response protein number, which should generally reflect the mechanisms of regulatory biology. Because proteins are created discretely, we quantify the probability for a single protein created in time Δt as $\omega(s)\Delta t$, which subsumes the protein-creation kinetics. We further assume that ω is time-independent (see below) and that Δt is chosen such that $\omega\Delta t \leq 1$, so that only one protein may be produced per time step. We similarly define $\nu\Delta t \leq 1$ to be the probability that a single protein is destroyed in one time step (either degraded by catabolic pathways or expelled from the cell by clearance mechanisms). Each response protein has an equal chance to degrade in a given time step, so whereas the ladder of "r"-states can only be climbed one rung at a time, it can be descended by any number of steps; however, the probability of descending n steps decreases with increasing n and decreasing Δt. These considerations lead to the following difference equation:

$$p(r|s; t + \Delta t) - p(r|s; t)$$
$$= \omega(s)\Delta t p(r-1|s; t) - \omega(s)\Delta t p(r|s; t)$$
$$+ \lim_{N \to \infty} \sum_{i=1}^{N} \binom{r+i}{i, r} (\nu\Delta t)^i (1 - \nu\Delta t)^r p(r+i|s; t)$$
$$- \sum_{i=1}^{r} \binom{r}{i, r-i} (\nu\Delta t)^i (1 - \nu\Delta t)^{r-i} p(r|s; t). \quad (3)$$

The first two terms on the right hand side of Eq. 3 account for gain and loss of a single protein molecule from a creation event; the third term accounts for a gain in probability due to an equivalent loss from states with more molecules; and the fourth term accounts for a loss of probability to states with fewer molecules driven by destruction events.

Although one advantage of Eq. 3 lies in its linearity, it reflects a large coupled system of difference equations, and is therefore difficult to solve exactly. To make the problem tractable, it is possible to choose Δt small enough that approximately no two molecules are destroyed simultaneously, which restricts transitions to the adjacent states. Expanding the destruction probabilities of Eq. 3 in a Taylor series
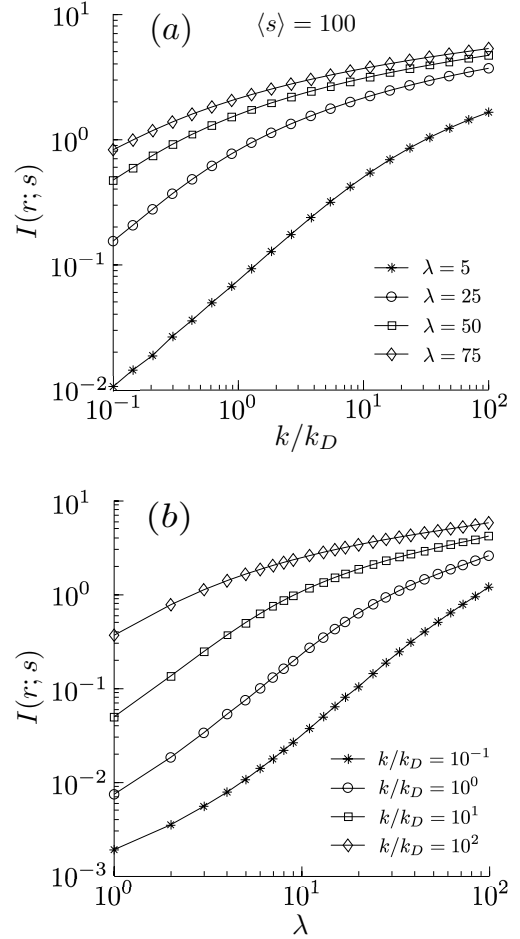

Figure 1: **Mutual information (measured in bits) plotted against (a) parameter values of the transcription-translation kinetics, k/k_D and (b) the width of the input distribution, λ.**

about $\Delta t = 0$, we find, in the limit $\Delta t \to 0$:

$$\frac{d}{dt} p(r|s; t) = \omega(s)p(r-1|s; t) - \omega(s)p(r|s; t)$$
$$+ \nu(r+1)p(r+1|s; t) - \nu r p(r|s; t). \quad (4)$$

The steady-state condition is $\lim_{t\to\infty} dp(r|s; t)/dt = 0$, and applying it to Eq. 4 yields:

$$\left(r + \frac{\omega(s)}{\nu}\right) p(r|s) = \frac{\omega(s)}{\nu} p(r-1|s) + (r+1)p(r+1|s).$$
$$(5)$$

This equation can be solved exactly to give a Poisson distribution:

$$p(r|s) = \frac{1}{r!}\left(\frac{\omega(s)}{\nu}\right)^r e^{-\omega(s)/\nu}, \qquad (6)$$

which depends on only a single dimensionless parameter, ω/ν, that contains the s-dependence of the protein kinetics.

2.2 Parameter values

Protein production kinetics can be related to the state-transition probabilities ω and ν by inspection with experimental re-

sults. Without losing any generality, consider an experiment wherein the mean protein concentration, $\bar{R}(t)$, is measured using a fluorescent reporter protein within a clonal population of prokaryotic cells (e.g., *E. coli* bacteria). The activity of these proteins changes in response to an up-regulatory interaction with a single transcription factor species present in concentration $\bar{S}(t)$, measured across the cell population. If $\bar{S}(0) = 0$ and $\bar{S}(t > 0) = S_0$ is constant–a good approximation to a bistable switch [3], then $\bar{R}(t)$ can be modeled dynamically with first-order reaction-limited chemical kinetics:

$$\frac{d}{dt}\bar{R}(t) = kS_0 - k_D\bar{R}(t). \tag{7}$$

Here, $1/k$ is the characteristic time to create a protein, and $\ln 2/k_D$ is the protein half-life. Equation 7 can be expected to hold if transcription achieves an mRNA steady-state much faster than the translational kinetics–a condition that holds approximately for some bacterial proteins [4]. The steady-state solution of Eq. 7 is given by $\lim_{t\to\infty}\bar{R}(t) = (k/k_D)S_0$, and the mean number of response proteins per cell, $\bar{r}$, is:

$$\bar{r} = (k/k_D)s_0, \tag{8}$$

wherein $s_0 = S_0 V_{cell}$. Here, V_{cell} is the volume of a typical *E. coli* bacterium.

Alternatively, Eq. 6 can be used to calculate the expectation value of r on a per-cell basis: $\langle r \rangle = \sum_r rp(r|s_0) = \omega(s_0)/\nu$. Because $\langle r \rangle$ should be comparable with the value measured from experiment, we equate it with Eq. 8:

$$\frac{\omega(s_0)}{\nu} = \frac{k}{k_D}s_0. \tag{9}$$

This identification can be used with Eq. 6 to give, generally:

$$p(r|s) = \frac{1}{r!}\left(\frac{k}{k_D}s\right)^r e^{-(k/k_D)s}. \tag{10}$$

While Eq. 10 reflects mass-action kinetics, alternative kinetic models, such as the Hill equation, can be accommodated in a similar manner.

3.   RESULTS AND DISCUSSION
The mutual information can be evaluated using Eq. 1 with the joint probability distribution determined by Eqs. 2, and 10:

$$p(r,s) = \frac{1}{(2\lambda + 1)\,r!}\left(\frac{k}{k_D}s\right)^r e^{-(k/k_D)s}, \tag{11}$$

for $\langle s \rangle - \lambda \leq s \leq \langle s \rangle + \lambda$, and $p(r,s) = 0$ otherwise. Results for the mutual information are plotted in Fig. 1 for fixed $\langle s \rangle = 100$.

Generally, the mutual information increases monotonically with both the dimensionless parameter k/k_D (Fig. 1(a)), which is related to the variance in the conditional probability, σ_r^2, by $\sigma_r^2 = (k/k_D)s_0$ (the same as the mean in this case), and the parameter λ (Fig. 1(b)), which is related to the variance in the signal distribution, $\sigma_s^2 = \lambda(\lambda + 1)/3$. This should be expected, because more information about the source should relate with a larger information transmission, and, therefore, larger mutual information. In contrast with some continuous models of biological communication

"channels" [13, 14], the channel noise, σ_r^2, is fixed in our model by biochemistry, rather than being a tunable parameter.

Some experimental data exists to support that the number of proteins per cell is approximately Poisson- distributed [4], as predicted by our model. However, phenomena such as transcriptional and translational "bursting" can lead to larger fluctuations than expected from a purely Poisson process [4, 8] and will consequently affect the mutual information. Our model could be readily extended to accommodate this phenomenon by incorporating an exponential wait-time distribution into the creation probability ω, making it a dynamic quantity. Although models that incorporate bursting are rather complicated [5], they otherwise result in approximately Poisson-distributed probabilities. It is therefore likely that protein-bursting processes will not substantially affect the qualitative features of the mutual information observed from our model.

The mutual information, Eq. 1, measures the logarithm of the number of input states that can be resolved by a receiver from a noisy communication channel [2]. In the biological setting of our model, a mutual information of 1 bit can be interpreted as the threshold to resolve whether the transcriptional activator is in a "high" (ON) or "low" (OFF) concentration state by examining the protein response at steady state. Figure 1 illustrates that 1 bit can be reached by either manipulating the input signal to allow for a larger number of signaling molecules (e.g., by increasing λ), or by choosing a transcriptional/translational system biased toward protein production (i.e., larger k/k_D).

4.   CONCLUSIONS
We investigated the qualities of the mutual information between a number of (source) transcription factors and the number of associated (response) proteins whose activity is regulated by a single, stimulatory transcriptional interaction. We developed a master equation to estimate the joint probability between source and response proteins, and solved it exactly in the long-time limit to reveal that the response follows a Poisson distribution. Other master-equation based models [15, 5] exhibit this feature, whereas continuous models exhibit mostly Gaussian behavior [13, 14]. This model requires values for only three parameters: the average (constant) number of source molecules, $\langle s \rangle$; the width of the source distribution, 2λ; and the ratio of creation to annihilation kinetic parameters, k/k_D. The former two quantities parameterize the source distribution, while the latter quantifies the protein kinetics. While parameters of the source distribution could be measured by single-cell fluorescent labeling techniques (e.g., [10]), rate constants are typically accessible through curve-fitting.

Numerical evaluation of the mutual information shows that it rises monotonically with λ and k/k_D (Fig. 1), because an increase in these parameters accompanies a respective increase in the information capacity of the source or "channel." A forthcoming paper will demonstrate, using a continuum model, that these findings persist for a regulatory chain of arbitrary length and do not depend upon the assumption of mass-action kinetics if the system is sufficiently close to steady state. The present work complements previous stud-

ies of intracellular molecular-transport channels [1, 6, 12], which together may enable the use of synthetic biological methods to engineer signaling pathways in novel ways, or to inspire improved "design principles" for current communication technologies, such as wireless sensor networks.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] B. Atakan and O. Akan. On channel capacity and error compensation in molecular communication. In *Transactions on computational systems biology X*, pages 59–80. Springer Berlin Heidelberg, 2008.

[2] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2012.

[3] T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[4] I. Golding, J. Paulsson, S. Zawilski, and E. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[5] A. Mugler, A. Walczak, and C. Wiggins. Spectral solutions to stochastic models of gene expression with bursts and regulation. *Phys. Rev. E.*, 80(4):041921, 2009.

[6] T. Nakano and J. Liu. Design and analysis of molecular relay channels: An information theoretic approach. *IEEE T. Nanobiosci.*, 9(3):213–221, 2010.

[7] K. Pilkiewicz, M. Mayo, et al., 2016. in preparation.

[8] A. Raj and A. van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Ann. Rev. Biophys.*, 38:255–270, 2009.

[9] J. Raser and E. O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309:2010, 2005.

[10] N. Rosenfeld, J. Young, U. Alon, P. Swain, and M. Elowitz. Gene regulation at the single-cell level. *Science*, 307:1962–1965, 2005.

[11] A. Sanchez, S. Choubey, and J. Kondev. Regulation of noise in gene expression. *Ann. Rev. Biophys.*, 42(1):469–491, 2013.

[12] K. Srinivas, A. Eckford, and R. Adve. Molecular communication in fluid media: The additive inverse gaussian noise channel. *IEEE T. Inform. Theory*, 58(7):4678–4692, 2012.

[13] G. Tkacik, C. C. Jr, and W. Bialek. Information capacity of genetic regulatory elements. *Phys. Rev. E*, 78(1):011910, 2008.

[14] G. Tkacik, C. C. Jr, and W. Bialek. Information flow and optimization in transcriptional regulation. *Proc. Natl. Acad. Sci. USA*, 105(34):12265–12270, 2008.

[15] A. Walczak, A. Mugler, and C. Wiggins. A stochastic spectral analysis of transcriptional regulatory cascades. *Proc. Natl. Acad. Sci. USA*, 106(16):6529–6534, 2009.