Mutual information and the fidelity of response of gene regulatory models

Omar P Tabbaa and C Jayaprakash

Department of Physics, The Ohio State University, 191 West Woodruff Avenue, Columbus, OH 43210-1117 USA

E-mail: tabbaa.2@osu.edu and jay@physics.osu.edu

Received 14 January 2014, revised 23 May 2014
Accepted for publication 10 June 2014
Published 22 July 2014

Abstract

We investigate cellular response to extracellular signals by using information theory techniques motivated by recent experiments. We present results for the steady state of the following gene regulatory models found in both prokaryotic and eukaryotic cells: a linear transcription-translation model and a positive or negative auto-regulatory model. We calculate both the information capacity and the mutual information exactly for simple models and approximately for the full model. We find that (1) small changes in mutual information can lead to potentially important changes in cellular response and (2) there are diminishing returns in the fidelity of response as the mutual information increases. We calculate the information capacity using Gillespie simulations of a model for the TNF-α-NF-κB network and find good agreement with the measured value for an experimental realization of this network. Our results provide a quantitative understanding of the differences in cellular response when comparing experimentally measured mutual information values of different gene regulatory models. Our calculations demonstrate that Gillespie simulations can be used to compute the mutual information of more complex gene regulatory models, providing a potentially useful tool in synthetic biology.

Keywords: gene regulatory models, stochastic modeling, information transmission

(Some figures may appear in colour only in the online journal)

1. Introduction

The viability of cells living in a complex environment depends on their ability to respond to a multiplicity of external signals by altering gene expression and thereby controlling concentrations of important proteins. Cells detect extracellular signals through a variety of receptors that are typically found on the cell surface or in the cytoplasm (e.g., the cytoplasmic receptor RIG-I); in turn, the receptors activate signal transduction pathways (a complex network of proteins) that transmit the information contained in these signals (typically as activated transcription factors) to the nucleus [1]. The cell then uses gene regulatory networks to process the information from the signals, resulting in changes in protein copy numbers, thus mounting a response to the environmental stimulus [2]. The signal transduction often occurs on fast time scales, as compared with those of gene regulatory networks. We focus on the gene regulatory network and view it as an input/output (I/O) system where the input is an external signal that is manifested as changes in the transcription factors and where the output is the copy number of proteins. The unavoidable stochastic fluctuations in biochemical reactions underlying gene networks can lead to signals that result in overlapping protein distributions, which limit the ability of cells to distinguish between these signals, thus restricting the fidelity of information transfer.

Information theory [3] provides the natural theoretical framework to analyze such systems and has been applied in...
diverse areas [4–7]. For example, in biology it has been used to study chemotaxis, vision, evolution, and neural networks. Information theory has also been applied to illuminate and quantitatively characterize some aspects of the transmission of information through gene regulatory networks in cellular systems [7–12].

Recently experimental measurements of mutual information were carried out for the bicoid gene in Drosophila [9] and for the transcription factor NF-κ B (essential in anti-viral immune response) in response to TNF-α (an activator of inflammatory response) [10]. The latter experiment provided measurements of the information capacity [13] between an extracellular signal and a protein characteristic of cellular response. Their results showed that the differences in the experimentally measured capacity of the gene regulatory networks are on average only fractions of bits (0.38 bit). This is surprising because many of these gene regulatory networks are expected to process complex outcomes. We theoretically address the question of whether such fractional bit differences in the mutual information of gene regulatory networks can have significant consequences in the ability of a population of cells to respond to extracellular signals.

We study the exact steady-state response of simple models of genetic regulatory networks to external signals that affect the transcription rate. We investigate models of transcription and translation, which include the effects of bursting and positive or negative auto-regulation. We quantify the efficacy of the information transfer by two measures: the fraction of cells that respond incorrectly to binary signals and how well the protein distribution across a population of cells reflects a continuous signal distribution to which it is exposed (quantified by the Kullback–Leibler (KL) divergence [13]). For simple gene regulatory models, we compute both the information capacity and the mutual information between a uniform input signal and the resultant protein distribution by using exact solutions and numerical simulations without a small-noise approximation that is often employed. We find that in these models, modest differences in the mutual information can imply substantial differences in the measures we employ to characterize the fidelity of cellular response. We explore the relative advantages of positive, negative, and no feedback in gene regulatory models from the perspective of information transfer. We highlight two results: (1) when small changes in mutual information can be potentially important and (2) when there are diminishing returns in the efficacy of cellular response as the mutual information is increased.

2. Methods

We model populations of cells that respond to an extracellular signal through the production of a protein whose copy number is a measure of that response. We assume that the signal modifies the rate of transcription of the mRNA that codes for the protein. We report results in the steady state for a variety of simple gene regulatory models, including the linear transcription-translation model and positive or negative auto-regulatory models defined in table 1 (see also figure 1(A)). In the linear transcription-translation model, the mRNA (m) is transcribed from the gene and is degraded at a rate proportional to m (it follows a birth–death process) and the protein (p) is translated from the mRNA and is degraded at a rate proportional to p. It has been shown that many genes in prokaryotes and eukaryotes are transcribed with bursting kinetics (i.e., periods of high transcription rates interspersed with silent or low transcription intervals) [14]. We model a gene that can switch between two states with high and low transcription rates. The switching rate is determined by the binding of the same protein to the promoter site of the gene. We study both positive and negative auto-regulatory cases by using a reduced model, for which we present exact calculations of the mutual information, and the complete model, where our results are obtained by simulations. In all the models, the extracellular signal is assumed to modify one of the rate constants in the network linearly (with the exceptions of a simplified model of the Gal-galactose network and a model of the TNF-α-NF-κ B network [15]; see tables S1 and S2).

We present results for our calculation of the mutual information (MI) between the extracellular input signal (s) and the output p defined by

\[ MI(s; p) = \sum_p \sum_s P(p, s) \log \left( \frac{P(p, s)}{Q(s)R(p)} \right) \] (1)

where \( P(p, s) \) is the joint probability distribution, \( Q(s) \) is the distribution of the input s, \( P(p|s) \) is the distribution of the output p conditioned on a specific value s of the input, and \( R(p) = \sum_p P(p, s) \) is the marginal distribution of the output. In our calculations, we calculate the information capacity (\( C = \max_{p,s} MI(s; p) \)) using the Blahut–Arimoto algorithm [13, 16, 17] and the MI using an extracellular signal distribution that is assumed to be a uniform distribution unless stated otherwise. The information capacity is the maximum amount of information that can be transmitted by a communication channel and is thus an intrinsic property of the channel, in contrast with MI, which depends on the signal distribution. The information capacity has been calculated for many biological systems [8–13]. On the other hand, for a fixed signal range, the flat distribution has a larger entropy (information) than any other distribution; this implies that in the deterministic limit the MI value calculated with a flat signal distribution will be maximal. Our results demonstrate that the mutual information calculated with a uniform input distribution, which is computationally convenient, is a good proxy for the information capacity. Due to intracellular noise, the response of a population of cells to a single s is a distribution of protein numbers (in this case, \( P(p|s) \)), displayed in figure 1(B) (i). The marginal protein distribution (in this case, \( R(p) \)) characterizes the response of a population of cells exposed to a continuous distribution of signals (in this case, \( Q(s) \)), and this is depicted in figure 1(B) (ii).

We calculate MI by evaluating the exact \( P(p|s) \) for three different models. Recently a closed-form expression for the generating function of the joint probability distribution for m
and $p$ in the steady state of the linear transcription-translation model was found [18]. We devise and use a different algorithm from the one employed by Bokes et al that allows us to numerically evaluate the complicated analytic expression to determine the probabilities for a range of $m$ and $p$ far beyond the low values ($p < 20$) in Bokes et al. This permits us to determine MI accurately. We study a reduced version of the auto-regulatory models in table 1, in which the protein is produced in bursts occurring at exponentially separated intervals with a random burst size, obeying a geometric distribution. The birth probability depends on the number of proteins through a Hill function that incorporates positive or negative auto-regulation. The steady-state master equation is given by

$$
p_{s}[p + 1]P(p + 1) - pP(p) = k_p \sum_{r=1}^{\infty} G(r) P(p - r)c(p - r) - P(p) \sum_{r=1}^{\infty} G(r)c(r)
$$

where $G(x) = (e^{k_x} - 1)e^{k_x x}$ is the burst size distribution, $c(x) = k_L k_H^{x} + k_H k_L^{x}$ (or $c(x) = k_L k_H^{x} + k_H k_L^{x}$) for the positive (or negative) auto-regulatory model, and $K = c_I/c_f$ is the affinity of the protein for its own DNA promoter site. This is a discrete version of the model with a continuous protein number, $p$, used earlier by Friedman et al to fit experimental data. In the case of no auto-regulation, we find a closed-form solution for $P(p|s)$. When auto-regulation is included, we solve the master equation by a recursive procedure that allows us to obtain the probability distribution for the protein number conditioned on the affinity $K$ with great accuracy (see text of S1). Our discrete solution is valid for both high and low $p$ values, whereas Friedman et al’s continuum solution is valid only for large values. Using a uniform distribution for the affinity $K$ that reflects the signal distribution, we calculate the MI. For the full model (see table 1), we perform simulations using the Gillespie algorithm [19]. We calculate the MI both by directly using the protein distributions generated by Gillespie and by finding the best fits to the protein distributions generated. We find that both of these methods agree. All the calculations and simulations are performed using Mathematica [20] or C++ codes on a PC.

3. Results

We present our results for the response to binary and continuous input signal distributions.
Table 1. Table of reactions for all four gene regulatory models. $k_b$ is the transcription rate of the mRNA, $p_b$ is the translation rate of the protein, and $k_d$ and $p_d$ are the degradation rates of the mRNA and the protein respectively. In the presence of bursting, the forward rate at which the gene switches from the low transcription state with rate $k_e$ to the high transcription state with rate $k_{id}$ is denoted by $c_f$ and the backward rate is $c_b$.

| Linear transcription-translation model | Gal-galactose | Positive auto-regulatory model | Negative Auto-regulatory model |
|--------------------------------------|--------------|--------------------------------|-------------------------------|
| $\emptyset \xrightarrow{k_b} m$ | $G \xrightarrow{k_1} G_1$ | $D + np \xrightarrow{c_f} D^n$ | $D^* \xrightarrow{c_f} np + D$ |
| $m \xrightarrow{\mu} m - \mu + p$ | $G_1 \xrightarrow{c_f} G_2$ | $D^* \xrightarrow{k_{id}} D + m$ | $D^* \xrightarrow{k_{id}} np + D + m$ |
| $p \xrightarrow{\mu_d} \emptyset$ | $G_2 \xrightarrow{c_f} G_1 + m$ | $m \xrightarrow{k_{id}} \emptyset$ | $m \xrightarrow{k_{id}} \emptyset$ |
|                                   | $m \xrightarrow{\mu} m + p$ | $m \xrightarrow{c_f} m + p$ | $m \xrightarrow{c_f} m + p$ |
|                                   | $p \xrightarrow{\mu_d} \emptyset$ | $p \xrightarrow{c_f} \emptyset$ | $p \xrightarrow{c_f} \emptyset$ |

3.1. Response to binary signal distributions

Cells often respond to signals that can be approximated by a binary distribution. It has been found that cellular response can be threshold-dependent and may be responsible for the diverse array of effector activities of T-cells [21–26]. For example, in the experiments of Cheong et al., gene regulatory networks exhibit a capacity of slightly less than 1 bit and can be threshold-dependent and may be responsible for the signal transduction (usually a fast process compared with biochemical reactions, the protein distribution can have numbers both above and below the threshold. This leads to an error proportional to the fraction of cells that respond below the threshold. In a natural cellular system, the fractional error might need to be below some limit for its viability. A limit on the error could be one of the design criteria in synthetic biology.

We consider a population of cells that are characterized by a gene regulatory model in which one of the parameters varies with the input signal. This is a simplified representation of signal transduction (usually a fast process compared with transcription and translation) that allows us to focus on the fidelity of the gene regulatory model. In the linear transcription-translation model, we vary the transcription rate linearly with the signal strength $s$. For the auto-regulatory models, we compute the mean protein number $\mu(s)$ as a function of the signal strength $s$. The dynamic range of the signal values, which the mean protein number is sensitive, and we define it as the range of signal values where $\frac{d\mu}{ds}$ is greater than or equal to 1/5 the largest value of $\frac{d\mu}{ds}$. In the regulatory models, we assume a flat distribution for the affinity of the protein for its promoter site across the dynamic range.

We consider the case where the cell has to respond to a fixed range of signal values $[s_L, s_H]$. Let a fraction of a population of cells, $f_L$, be exposed to a signal of value $s_L$, and the rest ($f_H = 1 - f_L$) to the value $s_H$. The two signals produce a distribution of proteins whose mean values are denoted by $\mu(s_L)$ and $\mu(s_H)$ respectively. We average the probabilities that cells subjected to the low or high values of the signal respond incorrectly with protein numbers respectively larger or smaller than the threshold value to define the fractional error,

$$f_{\text{error}} = f_L \left( \sum_{p=0}^{p_B} P(p|s_H) \right) + f_H \left( \sum_{p=0}^{p_H} P(p|s_L) \right)$$

where the threshold protein number is defined by $p_b = e^{\mu(s_{\text{th}})}$. We illustrate our results with this definition of the error; we check that other reasonable choices of the threshold yield similar results (figure S1).

We study gene regulatory networks that have $MI$ values less than 1 bit and compute the fractional error as previously defined. We vary the $MI$ of the system by varying the translation and degradation rate constants for the protein, keeping the curve $\mu(s)$ unchanged to better than 5 percent. Before discussing the results, we make a few pertinent comments. On a theoretical note, we observe that

$$P(p|m, k_s(s)) \neq P(p|k_s(s))$$

even for the linear transcription-translation model, as can be seen from the exact solution or, more simply, from a calculation of the variance of $P(p|m, k_s(s))$. Thus, the data processing inequality [13] does not hold in this case (indeed, it is violated in our calculations) and $MI(s; m)$ is not an upper bound for $MI(s; p)$. Therefore, $MI(s; m)$ is not a useful quantity, and we present results for $MI(s; p)$ only. Comparison of the results for different models is aided by the observation that the protein distribution has the most noise and largest tail for the positive auto-regulatory model when conditioned on $s_L$, whereas the negative auto-regulatory
model exhibits the least noise and a bias for smaller protein numbers when conditioned on $s_H$.

We find that $MI$ calculated with a uniform input distribution, which is a significantly easier quantity to compute, is a good proxy for $C$. The panels in figure 2 show $f_{error}$ as a function of $MI$ and $C$ for the linear transcription-translation model and that both of the measures provide a quantitative measure of the expected monotonic decrease. We choose parameters that were used to fit experimental data in *E. coli* or yeast [27, 28]; we have computed $f_{error}$ versus $MI$ for other sets of parameters, including those for human dendritic cells [29] and other gene regulatory models (see figure S2). We perform numerical simulations for both the full positive and negative auto-regulatory models with parameters used to fit experiments [30] and confirm that the qualitative results for the reduced model are reproduced. Because $MI$ is a good proxy for $C$, we mainly discuss $MI$ throughout the remainder of the paper.

Assuming that the different populations of cells (which have gene regulatory networks with differing $MI$ values) are all in the same static extracellular environment, our quantitative calculations show that a decrease in $MI$ as small as 0.2 bit can result in changes of more than twofold in $f_{error}$ from 10 percent to 25 percent, as seen in figure 2. This can also be seen by plotting the probability weight in the distributions of proteins that yield the wrong response, i.e., higher than the threshold response for the low signal and vice versa (see figure S3). This reinforces our conclusion that small changes in $MI$ can lead to potentially important changes in the response.

Figure 2 demonstrates that as $MI$ increases toward 1 bit, there are diminishing returns in the reduction of $f_{error}$ for the linear transcription-translation model. We find that this feature of diminishing returns in $f_{error}$ as $MI$ approaches 1 bit is also found in more complicated gene regulatory models with auto-regulation (see figure S2). A plausible explanation is that in the models we have studied (as in experimental measurements), the protein distributions often have exponential tails. To distinguish reasonably well between low and high signal values, the tails of the distributions given $s = s_L$ and $s = s_H$ should have modest overlap. Once the overlap is small, decreasing the exponential tails of the distributions any further leads to insubstantial changes in the overlap and correspondingly in the mutual information. Thus, we expect that there will be diminishing returns in $f_{error}$ as $MI$ approaches 1 bit when the protein distributions have exponentially (and faster than exponentially) decaying tails. The occurrence of diminishing returns emphasizes the fact that for populations of cells, both in nature and in synthetic biology, maximizing the mutual information may not be an optimal strategy. Indeed, increasing $MI$ can be expected to incur higher metabolic costs associated with transcription and translation. The form of the $f_{error}$ versus $MI$ curve displayed in figure 2 would suggest that intermediate values of $MI$ may be optimal because they can achieve substantial reduction in the error rate without hitting the barrier of diminishing returns.

The information capacity ($C$) of the TNF-$\alpha$-NF-$\kappa$-$B$ network was experimentally measured by Cheong et al to understand the transmission of information from an extracellular signal (TNF-$\alpha$) to multiple transcription factors (NF-$\kappa$-$B$ and ATF-2) that are activated by the extracellular signal. We calculated both $MI$ and $C$ (to compare with the experiment) using Gillespie simulations of a model of this network (assuming a linear relationship between TNF-$\alpha$ and IKK and a parameter set based on the model of Hoffmann et al [15] and Cheong et al’s experimental results). (See table S2 for more details.) Our *in silico* calculations yielded a $C$ of 0.9 bit, which is in good agreement with the experimentally measured value of 0.92 bit, and an $MI$ of 0.6 bit. From figure 2 (and figure S2), we see diminishing returns occur around an $MI$ of 0.6 bit, which is the value of $MI$ exhibited by the TNF-$\alpha$-NF-$\kappa$-$B$ network, providing further support for our claim that intermediate values of $MI$ may be optimal for biological systems.

We see from figure 2 that there is an asymmetry in $f_{error}$ as $f_H$ increases for a fixed $MI$, $f_{error}$ also increases. As the fraction of cells exposed to $s_L$ is changed from 0 to 1, the variation in $f_{error}$ is smallest in a positively auto-regulated gene (despite higher levels of noise), whereas for the

![Figure 2](image-url)

*Figure 2.* A plot of $f_{error}$ (defined in equation (3)) as a function of both (A) $MI$ and (B) the information capacity, $C$, for the no auto-regulation gene regulatory model (see table S1, columns LTTM3, for parameters used). In each panel, the orange dots represent $f_{error}$ for $f_L = 1$, and the black dots represent $f_{error}$ for $f_L = f_H = 0.5$. The panels in figure 2 show $f_{error}$ as a function of $MI$ and $C$ for the linear transcription-translation model and that both of the measures provide a quantitative measure of the expected monotonic decrease. We choose parameters that were used to fit experimental data in *E. coli* or yeast [27, 28]; we have computed $f_{error}$ versus $MI$ for other sets of parameters, including those for human dendritic cells [29] and other gene regulatory models (see figure S2). We perform numerical simulations for both the full positive and negative auto-regulatory models with parameters used to fit experiments [30] and confirm that the qualitative results for the reduced model are reproduced. Because $MI$ is a good proxy for $C$, we mainly discuss $MI$ throughout the remainder of the paper.

Assuming that the different populations of cells (which have gene regulatory networks with differing $MI$ values) are all in the same static extracellular environment, our quantitative calculations show that a decrease in $MI$ as small as 0.2 bit can result in changes of more than twofold in $f_{error}$ from 10 percent to 25 percent, as seen in figure 2. This can also be seen by plotting the probability weight in the distributions of proteins that yield the wrong response, i.e., higher than the threshold response for the low signal and vice versa (see figure S3). This reinforces our conclusion that small changes in $MI$ can lead to potentially important changes in the response.

Figure 2 demonstrates that as $MI$ increases toward 1 bit, there are diminishing returns in the reduction of $f_{error}$ for the linear transcription-translation model. We find that this feature of diminishing returns in $f_{error}$ as $MI$ approaches 1 bit is also found in more complicated gene regulatory models with auto-regulation (see figure S2). A plausible explanation is that in the models we have studied (as in experimental measurements), the protein distributions often have exponential tails. To distinguish reasonably well between low and high signal values, the tails of the distributions given $s = s_L$ and $s = s_H$ should have modest overlap. Once the overlap is small, decreasing the exponential tails of the distributions any further leads to insubstantial changes in the overlap and correspondingly in the mutual information. Thus, we expect that there will be diminishing returns in $f_{error}$ as $MI$ approaches 1 bit when the protein distributions have exponentially (and faster than exponentially) decaying tails. The occurrence of diminishing returns emphasizes the fact that for populations of cells, both in nature and in synthetic biology, maximizing the mutual information may not be an optimal strategy. Indeed, increasing $MI$ can be expected to incur higher metabolic costs associated with transcription and translation. The form of the $f_{error}$ versus $MI$ curve displayed in figure 2 would suggest that intermediate values of $MI$ may be optimal because they can achieve substantial reduction in the error rate without hitting the barrier of diminishing returns.

The information capacity ($C$) of the TNF-$\alpha$-NF-$\kappa$-$B$ network was experimentally measured by Cheong et al to understand the transmission of information from an extracellular signal (TNF-$\alpha$) to multiple transcription factors (NF-$\kappa$-$B$ and ATF-2) that are activated by the extracellular signal. We calculated both $MI$ and $C$ (to compare with the experiment) using Gillespie simulations of a model of this network (assuming a linear relationship between TNF-$\alpha$ and IKK and a parameter set based on the model of Hoffmann et al [15] and Cheong et al’s experimental results). (See table S2 for more details.) Our *in silico* calculations yielded a $C$ of 0.9 bit, which is in good agreement with the experimentally measured value of 0.92 bit, and an $MI$ of 0.6 bit. From figure 2 (and figure S2), we see diminishing returns occur around an $MI$ of 0.6 bit, which is the value of $MI$ exhibited by the TNF-$\alpha$-NF-$\kappa$-$B$ network, providing further support for our claim that intermediate values of $MI$ may be optimal for biological systems.

We see from figure 2 that there is an asymmetry in $f_{error}$ as $f_H$ increases for a fixed $MI$, $f_{error}$ also increases. As the fraction of cells exposed to $s_L$ is changed from 0 to 1, the variation in $f_{error}$ is smallest in a positively auto-regulated gene (despite higher levels of noise), whereas for the
negatively auto-regulated gene, it is constant for a range of MI values (see figure S2). This would indicate that the design criterion should include the probability of exposure to high versus low values of signals and should depend on the specific details of the system.

3.2. Response to continuous signal distributions

We present results for the response to continuous signal distributions of populations of cells with the gene regulatory models considered earlier. Cellular populations can be required to exhibit a proportionate response to extracellular stresses [22–24, 31]. Proportionate response may be important in the case of the innate immune system to prevent a cytokine storm [32, 33], although in this case communication between cells is important. Proportionate response to environmental signals at the population level can be measured by how well the protein distribution across the cell population reflects the input signal distribution.

We study how well the marginal protein distribution reflects the input signal distribution for different gene regulatory models and determine how well this is captured by MI. We provide both qualitative comparisons for input signals containing two or three peaks and a quantitative measure, the KL divergence. When the mean cellular response is graded decreasing (see figure S2). This would indicate that the design criterion should include the probability of exposure to high versus low values of signals and should depend on the specific details of the system.

Figure 3. (A) The marginal protein distribution \( R(p) \) for the GAL1 gene regulatory network in yeast exposed to a bimodal beta galactose distribution for the \( MI \) values of 0.97 bit (black; the native parameter set of the Gal-galactose network), 0.77 bit (red; increasing the translation and degradation rate constants by a factor of 2 from the native Gal-galactose network parameter set), and 1.21 bits (green; decreasing the translation and degradation rate constants by a factor of 2 from the native galactose network parameter set). (B) \( R(p) \) for a population of cells with the full positive auto-regulatory model exposed to a noisy three-peak signal distribution with an \( MI \) of 1.2 bits (red), 1.38 bits (black), and 1.56 bits (green; see inset) (see table S1, column PAR1).

study it as a function of \( MI \). The \( D_{KL} \) is defined as

\[
D_{KL} = \sum_p R_o(p) \log \left( \frac{R_o(p)}{R(p)} \right)
\]

where \( R(p) \) is the protein distribution for a fixed \( MI \). We point out that \( D_{KL} \) provides a comparison only on the interval of \( p \) values where \( R_o(p) \) is non-zero. For the biologically relevant parameters chosen in our studies, we find that the \( D_{KL} \) between a \( R_o(p) \) distribution resulting from a uniform signal distribution and a uniform protein number distribution across the same range of mean protein numbers is less than 0.1 bit. We also calculate the Kolmogorov–Smirnov statistic, which, in contrast with \( D_{KL} \), is a better measure of the distance when \( R(p) \) deviates significantly from \( R_o(p) \) outside the support of \( R_o(p) \); and we find that it has the same qualitative behavior as \( D_{KL} \) (see figure S4).

As expected, we find that \( D_{KL} \) is a monotonically decreasing function of \( MI \), rendering \( MI \) a good measure of how well the protein distribution of a population of cells reflects the signal distribution to which it is exposed. What is not obvious without calculations is that the trend of diminishing returns observed earlier is also seen in \( D_{KL} \) as the \( MI \) increases (see figure S4A). In figure S4D, we display how well \( R(p) \) reflects the deterministic response \( R_o(p) \) as the mutual information is varied, providing a qualitative picture of the improvement in performance revealed by the KL divergence.

In the first panel of figure 3, we present results for a simple model developed for the Gal-galactose gene regulatory network of yeast in a 0 percent glucose environment with the parameters in the literature [28]. We plot the protein distribution across a population of cells for an input signal described by a (bimodal) beta distribution with peaks at the ends. In this model, the signal (percent galactose in a 0 percent glucose environment) modifies both the forward and
reverse rates of transition of the Gal1 gene from an inactive state to an inactive intermediate state (see table S1). We calculated the MI between the protein and the signal using a flat signal strength distribution over a range of signal values set by the dynamic range. For the native parameter set, we obtained an MI value of 0.97 bit (D_{KL} = 5.2). Tuning the MI to be 0.2 bit lower (D_{KL} = 5.7) with the mean fixed degraded the resolution significantly (see the inset in figure 3(A)), whereas increasing the MI by 0.2 bit (D_{KL} = 5.2), on the other hand, only slightly increased the performance. This agrees with our earlier observation of the asymmetric effect of fractional bit changes in MI and diminishing returns.

In the next panel, we present results for a noisy three-peak input signal distribution: we plot the protein distribution R(p) for a population of cells with the full positive auto-regulatory model. The results from the calculations, performed using a Gillespie simulation, demonstrate that an increase in the mutual information from 1.2 bits (red) to 1.38 bits (black) shows a considerably improved response, whereas an increase in the mutual information from 1.38 bits (black) to 1.56 bits (red) shows diminishing returns in the improvement of response. Note that the transition from substantial gains to diminishing returns in the fidelity of response as MI increases occurs when the mean protein lifetime becomes much larger (approximately five times larger) than the mean time between transcriptional bursts. However, the value of the ratio of the mean protein lifetime to the mean time between bursts where this transition occurs must be calculated. We obtain similar results for different sets of parameters that fit experiments for eukaryotic and prokaryotic gene networks. We confine our attention to parameter values for which there is no bi-stability in the positive or negative auto-regulatory network because a graded mean response at the cellular level is required for a proportionate response. When bi-stability does arise, the cellular response becomes switch like and the deterministic protein distribution is not a good representation of the input signal distribution.

In figure 4, we overlay the results from three reduced models: (1) with no auto-regulation, (2) with negative auto-regulation, and (3) with positive auto-regulation. The protein birth probability in these models depends on only the extracellular signal (on both the extracellular signal and the protein number) through a Hill function for the no (positive or negative) auto-regulatory model. From the first panel, it is clear that the model with no auto-regulation outperforms both the positive and negative auto-regulatory models and that as MI increases, the positive auto-regulatory model performs the worst. This is surprising because one might have expected MI to be more sensitive to either the dynamic range or the noise, but from panels B and C it is clear that even though positive (negative) auto-regulation models have the largest (smallest) dynamic range and the most (least) noisy protein distributions, neither positive nor negative are the best performing gene regulatory networks. This makes a direct comparison of the relative merits of the different circuits difficult, as the mutual information is a complicated function of the noise, the dynamic range, and nonlinearities of the mean protein number as a function of the signal strength in the models.

4. Discussion

In this paper, we have calculated both the channel capacity and the mutual information between a uniformly distributed signal and the resultant protein distribution in simple gene regulatory models. In nature, populations of cells have to respond to environmental stressors—for example, to extracellular signals during embryonic development [34, 35]. In some systems, a proportionate response of a population of cells to an external distribution of inputs may be needed, as in the case of the innate immune system, where an overreaction can lead to a cytokine storm with potentially serious consequences [33]. Our study of the quantitative connection between MI and the fraction of cells that respond incorrectly to binary signals and how well the protein distribution across a population of cells reflects a continuous input signal can be useful in understanding interactions between cells and their environment. Our demonstration that fractional bit differences in mutual information can result in substantial differences in the capabilities of gene networks can allow for a qualitative interpretation of experimental data. For example, although the measured values of the EGF-Erk (0.6 bit) and TNF-ATF-2 (0.85 bit) networks have a difference in information capacities of only 0.25 bit, our results imply that the fraction of cells of EGF-Erk responding incorrectly can be larger than those of TNF-ATF-2 by greater than twofold. We varied the mutual information of simple models by altering the translation and degradation rates of the response protein, keeping the mean protein fixed in our study. This is an accessible way for tuning mutual information in synthetic biology where the engineered genetic circuits’ noise can be tuned while keeping its mean fixed [35]. Thus, calculations of mutual information may prove helpful in designing circuits that respond with required fidelity to external signals.

We caution that our interpretation of the connection of mutual information of gene regulatory models to the effectiveness of their response to a continuous signal distribution holds only when the response at the cellular level to extracellular signals is graded. However, this type of response is common in nature for cells that are responding under stress, such as yeast in 0 percent glucose, which we used as an example in this work. An obvious limitation of our calculation is that we have studied gene regulatory networks in the steady state, and hence, the dynamic features of information processing are not included [36–38]. The cellular machinery consists of complex networks of motifs, and the effect on information processing of embedding simple genetic regulatory systems such as those studied here in large networks needs elucidation. The effect of intercellular communication that would be relevant, for example, in the study of the innate immune system is not well understood.

In conclusion, we have demonstrated that modest increases in the calculated mutual information of simple gene regulatory models can result in significant differences in their...
ability to encode input signal distributions. Although this occurs at lower values of mutual information, there is a saturation effect that leads to diminishing returns at higher values of mutual information. Given the metabolic costs of increasing mutual information, intermediate values would be the best compromise. Our results provide a way of using experimentally measurable differences in mutual information values to obtain a qualitative understanding of cell response to external signals. Our calculations demonstrate that Gillespie simulations can be used to calculate the mutual information of gene regulatory models, which may serve as a useful tool in synthetic biology.

Acknowledgments

We thank Aleya Dhanji for helpful discussions. This work was supported by the National Institute of Allergy and Infectious Diseases (contract HHSN 272201000054C).

References

[1] Ptashne M and Gann A 2002 Genes and Signals (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press)
[2] Wagner A 2000 Inferring life style from gene expression patterns Mol. Biol. Evol. 17 1958–87
[3] Shannon C E 1948 A mathematical theory of communication Bell. Syst. Tech. J. 27 379–423
[4] Quastler H 1953 Information Theory in Biology (Urbana, IL: University of Illinois Press)
[5] Yockey H P 1974 An application of information theory to the central dogma and the sequence hypothesis J. Theor. Biol. 46 369–406
[6] Gatlin L L 1972 Information Theory and the Living System (New York: Columbia University Press)
[7] Rhee A, Cheong R and Levchenko A 2012 The application of information theory to biochemical signaling systems Phys. Biol. 9 045011
[8] Mugler A, Tostevin F and ten Wolde P R 2012 Spatial partitioning improves the reliability of biochemical signaling Proc. Natl Acad. Sci. USA 110 5927–32
[9] Tkacik G, Callan C G Jr and Bialek W 2008 Information flow and optimization in transcriptional regulation Proc. Natl Acad. Sci. USA 105 12265–70
