Information capacity of genetic regulatory elements

Gašper Tkačik, Curtis G. Callan, Jr.* and William Bialek*
Joseph Henry Laboratories of Physics, Lewis–Sigler Institute for Integrative Genomics,
*and *Princeton Center for Theoretical Physics, Princeton University, Princeton, New Jersey 08544*

(Dated: February 5, 2008)

Changes in a cell's external or internal conditions are usually reflected in the concentrations of the relevant transcription factors. These proteins in turn modulate the expression levels of the genes under their control and sometimes need to perform non-trivial computations that integrate several inputs and affect multiple genes. At the same time, the activities of the regulated genes would fluctuate even if the inputs were held fixed, as a consequence of the intrinsic noise in the system, and such noise must fundamentally limit the reliability of any genetic computation. Here we use information theory to formalize the notion of information transmission in simple genetic regulatory elements in the presence of physically realistic noise sources. The dependence of this "channel capacity" on noise parameters, cooperativity and cost of making signaling molecules is explored systematically. We find that, at least in principle, capacities higher than one bit should be achievable and that consequently genetic regulation is not limited the use of binary, or "on-off", components.

PACS numbers: 87.16.Yc, 87.16.Xa, 89.70.+c

I. INTRODUCTION

Networks of interacting genes coordinate complex cellular processes, such as responding to stress, adapting the metabolism to a varying diet, maintaining the circadian cycle or producing an intricate spatial arrangement of differentiated cells during development [1, 2, 3, 4]. The success of such regulatory modules is at least partially characterized by their ability to produce reliable responses to repeated stimuli or changes in the environment over a wide dynamic range, and to perform the genetic computations reproducibly, either on a day-by-day or generation timescale. In doing so the regulatory elements are confronted by noise arising from physical processes that implement such genetic computations, and this noise ultimately traces its origins back to the fact that the state variables of the system are concentrations of chemicals and "computations" are really reactions between individual molecules, usually present at low copy numbers [5, 6].

It is useful to picture the regulatory module as a device that, given some input, computes an output, which in our case will be a set of expression levels of regulated genes. Sometimes the inputs to the module are easily identified, such as when they are the actual chemicals that a system detects and responds to, for example chemoattractant molecules, hormones or transcription factors (TFs). There are cases, however, when it is beneficial to think about the inputs on a more abstract level: in embryonic development we talk of "positional information" and think of the regulatory module as trying to produce a different gene expression footprint at each spatial location [7]; alternatively, circadian clocks generate distinguishable gene expression profiles corresponding to various phases of the day [1]. Regardless of whether we view the input as a physical concentration of some transcription factor or perhaps a position within the embryo, and whether the computation is complicated or as simple as an inversion produced by a repressor, we want to quantify its reliability in the presence of noise, and ask what the biological system can do to maximize this reliability.

If we make many observations of a genetic regulatory element in its natural conditions we are collecting samples drawn from a distribution $p(\mathcal{I},\mathcal{O})$, where $\mathcal{I}$ describes the state of the input and $\mathcal{O}$ the state of the output. Saying that the system is able to produce a reliable response $\mathcal{O}$ across the spectrum of naturally occurring input conditions $p(\mathcal{I})$ amounts to saying that the dependency – either linear or strongly non-linear – between the input and output is high, i.e. far from random. Shannon has shown how to associate a unique measure, the mutual information I, with the notion of dependency between two quantities drawn from a joint distribution [8, 9, 10]:

$$I(\mathcal{I};\mathcal{O}) = \iint d\mathcal{I}\, d\mathcal{O}\, p(\mathcal{I},\mathcal{O}) \log_2 \frac{p(\mathcal{I},\mathcal{O})}{p(\mathcal{I})p(\mathcal{O})}. \qquad (1)$$

The resulting quantity is a measure in bits and is essentially the logarithm of the number of states in the input that produce distinguishable outputs given the noise. A device that has one bit of capacity can be thought of as an "on-off" switch, two bits correspond to four distinguishable regulatory settings, and so on. Although the input is usually a continuous quantity, such as nutrient concentration or phase of the day, the noise present in the regulatory element corrupts the computation and does not allow the arbitrary resolution of a real-valued input to propagate to the output; instead, the mutual information tells us how precisely different inputs are distinguishable to the organism.

Experimental or theoretical characterization of the joint distribution, $p(\mathcal{I},\mathcal{O})$, for a regulatory module can be very difficult if the inputs and outputs live in a highdimensional space. We can proceed, nevertheless, by remembering that the building blocks of complex modules are much simpler, and finally must reduce to the point where a single gene is controlled by transcription factors that bind to its promoter region and tune the level of its expression. While taking a simple element out of its network will not be illuminating about how the network as a whole behaves in general – especially if there are feedback loops – there may be cases where the information flow is "bottlenecked" through a single gene, and its reliability will therefore limit that of the network. In addition, the analysis of a simple regulatory element will provide directions for taking on more complicated systems; see Ref [11] for a recent related analysis.

Our aim is therefore to understand the reliability of a simple genetic regulatory element, that is, of a single activator or repressor transcription factor controlling the expression level of its downstream gene. We will identify the concentration c of the transcription factor as the only input, $\mathcal{I} \equiv \{c\}$, and the expression level of the downstream gene g as the only relevant output, $\mathcal{O} \equiv \{g\}$. The regulatory element itself will be parametrized by input/output kernel, $p(g|c)$, i.e. the distribution (as opposed to a "deterministic" function $g = g(c)$ in case of a noiseless system) of possible outputs given that the input is fixed to some particular level c. For each such kernel, we will then compute the maximum amount of information, $I(c; g)$, that can be transmitted through it, and examine how this capacity depends on the properties of the kernel.

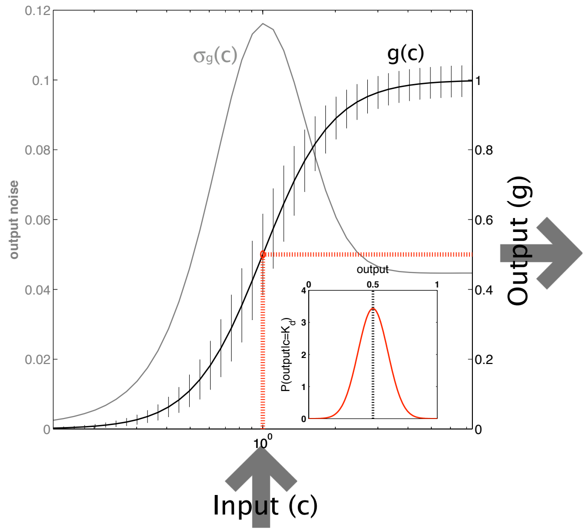

FIG. 1: (Color online) A schematic diagram of a simple regulatory element. Each input is mapped to a mean output according to the input/output relation (thick sigmoidal black line). Because the system is noisy, the output fluctuates about the mean. This noise is plotted in gray as a function of the input and shown in addition as error bars on the mean input/output relation. Inset shows the probability distribution of outputs at half saturation, $p(g|c = K_d)$ (red dotted lines); in this simple example we assume that the distribution is Gaussian and therefore fully characterized by its mean and variance.

II. MAXIMIZING INFORMATION TRANSMISSION

The input/output kernel of a simple regulatory element, $p(g|c)$, is determined by the biophysics of transcription factor-DNA interaction, transcription and translation. In contrast, the distribution of inputs, $p(c)$, that the cell uses during its "typical" lifetime, is free for the cell to change. The cell's transcription factor expression footprint is its representation of the environment and internal state, and the form of this representation can be the target of adaptation or evolutionary processes. Together, the input/output kernel and the distribution of inputs define the joint distribution, $p(c, g) = p(g|c)\,p(c)$, and consequently the mutual information of Eq (1) between the input and the output, $I(c; g)$.

Maximizing the information between the inputs and outputs, which corresponds to our notions of reliability in representation and computation, will therefore imply a specific matching between the given input/output kernel and the distribution of inputs, $p(c)$, that is being optimized. If one believes that a specific regulatory element has been tuned for maximal information transmission, then the optimal solution for the inputs, $p^*(c)$, and the resulting optimal distribution of outputs, $p^*(g) = \int dc\, p(g|c)p^*(c)$, become experimentally verifiable predictions. If, on the other hand, the system is not really maximizing information transmission, then the capacity achievable with a given kernel and its optimal input distribution, $I[p(g|c), p^*(c)]$, can still be regarded as a (hopefully revealing) upper bound on the true information transmission of the system.

During the past decades the measurements of regulatory elements have focused on recovering the mean response of a gene under the control of a transcription factor that had its activity modulated by experimentally adjustable levels of inducer or inhibitor molecules [12]. Typically, a sigmoidal response is observed with a single regulator, as in Fig 1, and more complicated regulatory "surfaces" are possible when there are two or more simultaneous inputs to the system [13, 14]. In our notation, these experiments measure the conditional average over the distribution of outputs, $\bar{g}(c) = \int dg\, g\, p(g|c)$. Developments in flow cytometry and single-cell microscopy enabled the experimenters to start tracking in time and across the population of cells the expression levels of fluorescent reporter genes and thus open a window into the behavior of fluctuations. Consequently, work exploring the noise in gene expression, or $\sigma_g^2(c) = \int dg\, (g - \bar{g})^2 p(g|c)$, has begun to accumulate, on both the experimental and biophysical modeling side [15, 16, 17, 18]. The efforts to further characterize and understand this noise were renewed by theoretical work by Swain and coworkers [19] that has shown how to separate intrinsic and extrinsic components of the noise, i.e. the noise due to the stochasticity of the observed regulatory process in a single cell, and the noise contribution that arises because typical experiments make many single-cell measurements and the internal chemical environments of these cells differ across the population.

A. Small noise approximation

We start by showing how the optimal distributions can be computed analytically if the input/output kernel is Gaussian and the noise is small, and proceed by presenting the exact numerical solution later. Let us assume then that the first and second moments of the conditional distribution are given, and write the input/output kernel as a set of Gaussian distributions $\mathcal{G}(g;\bar{g}(c),\sigma_g(c))$, or explicitly:

$$p(g|c)=\frac{1}{\sqrt{2\pi\sigma_g^2(c)}}\exp\left\{-\frac{[g-\bar{g}(c)]^2}{2\sigma_g^2(c)}\right\}, \qquad (2)$$

where both the mean response, $\bar{g}(c)$, and the noise, $\sigma_g(c)$, depend on the input, as illustrated in Fig 1.

We rewrite the mutual information between the input and the output of Eq (1) in the following way:

$$\begin{aligned} I(c;g) &= \int dc\, p(c)\int dg\, p(g|c)\log_2 p(g|c) - \\ &- \int dc\, p(c)\int dg\, p(g|c)\log_2 p(g). \end{aligned} \qquad (3)$$

The first term can be evaluated exactly for Gaussian distributions, $p(g|c)=\mathcal{G}(g;\bar{g}(c),\sigma_g(c))$. The integral over g is just the calculation of the (negative of the) entropy of the Gaussian, and the first term therefore evaluates to $-\langle S\left[\mathcal{G}(g;\bar{g},\sigma_g]\right\rangle_{p(c)}=-\frac{1}{2}\langle\log_2 2\pi e\sigma_g^2(c)\rangle_{p(c)}$.

In the second term of Eq (3), the integral over g can be viewed as calculating $\langle\log_2 p(g)\rangle$ under the distribution $p(g|c)$. For an arbitrary continuous function $f(g)$ we can expand the integrals with the Gaussian measure around the mean:

$$\begin{aligned} \langle f(g)\rangle_{\mathcal{G}(g;\bar{g},\sigma_g)} &= \int dg\,\mathcal{G}(g)f(\bar{g}) + \\ &+ \int dg\,\mathcal{G}(g)\frac{\partial f}{\partial g}\Big|_{\bar{g}}(g-\bar{g}) + \\ &+ \frac{1}{2}\int dg\,\mathcal{G}(g)\frac{\partial f^2}{\partial g^2}\Big|_{\bar{g}}(g-\bar{g})^2+\cdots \end{aligned} \qquad (4)$$

The first term of the expansion simply evaluates to $f(\bar{g})$. The series expansion would end at the first term if we were to take the small noise limit, $\lim_{\sigma_g\to 0}\mathcal{G}(g;\bar{g},\sigma_g)=\delta(g-\bar{g})$. The second term of the expansion is zero because of symmetry, and the third term evaluates to $\frac{1}{2}\sigma_g^2 f''(\bar{g})$. We apply the expansion of Eq (4) and compute the second term in the expression for the mutual information, Eq (3), with $f(g)=\log_2 p(g)$. Taking only the zeroth order of the expansion, we get

$$I(c;g)=-\int dc\, p(c)\left[\log_2\sqrt{2\pi e}\sigma_g(c)+\log_2 p(\bar{g}(c))\right]; \qquad (5)$$

we can rewrite the probability distributions in terms of $\bar{g}$, using $p(c)\,dc=p(\bar{g})\,d\bar{g}$. To maximize the information transmission we form the following Lagrangian and introduce the multiplier Λ that keeps the resulting distribution normalized:

$$\mathcal{L}[p(\bar{g})]=-\int d\bar{g}\,p(\bar{g})\log_2\left(\sqrt{2\pi e}\sigma_g(\bar{g})p(\bar{g})\right)-\Lambda\int d\bar{g}\,p(\bar{g}). \qquad (6)$$

The optimal solution is obtained by taking a variational derivative with respect to $p(\bar{g})$, $\frac{\delta\mathcal{L}[p(\bar{g})]}{\delta p(\bar{g})}=0$. The solution is

$$p^*(\bar{g})=\frac{1}{Z}\cdot\frac{1}{\sigma_g(\bar{g})}. \qquad (7)$$

By inserting the optimal solution, Eq (7), into the expression for mutual information, Eq (3), we get an explicit result for the capacity:

$$I_{\rm opt}(c;g)=\log_2\left[\frac{Z}{\sqrt{2\pi e}}\right], \qquad (8)$$

where Z is the normalization of the optimal solution in Eq (7):

$$Z=\int_0^1\frac{d\bar{g}}{\sigma_g(\bar{g})}. \qquad (9)$$

The optimization with respect to the distribution of inputs, $p(c)$, has led us to the result for the optimal distribution of *mean outputs*, Eq (7). We had to assume that the input/output kernel is Gaussian and that the noise is small, and we refer to this result as the small-noise approximation (SNA) for channel capacity. Note that in this approximation only the knowledge of the noise in the output as a function of mean output, $\sigma_g(\bar{g})$, matters for capacity computation and the direct dependence on the input c is irrelevant. This is important because the behavior of intrinsic noise as a function of the mean output is an experimentally accessible quantity [16]. Note also that for big enough noise the normalization constant Z will be small compared to $\sqrt{2\pi e}$, and the small-noise capacity approximation of Eq (8) will break down by predicting negative information values.

B. Large noise approximation

Simple regulatory elements usually have a monotonic, saturating input/output relation, as shown in Fig 1, and

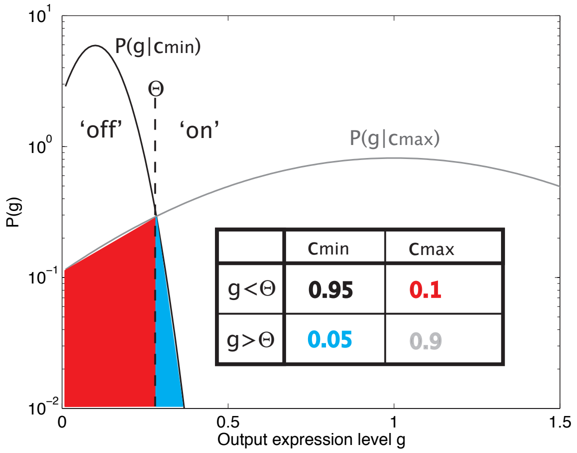

FIG. 2: (Color online) An illustration of the large noise approximation. We consider distributions of the output at minimal ($c_{\rm min}$) and full ($c_{\rm max}$) induction as trying to convey a single binary decision, and construct the corresponding encoding table (inset) by discretizing the output using the threshold Θ. The capacity of such an asymmetric binary channel is degraded from the theoretical maximum of 1 bit, because the distributions overlap (blue and red). For unclipped Gaussians the optimal threshold Θ is at the intersection of two alternative pdfs, but in general one searches for the optimal Θ that maximizes information in Eq (11).

(at least) a shot noise component whose variance scales with the mean. If the noise strength is increased, the information transmission must drop and, even with the optimally tuned input distribution, eventually yield only a bit or less of capacity. Intuitively, the best such a noisy system can do is to utilize only the lowest and highest achievable input concentrations, and ignore the continuous range in between. Thus, the mean responses will be as different as possible, and the noise at low expression will also be low because it scales with the mean. More formally, if only $\{c_{\rm min}, c_{\rm max}\}$ are used as inputs, then the result is either $p(g|c_{\rm min})$ or $p(g|c_{\rm max})$; the optimization of channel capacity reduces to finding $p(c_{\rm min})$, with $p(c_{\rm max}) = 1 - p(c_{\rm min})$. This problem can be solved by realizing that each of the two possible input concentrations produces their respective Gaussian output distributions, and by maximizing information by varying $p(c_{\rm min})$. Simplifying even further, we can threshold the outputs and allow g to take on only two values instead of a continuous range; then, each of the two possible inputs, "min" and "max", maps into two possible outputs, "on" and "off", and confusion in the channel arises because "min" input might be misunderstood as "on" output and vice versa with probabilities given by the output distribution overlaps, as shown schematically in Fig 2.

In the latter case we can use the analytic formula for the capacity of the binary asymmetric channel. If η is the probability of detecting an "off" output if "max" input was sent, and ξ is a probability of receiving an "off" output if "min" input was sent, and $H(\cdot)$ is a binary entropy function:

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p), \qquad (10)$$

then the capacity of such asymmetric channel is [20]:

$$I(c;g) = \frac{-\eta H(\xi) + \xi H(\eta)}{\eta - \xi} + \log_2(1 + 2^{\frac{H(\xi)-H(\eta)}{\eta-\xi}}). \qquad (11)$$

Because this approximation reduces the continuous distribution of outputs to only two choices, "on" or "off", it can underestimate the true channel capacity and is therefore a lower bound.

C. Exact solution

The information between the input and output in Eq (3) can be maximized numerically for any input/output kernel, $p(g|c)$, if the variables c and g are discretized, making the solution space that needs to be searched, $p(c_i)$, finite. One possibility is to use a gradient descent-based method and make sure that the solution procedure always stays within the domain boundaries $\sum_i p(c_i) = 1, p(c_j) \geq 0$ for every j. Alternatively, a procedure known as Blahut-Arimoto algorithm has been derived specifically for the purpose of finding optimal channel capacities [21]. Both methods yield consistent solutions, but we prefer to use the second one because of faster convergence and convenient inclusion of constraints on the cost of coding (see Appendix A for details).

One should be careful in interpreting the results of such naive optimization and worry about the artifacts introduced by discretization of input and output domains. After discretization, the formal optimal solution is no longer required to be smooth and could, in fact, be composed of a collection of Dirac-delta function spikes. On the other hand, the real, physical concentration c cannot be tuned with arbitrary precision in the cell; it is a result of noisy gene expression, and even if this noise source were removed, the *local* concentration at the binding site would still be subject to fluctuations caused by randomness in diffusive flux [22, 23]. The Blahut-Arimoto algorithm is completely agnostic as to which (physical) concentrations belong to which bins after concentration has been discretized, and so it could assign wildly different probabilities to concentration bins that differ in concentration by less than σ_c (i.e. the scale of local concentration fluctuations), making such a naive solution physically unrealizable. In Appendix A we describe how to properly use Blahut-Arimoto algorithm despite the difficulties induced by discretization.

III. A MODEL OF SIGNALS AND NOISE

If enough data were available, one could directly sample $p(g|c)$ and proceed by calculating the optimal solutions as described previously. Here we start, in contrast, by assuming a Gaussian model of Eq (2), in which the mean, $\bar{g}(c)$, and the output variance, $\sigma_g^2(c)$, are functions of the transcription factor concentration, c. Our goal for this section is to build an effective microscopic model of transcriptional regulation and gene expression, and therefore define both functions with a small number of biologically interpretable parameters. In the subsequent discussion we plan to vary those and thus systematically observe the changes in information capacity.

In the simplest picture, the interaction of the TF with the promoter site consists of binding with a (second order) rate constant k_+ and unbinding at a rate k_-. In a somewhat more complicated case where h TF molecules cooperatively activate the promoter, the analysis still remains simple as long as the favorable interaction energy between the TFs is sufficient to make only the fully occupied (and thus activated) and the empty (and thus inactivated) states of the promoter likely; this effective two-state system is once more describable with a single rate for switching off the promoter, k_-, and the corresponding activation rate has to be $\propto c^h$ (see Ref [23], in particular Appendix B). Generally, therefore, the equilibrium occupancy of the site will be:

$$n = \frac{c^h}{c^h + K_d^h}, \tag{12}$$

where the Hill coefficient, h, captures the effects of cooperative binding, and K_d is the equilibrium constant of binding. The mean expression level g is then:

$$g(c) = g_0 \bar{g} = g_0 \left\{ \begin{array}{cc} n & \text{activator} \\ 1-n & \text{repressor} \end{array} \right. , \tag{13}$$

where $\bar{g}$ has been normalized to vary between 0 and 1, and g_0 is the maximum expression level. In what follows we will assume the activator case, where $\bar{g} = n$, and present the result for the repressor at the end.

The fluctuations in occupancy have a (binomial) variance $\sigma_n^2 = n(1-n)$ and a correlation time $\tau_c = 1/(k_+c^h + k_-)$ [23]. If the expression level of the target gene is effectively determined by the average of the promoter site occupancy over some window of time $\tau_{\rm int}$, then the contribution to variance in the expression level due to the "on-off" promoter switching will be:

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \sigma_n^2 \frac{\tau_c}{\tau_{\rm int}} = \frac{n(1-n)}{(k_+c^h + k_-)\tau_{\rm int}} = \frac{n(1-n)^2}{k_-\tau_{\rm int}}, \tag{14}$$

where in the last step we use the fact that $k_+c^h(1-n) = k_-n$.

At low TF concentrations the arrival times of single transcription factor molecules to the binding site are random events. Recent measurements [24] seem to be consistent with the hypothesis that this variability in diffusive flux contributes an additional noise term [22, 23], similar to the Berg-Purcell limit to chemoattractant detection in chemotaxis. The noise in expression level due to fluctuations in the binding site occupancy, or the total *input* noise, is therefore a sum of this diffusive component (see Eq (11) of Ref [23]) and the switching component of Eq (14):

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\rm input} = \frac{n(1-n)^2}{k_-\tau_{\rm int}} + \frac{h^2(1-n)^2n^2}{\pi D a c \tau_{\rm int}}, \tag{15}$$

where D is the diffusion constant for the TF and a is the receptor site size, $a \sim 3$ nm for a typical binding site on the DNA.

To compute the information capacity in the small noise limit using the simple model developed so far we need the constant Z from Eq (9), which is defined as an integral over expression levels. As both input noise terms are proportional to $(1-\bar{g})^2$, the integral must take the form:

$$Z \propto \int_0^1 \frac{d\bar{g}}{(1-\bar{g})\,F(\bar{g})}, \tag{16}$$

where $F(\bar{g})$ is a function that approaches a constant as $\bar{g} \to 1$. Strangely, we see that this integral diverges near full induction ($\bar{g} = 1$), which means that the information capacity also diverges.

Naively we expect that modulations in transcription factor concentration are *not* especially effective at transmitting regulatory information once the relevant binding sites are close to complete occupancy. More quantitatively, the sensitivity of the site occupancy to changes in TF concentration, $\partial n/\partial c$, vanishes as $n \to 1$, and hence small changes in TF concentration will have vanishingly small effects. Our intuition breaks down, however, because in thinking only about the mean occupancy we forget that even very small changes in occupancy could be effective if the noise level is sufficiently small. As we approach complete saturation, the variance in occupancy decreases, and the correlation time of fluctuations becomes shorter and shorter; together these effects cause the standard deviation as seen through an averaging time $\tau_{\rm int}$ to decrease faster than $\partial n/\partial c$, and this mismatch is the origin of the divergence in information capacity. Of course the information capacity of a physical system can't really be infinite; there must be an extra source of noise (or reduced sensitivity) that becomes limiting as $n \to 1$.

The noise in Eq (15) captures only the *input* noise, i.e. the noise in the protein level caused by the fluctuations in the occupancy of the binding site. In contrast, the *output* noise arises even when the occupancy of the binding site is fixed (for example, at full induction), and originates in the stochasticity in transcription and translation. The simplest model postulates that when the activator binding site is occupied with fractional occupancy n, mRNA molecules are synthesized in a Poisson process at a rate R_e that generates $R_e\tau_e n$ mRNA molecules on average during the lifetime of a single mRNA molecule, τ_e. Every message is a template for the production of proteins, which is another Poisson process with rate R_g.

Parameter	Value	Description
α	$(1+b)/g_0$	Output noise strength
β	$h^2/\pi D a K_d \tau_{\rm int}$	Diffusion input noise strength
γ	$(k_-\tau_{\rm int})^{-1}$	Switching input noise strength
h		Cooperativity (Hill coefficient)

TABLE I: Gaussian noise model parameters. Note that if burst size $b \gg 1$, then the output noise is determined by the average number of mRNA molecules, $\alpha \sim (\langle \text{mRNA} \rangle)^{-1}$. Note further that if the on-rate is diffusion limited, i.e. $k_+ = 4\pi D a$, then both input noise magnitudes, β and γ, are proportional and decrease with increasing k_-, or alternatively, with increasing $K_d = k_-/k_+$.

If the integration time is larger than the lifetime of single mRNA molecules, $\tau_{\rm int} \gg \tau_e$, the mean number of proteins produced is $g = R_g \tau_{\rm int} R_e \tau_e n = g_0 n$, and the variance associated with both Poisson processes is [23]:

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\rm output} = \frac{1 + R_g \tau_e}{g_0} n, \tag{17}$$

where $b = R_g \tau_e$ is the burst size, or the number of proteins synthesized per mRNA.

We can finally put the results together by adding the input noise Eq (15) and the output noise Eq (17), and expressing both in terms of the normalized expression level $\bar{g}(c)$:

$$\begin{aligned} \left(\frac{\sigma_g}{g_0}\right)^2_{\rm act} &= \alpha\,\bar{g} + \\ &+ \beta(1-\bar{g})^{2+\frac{1}{h}}\bar{g}^{2-\frac{1}{h}} + \gamma\bar{g}(1-\bar{g})^2, \quad (18) \\ \left(\frac{\sigma_g}{g_0}\right)^2_{\rm rep} &= \alpha\,\bar{g} + \\ &+ \beta(1-\bar{g})^{2-\frac{1}{h}}\bar{g}^{2+\frac{1}{h}} + \gamma\bar{g}^2(1-\bar{g}), \quad (19) \end{aligned}$$

with the relevant parameters $\{\alpha, \beta, \gamma, h\}$ explained in Table I. Note that both repressor and activator cases differ only in the shape of the input noise contributions (especially for low cooperativity h). Note further that the output noise increases monotonically with mean expression $\bar{g}$, while the input noise peaks at the intermediate levels of expression. To make the examination of the parameter space in the next section feasible, we set $\gamma = 0$; models with switching noise instead of diffusive noise produce qualitatively similar results.

IV. RESULTS

A. Capacity of simple regulatory elements

Having at our disposal both a simple model of signals and noise and a numerical way of finding the optimal solutions given an arbitrary input/output kernel, we are now ready to examine the channel capacity as a function of the noise parameters from Table I. Our first result, shown in Fig 3, concerns the simplest case of an activator with no cooperativity, $h = 1$; for this case, the noise in Eq (18) simplifies to:

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \alpha\bar{g} + \beta(1-\bar{g})^3\bar{g}. \tag{20}$$

Here we have assumed that there are two relevant sources of noise, i.e. the output noise (which we parametrize by α and plot on the horizontal axis) and the input diffusion noise (parametrized by β, vertical axis). Each point of the noise plane in Fig 3a therefore represents a system characterized by a Gaussian noise model, Eq (2), with variance given by Eq (20) above.

As expected, the capacity increases most rapidly when the origin of the noise plane is approached approximately along its diagonal, whereas along each of the edges one of the two noise sources effectively disappears, leaving the system dominated by either output or input noise alone. We pick two illustrative examples, the blue and the red systems of Figs 3b and 3c, that have realistic noise parameters. The blue system has, apart for the decreased cooperativity ($h = 1$ instead of $h = 5$), the characteristics of the Bicoid-Hunchback regulatory element in *Drosophila melanogaster* [23, 25]; the red system is dominated by output noise with characteristics measured recently for about 40 yeast genes [27]. We would like to emphasize that both the small-noise approximation and the exact solution predict that these realistic systems are capable of transmitting more than 1 bit of regulatory information and that they, indeed, could transmit up to about 2 bits. In addition, we are also reminded that while the distributions (for example the optimal output distribution in Fig 3b) can look bimodal and this has often been taken as an indication that there are two relevant states of the output, such distributions really can have capacities above 1 bit; similarly, distributions without prominent features, such as monotonically decreasing optimal output distribution of Fig 3c, should also not be expected necessarily to have low capacities.

A closer look at the overall agreement between the small-noise approximation (dashed lines in Fig 3a) and the exact solution (thick lines) shows that the small-noise approximation underestimates the true capacity, consistent with our remark that for large noise the approximation will incorrectly produce negative results; at the 2-bit information contour the approximation is about $\sim 15\%$ off but improves as the capacity is increased.

In the high noise regime we are making yet another approximation, the validity of which we now need to examine. In our discussion about the models of signals and noise we assumed that we can talk about the fractional occupancy of the binding site and the *continuous* concentrations of mRNA, transcription factors and protein, instead of counting these species in discrete units, and

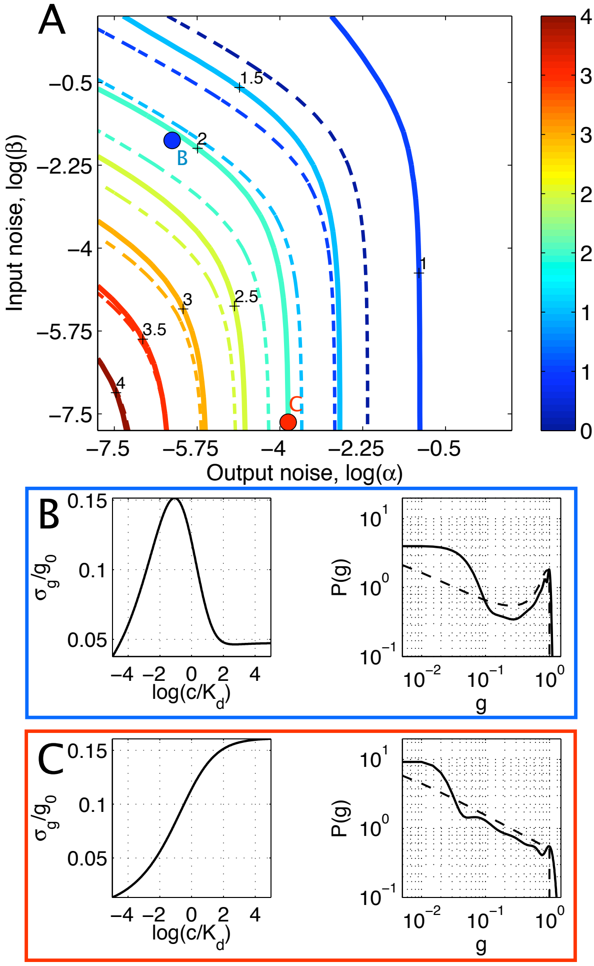

FIG. 3: (Color online) Information capacity (color code, in bits) as a function of input and output noise using the activator input/output relation with Gaussian noise given by Eq (20) and no cooperativity ($h = 1$). Panel A shows the exact capacity calculation (thick line) and the small noise approximation (dashed line). Panel B displays the details of the blue point in A: the noise in the output is shown as a function of the input, with a peak being characteristic of a dominant input noise contribution; also shown is the exact solution (thick black line) and the small-noise approximation (dashed black line) to the optimal distribution of output expression levels. Panel C similarly displays details of the system denoted by a red dot in A; here the output noise is dominant and both approximate and exact solutions for the optimal distribution of outputs show a trend monotonically decreasing with the mean output.

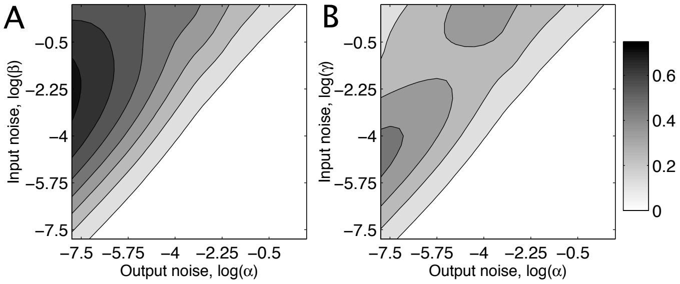

FIG. 4: Difference in the information capacity between the repressors and activators (color code in bits). Panel A shows $I_{\rm rep}(h = 1) - I_{\rm act}(h = 1)$, with the noise model that includes output (α) and input diffusion noise (β) contributions (see Fig 3 for absolute values of $I_{\rm act}(h = 1)$). Panel B shows $I_{\rm rep} - I_{\rm act}$ for the noise model that includes output noise (α) and input switching noise (γ) contributions; this plot is independent of cooperativity, h.

that noise can effectively be treated as Gaussian. Both of these assumptions are the cornerstones of the Langevin approximation for calculating the noise variance [26]. If parameters α and β actually arise due to the underlying microscopic mechanisms described in the Section III on signals and noise, we expect that at least for some large-noise regions of the noise plane the discreteness in the number of mRNA molecules will become important and the Langevin approximation will fail. In such cases (a much more time-consuming) exact calculation of the input/output relations using the Master equation is possible for some noise models (see Appendix B); we show that in the region where $\log\alpha > -2$ the channel capacities calculated with Gaussian kernels can be overestimated by $\sim$ 10% or more; there the Langevin calculation gives the correct second moment, but misses the true shape of the distribution. Although both examples with realistic noise parameters, B and C, of Fig 3 lie safely in the region where Langevin approximation is valid, care should be used whenever both output noise α and burst size are large, and α is consequently dominated by small number of transcripts.

Is there any difference between activators and repressors in their capacity to convey information about the input? We concluded Section III on the noise models with separate expressions for activator noise, Eq (18), and repressor noise, Eq (19); focusing now on the repressor case, we recompute the information in the same manner as we did for the activator in Fig 3a, and display the difference between the capacities of the repressor and activator with the same noise parameters in Fig 4. As expected, the biggest difference occurs above the main diagonal, where the input noise dominates over the output noise. In this region the capacity of the repressor can be bigger by as much as third than that of the corresponding activator. Note that as $h \rightarrow \infty$, the activator and repressor noise expressions become indistinguishable and the difference in capacity vanishes for the noise models with output and input diffusion noise contributions, Eqs (18, 19).

The behavior of the regulatory element is conveniently visualized in Fig 5 by plotting a cut through the noise plane along its main diagonal. Moving along this cut scales the total noise variance of the system up or down by a multiplicative factor, and allows us to observe the overall agreement between the exact solution and small- and large-noise approximations. In addition we point

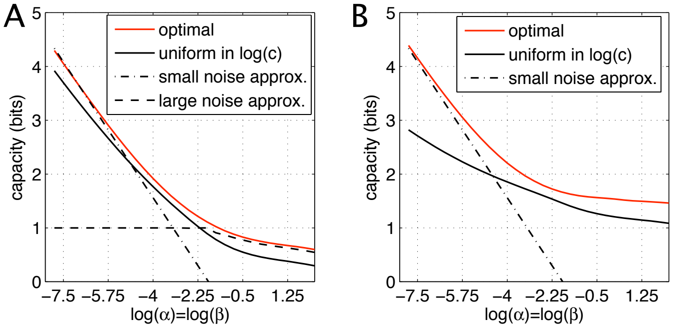

FIG. 5: (Color online) Comparison of exact channel capacities and various approximate solutions. For both panels (panel A, no cooperativity, $h = 1$; panel B, strong cooperativity, $h = 3$) we take a cross-section through the noise plane in Fig 3 along the main diagonal, where the values for noise strength parameters α and β are equal. The exact optimal solution is shown in red. By moving along the diagonal of the noise plane (and along the horizontal axis in the plots above) one changes both input and output noise by the same multiplicative factor s, and since, in small-noise approximation, $I_{\rm SNA} \propto \log Z$, $Z = \int \sigma_g(\bar{g})^{-1}\, d\bar{g}$, that factor results in an additive change in capacity by $\log_2 s$. We can use the large noise approximation lower bound on capacity for the case $h = 1$, in the parameter region where capacities fall below 1 bit.

out the following interesting features of Fig 5 that will be examined more closely in subsequent sections.

First, the parameter region in non-cooperative case, in which the capacity falls below one bit and the large noise approximation is applicable, is small and shrinks further at higher cooperativity. This suggests that a biological implementation of a reliable binary channel could be relatively straightforward, assuming our noise models are appropriate. Moreover, there exist distributions not specifically optimized for the input/output kernel, such as the input distribution uniform in $\log(c/K_d)$ that we pick as illustrative example in Fig 5 (thick black line); we find that this simple choice can achieve considerable information transmission, and are therefore motivated to raise a more general question about the sensitivity of channel capacity with respect to perturbations in the optimal solution, $p^*(c)$. We revisit this idea more systematically in the next section.

Second, it can be seen from Fig 5 that at small noise the cooperativity has a minor effect on the channel capacity. This is perhaps unexpected as the shape of the mean response $\bar{g}(c)$ strongly depends on h. We recall, however, that mutual information $I(c; g)$ is invariant to any invertible reparametrization of either g or c. In particular, changing the cooperativity or the value of the equilibrium binding constant, K_d, in theory only results in an invertible change in the input variable c, and therefore the change in the steepness or midpoint of the mean response must not have any effect on $I(c; g)$. This argument does break down in the high noise regime, where the cooperative system achieves capacities above one bit while the non-cooperative system fails to do so. Reparametrization invariance would work only if the input concentration could extend over the whole positive interval, from zero to infinity. The substantial difference between capacities of cooperative and non-cooperative systems in Fig 5 at *low capacity* stems from the fact that in reality the cell (and our computation) is limited to a finite range of concentrations, $c \in [c_{\rm min}, c_{\rm max}]$, instead of the whole positive half-axis, $c \in [0, \infty)$. We explore the issue of limited input dynamic range further in the next section.

Finally, we draw attention to the simple linear scaling of the channel capacity with the logarithm of the total noise strength in small noise approximation, as explained in the caption of Fig 5. In general, increasing the number of input and output molecules by a factor of four will decrease the relative input and output noise by a factor of $\sqrt{4} = 2$, and therefore, in the small noise approximation, increase the capacity by $\log_2 2 = 1$ bit. If one assumes that the cell can make transcription factor and output protein molecules at no cost, then scaling of the noise variance along the horizontal axis of Fig 5 is inversely proportional to the total number of signaling molecules used by the regulatory element, and its capacity can grow without bound as more and more signaling molecules are used. If, however, there are metabolic or time costs to making more molecules, our optimization needs to be modified appropriately, and we present the relevant computation in Section IV C on the costs of coding.

B. Cooperativity, dynamic range and the tuning of solutions

In the analysis presented so far we have not paid any particular attention to the question of whether the optimal input distributions are biologically realizable or not. We will proceed to relax some of the idealizations made until now and analyze the corresponding changes in the information capacity.

We start by considering the impact on channel capacity of changing the allowed dynamic range to which the input concentration is restricted. Figure 6a displays the capacity as a function of the dynamic range, output noise and cooperativity. The main feature of the plot is the difference between the low and high cooperativity cases at each noise level; regardless of cooperativity the total information at infinite dynamic range would saturate at approximately the same value (which depends on the output noise magnitude). However, highly cooperative systems manage to reach a high fraction of 80% or more of their saturated information capacity even at reasonable dynamic ranges of 25 to 100-fold (meaning that the input concentration varies between $[\frac{1}{5}K_d, 5K_d]$ or $[\frac{1}{10}K_d, 10K_d]$, respectively), whereas low cooperativity systems require a much bigger dynamic range for the same effect. The decrease in capacity with decreasing

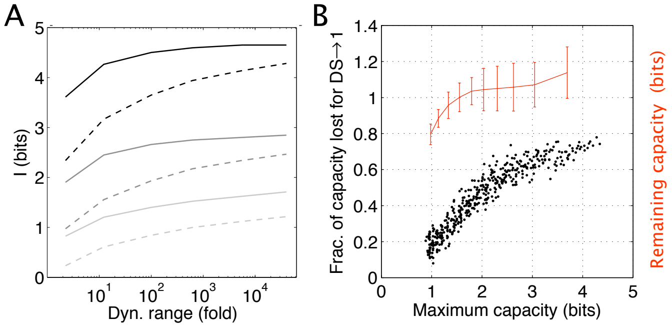

FIG. 6: (Color online) Effects of imposing realistic constraints on the space of allowed input distributions. Panel A shows the change in capacity if the dynamic range of the input around K_d is changed ("25-fold range" means $c \in \left[\frac{1}{5}K_d, 5K_d\right]$). The regulatory element is a repressor with either no cooperativity (dashed line) or high cooperativity, $h = 3$ (thick line). We plot three high-low cooperativity pairs for different choices of the output noise magnitude (high noise in light gray, $\log\alpha \approx -2.5$; medium noise in dark gray, $\log\alpha \approx -5$; low noise in black, $\log\alpha \approx -7.5$). Panel B shows the sensitivity of channel capacity to perturbations in the optimal input distribution. For various systems from Fig 3 we construct suboptimal input distributions, as described in the text, compute the fraction of capacity lost relative to the unperturbed optimal solution and plot this fraction against the optimal capacity of that system (black dots); extrapolated absolute capacity left when the input tends to be very different from optimal, i.e. $D_{\rm JS} \to 1$, is plotted in red.

dynamic range is a direct consequence of the nonlinear relationship between the concentration and occupancy, Eq (12), and for low cooperativity systems means being unable to fully shut down or fully induce the promoter. In theory, Eq (18) predicts that $\sigma_g(\bar{g}(c) \to 0) = 0$, making the state in which the gene is "off" very informative about the input. If, however, the gene cannot be fully repressed either because there is always some residual input, $c_{\rm min}$, or because there is leaky expression even when the input is exactly zero, then at any biologically reasonable input dynamic range some capacity will be lost.

Next, we briefly discuss how precisely tuned the resulting optimal distributions have to be to take full advantage of the regulatory element's capacity. For each point in the noise plane of Fig 3a the optimal input distribution $p^*(c)$ is perturbed many times to create an ensemble of suboptimal inputs $p_i(c)$ (see Appendix C). For each $p_i(c)$, we compute, first, its distance away from the optimal solution by means of Jensen-Shannon divergence, $d_i = D_{\rm JS}(p_i, p^*)$ [29]; next, we use the $p_i(c)$ to compute the suboptimal information transmission I_i. The divergence d_i is a measure of similarity between two distributions and ranges between 0 (distributions are the same) and 1 (distributions are very different); $1/d_i(p_i, p^*)$ approximately corresponds to the number of samples one would have to draw to say with confidence that they were selected either from p_i or p^*. A scatter plot of many such pairs (d_i, I_i) obtained with various perturbations $p_i(c)$ for each system of the noise plane characterizes the sensitivity of the optimal solution for that system; the main feature of such a plot, Fig 9, is the linear (negative) slope that describes the fraction of channel capacity lost for a unit of Jensen-Shannon distance away from the optimal solution. Figure 6b displays these fractions as a function of the optimal capacity, and each system from the noise plane shown in Fig 3 is represented by a black dot. We note that systems with higher capacities require more finely tuned solutions and suffer a larger fractional (and thus absolute) loss if the optimal input distribution is perturbed. Importantly, if the linear slopes are taken seriously and are used to extrapolate towards distributions that are very different from optimal, $D_{\rm JS} \to 1$, we observe that for most of the noise plane the leftover capacity still remains about a bit, indicating that biological regulatory elements capable of transmitting an "on-off" decision perhaps are not difficult to construct. On the other hand, transmitting significantly more than one bit requires some degree of tuning that matches the distribution of inputs to the characteristics of the regulatory element.

C. Costs of higher capacity

Real regulatory elements must balance the pressure to convey information reliably with the cost of maintaining the cell's internal state, represented by the expression levels of transcription factors. The fidelity of the representation is increased (and the fractional fluctuation in their number is decreased) by having more molecules "encode" a given state. On the other hand, making or degrading more transcription factors puts a metabolic burden on the cell, and frequent transitions between various regulatory states could involve large time lags as, for example, the regulation machinery attempts to keep up with a changed environmental condition, by accumulating or degrading the corresponding TF molecules. In addition, the output genes themselves that get switched on or off by transcription factors and therefore "read out" the internal state must not be too noisy, otherwise the advantage of maintaining precise transcription factor levels is lost.

Suppose that there is a cost to the cell for each molecule of output gene that it needs to produce, and that this incremental cost per molecule is independent of the number of molecules already present. Then, on the output side, the cost must be proportional to $\langle g \rangle = \int dg\, g\, p(g)$. We remember that in optimal distribution calculations g is expressed as relative to the maximal expression, such that its mean is between zero and one. To get an absolute cost in terms of the number of molecules, this normalized $\bar{g}$ therefore needs to be multiplied by the inverse of the output noise strength, α^{-1}, as the latter scales with g_0 (see Table I). The contribution of the output cost is thus $\propto \alpha^{-1}\bar{g}$.

On the input side, the situation is similar: the cost must be proportional to $K_d\langle\tilde{c}\rangle = K_d\int d\tilde{c}\,\tilde{c}\,p(\tilde{c})$, where our optimal solutions are expressed, as usual, in dimensionless concentration units, $\tilde{c} = c/K_d$. In either of the two input noise models (i.e. diffusion or switching input noise), with diffusion constant held fixed, $K_d \propto \beta^{-1}$ or $K_d \propto \gamma^{-1}$. See Appendix D for the notes on the effects of non-specific binding of transcription factors to the DNA.

Collecting all our thoughts on the costs of coding, we can write down the "cost functional" as the sum of input and output cost contributions:

$$\langle\mathcal{C}[p(c)]\rangle = \frac{v_1}{\beta}\int dc\,p(c)\,c + \frac{v_2}{\alpha}\int dc\,p(c)\int dg\,p(g|c)\,g, \tag{21}$$

where v_1 and v_2 are proportional to the unknown costs per molecule of input or output, respectively, and α and β are noise parameters of Table I. This ansatz captures the intuition that while decreasing noise strengths will increase information transmission, it will also increase the cost. Instead of maximizing the information without regard to the cost, the new problem to extremize is:

$$\mathcal{L}[p(c)] \;=\; I[p(c)] - \Phi\langle\mathcal{C}[p(c)]\rangle - \Lambda\int dc\,p(c), \tag{22}$$

and the Lagrange multiplier Φ has to be chosen so that the cost of the resulting optimal solution $\langle\mathcal{C}[p^*(c)]\rangle$ equals some predefined cost C_0 that the cell is prepared to pay.

We now wish to recreate the noise plane of Fig 3, while constraining the total cost of each solution to C_0. To be concrete and pick the value for the cost and proportionality constants in Eq (21), we use the estimates from *Drosophila* noise measurements and analysis in [23, 24], which assign to the system denoted by a blue dot in Fig 3a, the values of ~ 800 Bicoid molecules of input at K_d, and a maximal induction of $g_0 \sim 4000$ Hunchback molecules if the burst size b were 10. Figure 7a is the noise plane for an activator with no cooperativity, as in Fig 3, but with the cost limited to an average total of $C_0 \sim 7000$ molecules of input and output per nucleus. There is now one optimal solution denoted by a green dot (with a dominant input noise contribution); if one tries to choose a system with lower input or output noise, the cost constraint forces the input distribution, $p(c)$, and the output distribution, $p(g)$, to have very low probabilities at high induction, consequently limiting the capacity.

Clearly, a different system will be optimal if another total allowed cost C_0 is selected. The dark green line on the noise plane in Fig 7a corresponds to the flow of the optimal solution for an activator with no cooperativity if the allowed cost is increased, and the corresponding cost-capacity curve is shown in Fig 7b. The light green line is the trajectory of the optimal solution in the noise plane of the activator system with cooperativity $h = 3$, and the dark and light red trajectories are shown for the repressor with $h = 1$ and $h = 3$, respectively. We note first that the behavior of the cost function is quite different for the activator (where low input implies low output and therefore low cost; and conversely high input means high output and also high cost) and the repressor (where input and output are mutually exclusively high or low and the cost is intermediate in both cases). Secondly, in Fig 7b we observe that the optimal capacity as a function of cost is similar for the activators and repressors, in contrast to the comparison of Fig 4, where repressors provided higher capacities. Thirdly, we note in the same figure that increasing the cooperativity at fixed noise strength β brings a substantial increase, of almost a bit over the whole cost range, in the channel capacity, in agreement with our previous observations about the interaction between capacity and the dynamic

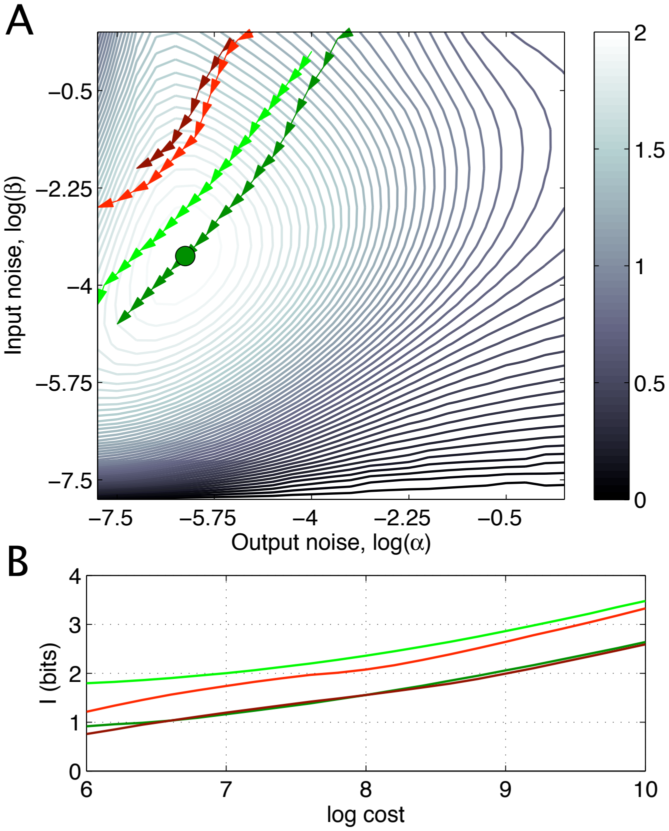

FIG. 7: (Color online) The effects of metabolic or time costs on the achievable capacity of simple regulatory elements. Contours in panel A show the noise plane for non-cooperative activator from Fig 3, with the imposed constraint that the average total (input + output) cost is fixed to some C_0; as the cost is increased, the optimal solution (green dot) moves along the arrows on a dark green line (the contours change correspondingly, not shown). Light green line shows activator with cooperativity $h = 3$, dark and light red lines show repressors without and with cooperativity ($h = 3$). Panel B shows the achievable capacity as a function of cost for each line in panel A.

range. The last and perhaps the most significant conclusion is that even with input distributions matched to maximize the transmission at a fixed cost, the capacity still only scales roughly linearly with the logarithm of the number of available signaling molecules, and this fact must ultimately be limiting in a single regulatory element.

V. DISCUSSION

We have tried to analyze a simple regulatory element as an information processing device. One of our major results is that one cannot discuss an element in isolation from the statistics of the input that it is exposed to. Yet in cells the inputs are often transcription factor concentrations that "encode" the state of various genetic switches, from those responsible for cellular identity to those that control the rates of metabolism and cell division, and the cell exerts control over these concentrations. While it *could* use different distributions to represent various regulatory settings, we argue that the cell *should* use the one distribution that allows it to make the most of its genetic circuitry – the distribution that maximizes the dependency, or mutual information, between inputs and outputs. Mutual information can then be seen both as a measure of how well the cell is doing by using *its* encoding scheme, and the best it could have done using the *optimal* scheme, which we can compute; comparison between the optimal and measured distributions gives us a sense of how close the organism is to the achievable bound [25]. Moreover, mutual information has absolute units, i.e. bits, that have a clear interpretation in terms of the number of discrete distinguishable states that the regulatory element can resolve. This last fact helps clarify the ongoing debates about what is the proper noise measure for genetic circuits, and in what context a certain noise is either "big" or "small" (as it is really a function of the inputs). Information does not replace the standard *noise-over-the-mean* measure – noise calculations or measurements are still necessary to compute the element's capacity – but does give it a functional interpretation.

We have considered a class of simple parametrizations of signals and noise that can be used to fit measurements for several model systems, such as Bicoid-Hunchback in the fruit fly, a number of yeast genes, and the *lac* repressor in *Escherichia coli* (see Ref [25] for the latter). We find that the capacities of these realistic elements are generally larger than 1 bit, and can be as high as 2 bits. By simple inspection of optimal output distributions in Figs 3b or 3c it is difficult to say anything about the capacity: the distribution might look bimodal yet carry more than one bit, or might even be a monotonic function without any obvious structure, indicating that the information is encoded in the graded response of the element. When the noise is sufficiently high, on the other hand, the optimal strategy is that of achieving one bit of capacity and only utilizing maximum and minimum available levels of transcription factors for signaling. The set of distributions that achieve capacities close to the optimal one is large, suggesting that perhaps one-bit switches are not difficult to implement biologically, while in contrast we find that transmission of much more than one bit requires some tuning of the system.

Finally, we discussed how additional biophysical constraints can modify the channel capacity. By assuming a linear cost model for signaling molecules and a limited input dynamic range, the capacity and cost couple in an interesting way and the maximization principle allows new questions to be asked. For example, increasing the cooperativity reduces the cost, as we have shown; on the other hand, it increases the sensitivity to fluctuations in the input, because the input noise strength β is proportional to the square of Hill's coefficient, h^2. In a given system we could therefore predict the optimal effective cooperativity, if we knew the real cost per molecule. Further work is needed to tease out the consequences of cost (if any) from experimental data.

The principle of information maximization clearly is not the only possible lens through which regulatory networks are to be viewed. One can think of examples where there are constraints on the *dynamics*, something that our analysis has ignored by only looking at steady state behavior; for instance, the chemotactic network of *Escherichia coli* has to perfectly adapt in order for the bacterium to be able to climb the attractant gradients. Alternatively, suppose that a system needs to convey only a single bit, but it has to be done reliably in a fluctuating environment, perhaps by being *robust* to the changes in outside temperature. In this case it seems that both concepts, that of maximal information transmission and the robustness to fluctuations in certain auxiliary variables which also influence the noise, could be included into the same framework, but the issue needs further work. More generally, however, these and similar examples assume that one has identified in advance the biologically relevant features of the system, e.g. *perfect adaptation* or *robustness*, and that there exists a problem-specific error measure which the regulatory network is trying to minimize. Such a measure could then either replace or complement the assumption-free information theoretic approach presented here.

We emphasize that the kind of analysis carried out here is not restricted to a single regulatory element. As was pointed out in the introduction, the inputs $\mathcal{I}$ and the outputs $\mathcal{O}$ of the regulatory module can be multidimensional, and the module could implement complex internal logic with multiple feedback loops. It seems that especially in such cases, when our intuition about the noise – now a function of multiple variables – starts breaking down, the information formalism could prove to be helpful. Although the solution space that needs to be searched in the optimization problem grows exponentially in the inputs, there are biologically relevant situations that nevertheless appear tractable: for example, when there are multiple readouts of the same input, or combinatorial regulation of a single output by a pair of inputs; in addition, knowing that the capacities of a single input/output chain are on the order of a few bits also means that only a small number of distinct input levels for each input need to be considered. Some cases of interest therefore appear immediately amenable to biophysical modeling approaches and the computation of channel capacities, as presented in this paper.

We have focused here on the theoretical exploration of information capacity in simple models. It is natural to ask how our results relate to experiment. Perhaps the strongest connection would be if biological systems really were selected by evolution to optimize information flow in the sense we have discussed. If this optimization principle applies to real regulatory elements, then, for example, given measurements on the input/output relation and noise in the system we can make parameter free predictions for the distribution of expression levels that cells will use. Initial efforts in this direction, using the Bicoid-Hunchback element in the *Drosophila* embryo as an example, are described in Ref [25]. It is worth noting that a parallel discussion of optimization principles for information transmission has a long history in the context of neural coding, where we can think of the distribution on inputs as given by the sensory environment and optimization is used to predict the form of the input/output relation [30, 31, 32, 33]. Although there are many open questions, it would be attractive if a single principle could unify our understanding of information flow across such a wide range of biological systems.

Acknowledgments

We thank T Gregor, JB Kinney, DW Tank and EF Wieschaus for helpful discussions. This work was supported in part by NIH grants P50 GM071508 and R01 GM077599, NSF grant PHY-0650617, by the Swartz Foundation, the Burroughs Wellcome Fund Program in Biological Dynamics (GT) and by US Department of Energy grant DE-FG02-91ER40671 (CC).

APPENDIX A: FINDING OPTIMAL CHANNEL CAPACITIES

If we treat the kernel on a discrete (c, g) grid we can easily choose such $p(c)$ as to maximize the mutual information $\mathrm{I}(c; g)$ between the expression level and the concentration. The problem can be stated in terms of the following variational principle:

$$\mathcal{L}[p(c)] = \sum_{c,g} p(g|c)p(c) \log_2 \frac{p(g|c)}{p(g)} - \Lambda \sum_c p(c), \quad \text{(A1)}$$

where the multiplier Λ enforces the normalization of $p(c)$, and $p(g)$ itself is a function of the unknown distribution (since $p(g) = \sum_c p(g|c)p(c)$). The solution $p^*(c)$ of this problem achieves the capacity, $\mathrm{I}(c; g)$, of the channel.

The original idea behind the Blahut-Arimoto approach [21] was to understand that the maximization of Eq (A1) using variational objects $p(c_i)$ is equivalent to the following maximization:

$$\max_{p(c)} \mathcal{L}[p(c)] \sim \max_{p(c)} \max_{p(c|g)} \mathcal{L}'[p(c), p(c|g)], \quad \text{(A2)}$$

where

$$\mathcal{L}'[p(c), p(c|g)] = \sum_{g,c} p(c)p(g|c) \log \frac{p(c|g)}{p(c)} - \Lambda \sum_c p(c). \quad \text{(A3)}$$

In words, finding the extremum in variational object $p(c)$ is equivalent to a double maximization of a modified Lagrangian, where both $p(c)$ and $p(c|g)$ are treated as independent variational objects. The extremum of the modified Lagrangian is achieved exactly when the consistency condition $p(c|g) = \frac{p(g|c)p(c)}{\sum_c p(g|c)p(c)}$ holds. This allows us to make an iterative algorithm that we detail below, where Eq (A3) is solved for the optimal $p(c)$ and evaluated at some "known" $p(c|g)$, which is in turn updated with the newly obtained estimate of $p(c)$.

Before describing the algorithm let us also suppose that each input signal c carries some metabolic or time cost to the cell. Then we can introduce a cost vector $v(c)$ that assigns a cost to each codeword c, and require of the solution the following:

$$\sum_c p(c)v(c) \le C_0, \quad \text{(A4)}$$

where C_0 is the maximum allowed expense. The constraint can be introduced into the functional, Eq (A1) or Eq (A3), through an appropriate Lagrange multiplier; the same approach can be taken to introduce the cost of coding for the output words, $\sum_g \sum_c p(g|c)p(c)v(g)$, because it reduces to an additional "effective" cost for the input, $\tilde{v}(c) = \sum_g p(g|c)v(g)$.

As was pointed out in the main text, after discretization we have no guarantees that the optimal distribution $p(c_i)$ is going to be smooth. One way to address this problem is to enforce the smoothness on the scale set by the precision at which the input concentration can be controlled by the cell, $\sigma_c(\bar{c})$, by penalizing big derivatives in the Lagrangian of Eq (A3). An alternative way is to find the spiky solution (without imposing any direct penalty term), but interpret it not as a real, "physical" concentration distribution, but rather as the distribution of concentrations that the cell attempts to generate, c^*. In this case, however, the limited resolution of the input, $\sigma_c(\bar{c})$, must be referred to the output as an additional effective noise in gene expression, $\sigma_g^2 = \sigma_c^2 \left(\frac{\partial \bar{g}}{\partial c}\right)^2$. The optimal solution $p(c^*)$ is therefore the distribution of the levels that the cell would use if it had infinitely precise control over choosing various c^* (i.e. if the input noise were absent), but the physical concentrations are obtained by convolving this optimal result $p(c^*)$ with a Gaussian of width $\sigma_c(c^*)$. Although we chose to use the second approach to compute the results of this paper, we will, for completeness, describe next how to include the smoothness constraint into the functional explicitly.

If the smoothness of the input distribution $p(c)$ is explicitly constrained in the optimization problem, then it will be controllable through an additional Lagrange multiplier, and both ways of computing the capacity – that of referring the limited input resolution $\sigma_c(\bar{c})$ to the noise in the output, and that of including it as a smoothness constraint on the input distribution – will be possible within a single framework. We proceed by analogy to field theories in which the kinetic energy terms of the form $\int |\nabla f(x)|^2\, dx$ constrain the gradient magnitude, and form the following functional:

$$\begin{aligned} \mathcal{L}[p(c)] &= \mathrm{I}(c;g) - \lambda_0 \sum_c p(c) - & \text{(A5)} \\ &- \Phi_1 \sum_c p(c) v_1(c) - \Phi_2 \sum_g p(g) v_2(g) - & \text{(A6)} \\ &- \Theta \sum_c \left(\frac{\Delta p}{\Delta c} \sigma(c) \right)^2 . & \text{(A7)} \end{aligned}$$

Eq (A5) maximizes the capacity with respect to variational objects $p(c)$ while keeping the distribution normalized; Eq (A6) imposes cost $v_1(c)$ on input symbols and cost $v_2(g)$ on output symbols; finally, Eq (A7) limits the derivative of the resulting solution. The difference operator Δ is defined for an arbitrary function $f(c)$:

$$\Delta f(c) = f(c_{i+1}) - f(c_i). \tag{A8}$$

$\sigma(c)$ assigns a different weight to various intervals on the input axis, c. If the input cannot be precisely controlled, but has an uncertainty of $\sigma(c)$ at mean input level c, we require that the optimal probability distribution must not change much as the input fluctuates on the scale $\sigma(c)$. In other words, we require for each input concentration that:

$$\delta p = \frac{\Delta p}{\Delta c} \sigma(c) \ll 1; \tag{A9}$$

the term in Eq.A7 constrained by Lagrange multiplier Θ can be seen as the sum of squares of such variations over all possible values of the input.

By differentiating the functional, Eq (A3), that includes the relevant constraints, with respect to $p(c_i)$ we get the following equation:

$$\begin{aligned} 0 &= \sum_g p(g|c_i) \log p(c_i|g) - \log p(c_i) - \lambda - \Phi_1 v_1(c_i) - \Phi_2 \sum_g p(g|c_i) v_2(g) + & \text{(A10)} \\ &+ \Theta \left\{ [p(c_{i+1}) - p(c_i)] \frac{\sigma^2(c_i)}{(c_{i+1} - c_i)^2} - [p(c_i) - p(c_{i-1})] \frac{\sigma^2(c_{i-1})}{(c_i - c_{i-1})^2} \right\} . & \text{(A11)} \end{aligned}$$

Let us denote by $F(c, p(c)) = \Delta \frac{\Delta p}{(\Delta c)^2} \sigma^2$ the term in braces. The solution for $p(c)$ is therefore given by:

$$p(c) = \frac{1}{Z} \exp \left\{ \sum_g p(g|c) \log p(c|g) - \Phi_1 v_1(c) - \Phi_2 \sum_g p(g|c) v_2(g) + \Theta F(c, p(c)) \right\} . \tag{A12}$$

We can now continue to use the Blahut-Arimoto trick of pretending that $p(c|g)$ is an independent variational object, and that $p(c)$ has to be solved with $p(c|g)$ held fixed; however, even in that case, Eq (A12) is an implicit equation for $p(c)$ which needs to be solved by numerical means. The complete iterative prescription is therefore as follows:

$$\begin{aligned} p^n(g) &= \sum_c p(g|c) p^n(c) & \text{(A13)} \\ p^n(c|g) &= \frac{p(g|c) p^n(c)}{p^n(g)} & \text{(A14)} \\ p^{n+1}(c) &= \frac{1}{Z} \exp \left\{ \sum_g p(g|c) \log p^n(c|g) - \Phi_1 v_1(c) - \Phi_2 \sum_g p(g|c) v_2(g) + \Theta F(c, p^{n+1}(c)) \right\} . & \text{(A15)} \end{aligned}$$

Again, Eq (A15) has to be solved on its own by numerical means as the variational objects for iteration $(n+1)$ appear both on the left- and right-hand sides. The input and output costs of coding are neglected if one sets $\Phi_1 = \Phi_2 = 0$; likewise, smoothness constraint is ignored for $\Theta = 0$, in which case Eq (A15) is the same as in the original Blahut-Arimoto derivation and it gives the value of $p^{n+1}(c)$ explicitly.

For the capacities computed in this paper we have calculated the effective output noise that includes the intrinsic output noise as well as the input noise that has been referred to the output (see Section III); we can therefore set $\Theta = 0$. This approach treats all sources of noise on the same footing and allows us to directly compare the magnitudes of noise sources at the input and the output. We also note that it makes sense to compute and compare the optimal distribution of outputs rather than inputs: the input/output kernels are degenerate and there are various input distributions (differing either in the regions that give saturated or zero response, or by having variations on a scale below σ_c) that will yield essentially the same distribution of outputs.

APPENDIX B: VALIDITY OF LANGEVIN APPROXIMATIONS

Langevin approximation assumes that the fluctuations of a quantity around its mean are Gaussian and proceeds to calculate their variance [26]. For the calculation of exact channel capacity we must calculate the full input/output relation, $p(g|c)$. Even if Langevin approach ends up giving the correct variance as the function of the input, $\sigma_g(c)$, the shape of the distribution might be far from Gaussian. We expect such a failure when the number of mRNA is very small: the distribution of expression levels might be then multi-peaked, with peaks corresponding to $b, 2b, 3b, \ldots$ proteins, where b is the burst size (number of proteins produced per transcript lifetime).

In the model used in Eq (18), parameter $\alpha = (1+b)/g_0$ determines the output noise; $g_0 = b\bar{m}$, where $\bar{m}$ is the average number of transcripts produced during the integrating time (i.e. the longest averaging timescale in the problem, for example the protein lifetime or cell doubling time). If $b \gg 1$, then the output noise is effectively determined only by the number of transcripts, $\alpha \approx 1/\bar{m}$. We should therefore be particularly concerned what happens as $\bar{m}$ gets small.

Our plan is therefore to solve for $p(g|c)$ exactly by finding the stationary solution of the Master equation in the case where the noise consists of the output and switching input contributions. In this approach, we explicitly treat the fact that the number of transcribed messages, designated by m, is discrete. We start by calculating $p_i(m|c,t)$. The state of the promoter is described index i, which can be 0 or 1, depending on whether the promoter is bound by the transcription factor or not, respectively. Normalization requires that for each value of c:

$$\sum_{i=0,1} \sum_m p_i(m|c,t) = 1. \tag{B1}$$

The time evolution of the system is described by the following set of equations for an activator:

$$\begin{aligned}
\frac{\partial p_0(m|c,t)}{\partial t} &= R_e\left(p_0(m-1|c,t) - p_0(m|c,t)\right) \\
&- \frac{1}{\tau}\left(m p_0(m|c,t) - (m+1)p_0(m+1|c,t)\right) \\
&- k_- p_0(m|c,t) + k_+ c p_1(m|c,t), \qquad \text{(B2)} \\
\frac{\partial p_1(m|c,t)}{\partial t} &= -\frac{1}{\tau}\left(m p_1(m|c,t) - (m+1)p_1(m+1|c,t)\right) \\
&+ k_- p_0(m|c,t) - k_+ c p_1(m|c,t), \qquad \text{(B3)}
\end{aligned}$$

where τ is the integrating time, k_- is the rate for switching into the inactive state (off-rate of the activator), k_+ is the second-order on-rate, and R_e is the rate of mRNA synthesis. These constants combine to give $\bar{m} = R_e\tau$ and the input switching noise strength $\gamma = (k_-\tau)^{-1}$, see Table I. This set of equations is supplemented by appropriate boundary conditions for $m = 0$. To find a steady state distribution $p(m|c, t \to \infty) = p(m|c)$, we set the left-hand side to zero and rewrite the set of equations (with high enough cutoff value of $m_{\max}$) in matrix form:

$$\mathbf{M}(c)\mathbf{p}(c) = \mathbf{b}, \tag{B4}$$

where $\mathbf{p} = (p_0(0|c), p_1(0|c), p_0(1|c), p_1(1|c), \cdots)$ and $\mathbf{b} = (0, 0, \cdots, 0, 1)$. Matrix $\mathbf{M}$ (of dimension $2(m_{\max}+1)+1$ rows and $2(m_{\max}+1)$ columns) contains, in its last row, only ones, which enforces normalization. The resulting system is a non-singular band-diagonal system that can be easily inverted. The input/output relation for the number of messages is given by taking $p(m|c) = p_0(m|c) + p_1(m|c)$.

Having found the distribution for the number of transcripts we then convolve it another Poisson process, $p(g|\langle g\rangle = b\,m)$, i.e. $p(g|c) = \sum_m p(m|c)p(g|\langle g\rangle = b\,m)$. Finally, the result is rediscretized such that mean expression $\bar{g}$ runs from 0 to 1.

Note that the Langevin approximation only depends on the combination of the burst size b and the mean number of transcripts $\bar{m}$ through α; in contrast, the Master equation solution depends on both b and $\bar{m}$ independently. The generalization of this calculation to repressors or Hill-coefficient-type cooperativity is straightforward.

Fig 8c shows that the Langevin approximation yields correct second moments of the output distribution; however, Gaussian distributions themselves are, for large burst sizes and small number of messages, inconsistent with the exact solutions, as can be seen in Fig 8a. In the opposite limit where the number of messages is increased and burst size kept small, see Fig 8b, normal distributions are an excellent approximation. Despite these difficulties the information capacity calculated with either Gaussian or Master input/output relations differs by at most 12% over a large range of burst sizes b and values for α, illustrated by Fig 8d.

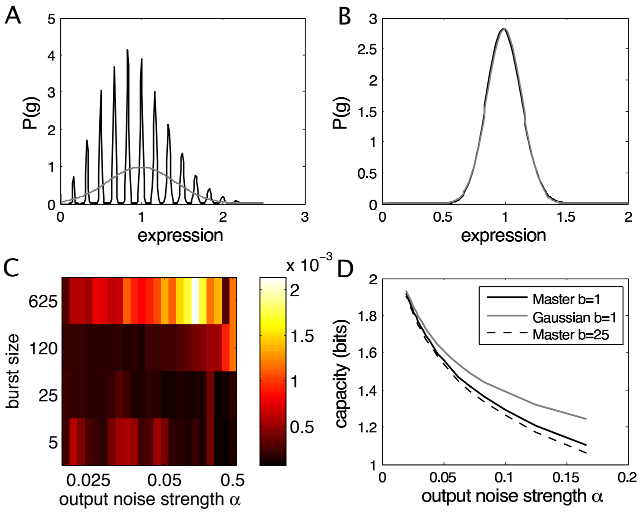

FIG. 8: (Color online) Exact solutions (black) for input/output relations, $p(g|c)$, compared to their Gaussian approximations (gray). Panel A shows the distribution of outputs at maximal induction, $p(g|c_{\max})$ for a system with a large burst size, $b = 5^4$ and a large output noise $\alpha = \frac{1}{6}$ (i.e. the average number of messages is 6, as is evident from the number of peaks, each of which corresponds to a burst of translation at different number of messages). Panel B shows the same distribution for smaller output noise, $b = 5^2$ and $\alpha = \frac{1}{50}$; here Gaussian approximation performs well. Both cases are computed with switching noise parameter $\gamma = \frac{1}{50}$, and cooperativity of $h = 2$. Panel C shows in color-code the error made in computing the standard deviation of the output given c; the error measure we use is the maximum difference between the exact and Gaussian results over the full range of concentrations: $\max_c \text{abs}\left\{[\sigma_g(c)/g_0]_{\text{Master}} - [\sigma_g(c)/g_0]_{\text{Gaussian}}\right\}$. As expected the error decreases with decreasing output noise. Panel D shows that the capacity is overestimated by using an approximate kernel, but the error again decreases with decreasing noise as Langevin becomes an increasingly good approximation to the true distribution. In the worst case the approximation is about 12% off. Gaussian computation only depends on α and not separately on burst size, so we plot only one curve for $b = 1$.

APPENDIX C: FINE-TUNING OF OPTIMAL DISTRIBUTIONS

To examine the sensitivity to the perturbations in the optimal input distributions for Fig 6 we need to generate an ensemble of perturbations. We pick an *ad hoc* prescription, whereby the optimal solution is taken, and we add to it 5 lowest harmonic modes on the input domain, each with a weight that is uniformly distributed on some range. The range determines whether the perturbation is small or not. The resulting distribution is clipped to be positive and renormalized. This choice was made to induce low-frequency perturbations (high frequency perturbations get averaged out because the kernel is smooth). Then, for an ensemble of 100 such perturbations, $p_i(c)$, $i = 1, \dots, 100$, and for every system of the noise plane in Fig 3a, the divergence of the perturbed input distribution to the true solution, $d_i = D_{\text{JS}}(p_i(c), p^*(c))$, is computed, as well as the information transmission, $I_i = I[p(g|c), p_i(c)]$. Figure 9 plots the (d_i, I_i) scatter plots for 3×3 representative systems with varying amounts of output (α) and input (β) noise, taken from Fig 3a uniformly along the horizontal and vertical axes.

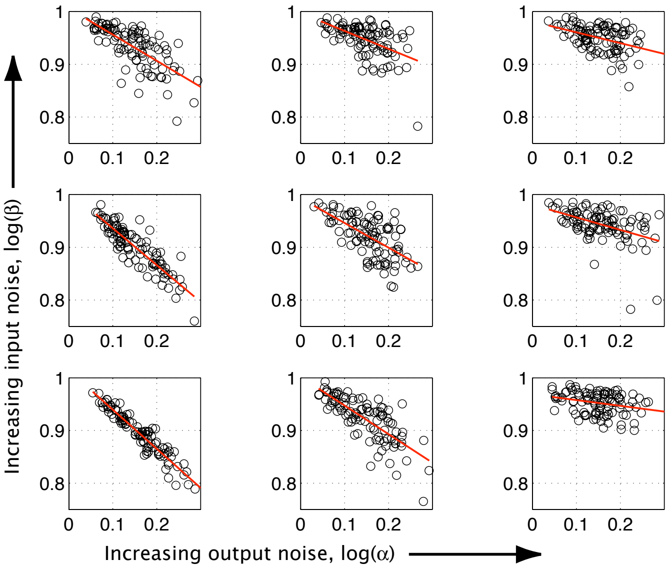

FIG. 9: Robustness of the optimal solutions to perturbations in the input distribution. Activator systems with no cooperativity are plotted; their parameters are taken from an uniformly spaced, 3×3 grid of points in the noise plane of Fig 3a, such that the output noise increases along the horizontal edge of the figure and the input noise along the vertical edge. Each subplot shows a scatter plot of 100 perturbations from the ideal solution; the Jensen-Shannon distance from the optimal solution, d_i, is plotted on the horizontal axis and the channel capacity (normalized to maximum when there is no perturbation), $I_i/I_{\max}$, on the vertical axis. Red lines are best linear fits.

Figure 9 shows that as we move towards systems with higher capacity (lower left corner), perturbations to the optimal solution that are at the same distance from the optimum as in the low capacity systems (upper right corner), will cause greater relative loss (and therefore an even greater absolute loss) in capacity. As expected, higher capacity systems must be better tuned, but even for the highest capacity system considered, a perturbation of around $d_{\text{JS}} \approx 0.2$ will only cause an average 15% loss in capacity. We also note that for systems with high capacity the linear relationship between the the divergence d_i and capacity I_i provides a better fit than for systems with small capacity.

APPENDIX D: NONSPECIFIC BINDING

One needs to make a careful distinction between the total concentration of the input transcription factors, c_t, and the free concentration c_f, diffusing in solution in the nucleus. We imagine the true binding site embedded in a pool of non-specific binding sites – perhaps all other short fragments of DNA – and there being an ongoing competition between one functional site (with strong affinity) and large number of weaker non-specific sites. If these non-specific sites are present at concentration ρ in the cell, and have affinities drawn from some distribution $p(K)$, the relationship between the free and the total concentration of the input is:

$$c_t = c_f + \rho \int dK\, p(K)\, \frac{c_f}{c_f + K}. \tag{D1}$$

Importantly, the concentration that enters all information capacity calculations is the *free* concentration c_f, because it directly determines both the promoter occupancy in Eq (12) as well as diffusive noise; on the other hand, the cell can influence the free concentration only by producing more or less of the transcrption factor, i.e. by varying (and paying for) the *total* concentration. If the free concentration is well below the strength of the non-specific binding, $\langle K\rangle$, Eq (D1) can be approximated by $c_t \approx c_f(1+\rho/\langle K\rangle)$, with the total and free concentrations being proportional to each other. Because the cost functional, Eq (21), is only determined to within a factor anyway, the presence of non-specific sites will effectively just rescale the cost per free molecule of transcription factor. A separate calculation is needed to show that the presence of non-specific binding does not appreciably increase the noise in gene expression (to be presented elsewhere).

[1] JC Leloup & A Goldbeter, Toward a detailed computational model for the mammalian circadian clock. *Proc Nat'l Acad Sci USA* **100,** 7051–7056 (2003).

[2] PA Lawrence, *The Making of a Fly: The Genetics of Animal Design* (Blackwell, Oxford, 1992).

[3] M Levine & EH Davidson, Gene regulatory networks for development. *Proc Nat'l Acad Sci USA* **102,** 4936–4942 (2005).

[4] AP Gasch, PT Spellman, CM Kao, O Carmel-Harel, MB Eisen, G Storz, D Botstein & PO Brown, Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11,** 4241–4257 (2000).

[5] HH McAdams & A Arkin, Stochastic mechanisms in gene expression. *Proc Nat Acad Sci USA* **94,** 814–819 (1997).

[6] TB Kepler & TC Elston, Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J* **81,** 3116–3136 (2001).

[7] L Wolpert, Positional information and the spatial pattern of cellular differentiation. *J Theor Biol* **25,** 1–47 (1969).

[8] CE Shannon, A mathematical theory of communication. *Bell Sys. Tech. J.* **27,** 379–423 & 623–656 (1948). Reprinted in CE Shannon & W Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

[9] CE Shannon, Communication in the presence of noise. *Proc IRE* **37,** 10–21 (1949).

[10] TM Cover & JA Thomas, *Elements of Information Theory* (John Wiley, New York, 1991).

[11] E Ziv, I Nemenman & C Wiggins, Optimal signal processing in small stochastic biochemical networks. arXiv:q-bio.MN/0612041 (2006).

[12] JD Watson, NH Hopkins, JW Roberts, JA Steitz & AM Weiner, *Molecular Biology of the Gene, 4th Edition* (Benjamin/Cummings Publishing, Menlo Park, 1988), pg. 465–502.

[13] Y Setty, AE Mayo, MG Surette & U Alon, Detailed map of a cis-regulatory input function. *Proc Nat'l Acad Sci USA* **100,** 7702–7707 (2003).

[14] T Kuhlman, Z Zhang, MH Saier Jr & T Hwa, Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proc Nat'l Acad Sci USA* **104,** 6043–6048 (2007).

[15] M Thattai & A van Oudenaarden, Intrinsic noise in gene regulatory networks. *Proc Nat'l Acad Sci USA* **98,** 8614–8619 (2001).

[16] MB Elowitz, AJ Levine, ED Siggia & PD Swain, Stochastic gene expression in a single cell. *Science* **207,** 1183–1186 (2002).

[17] E Ozbudak, M Thattai, I Kurtser, AD Grossman & A van Oudenaarden, Regulation of noise in the expression of a single gene. *Nature Gen.* **31,** 69–73 (2002).

[18] WJ Blake, M Kaern, CR Cantor & JJ Collins, Noise in eukaryotic gene expression. *Nature* **422,** 633–637 (2003).

[19] PS Swain, MB Elowitz & ED Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Nat'l Acad Sci USA* **99,** 12795–12800 (2002).

[20] R Silverman, On binary channels and their cascades. *IRE Trans Info Th* **IT-1,** 19–27 (1955).

[21] RE Blahut, Computation of channel capacity and rate-distortion functions. *IEEE Trans Info Th* **4,** 460–473 (1972).

[22] W Bialek & S Setayeshgar, Physical limits to biochemical signaling. *Proc Nat'l Acad Sci USA* **102,** 10040–10045 (2005); arXiv:physics/0301001.

[23] G Tkačik, T Gregor & W Bialek, The role of input noise in transcriptional regulation. arXiv:q-bio/0701002.

[24] T Gregor, DW Tank, EF Wieschaus & W Bialek, Probing the limits to positional information, *Cell* **130,** 153–164 (2007).

[25] G Tkačik, CG Callan Jr & W Bialek, Information flow and optimization in transcriptional control. arXiv/0705.0313 (2007).

[26] NG van Kampen, *Stochastic Processes in Physics and Chemistry, Third Edition*, (North Holland, 2007).

[27] A Bar–Even, J Paulsson, N Maheshri, M Carmi, E O'Shea, Y Pilpel & N Barkai, Noise in protein expression scales with natural protein abundance. *Nature Genetics* **38,** 636–643 (2006).

[28] W Bialek & S Setayeshgar, Cooperativity, sensitivity and noise in biochemical signaling. arXiv:q-bio/0601001 (2005).

[29] J Lin, Divergence measures based on the Shannon entropy. *IEEE Trans Info Th* **37**, 145–151 (1991).

[30] HB Barlow, Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, W Rosenblith, ed, pp 217–234 (MIT Press, Cambridge, 1961).

[31] SB Laughlin, A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch* **36c,** 910–912 (1981).

[32] JJ Atick & AN Redlich, Towards a theory of early visual processing. *Neural Comp* **2,** 308–320 (1990).

[33] N Brenner, W Bialek & R de Ruyter van Steveninck, Adaptive rescaling optimizes information transmission. *Neuron* **26,** 695–702 (2000).