

Information flow and optimization in transcriptional regulation

Gašper Tkačik^{1‡}, Curtis G. Callan, Jr.^{2§}, and William Bialek⁵

¹Joseph Henry Laboratories of Physics, Lewis–Sigler Institute for Integrative Genomics, and ⁵Princeton Center for Theoretical Physics, Joseph Henry Laboratories of Physics, Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

Contributed by Curtis G. Callan, Jr., June 24, 2008 (sent for review April 3, 2008)

In the simplest view of transcriptional regulation, the expression of a gene is turned on or off by changes in the concentration of a transcription factor (TF). We use recent data on noise levels in gene expression to show that it should be possible to transmit much more than just one regulatory bit. Realizing this optimal information capacity would require that the dynamic range of TF concentrations used by the cell, the input/output relation of the regulatory module, and the noise in gene expression satisfy certain matching relations, which we derive. These results provide parameter-free, quantitative predictions connecting independently measurable quantities. Although we have considered only the simplified problem of a single gene responding to a single TF, we find that these predictions are in surprisingly good agreement with recent experiments on the Bicoid/Hunchback system in the early *Drosophila* embryo and that this system achieves ~90% of its theoretical maximum information transmission.

gene regulatory networks | information theory

Cells control the expression of genes in part through transcription factors (TFs), proteins which bind to particular sites along the genome and thereby enhance or inhibit the transcription of nearby genes. We can think of this control process as an input/output device in which the input is the concentration of TF and the output is the concentration of the gene product. Although this qualitative picture has been with us for ~40 years (1), only recently have there been quantitative measurements of *in vivo* input/output relations and of the noise in output level when the input is fixed (2–11). Because these input/output relations have a limited dynamic range, noise limits the “power” of the cell to control gene expression levels. In this article, we quantify these limits and derive the strategies that cells could use to take maximum advantage of the available regulatory power.

To make precise our intuition about regulatory power, we need to quantify the number of reliably distinguishable regulatory settings of the transcription apparatus, a task to which Shannon’s mutual information (12, 13) is ideally suited. Although there are many ways to associate a scalar measure of correlation or control with a joint distribution of input and output signals, Shannon proved that mutual information is the only such quantity that satisfies certain plausible general requirements, independent of the details of the underlying distributions. We can then show that maximizing the mutual information between the input and output of a regulatory element—in effect, maximizing the control that the cell can exert over the expression level of a gene—requires a quantitative matching among the input/output relation, the noise level, and the distribution of TF concentrations used during the life of the cell. If the regulation of gene expression has been optimized, these matching conditions provide parameter-free predictions that connect several independently measurable quantities.

The general problem of optimizing information flow in regulatory networks is difficult. We begin here with the simplest case, where a single TF controls the expression of a single target gene; see also refs 14 and 15. In this case, our optimization

problem is very similar to that discussed by Laughlin (16), who considered the optimization of information transmission from light intensity to intracellular voltage in the fly retina. More generally, ideas from information theory, including optimization and matching, have been productive in analyzing many aspects of neural coding (17–20), and it is attractive to think that similar theoretical principles could apply to the regulation of gene expression.

Although we have treated only the simplest version of the general problem, we try to compare our theoretical results with experiment. Recent work has characterized the input/output relations and noise in the transformation between the Bicoid and Hunchback morphogens in the early *Drosophila* embryo (11, 21). We suggest that, quite generally, the regulation of gene expression in the developing embryo provides an interesting testing ground for our ideas, because the information that is transmitted in this case is precisely the “positional information” (22) that drives the formation of spatial patterns. Using the measured input/output relations and noise in the Bicoid/Hunchback system, our theoretical matching relations provide a parameter-free prediction for the distribution of Hunchback expression levels that we expect to see across the embryo, and the observed distributions have a nontrivial structure that is in good agreement with theory. We also reanalyze measurements of the nucleus-by-nucleus relationship between Bicoid and Hunchback levels (11) to show that the mutual information between these two variables is ~90% of the theoretical maximum. Although there are many caveats, we view these results as strong support for the idea that, in this system at least, genetic regulatory mechanisms provide for optimal information flow.

Setting up the Problem

Gene expression levels (g) change in response to changes in the concentrations of the relevant TFs (c_i). In general, the gene regulatory network is a noisy dynamical system, where TFs can regulate other genes (including other TFs) or they can autoregulate themselves, and this makes both theoretical attempts and experimental approaches to understanding the network of regulatory interactions difficult. Some progress can be achieved, however, by selecting biological systems where one gene responds mainly to a single primary determinant, a TF present at concentration c , and by focusing on the steady-state response of the gene of interest to its input. Although this may seem a drastic approximation, it is also the framework within which most of the recent measurements of noise in gene expression have been performed (2–11).

The changes in the regulated gene often are summarized by an input/output relation $\bar{g}(c)$ in which the mean expression level is plotted as a function of TF concentration (Fig. 1). This average

Author contributions: G.T., C.G.C., and W.B. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

[‡]To whom correspondence may be addressed. E-mail: gtkacik@sas.upenn.edu or ccallan@princeton.edu.

© 2008 by The National Academy of Sciences of the USA

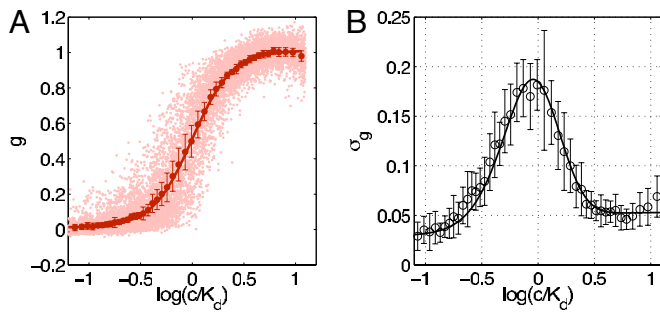


Fig. 2. The Bcd/Hb input/output relationship in the *Drosophila melanogaster* syncytium at early nuclear cycle 14 (see ref. 11). (A) Each point marks the Hb (g) and Bcd (c) concentration in a single nucleus, as inferred from immunofluorescent staining; data are from $\sim 11 \times 10^3$ individual nuclei across nine embryos. Hb expression levels g are normalized so that the maximum and minimum mean expression levels are 1 and 0 respectively; small errors in the estimate of background fluorescence result in some apparent expression values being slightly negative. Bcd concentrations c are normalized by K_d , the concentration of Bcd at which the mean Hb expression level is half-maximal. For details of normalization across embryos, see ref. 11. Solid red line is $\bar{g}(c) = c^n / (c^n + K_d^n)$, with $n = 5$, and error bars are ± 1 SEM. (B) Standard deviation of the noise in Hb expression level as a function of Bcd concentration; error bars are ± 1 SD across embryos. The curve is a fit to $\sigma_g^2(\bar{g}(c)) = \alpha \bar{g} + \beta \bar{g}^{1.8} (1 - \bar{g})^{2.2} + \delta$, with $\alpha \sim 2.5 \times 10^{-3}$, $\beta \sim 0.5$, and $\delta \sim 4 \times 10^{-4}$. This functional form has a microscopic motivation discussed in ref. 30, but note that any smooth phenomenological fit to the data would suffice.

directly to make phenomenological predictions about maximum available regulatory power and optimal distribution of expression levels. Caveats in the connection of theory with experiment are collected at the end of this section. We emphasize at the outset, however, that our goals are rather different from current discussions of models of spatial patterning; rather than trying to trace the pattern of expression levels down to specific molecular mechanisms, we are trying to see whether we can predict at least some features of these patterns by assuming that they reflect optimal solutions to the problem facing the organism.

Maximum Information Transmission. Given the measurements of the mean input/output relation $\bar{g}(c)$ and noise $\sigma_g(c)$ shown in Fig. 2, we can calculate the maximum mutual information between Bcd and Hb concentrations by following the steps outlined above; we find $I_{\text{opt}}(c; g) = 1.7$ bits. To place this result in context, we imagine a system that has the same mean input/output relation, but the noise variance is scaled by a factor F , and ask how the optimal information transmission depends on F . This is not just a mathematical trick: For most physical sources of noise, the relative variance is inversely proportional to the number of signaling molecules, and so scaling the expression noise variance down by a factor of 10 is equivalent to assuming that all relevant molecules are present in 10 times as many copies. We see in Fig. 3 that there is a large regime in which the regulatory power is well approximated by the small noise approximation. In the opposite extreme, at large noise levels, we expect that there are (at best) only two distinguishable states of high and low expression, so that our problem approaches the asymmetric binary channel (31). The exact result interpolates smoothly between these two limiting cases with the real system ($F = 1$) lying closer to the small noise limit, but deviating from it significantly. In particular, it is interesting to note that in this regime, increasing the capacity from the optimum achievable at $F = 1$ by 1 bit would require a substantial increase (of 6-fold), in the number of available signaling molecules, whereas doubling it would require ~ 20 times as many molecules.

In the embryo, maximizing information flow from TF to target gene has a very special meaning. Cells acquire “positional

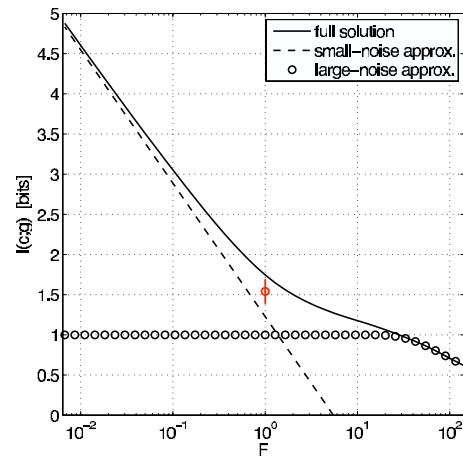


Fig. 3. Optimal information transmission for the Bcd/Hb system as a function of the noise variance rescaling factor F . The factor by which the number of input and output signaling molecules has to be increased for the corresponding gain in capacity is $\sim 1/F$. Dashed and dotted curves show the solutions in the small-noise and large-noise approximations, respectively. The real system, $F = 1$, lies in an intermediate region where neither the small- nor the large-noise approximation is valid. Measured information $I_{\text{data}}(c; g)$ shown in red (error bar is SD over nine embryos).

information,” and thus can take actions that are appropriate to their position in the embryo by responding to the local concentration of morphogen molecules (22). In the original discussions, “information” was used colloquially. But in the simplest picture of *Drosophila* development (25, 32), information in the technical sense really does flow from physical position along the anterior–posterior axis to the concentration of the primary maternal gradients (such as Bcd) to the expression level of the gap genes (such as Hb). Maximizing the mutual information between Bcd and Hb thus maximizes the positional information that can be carried by the Hb expression level.

More generally, rather than thinking of each gap gene as having its own spatial profile, we can think of the expression levels of all of the gap genes together as a code for the position of each cell. In the same way that the 4 bases (2 bits) of DNA must code in triplets to represent arbitrary sequences of 20 aa, we can ask how many gap genes would be required to encode a unique position in the $N_{\text{rows}} \sim 100$ rows of nuclei along the anterior–posterior axis. If the regulation of Hb by Bcd is typical of what happens at this level of the developmental cascade, then each letter of the code is limited to less than two bits ($I_{\text{opt}} = 1.7$ bits) of precision; because $\log_2(N_{\text{rows}})/I_{\text{opt}} = 3.9$, the code would need to have at least four letters. It is interesting to note that there are four known gap genes—*hunchback*, *krüppel*, *giant*, and *knirps* (32)—which provide the initial readout of the maternal anterior–posterior gradients.

We emphasize that in comparing the information capacity of the Bcd/Hb system with the overall information needed for anterior–posterior fate determination, we are making a suggestion rather than drawing a conclusion. Although we tend to think of each row of cells as adopting a unique and largely deterministic fate, which we can identify from the expression levels of pair rule and other later genes in the developmental cascade (see, for example, ref. 33), it is not known whether the gap genes convey enough information to specify this fate, or whether other inputs are essential. Our calculation does indicate, however, that the limits to information transmission in transcriptional regulation are significant on the scale of the information needed for embryonic development, suggesting that the optimization of information transmission is of direct biological relevance.

Thus far, we have emphasized the theoretical maximum

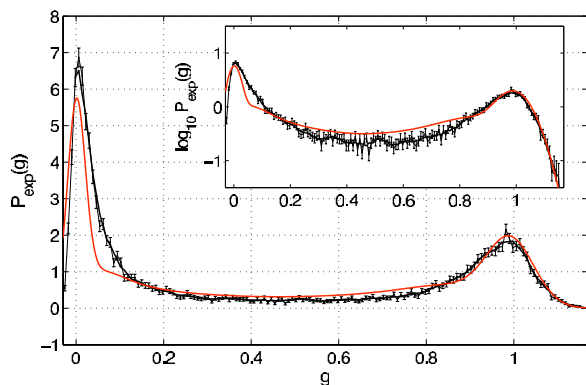


Fig. 4. The measured (black) and optimal (red) distributions of Hunchback expression levels. The measured distribution is estimated from data of ref. 11 by making a histogram of the g values for each data point in Fig. 2. The optimal solution corresponds to the capacity of $I_{opt}(c; g) = 1.7$ bits. The same plot is shown on logarithmic scale in the *Inset*.

information transmission, given the measured noise levels. But does the embryo actually reach this optimum? The experiments of ref. 11 can be thought of as sampling the joint distribution of Bicoid and Hunchback concentrations in the (many) nuclei of the embryo, $P(c, g)$. From such samples, we can estimate the distribution itself and, hence, the actual mutual information between the Bicoid and Hunchback levels; because it is possible to obtain $>10^3$ samples from a single embryo, the often challenging problems of finite sample size can be brought under control by following the strategies described in refs. 34 and 35. We find $I_{data}(c; g) = 1.5 \pm 0.15$ bit, where the error bar is a standard deviation across nine embryos. This represents $\sim 90\%$ (0.88 ± 0.09) of the theoretical maximum.

One might worry that $\sim 90\%$ of the maximum information transmission is easy to achieve; after all, the input/output relation is nearly switch-like, and many different inputs thus are mapped to nearly identical outputs. In fact, were we interested in transmitting just 1 bit, this intuition would be correct, and almost any randomly chosen distribution of inputs would be sufficient. However, a detailed analysis (15) shows that transmission of >1 bit is quite difficult, and requires something close to the optimal matching discussed here. Quantitatively, we find that if we perturb about an optimum distribution and select perturbed distributions that are readily distinguishable from the optimum (technically, the Jensen–Shannon divergence between the two approaches 1 bit), the information capacity of the perturbed distributions rarely exceeds 1 bit, even if the optimal distribution has a capacity of as many as 3 bits. Thus, the fact that the data of ref. 11 demonstrate transmission of 1.5 ± 0.15 bits, which is $\sim 90\%$ of the optimum, means that the embryo must generate distributions of expression levels whose detailed structure is close to the predicted optimum. We will now test this prediction directly.

The Distribution of Expression Levels. In *Drosophila* embryos, Hunchback in each nucleus is responding to its local Bicoid concentration, set during the natural process of morphogen gradient establishment. If we picture this as a process where a single nucleus with its Bcd/Hb regulatory element probes different Bcd concentrations along the anterior–posterior axis, we can pose several deeper questions. Instead of plotting Hunchback expression levels vs. either position or Bcd concentration as is customary, we can ask about the distribution of expression levels seen across all nuclei, $P_{exp}(g)$, as shown in Fig. 4. The distribution is bimodal, so that large numbers of nuclei have near zero or near maximal Hb, consistent with the idea that there is

an expression boundary—cells in the anterior of the embryo have Hb “on” and cells in the posterior have Hb “off.” But intermediate levels of Hunchback expression also occur with nonzero probability, and the overall distribution is quite smooth.

We can compare the experimentally measured distribution of Hb expression levels with the distribution predicted if the system maximizes information flow, and we see from Fig. 4 that the agreement is quite good. The optimal distribution reproduces the bimodality of the real system, hinting in the direction of a simple on/off switch, but also correctly predicts that the system makes use of intermediate expression levels. In particular, the matching of the probability weights in the “on” and “off” state as well as the nonnegligible number of nuclei ($\sim 20\%$) experiencing the intermediate state of induction are nontrivial predictions of our calculation.

The smooth distribution of expression levels is at variance to the common view of the Bcd/Hb system as serving only to delineate a sharp spatial boundary, for which a single bit of capacity would suffice. As noted above, direct computations from the data show that information beyond 1 bit is present, in amounts very close to the maximum possible value. This extra information depends upon the detailed structure of the distribution, which we see is correctly predicted by the theory. More precisely, the intermediate levels of Hb can have meaning only if the noise at those levels is sufficiently low, and it is this intuitive condition that leads to the predicted matching of the expression level distribution to the noise levels. These details are important, because it is precisely in this concentration interval where the embryo ultimately partitions the anterior–posterior axis with a precision of one nuclear row.

Caveats. Our results on the Bcd/Hb system are extremely encouraging. The real system achieves $\sim 90\%$ of the maximum information transmission, and the observed distribution of Hunchback expression levels is in rather good agreement with the distribution predicted from our optimization principle, with no adjustable parameters. To be fair, however, we collect here several caveats.

We have treated the simplest version of our theoretical problem, considering one input and one output, with no feedback. In fact, Hb activates its own expression (36), and this must contribute to the shape of the input/output relation and perhaps also to the noise level. But an important aspect of our analysis is that the maximum mutual information depends on the observed input/output relation and noise in the system, and not directly on the molecular mechanisms that generate these characteristics. Thus, the information capacity is the same no matter whether (for example) the steepness of the input/output relation is the result of intrinsic cooperativity among Bcd molecules or a self-activating feedback loop, assuming that both mechanisms also account correctly for the observed noise level. On the other hand, if the gradient of maternally expressed Hb provides a separate path for transmission of positional information to the final zygotic expression level, then our discussion of one input and one output may be too drastic a simplification.

Bicoid has multiple targets and many of these genes have multiple inputs (37), so to fully optimize information flow, we need to think about a more complex problem than the single input, single output system considered here. How does this affect, for example, our discussion of coding by combinations of gap gene expression levels? Because Bcd acts as an activator for all of the gap genes, their expression levels would tend to provide redundant information about the local Bcd level, reducing the available positional information below the nominal capacity estimated above. On the other hand, several of the gap genes are mutually repressive, and [as with lateral inhibition in the visual system (17)] this serves to remove redundancy and increase information transmission. It would be attractive if these inter-

actions within the gap gene network could also be seen as solutions to an optimization problem.

Clearly, there are more steps to the developmental cascade than the primary gradients and the gap genes, and several interacting genes comprise each step. We emphasize that, despite this complexity, information theory tells us that information cannot spontaneously be created as it propagates through a gene cascade; rather, information can only be lost due to noisy processing. If maternal morphogens did not, to some extent, feed directly into subsequent layers, i.e., the pair-rule or segment polarity genes, in addition to controlling their primary gap gene targets, the information transmission from the maternal gradients to the ultimate nuclear identities would be bounded from above precisely by the transmission between maternal gradients and the gap genes.

One can also raise concerns about the experiments with which we are comparing. Measurement of the distribution of expression levels requires a fair sampling of all of the nuclei in the embryo, and this was not the intent of the experiments of ref. 11. Similarly, the theoretical predictions depend somewhat on the behavior of the input/output relation and noise at low expression levels, which are difficult to characterize experimentally, as well as the (possible) deviations from Gaussian noise. A complete test of our theoretical predictions will thus require a new generation of experiments.

Concluding Remarks

The functionality of a transcriptional regulatory element is determined by a combination of its input/output relation, the

noise level, and the dynamic range of TF concentrations used by the cell. In parallel to discussions of neural coding (16, 19), we have suggested that organisms can make maximal use of the available regulatory power by achieving consistency among these three different ingredients; in particular, if we view the input/output relation and noise level as fixed, then the distribution of TF concentrations or expression levels is predicted by the optimization principle. Although many aspects of transcriptional regulation are well studied, especially in unicellular organisms, these distributions of protein concentrations have not been investigated systematically. In embryonic development, by contrast, the distributions of expression levels can literally be read out from the spatial gradients in morphogen concentration. We have focused on the simplest possible picture, in which a single input TF regulates a single target gene, but nonetheless find encouraging agreement between the predictions of our optimization principle and the observed distribution of the Hunchback morphogen in *Drosophila*. We emphasize that our prediction is not the result of a model with many parameters; instead we have a theoretical principle for how the system should behave so as to maximize its performance and no free parameters.

ACKNOWLEDGMENTS. We thank T. Gregor, D. W. Tank, and E. F. Wieschaus for many helpful discussions, as well as for sharing the raw data from ref. 11. This work was supported in part by National Institutes of Health Grants P50 GM071508 and R01 GM077599, by the Burroughs Wellcome Fund Program in Biological Dynamics (G.T.), by U.S. Department of Energy Grant DE-FG02-91ER40671(to C.G.C.), and by National Science Foundation Grant PHY-0650617.

- Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356.
- Elowitz MB, Levine AJ, Siggia ED, Swain PD (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186.
- Ozbudak E, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31:69–73.
- Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422:633–637.
- Setty Y, Mayo AE, Surette MG, Alon U (2003) Detailed map of a *cis*-regulatory input function. *Proc Natl Acad Sci USA* 100:7702–7707.
- Raser JM, O’Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304:1811–1814.
- Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single cell level. *Science* 307:1962–1965.
- Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. *Science* 307:1965–1969.
- Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036.
- Kuhlman T, Zhang Z, Saier MH, Jr, Hwa T (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proc Natl Acad Sci USA* 104:6043–6048.
- Gregor T, Tank DW, Wieschaus EF, Bialek W (2007) Probing the limits to positional information. *Cell* 130:153–164.
- Shannon CE (1948) A mathematical theory of communication. *Bell Sys Tech J* 27:379–423 and 623–656. Reprinted in Shannon CE, Weaver W (1949) *The Mathematical Theory of Communication* (Univ of Illinois Press, Urbana, IL).
- Cover TM, Thomas JA (1991) *Elements of Information Theory* (Wiley, New York).
- Ziv E, Nemenman I, Wiggins C (2007) Optimal signal processing in small stochastic biochemical networks. *PLoS One* 2:e1077.
- Tkačik G, Callan CG, Jr, Bialek W (2008) Information capacity of genetic regulatory elements. *Phys Rev E* 78:011910–011926.
- Laughlin SB (1981) A simple coding procedure enhances a neuron’s information capacity. *Z Naturforsch* 36C:910–912.
- Barlow HB (1961) Possible principles underlying the transformation of sensory messages. *Sensory Communication*, ed Rosenblith W (MIT Press, Cambridge, MA), pp 217–234.
- Atick JJ, Redlich AN (1990) Towards a theory of early visual processing. *Neural Comp* 2:308–320.
- Brenner N, Bialek W, de Ruyter van Steveninck R (2000) Adaptive rescaling optimizes information transmission. *Neuron* 26:695–702.
- Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
- Gregor T, Wieschaus EF, McGregor AP, Bialek W, Tank DW (2007) Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* 130:141–152.
- Wolpert L (1969) Positional information and the spatial pattern of cellular differentiation. *J Theor Biol* 25:1–47.
- Detwiler PB, Ramanathan S, Sengupta A, Shraiman BI (2000) Engineering aspects of enzymatic signal transduction: Photoreceptors in the retina. *Biophys J* 79:2801–2817.
- Blahut RE (1972) Computation of channel capacity and rate-distortion functions. *IEEE Trans Info Theory* 4:460–473.
- Lawrence PA (1992) *The Making of a Fly: The Genetics of Animal Design* (Blackwell, Oxford).
- Driever W, Nüsslein-Volhard V (1988) A gradient of bicoid protein in *Drosophila* embryos. *Cell* 54:83–93.
- Driever W, Nüsslein-Volhard C (1988) The bicoid protein determines position in the *Drosophila* embryo. *Cell* 54:95–104.
- Driever W, Nüsslein-Volhard C (1989) The bicoid protein is a positive regulator of *hunchback* transcription in the early *Drosophila* embryo. *Nature* 337:138–143.
- Struhl G, Struhl K, Macdonald PM (1989) The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell* 57:1259–1273.
- Tkačik G, Gregor T, Bialek W (2008) The role of input noise in transcriptional regulation. *PLoS One* 3:e2774.
- Silverman R (1955) On binary channels and their cascades. *IEEE Trans Info Theory* 1:19–27.
- Rivera-Pomar R, Jäckle H (1996) From gradients to stripes in *Drosophila* embryogenesis: Filling in the gaps. *Trends Genet* 12:478–483.
- Gergen JP, Coulter D, Wieschaus EF (1986) Segmental pattern and blastoderm cell identities. *Gametogenesis and the Embryo*, ed Gall JG (Liss, New York), pp 195–220.
- Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* 80:197–200.
- Slonim N, Atwal GS, Tkačik G, Bialek W (2005) Information-based clustering. *Proc Natl Acad Sci USA* 102:18297–18302.
- Simpson-Brose M, Treisman J, Desplan C (1994) Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila*. *Cell* 78:855–865.
- Ochoa-Espinosa A, et al. (2005) The role of binding site cluster strength in Bicoid-dependent patterning in *Drosophila*. *Proc Natl Acad Sci USA* 102:4960–4965.