Scaling of Gene Expression with Transcription-Factor Fugacity

Franz M. Weinert
Department of Applied Physics, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA

Robert C. Brewster
Department of Applied Physics, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA and Division of Biology and Biological Engineering, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA

Mattias Rydenfelt
Department of Physics, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA and Integrative Research Institute for the Life Sciences and Institute for Theoretical Biology, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany

Rob Phillips[*]
Department of Applied Physics, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA and Division of Biology and Biological Engineering, California Institute of Technology, 1200 E. California Boulevard, Pasadena 91125, California, USA

Willem K. Kegel[†]
Van 't Hoff Laboratory for Physical and Colloid Chemistry, Debye Research Institute, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

(Received 18 September 2014; published 16 December 2014)

The proteins associated with gene regulation are often shared between multiple pathways simultaneously. By way of contrast, models in regulatory biology often assume these pathways act independently. We demonstrate a framework for calculating the change in gene expression for the interacting case by decoupling repressor occupancy across the cell from the gene of interest by way of a chemical potential. The details of the interacting regulatory architecture are encompassed in an effective concentration, and thus, a single scaling function describes a collection of gene expression data from diverse regulatory situations and collapses it onto a single master curve.

DOI: 10.1103/PhysRevLett.113.258101 PACS numbers: 87.18.Vf, 05.20.Gg, 87.16.Yc, 87.18.Cf

Cells undertake multiple signaling, regulatory, and biochemical tasks simultaneously, and typically the proteins engaged in these pathways are multipurposed [1]. In the gene regulatory setting, this leaves each gene to compete for regulatory proteins (transcription factors) with an array of binding sites across the genome. In addition, genes often exist in multiple, identical copies due to being carried on plasmid or viral vectors or simply from chromosomal replication as a natural part of the cell cycle. As a result, it is of great interest to predict the quantitative effect of this competition on the regulation of gene expression as a function of parameters such as the transcription factor copy number, the arrangement of binding sites on the gene of interest, and the total number and binding strengths of the array of binding sites available to the transcription factor across the genome.

However, systematic studies of gene expression often measure expression from genes that are isolated from the remaining genes in the cell [2–8]. Recently, there has been great progress in understanding and predicting the consequences for gene expression of competition from other genes in the cell [9–12]. Furthermore, it has been shown that a simple extension of the thermodynamic models of gene expression [13–15] can predict gene expression in a wide array of situations where a transcription factor is shared between either multiple identical copies of a gene or a single copy of a gene competing with other unrelated binding sites [16,17].

The theory used to predict and interpret expression, derived in the canonical ensemble, is powerful in the sense that it can be used to make predictions for any competition scenario, assuming that the various states of the system can be enumerated, i.e., all the ways the R transcription factors can be distributed amongst N binding sites with binding energy ϵ. Figures 1(a)–1(c) show theoretical predictions for how changing key regulatory parameters results in unique gene expression profiles as a function of repressor copy number for the case of simple repression in which a gene is under negative control by the action of a repressor molecule. However, an unwieldy facet of the theory is that each curve, though derived from the same core principle, appears to imply a unique and unrelated response curve.

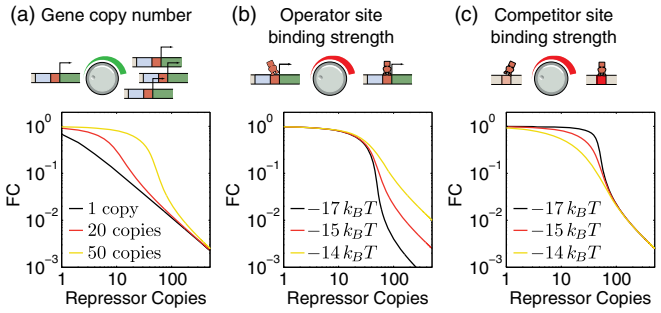

FIG. 1 (color online). Predicted regulatory response. Examples of parameters to tune the competition for transcription factors for the case of a simple repression regulatory architecture and the predicted fold change in gene expression (FC) as a function of the repressor copy number. (a) The gene copy number determines at what value of the repressor copy number the gene shifts from being unrepressed to being repressed. (b) The operator site strength effects the fold change in expression at high repressor copy numbers in the presence of a fixed number (50) of identical genes. (c) The binding strength of competing binding sites effects the sharpness of the transition between unrepressed and repressed state for a fixed operator site binding strength of $-15k_BT$.

In this Letter, we show that when the target sites for repressor molecules are decoupled using the grand canonical ensemble, the predicted transcription of all competition scenarios collapses onto the same simple scaling function for a given promoter architecture. The parameter that fully determines the response is simply $e^{-\beta(\epsilon-\mu)}$, where μ is the chemical potential of the transcription factor, ϵ is the interaction strength between transcription factor and its binding site at the promoter and $\beta = 1/k_BT$ where k_B is Boltzmann's constant and T is the absolute temperature. This formulation has the added benefit that it can be solved analytically very simply for a number of competition scenarios, which alternatively, in the canonical ensemble lead to challenging calculations. In this work, we calculate the fold change in gene expression (FC), which is defined as the gene expression in the presence of a given number of transcription factors divided by the gene expression in the absence of those transcription factors, as a way to measure the level of regulation from systematically tuning the parameters of that regulation (number of transcription factors, binding strength, etc.) [6,8,18–21]. In the remainder of this Letter, we examine the general framework of the thermodynamic theory in the grand canonical ensemble and work through several examples of transcription factor competition. Most importantly, we show how this new approach to thermodynamic descriptions of gene expression suggests what one might call the "natural variable" for the scaling of fold change in expression. When plotted against this natural variable for the system in question, a broad spectrum of regulatory data from diverse experimental situations is shown to collapse onto a single master curve, indicating that although these different regulatory scenarios appear superficially different, the underlying structure of the regulatory response is the same.

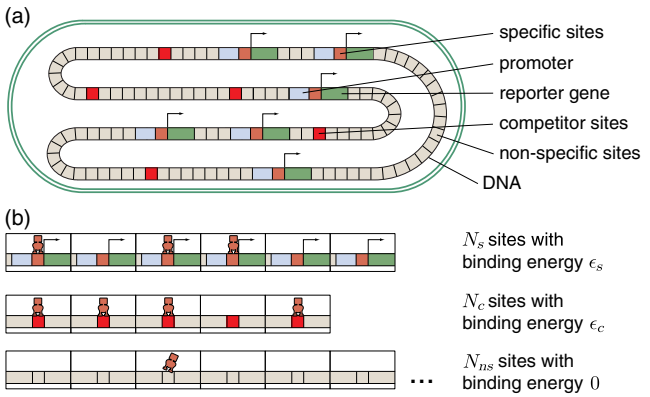

FIG. 2 (color online). (a) Schematic of chromosome when viewed as a lattice of possible binding sites that can be occupied (or not) by a repressor. Within the cell are multiple identical, regulated promoters (that produce a measurable gene product), "competitor sites" that bind the repressor stronger than a nonspecific interaction but do not regulate a gene, and nonspecific sites that each bind the repressor weakly. (b) In the grand canonical framework, each of these types of binding site is treated as a lattice of possible binding sites, characterized by the number of sites in the lattice N and the energy with which each site binds the transcription factor ϵ, with a chemical potential responsible for maintaining balance between the number of molecules bound on each lattice.

For this work, we focus on the familiar promoter architecture of simple repression (see Fig. 1) [8,15,22] consisting of a single repressor binding site capable of halting transcription by RNA polymerase (RNAP) when a repressor is bound. Note, however, that the framework developed here can be applied to any regulatory architecture (see table 1 in Ref. [15]). In order to derive an expression for FC in the limit where RNAP binding is weak (the promoter is typically not occupied by RNAP [8]), consider a cell with N_s promoter sites. The subscript "s" stands for "specific," in contrast to the nonspecific sites with subscript "ns" or competitor sites with subscript "c," which are introduced later (shown schematically in Fig. 2). Uncorrelated binding of repressors, with copy number R, and RNAP, with copy number P, may occur on the promoter sites. If a promoter site is occupied by repressor, the RNAP cannot bind and the gene is inactive. Let the repressor binding energy to its binding site at the promoter sites be ϵ_s, and the RNAP binding energy to the promoter sites be ϵ_p. The grand partition function of this binary lattice is given by

$$\Xi_s = \sum_{\tilde{p}=0}^{N_s} \left[\sum_{\tilde{r}=0}^{N_s-\tilde{p}} \binom{N_s}{\tilde{p},\tilde{r}} \lambda_p^{\tilde{p}} e^{-\beta\tilde{p}\epsilon_p} \lambda_r^{\tilde{r}} e^{-\beta\tilde{r}\epsilon_s} \right] \quad (1)$$

$$= (1 + \lambda_p e^{-\beta\epsilon_p} + \lambda_r e^{-\beta\epsilon_s})^{N_s}. \quad (2)$$

In this equation, $\tilde{p}$ is the number of adsorbed RNAP molecules onto the promoter sites, and $\tilde{r}$ is the number of repressors adsorbed onto their promoter binding sites. The multinomial coefficient is

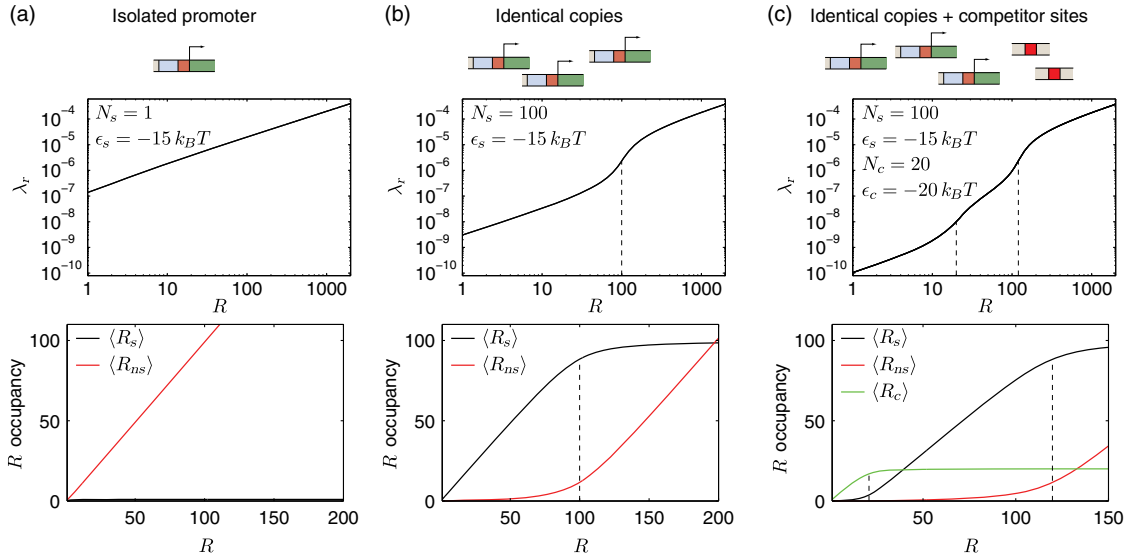

FIG. 3 (color online). Functional form of the fugacity λ_r and occupancies of repressor binding sites for different situations. (a) A single isolated promoter. The single specific repressor binding site in the promoter region is filled up immediately, and almost all repressors are bound to the nonspecific sites. (b) Multiple identical copies of the promoter. The specific repressor binding sites are filled up first, before the repressors bind to the nonspecific sites with a $15k_BT$ higher binding energy. The fugacity of the repressors increases abruptly at $R = N_s$, marked by dashed vertical lines. (c) Multiple identical copies of the promoter and multiple competitor sites. The repressors fill up the competitor binding sites with the lowest repressor binding energy of $\epsilon_c = -20k_BT$, before binding to the specific binding sites and finally to the nonspecific sites. The fugacity increases abruptly at $R = N_s$ and $R = N_s + N_c$, marked by the dashed vertical lines.

$$\binom{N_s}{\tilde{p}, \tilde{r}} = \frac{N_s!}{\tilde{p}!\tilde{r}!(N_s - \tilde{p} - \tilde{r})!}.$$

The fugacities are $\lambda_p = e^{\beta\mu_p}$, where μ_p is the chemical potential of the RNAP, and $\lambda_r = e^{\beta\mu_r}$, where μ_r is the chemical potential of the repressor molecule. The average number of RNAP molecules adsorbed onto the N_s promoter sites is given by

$$\langle P_s \rangle = \lambda_p \frac{\partial \ln \Xi_s}{\partial \lambda_p} = N_s \frac{\lambda_p e^{-\beta\epsilon_p}}{1 + \lambda_p e^{-\beta\epsilon_p} + \lambda_r e^{-\beta\epsilon_s}}. \quad (3)$$

The fold change, FC, is given by the average number of adsorbed RNAP molecules in the presence of the repressors divided by the average number of adsorbed RNAP molecules in the absence of the repressors yielding

$$\mathrm{FC} = \frac{\langle P_s \rangle}{\langle P_s(R = 0) \rangle}, \quad (4)$$

where $\langle P_s(R = 0) \rangle$ follows from Eq. (3) with $\lambda_r e^{-\beta\epsilon_s} = 0$. In the weak promoter limit, which is defined by $\lambda_p e^{-\beta\epsilon_p} \ll 1$ [8,15,22], we have

$$\mathrm{FC} = \frac{1}{1 + \lambda_r e^{-\beta\epsilon_s}}. \quad (5)$$

In other words, in the weak promoter limit only the repressor properties are relevant, and we may ignore all properties of the RNAP in the analyses. Another way of looking at Eq. (5) is that the fraction of promoter sites available for the RNAP is proportional to the fraction of promoter sites that are not covered by repressors: repressors are the "masters" and RNAPs are "followers". We now have a general expression for the fold change in expression, FC, for the simple repressor promoter architecture with a parameter λ_r that contains information regarding the availability of the repressor in a specific competitive environment (the number of repressors and the strength and copy number of identical or competing binding sites). In the following section, we will derive λ_r for several important and common regulatory scenarios.

We now wish to derive an expression for the fugacity which tells us about the relative availability of repressors (given a total number of repressors R) to our binding sites of interest. In this way, λ_r contains information of alternative binding reservoirs for repressors such as number of binding sites and binding affinity. Figure 2 shows a schematic of a cell that contains three options for repressor binding: (1) N_s specific binding sites representing repressor binding with energy ϵ_s and regulating a gene copy, (2) N_c competitor binding sites representing specific binding with energy ϵ_c to a binding site whose occupancy does not regulate expression, and (3) N_{ns} nonspecific binding sites representing repressor binding to the nonspecific genomic background (taken to be 5×10^6, the size of the *E. coli* genome). We make the approximation that the binding energies of all the nonspecific sites have the same value and set this value as zero (such that all energies are measured with respect to this nonspecific binding). Each reservoir is

characterized by the number of available sites and the energy with which a repressor binds one of these sites. The average number of repressors bound to the specific lattices is

$$\langle R_s \rangle = \lambda_r \frac{\partial \ln \Xi_s}{\partial \lambda_r} = N_s \frac{\lambda_r e^{-\beta\epsilon_s}}{1 + \lambda_r e^{-\beta\epsilon_s}}. \quad (6)$$

Analogous to Eq. (2) the grand partition energy of either of the two additional, single species binding lattices (nonspecific or competitor) is generically $\Xi = \sum_{\tilde{r}=0}^{N} \binom{N}{\tilde{r}} \lambda_r^{\tilde{r}} e^{-\beta \tilde{r} \epsilon} = (1 + \lambda_r e^{-\beta\epsilon})^N$, which leads to the average number of adsorbed repressors on nonspecific DNA sites,

$$\langle R_{ns} \rangle = N_{ns} \frac{\lambda_r}{1 + \lambda_r}, \quad (7)$$

and similarly for the competitor sites,

$$\langle R_c \rangle = N_c \frac{\lambda_r e^{-\beta\epsilon_c}}{1 + \lambda_r e^{-\beta\epsilon_c}}. \quad (8)$$

These reservoirs are connected by the constraint that the total number of repressors bound between them is equal (on average) to the total number of repressors in the cell,

$$R = \langle R_s \rangle + \langle R_c \rangle + \langle R_{ns} \rangle. \quad (9)$$

In principle, more unique reservoirs can be added to the conservation equation to account for each specific binding energy available to the molecule of interest; each unique binding energy adds one more reservoir to the problem whose chemical potential must be considered. The substitution of Eqs. (6)–(8) into Eq. (9) leads to a cubic equation for λ_r of the form $a\lambda_r^3 + b\lambda_r^2 + c\lambda_r - R = 0$, which yields the positive real root

$$\lambda_r = \Delta_+ + \Delta_- - \frac{b}{3a}, \quad (10)$$

with $\Delta_\pm = (C_2 \pm \sqrt{C_1^3 + C_2^2})^{1/3}$, $C_1 = (c/3a) - (b/3a)^2$, and $C_2 = (bc/6a^2) + (R/2a) - (b/3a)^3$. The coefficients a, b, and c are derived under the conditions that $N_{ns} \gg (R, N_s, N_c)$, given by $a = e^{\beta\epsilon_c} e^{\beta\epsilon_s} N_{ns}$, $b = (e^{\beta\epsilon_c} + e^{\beta\epsilon_s}) N_{ns} + e^{\beta\epsilon_c} e^{\beta\epsilon_s} (N_s + N_c - R)$, and $c = N_{ns} + e^{\beta\epsilon_c}(N_c - R) + e^{\beta\epsilon_s}(N_s - R)$. When taken with Eq. (5), we now have a closed equation for FC as a function of total repressor copy number R, number of specific (N_s), competitor (N_c), nonspecific (N_{ns}), and binding energies to each of these types of sites. In the limit of no competing sites, i.e., $N_c = 0$ and $e^{-\beta\epsilon_c} = 0$, Eq. (10) simplifies to the root of a quadratic equation. In the limit of an isolated promoter, where $N_s = 1$, we recover the canonical expression for FC derived previously, that is, Eq. (5) with $\lambda_r = R/N_{ns}$ when $R \gg 1$ such that $R \approx \langle R_{ns} \rangle$ [8,15]; however, in the limit of small R the predictions differ slightly because

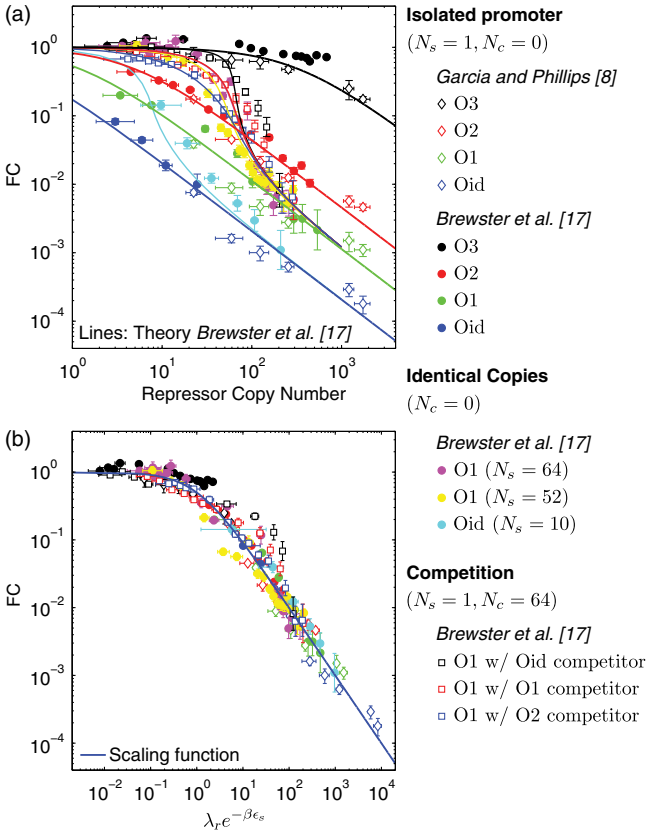

FIG. 4 (color online). Gene expression data by Garcia and Phillips [8] and Brewster *et al.* [17] for various regulatory scenarios. (a) The data and theory plotted versus the repressor copy number R shows a variety of functional forms. (b) The data rescaled to collapse to the same functional form. The blue solid line is the prediction from Eq. (5) without fitting parameter. The repressor binding energies are taken from Ref. [8] as $\epsilon = -9.7k_BT$ for O3, $\epsilon = -13.9k_BT$ for O2, $\epsilon = -15.3k_BT$ for O1, $\epsilon = -17.0k_BT$ for Oid. Values for copy numbers of promoters N_s or competitor binding sites N_c are measured in Ref. [17] by qPCR. For each data set, λ_r is calculated using these parameters and Eq. (10).

of the different constraints imposed by the models. Figures 3(a)–3(c) shows the fugacity and average occupancy of each lattice vs the number of repressors for each of these cases: isolated promoter, identical promoters, and identical promoters with competitors. The primary features in the occupancy and λ_r curves occur whenever R becomes large enough to saturate one of the binding lattices; for instance, in Fig. 3(c), first the competitor and then the specific lattice saturate as R becomes larger than N_c and then larger than $N_c + N_s$.

The theoretical ideas developed above really demonstrate their power when used as a prism through which to view a broad spectrum of gene regulation data. Much recent effort has gone into careful measurement of gene expression as a function of key tunable parameters such as the number of transcription factors, the transcription factor binding site strength, the number of gene copies, and the number and

strength of competing binding sites. These various scenarios, however, give the superficial appearance of each being a separate and unique quantitative story. To that end, Fig. 4(a) shows the data by Garcia and Phillips [8] and Brewster *et al.* [17] plotted as the fold change in gene expression FC versus the repressor number R. The common feature between each of these data sets is that the expressed gene is regulated by simple repression; however, each distinct symbol represents a unique repressor binding energy, different number of promoters or competitor sites all with a unique functional form, which are quantitatively described by the theory curves derived in the canonical ensemble and reported in Ref. [17]. Here, we demonstrate that, in fact, within the grand canonical approach Eq. (5) directly reveals the relevant parameter for data collapse, with the same functional form for FC. That is, for any scenario, be it single or several promoters, presence or absence of competitor sites, etc., a data point is uniquely determined by λ_r and ϵ_s. If plotted as FC versus $\lambda_r e^{-\beta\epsilon_s}$, data collapse should occur and obey Eq. (5). As can be seen in Fig. 4(b) this is indeed the case; over several decades and without adjustable parameters, the data for all of these seemingly distinct regulatory scenarios falls on a single universal curve.

In conclusion, through the grand canonical formalism described here, we are able to predict the fold change in gene expression for a gene regulated by a transcription factor protein that also binds at other unrelated sites within the cell. Conveniently, the effect of this sharing is totally encapsulated in the fugacity parameter λ, which acts as an effective concentration of the transcription factor for a given regulatory scenario, i.e., the spectrum of competing binding sites for the transcription factor present in the cell. As a result, a single scaling function describes the gene regulation for quite distinct competition scenarios, which highlights the fact that in some complex biological settings, distinct phenomena can be seen as reflecting similar underlying mechanisms when using the natural variables of the problem. Specifically, the data collapse in Fig. 4(b) tells us that the same statistical mechanical phenomena are at work in all cases; namely, binding and unbinding of proteins at their target sites. The occupancy of the specific binding sites, which is the relevant quantity that dictates the fold change in gene expression, is determined solely by the fugacity of the repressor and the binding energy of the repressor to the specific sites. It will be of great interest to apply these ideas to other regulatory architectures to see if this same kind of data collapse is able to link seemingly disparate regulatory phenomena.

We are grateful to Ned Wingreen, Sarah Marzen, Jane Kondev, Hernan Garcia, Al Sanchez, and Jan Groenewold for extremely helpful discussions. We are also grateful to the NIH for support through Grants No. DP1 0D000217 (Directors Pioneer Award) and No. R01 GM085286 and La Fondation Pierre Gilles de Gennes (R. P.). F. M. W. and R. C. B. contributed equally to this work.

*phillips@pboc.caltech.edu
†w.k.kegel@uu.nl

[1] H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muiz-Rascado, J. S. Garca-Sotelo, V. Weiss, H. Solano-Lira, I. Martnez-Flores, A. Medina-Rivera *et al.*, Nucleic Acids Res. **41**, D203 (2013).
[2] J. Müller, S. Oehler, and B. Müller-Hill, J. Mol. Biol. **257**, 21 (1996).
[3] M. B. Elowitz and S. Leibler, Nature (London) **403**, 335 (2000).
[4] L. Saiz, J. M. Rubi, and J. M. Vilar, Proc. Natl. Acad. Sci. U.S.A. **102**, 17642 (2005).
[5] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).
[6] T. Kuhlman, Z. Zhang, M. H. Saier, Jr., and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **104**, 6043 (2007).
[7] T. S. Moon, C. Lou, A. Tamsir, B. C. Stanton, and C. A. Voigt, Nature (London) **491**, 249 (2012).
[8] H. G. Garcia and R. Phillips, Proc. Natl. Acad. Sci. U.S.A. **108**, 12 173 (2011).
[9] D. Del Vecchio, A. J. Ninfa, and E. D. Sontag, Mol. Syst. Biol. **4**, 161 (2008).
[10] K. H. Kim and H. M. Sauro, Biophys. J. **100**, 1167 (2011).
[11] F. Ricci, A. Vallee-Belisle, and K. W. Plaxco, PLoS Comput. Biol. **7**, e1002171 (2011).
[12] T. H. Lee and N. Maheshri, Mol. Syst. Biol. **8**, 576 (2012).
[13] G. K. Ackers, A. D. Johnson, and M. A. Shea, Proc. Natl. Acad. Sci. U.S.A. **79**, 1129 (1982).
[14] N. E. Buchler, U. Gerland, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **100**, 5136 (2003).
[15] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Curr. Opin. Genet. Dev. **15**, 116 (2005).
[16] M. Rydenfelt, R. S. Cox, H. Garcia, and R. Phillips, Phys. Rev. E **89**, 012702 (2014).
[17] R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, Cell **156**, 1312 (2014).
[18] S. Oehler, E. R. Eismann, H. Kramer, and B. Muller-Hill, EMBO J. **9**, 973 (1990).
[19] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO J. **13**, 3348 (1994).
[20] S. Ryu, N. Fujita, A. Ishihama, and S. Adhya, Gene **223**, 235 (1998).
[21] R. Daber, M. A. Sochor, and M. Lewis, J. Mol. Biol. **409**, 76 (2011).
[22] R. C. Brewster, D. L. Jones, and R. Phillips, PLoS Comput. Biol. **8**, e1002811 (2012).