Gene regulation and noise reduction by coupling of stochastic processes

Alexandre F. Ramos[*]

Departamento de Radiologia, Faculdade de Medicina, Núcleo de Estudos Interdisciplinares em Sistemas Complexos, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, Avenida Arlindo Béttio 1000, CEP 03828-000, São Paulo, São Paulo, Brazil

José Eduardo M. Hornos[†]

Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil

John Reinitz

Department of Statistics, Department of Ecology and Evolution, Department of Molecular Genetics and Cell Biology, and the Institute of Genomics and Systems Biology, University of Chicago, 5734 South University Avenue, Chicago, Illinois 60637, USA

(Received 29 September 2014; published 26 February 2015)

Here we characterize the low-noise regime of a stochastic model for a negative self-regulating binary gene. The model has two stochastic variables, the protein number and the state of the gene. Each state of the gene behaves as a protein source governed by a Poisson process. The coupling between the two gene states depends on protein number. This fact has a very important implication: There exist protein production regimes characterized by sub-Poissonian noise because of negative covariance between the two stochastic variables of the model. Hence the protein numbers obey a probability distribution that has a peak that is sharper than those of the two coupled Poisson processes that are combined to produce it. Biochemically, the noise reduction in protein number occurs when the switching of the genetic state is more rapid than protein synthesis or degradation. We consider the chemical reaction rates necessary for Poisson and sub-Poisson processes in prokaryotes and eucaryotes. Our results suggest that the coupling of multiple stochastic processes in a negative covariance regime might be a widespread mechanism for noise reduction.

DOI: 10.1103/PhysRevE.91.020701 PACS number(s): 87.10.Mn, 87.10.Ca, 87.16.Yc, 87.18.Tt

Intrinsic fluctuations are an inherent feature of the intracellular environment of both prokaryotic and eukaryotic cells because critical regulatory molecules are present in very small numbers. The importance of such fluctuations and the Poissonian stochastic processes that govern them was pointed out over 70 years ago by Delbrück [1]. More recently, nondeterministic biological processes of fundamental importance such as infection by phage λ [2] and bacterial chemotaxis [3] have been treated by direct simulation of the master equation [4]. Intrinsic fluctuations have been directly observed experimentally in both prokaryotic [5,6] and eukaryotic [7,8] cells by fluorescence techniques. Nevertheless, a full understanding of the control of intrinsic fluctuations remains elusive, particularly in the metazoa.

The level of intrinsic fluctuations (or noise) for a stochastic process is frequently described in terms of the ratio between the variance and the mean, referred to as the Fano factor

$$\mathcal{F} = \frac{\langle n^2 \rangle - \langle n \rangle^2}{\langle n \rangle}, \tag{1}$$

where n indicates the number of molecules. A Poissonian (or Fano) distribution has variance equal to the mean and hence a Fano factor of one. More dispersed distributions, such as the geometric, have $\mathcal{F} > 1$ and are referred to as super-Fano. In prokaryotes the reaction mechanisms governing transcription and translation are reasonably well understood, such that transcription initiation is governed by a Poisson distribution and translation by a geometric distribution. These facts together with a wide variety of numerical simulations have established a widespread belief that fluctuations of gene products are typically super-Fano, with a Fano process as the lower limit. This picture is difficult to reconcile with the fact that developing organisms, such as *D. melanogaster*, exhibit strikingly precise spatiotemporal patterns of gene expression in the face of intracellular molecular numbers on the order of several hundred per cell [9–11]. In these organisms the detailed chemical reaction mechanisms underlying transcription and translation are complex and poorly understood and a direct stochastic simulation at the mechanistic level is not possible.

In this paper we probe the possible reduction of intrinsic fluctuations using a simple model [12] in which the master equation has exact solutions [13]. The loss of direct representation of chemical mechanism is compensated for by the existence of analytical solutions. These solutions reveal nonintuitive regimes of behavior far from equilibrium that have been overlooked in numerical experiments or theoretical analysis in the neighborhood of a steady state [14]. We will show that a simple model of a self-repressing gene [13] can function in a sub-Fano regime where $\mathcal{F} \approx 0.5$ over a wide range of parameter values and an infra-Fano regime where $\mathcal{F}$ can approach 0 arbitrarily closely in a particular intracellular situation. We have previously shown that the equations used in this work have super-Fano, Fano, and sub-Fano regimes for particular parameter values [15], but these parameter values lacked a biological interpretation. Here we supply such an interpretation and further show that the low-noise behavior is a consequence of negative correlation between two Poisson processes together with very rapid switching between the activated and repressed states of the gene.

[*] alex.ramos@usp.br

[†] Deceased.

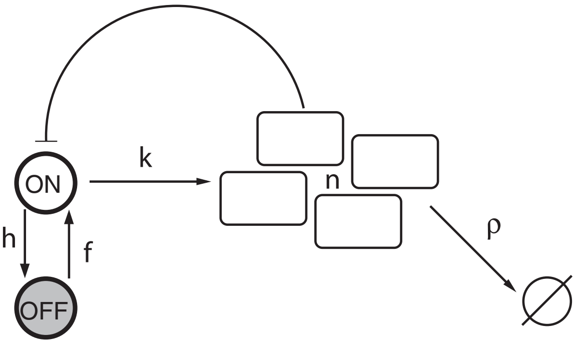

FIG. 1. Binary negative self-regulating gene. The white (gray) circle represents the gene at the on (off) state, also indicated by α (β). The arrows indicate the chemical reactions of protein synthesis, protein degradation, and off-on and on-off gene switching with rates as indicated in the kinetic scheme (3). Protein molecules are represented by the rounded rectangles, their number by n, and their destruction by $\varnothing$. The bar-terminated line denotes repression.

Stochastic model. We consider a stochastic model for a gene under negative self-regulation with transcription and translation treated as a single process (see Fig. 1). The state of the system is described by a joint probability distribution

$$\Pi(\alpha_n,\beta_n), \quad n = 0,1,\ldots, \tag{2}$$

where α_n and β_n denote, respectively, the probability of the gene being on (repressor not bound to the operator) and off (repressor bound to the operator) when there are n molecules of gene product present. This corresponds to the kinetic scheme

$$\begin{aligned} \varnothing &\overset{k}{\to} \mathcal{P}, \\ \mathcal{P} &\overset{\rho}{\to} \varnothing, \\ \mathcal{P} + \mathcal{O} &\overset{h}{\to} \mathcal{PO}, \\ \mathcal{PO} &\overset{f}{\to} \mathcal{P} + \mathcal{O}, \end{aligned} \tag{3}$$

where the first equation describes the synthesis of a single molecule of the product $\mathcal{P}$ of gene $\mathcal{P}$ with rate k. Without loss of generality, we refer to $\mathcal{P}$ as a protein in what follows, but the kinetic scheme (3) could also describe an autorepressing gene that only synthesized RNA. The second of Eqs. (3) describes the annihilation of $\mathcal{P}$ by a first-order process of rate ρ. The third equation describes the binding of protein $\mathcal{P}$ to the operator $\mathcal{O}$ with rate h, while the fourth equation describes the unbinding of $\mathcal{P}$ from $\mathcal{O}$ with rate f.

Given this kinetic scheme, the joint probability $\Pi(\alpha_n,\beta_n)$ and the probability of finding n molecules in the system $\phi_n = \alpha_n + \beta_n$ can be found by solving the master equation

$$\frac{d\alpha_n}{dt} = k(\alpha_{n-1}-\alpha_n)+\rho[(n+1)\alpha_{n+1}-n\alpha_n] - hn\alpha_n + f\beta_n, \tag{4}$$

$$\frac{d\beta_n}{dt} = \rho[(n+1)\beta_{n+1} - n\beta_n] + hn\alpha_n - f\beta_n, \tag{5}$$

where $n,\alpha_n,\beta_n,k,\rho,h,f$ are as defined above. The term $hn\alpha_n$ indicates the repressive action of the protein $\mathcal{P}$.

At the steady-state limit ϕ_n is given by [15]

$$\phi_n = C\frac{(a)_n}{(b)_n}\frac{(N_1 z_0)^n}{n!}\mathcal{M}\big(a+n,b+n,-N_1 z_0^2\big), \tag{6}$$

where $\mathcal{M}$ indicates the confluent hypergeometric function [16]. The symbol $(a)_n$ represents the Pochhammer function defined as $(a)_n = a(a+1)\cdots(a+n-1)$ and $(a)_0 = 1$. The constants a,b,N_1,z_0 are expressed in terms of the rate constants as

$$z_0 = \frac{\rho}{\rho+h}, \quad N_1 = \frac{k}{\rho}, \quad a = \frac{f}{\rho},$$
$$b = \frac{f}{\rho+h} + \frac{hk}{(\rho+h)^2}, \tag{7}$$

where C is a normalization constant with $C^{-1} = \mathcal{M}(a,b,N_1 z_0(1-z_0))$. The mean number of molecules is given by

$$\langle n\rangle = CN_1(az_0/b)\mathcal{M}(a+1,b+1,N_1 z_0(1-z_0)). \tag{8}$$

Biological interpretation. We now elucidate the biological significance of the exact solutions by considering specific values for the kinetic rates. Previously, we demonstrated that sub-Fano, Fano, and super-Fano regimes correspond to $a > b$, $a = b$, and $a < b$, respectively, where a and b are given in Eq. (7). A biologically realistic example of an autorepressing gene when $a = b$ is the synthesis of λ Cro protein from the P_R promoter under the control of an $O_R3^- O_R2^-$ operator. Reasonable estimates for the parameters considered here can be obtained from a much more detailed biophysical model, which implies values of about $h = 10^8$ min^{-1}, $f = 0.4$ min^{-1}, $\rho = 0.01$ min^{-1}, and $k = 4 \times 10^9$ min^{-1} [17]. For this set of values the Fano factor is equal to one and the mean number of proteins is 40. This appears to conform to the expectation that gene regulation processes have $\mathcal{F} \geqslant 1$.

Sub-Fano regime. A cell with *lac* operon in a repressed state contains approximately 20 molecules of *lac* repressor. A set of parameters corresponding to this mean protein number is given by $h = 10^8$ min^{-1}, $f = 4 \times 10^3$ min^{-1}, $\rho = 0.01$ min^{-1}, and $k = 1.0 \times 10^5$ min^{-1}. For this set of values the variance is 10.0, indicating a Fano factor of 0.5 and a signal-to-noise ratio 6.3 compared to a signal-to-noise ratio for a Poisson process of 4.5, an increase of $\sqrt{2}$ over the Poisson case. With respect to the parameters, ρ has a reasonable value, but k appears to be unbiologically large. In fact, the actual synthesis rate is much less than k. Here h and f describe very rapid transitions between the on and off states, but because $h \gg f$, the on state has a very low probability of occurring. For this combination of parameters the expectation of finding the gene on is $\sum_{n=0}^{\infty}\alpha_n = 2 \times 10^{-6}$, a high level of repression.

This sub-Fano behavior is quite generic. Figure 2 shows that as $\langle n\rangle$ increases, $\mathcal{F}$ tends to an asymptotic value of 0.5 for a wide variety of parameter values that obey the condition $a > b$.

Infra-Fano regime. We have discovered a regime, called infra-Fano, in which $\mathcal{F}$ can be arbitrarily small. Inspection of Fig. 2 reveals a minimum value for $\mathcal{F}$ when $\langle n\rangle = 1$. A set of parameters resulting in $\langle n\rangle = 1.0$ and $\mathcal{F} = 5.0 \times 10^{-3}$ is given by $h = 10^{10}$ min^{-1}, $f = 3$ min^{-1}, $\rho = 0.01$ min^{-1}, and $k = 1.0 \times 10^5$ min^{-1}. This represents a situation in which a

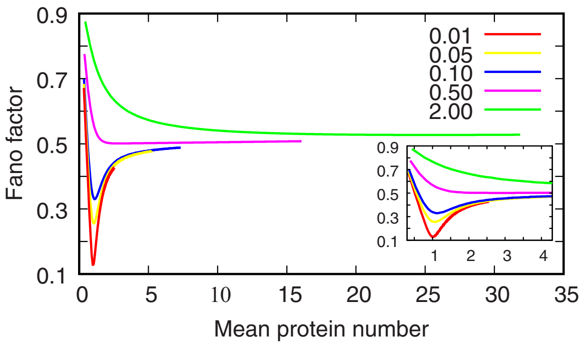

FIG. 2. (Color online) Fano factor versus the mean protein number. Here we fixed $a = 500$ with different colors standing for fixed values of b as indicated by the key. Each curve corresponds to a fixed value of b and variation of z_0. For fixed a and b, $\langle n \rangle$ depends only on z_0.

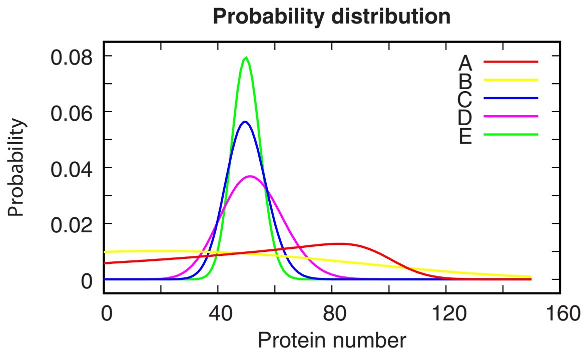

FIG. 3. (Color online) Probability distribution of the protein number. The probability distribution ϕ_n is shown for increasing levels of coupling. Curves A–E show probability distributions corresponding to increasing values of f and h. As the coupling gets stronger the variance of the distribution decreases. The parameter values (a,b,z_0) for each curve are A, (1.0,2.0,0.99); B, (1.0,15.0,0.95); C, (14.0,70.0,0.5); D, (50.0,50.0,0.5); and E,$(5 \times 10^3,1.0,10^{-4})$. Equation (7) shows that f is linearly dependent on a, while h is inversely proportional to z_0.

single protein molecule is bound to the operator and there are no protein molecules in the cytosol. On those rare occasions when the protein molecule dissociates from the operator it immediately binds again or, if it degrades, new protein is rapidly synthesized and binds. Note that ρ and k have the same value as the previous example with $\langle n \rangle = 20$ and $\mathcal{F} = 0.5$. The cause of the infra-Fano behavior is the ratio between f and h. At a thermodynamic level, this ratio is equal to the equilibrium binding constant K and represents extremely tight binding of the protein.

Mechanism of sub- and infra-Fano behavior. In order to understand the theoretical basis of sub-Fano and infra-Fano behavior, it turns out to be useful to consider the covariance between the state of the gene and the number of protein molecules present. The Fano factor can be written in a simple way using this covariance and the covariance can in turn be written in the terms of the exact solution. We do this by first defining the discrete two-valued random variable $N = \{N_1, 0\}$, with N_1 given by Eq. (7). Here N_1 and 0 represent the asymptotic mean protein number if the gene were entirely on or off, respectively. The probability that $N = N_1$ is given by $p_1 = \sum_{n=0}^{\infty} \alpha_n$. This probability coincides with the probability of the gene being on (or off). Hence, the mean value of N is

$$\langle N \rangle = p_1 N_1. \tag{9}$$

The covariance of n and N is

$$\xi = \langle nN \rangle - \langle n \rangle \langle N \rangle. \tag{10}$$

In Ref. [18] we show that

$$\mathcal{F} = 1 + \frac{\xi}{\langle n \rangle}, \tag{11}$$

where $\langle n \rangle$ is a strictly positive quantity and ξ may be positive, zero, or negative causing $\mathcal{F}$ to be greater than, equal to, or less than one, that is, super-Fano, Fano, or sub-Fano. We further show in Ref. [18] that the covariance can be written in terms of the exact solutions of the model as

$$\xi = a z_0 \frac{N_1 - \langle n \rangle}{1 - z_0} - \langle n \rangle^2, \tag{12}$$

where the formula for $\langle n \rangle$ is given in (8). The set of parameters of the example for the sub-Fano regime results in a covariance ξ of -10.0, while ξ is -0.995 for the set of values of the parameters resulting in an infra-Fano regime.

Fast on-off switching. We pointed out above that infra-Fano behavior arises in the case where a single repressor is bound to the operator. A further reason that sub- and infra-Fano behaviors arise can be seen by considering the switching rates of the gene between active and repressed states. The switching rates are high compared to the decay rate of the protein. Were that not the case, the repressor could bind for a period comparable to or longer than the protein half-life, resulting in an appreciable decline in protein number during the repressor's dwell time at the operator. The opposite is also true. In the sub-Fano example given above, $k = 1.0 \times 10^5$ min^{-1}, but the average expression rate was less than 10^{-6} of k. If the repressor was unbound for a period comparable to the protein half-life, large spikes of protein concentration would result, greatly increasing $\mathcal{F}$.

The biological interpretation of the sub- and infra-Fano behavior can be understood in terms of the values of the rate parameters for which the sub- Fano regime occurs. The value of the binding rate h is increased so that p_1 approaches 0.

During the small fraction of time when the gene is on, little protein synthesis occurs. As the degradation rate is also very small, the net variation in protein number is also small and its probability distribution becomes very sharp.

The reduction of the variance is shown in Fig. 3, where we have plotted the probability of finding n proteins inside the cell. It is evident that the variance of these distributions decreases greatly as switching rate increases over the sequence of curves from A to E. Curves A–C are super-Fano, D is Fano, and E is sub-Fano.

For dimensional reasons, noise is often characterized by the ratio of the variance to the squared mean, a quantity known as the coefficient of variation $CV^2 = \mathcal{F}/\langle n \rangle$. Our findings apply to CV^2 as well as $\mathcal{F}$. Inspection of Eq. (11) shows that for small values of $\langle n \rangle$ and for $\xi \sim -\langle n \rangle$ one may still obtain low values for CV^2. Because $\lim_{\langle n \rangle \to +\infty} CV^2 = 0$, for large values of $\langle n \rangle$ the Fano and sub-Fano regimes are rendered indistinguishable by the onset of deterministic behavior.

Coupling. Equations (4) and (5) are a coupling of two different Poissonian processes, each of them related to one of the gene states. Hence, we analyze the Fano factor in terms of the covariance between n and N. One Poisson process is the protein synthesis and degradation when the gene is on, while the other represents the same process when the gene is off. The second random variable is the gene state, which is the variable that couples the two processes.

The sub-Poissonian regime is the result of the combination of two noisier Poisson processes with negative ξ. One may conclude that negative covariance induces noise reduction in the composition of stochastic processes. This coupling regime may be the mechanism underlying the higher precision of the negative self-regulating gene [19,20].

Experimental implications. The kinetic scheme (3) leading to Eqs. (4) and (5) is a coarse graining of a more complex set of elementary reactions. In both prokaryotes and eucaryotes, all reactions taking place between the initiation of transcription and the binding of the repressor are coarse grained away. In a prokaryote both elongation and translation are neglected. The well known geometric distribution of the initiation of bacterial translation will greatly alter the statistics of the sub-Fano regime reported here in a real system. Coarse graining does not affect the interpretation of the infra-Fano regime in prokaryotes, however, because with a single protein bound to the operator, translation and transcription do not take place. Moreover, the equilibrium constant $K = f/h = 3 \times 10^{-11} M$ implied by the infra-Fano parameter set is well within the range of affinities for reversible binding reactions involving proteins [21]. For this reason we expect that real prokaryotic systems exist with operators that function in the infra-Fano regime.

With respect to the control of transcription initiation in eukaryotes, particularly the metazoa, the equations may have predictive value. In these organisms the elementary chemical reactions underlying transcription are not well understood because they are vectorial in that the spatial arrangement of reactants is important, they occur far from equilibrium, and they involve at least 58 polypeptides [22]. There is evidence [10] that *Drosophila* promoters have discrete on and off states as in the system considered here and that miRNA-protein complexes reduce noise in a manner not mechanistically well understood [23,24]. Suppose the substance $\mathcal{P}$ in kinetic scheme (3) were such a noise-reducing gene product and the chemical species $\mathcal{P} + \mathcal{O}$ and $\mathcal{OP}$ were interpreted as two allosteric states of a transcription complex, the former permitting and the latter forbidding initiation. Then the low level of noise seen in the sub-Fano and infra-Fano regimes described here would be a consequence of rapid state changes in the transcription complex that occur far from equilibrium. This amounts to a mechanism of stochastic focusing of gene expression, in which fast fluctuations in the state of a transcription complex reduce fluctuations in the synthesis of gene product.

Compared with prokaryotes, in eukaryotes many more steps occur between transcription initiation and the binding of repressor that are not represented in (4) and (5). These processes include RNA splicing, capping, and polyadenylation, as well as the transport of molecules between the nuclear and cytoplasmic compartments. This point is relevant to the interpretation of the results presented here. The sub- and infra-Fano regimes occur in the limit where the transition from on to off state is much faster than the transition from off to on. As a result, the gene spends most of its time in the off state with a small effective synthesis rate that, when combined with slow protein degradation, reduces fluctuations. We point out that this low synthesis rate could also arise from any postinitiation event in protein synthesis, including translation. The decrease of noise with decreased rates of transcription or translation has previously been predicted and experimentally confirmed in yeast [7].

Theoretical implications. Our discovery of the infra-Fano regime demonstrates the importance of exact solutions in providing physical insight into the behavior of stochastic systems far from equilibrium. Exact solutions of the Fokker-Planck equation provided important insights into the bistability of λ lysogens [25,26]. These approaches did not reveal the infra-Fano regime because the Langevin–Fokker-Planck approximation breaks down as molecular numbers approach one [27]. A seminal work exploring the relative noise contributions of transcription and translation used exact solutions of a system in which regulation is represented as a linear dependence on the synthesis rate [14]. Here we represent regulation as an inducer of the switching from the on to the off state of protein synthesis. This formulation permits multiple transitions of the gene state without significant changes of protein numbers. Rapid switching alone is not sufficient to ensure noise reduction without negative correlation. Peccoud and Ycart's approach [28], based on a model similar to ours but with external regulation, did not produce a low-noise regime because this regulation caused the covariance between the protein number and the gene state to obey $\xi \geqslant 0$.

Acknowledgments. A.F.R. was supported by CNPq, FAPESP, and NIH Grant No. RO1 OD010936 (formerly RR07801). J.R. was supported by NIH Grant No. RO1 OD010936 (formerly RR07801). J.E.M.H. was supported by CNPq and FAPESP. J.R. is grateful to the Graduate School on Modeling of Complex Systems, EACH, USP for hospitality.

[1] M. Delbrück, J. Chem. Phys. **8**, 120 (1940).

[2] A. Arkin, J. Ross, and H. H. McAdams, Genetics **149**, 1633 (1998).

[3] D. B. C. J. Morton-Firth and T. S. Shimizu, J. Mol. Biol. **286**, 1059 (1999).

[4] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).

[5] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).

[6] L. Cai, N. Friedman, and X. S. Xie, Nature (London) **440**, 358 (2006).

[7] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, Nature (London) **422**, 633 (2003).

[8] W. J. Blake, G. Balázi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins, Mol. Cell **24**, 853 (2006).

[9] T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek, Cell **130**, 153 (2007).

[10] A. N. Boettiger and M. Levine, Science **325**, 471 (2009).

[11] Q. Wills, K. Livak, A. Tipping, T. Enver, A. Goldson, D. Sexton, and C. Holmes, Nat. Biotechnol. **31**, 748 (2013).

[12] M. Sasai and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **100**, 2374 (2003).

[13] J. E. Hornos, D. Schultz, G. C. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Phys. Rev. E **72**, 051907 (2005).

[14] M. Thattai and A. van Oudenaarden, Proc. Natl. Acad. Sci. USA **98**, 8614 (2001).

[15] A. F. Ramos and J. E. M. Hornos, Phys. Rev. Lett. **99**, 108103 (2007).

[16] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (U.S. GPO, Washington, DC, 1972).

[17] J. Reinitz and J. R. Vaisnys, J. Theor. Biol. **145**, 295 (1990).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.91.020701 for demonstration.

[19] A. Becskei and L. Serrano, Nature (London) **405**, 590 (2000).

[20] D. Nevozhay, R. M. Adams, K. F. Murphy, K. Josic, and G. Balázsi, Proc. Natl. Acad. Sci. USA **106**, 5123 (2009).

[21] N. M. Green, Biochem. J. **89**, 585 (1963).

[22] R. Kornberg, Proc. Natl. Acad. Sci. USA **104**, 12955 (2006).

[23] E. Hornstein and N. Shomron, Nat. Genet. Suppl. **38**, S20 (2006).

[24] H. Herranz and S. M. Cohen, Genes Dev. **24**, 1339 (2010).

[25] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins, Proc. Natl. Acad. Sci. USA **97**, 2075 (2000).

[26] F. J. Isaacs, J. Hasty, C. R. Cantor, and J. J. Collins, Proc. Natl. Acad. Sci. USA **100**, 7714 (2003).

[27] D. T. Gillespie, J. Chem. Phys. **113**, 297 (2000).

[28] J. Peccoud and B. Ycart, Theor. Popul. Biol. **48**, 222 (1995).