Self-regulating gene: An exact solution

J. E. M. Hornos,[1,2] D. Schultz,[1] G. C. P. Innocentini,[2] J. Wang,[3] A. M. Walczak,[1] J. N. Onuchic,[1] and P. G. Wolynes[1]
[1]*Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92093-0374, USA*
[2]*Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, BR-13560-970 São Carlos, S.P., Brazil*
[3]*Department of Chemistry, State University of New York, Stony Brook, New York 11794, USA*
(Received 22 December 2004; published 4 November 2005)

An exact steady-state solution of the stochastic equations governing the behavior of a gene regulated by a self-generated proteomic atmosphere is presented. The solutions depend on an adiabaticity parameter measuring the relative rate of DNA-protein unbinding and protein degradation. The steady-state solution reveals deviations from the commonly used Ackers *et al.* approximation based on the equilibrium law of mass action, allowing anticooperative behavior in the "nonadiabatic" limit of slow binding and unbinding rates. Noise from binding and unbinding events dominates the shot noise of protein synthesis and degradation up to quite high values of the adiabaticity parameter.

DOI: 10.1103/PhysRevE.72.051907 PACS number(s): 87.16.Ac, 87.10.+e, 87.16.Yc

INTRODUCTION

Production of functional biomolecules in the cell is governed by a complex and diverse genetic network involving an intricate set of biochemical reactions. The mathematical description of this network is intrinsically nonlinear because the transcription of DNA is regulated by the binding reactions with the very protein products of the decoding process itself [1]. This description must also be stochastic because the genes are single molecules of DNA and their regulatory proteins are also present often in small numbers. The average behavior of a nonlinear, stochastic system cannot be inferred from macroscopic chemical rate laws alone [2–13]. In this paper we examine the simplest model of an element of a gene regulatory network and show that its master equation admits an exact solution. In regimes where the binding and unbinding process is not significantly faster than the synthesis and degradation of the proteins, this solution is quantitatively different from the deterministic description [15–17].

In deterministic models of gene expression the concentration of various transcription factors controls the rate of protein production for a particular gene [17–19]. The stochastic analysis of gene switches treats the numbers of these various proteins, $n_1,\ldots,n_N$, in a given cell as random variables [20–24]. If we ignore the mechanistic details of protein biosynthesis with their resulting time delays and mRNA fluctuations [12,13], we can model each gene as a two-state stochastic system. A single gene can then be described by a two-component master equation with one probability distribution $\alpha(n_1,\ldots)$ corresponding to situations where the DNA is free (*on* state) and a second component $\beta(n_1,\ldots)$ describing the distribution when the DNA has a repressing protein bound to it (*off* state). The dynamics of these genetic expression probabilities is described by coupled birth-death processes. Birth corresponds to protein synthesis while death occurs via degradation. The rates for protein production g_α and g_β are different for the free and bound states of the DNA. The rate for protein degradation is k, varying linearly with n. If the binding state of the DNA did not change, the stationary probabilities α and β would be described by Poisson distributions with mean values at g_α/k and g_β/k. We show that the time evolution from any initial state of this simple self-repressing switch to the stationary configuration can be written explicitly in terms of hypergeometric functions as in the theory of the threshold switch [3].

STOCHASTIC FORMULATION

In the present model a single gene produces the same protein that represses its own activity. While not often found as an isolated entity, the self-regulating gene is a very common element of biological networks; for example, 40% of *E. Coli* transcription factors negatively regulate their own transcription [14]. The master equations for this case are explicitly

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

$$\frac{d\beta_n}{dt} = g_\beta[\beta_{n-1} - \beta_n] + k[(n+1)\beta_{n+1} - n\beta_n] + hn\alpha_n - f\beta_n \text{ for } n \geqslant 2, \tag{2}$$

where α_n and β_n are the individual probabilities that the DNA is unbound and bound, respectively, while immersed in n proteins. h is the bimolecular rate describing the process of repressor binding to the DNA, and f is the unimolecular rate describing release of the repressor protein from the repressor site. More generally, h can be a more complicated function of n if, for example, proteins bind as oligomers [2]. In this case, we consider a mechanism of monomer binding. The binding and unbinding process does not alter the total number of proteins. Since a bound protein is still included in n, there is a need to modify the master equation for the states near $n=0$. The gene cannot be in a bound state in which there are no proteins in the system ($\beta_0=0$). Thus we will use a set of equations in which a degradation reaction will transform the state where the only existing protein is bound (β_1) into the unbound state α_0:

$$\frac{d\alpha_0}{dt} = -g_\alpha\alpha_0 + k[\alpha_1 + \beta_1], \tag{3}$$

$$\frac{d\beta_1}{dt} = -g_\beta\beta_1 + k[2\beta_2 - \beta_1] + h\alpha_1 - f\beta_1, \tag{4}$$

$$\frac{d\alpha_1}{dt} = g_\alpha[\alpha_0 - \alpha_1] + k[2\alpha_2 - \alpha_1] - h\alpha_1 + f\beta_1. \tag{5}$$

AN EXACT SOLUTION

The master equations are differential and difference equations for t and n, respectively. The two sets of master equations need to be solved in the appropriate subspaces of n. The general solution may then be determined using the continuity condition at $n=2$. The solution of Eqs. (1) and (2) can be described in terms of the generating functions $\alpha(z) = \Sigma_{n=0}^{\infty}\alpha_n z^n$ and $\beta(z) = \Sigma_{n=0}^{\infty}\beta_n z^n$, where z lies in the complex unitary circle. The original probabilities for $n \geqslant 2$ can be recovered as derivatives of these generating functions at $z=0$:

$$\alpha(n) = \frac{1}{n!}\frac{d^n}{dz^n}\alpha(t,z)$$

and

$$\beta(n) = \frac{1}{n!}\frac{d^n}{dz^n}\beta(t,z).$$

The correct probabilities for the states in which $n<2$ are calculated by using α_2 and β_2 derived from the generating functions in the modified master equations (3)–(5). Various moments of the distribution, including the average number of proteins, can still be expressed in terms of derivatives of these generating functions $(\partial/\partial z)\alpha(z,t)$ and $(\partial/\partial z)\beta(z,t)$ evaluated at $z=1$. Before taking into account the boundary behavior the generating functions satisfy the first-order partial differential equations

$$\frac{\partial\alpha(z,t)}{\partial t} = (z-1)\left[g_\alpha\alpha(z,t) - k\frac{\partial\alpha(z,t)}{\partial z}\right] - hz\frac{\partial\alpha(z,t)}{\partial z} + f\beta(z,t), \tag{6}$$

$$\frac{\partial\beta(z,t)}{\partial t} = (z-1)\left[g_\beta\beta(z,t) - k\frac{\partial\beta(z,t)}{\partial z}\right] + hz\frac{\partial\alpha(z,t)}{\partial z} - f\beta(z,t). \tag{7}$$

The stationary solution of this system of equations is easily obtained. From Eq. (6), we can find β as a function of α and $d\alpha(z)/dz$. Substituting this expression for β in Eq. (7), a second-order differential equation is obtained:

$$\frac{d^2\alpha(z)}{dz^2} + p\frac{d\alpha(z)}{dz} + q\alpha(z) = 0, \tag{8}$$

where the coefficients p and q are

$$p = \frac{g_\alpha + g_\beta + f + h + k - z(g_\beta(1+h/k) + g_\alpha)}{(k+h)z - k}, \tag{9}$$

$$q = \frac{g_\alpha g_\beta z - g_\alpha(g_\beta + f + k)}{k(k+h)z - k^2}. \tag{10}$$

Noting that p and q are rational functions of z with a simple pole at $z_0 = k/(k+h)$ and an irregular singularity at $z=\infty$, we see that the structure of this equation corresponds to the confluent hypergeometric equation.

The dependence on z in the numerator of q can be eliminated by making the transformation

$$\alpha(z) = A\exp(zg_\beta/k)M(a,b,\eta), \tag{11}$$

which leads to the confluent hypergeometric equation in a canonical form. The normalization constant A guarantees that the sum of the probabilities is 1. The solutions are linear combinations of the Kummer functions M and U. The irregular function U does not satisfy the condition $\alpha_n \to 0$ when $n \to \infty$ and therefore is discarded. The resulting generating function α has the Kummer $M(a,b,\eta)$ parameters

$$a = 1 + \frac{f}{k+h}\left(1 + \frac{hg_\alpha}{kg_\alpha - (k+h)g_\beta}\right), \tag{12}$$

$$b = 1 + \frac{f}{k+h} + \frac{hg_\alpha}{(k+h)^2}, \tag{13}$$

and the argument of the function is

$$\eta = -\frac{[g_\beta(1+h/k) - g_\alpha][(k+h)z - k]}{(k+h)^2}. \tag{14}$$

As described above, α_n's for $n \geqslant 2$ can be calculated from the derivatives at $z=0$. Explicitly these are [25]

$$\alpha_n = \frac{A}{n!}\sum_{s=0}^{n}\binom{n}{s}(g_\beta)^{n-s}\frac{d\eta_s}{dz}\frac{(a)_s}{(b)_s}M(a+s,b+s,\eta_0). \tag{15}$$

$\beta(z)$ can be calculated directly from Eq. (6), and the probabilities β_n for $n \geqslant 2$ are again derivatives at $z=0$. It is worth noticing that in the limit where there is no protein synthesis at all in the *off* state $(g_\beta = 0)$, there is only one nonzero term in the series for α_n $(s=n)$. This leads to a simple expression for

$$\alpha_n = \frac{A}{n!}\frac{d\eta_n}{dz}\frac{(a)_n}{(b)_n}M(a+n,b+n,\eta_0)$$

and

$$\beta_n = \frac{A(k+h)}{fn!}\frac{d\eta_n}{dz}\left[\left(\frac{hg_\alpha}{(k+h)^2} + b - 1\right)\frac{(a)_n}{(b)_n}M(a+n,b+n,\eta_0) - (b-1)\frac{(a-1)_n}{(b-1)_n}M(a-1+n,b-1+n,\eta_0)\right].$$

The normalization constant A is determined by $\Sigma_{n=0}^{\infty}\alpha_n + \Sigma_{n=0}^{\infty}\beta_n = 1$. These sums can be expressed in terms of $\alpha(1)$ and $\beta(1)$ and appropriate corrections to account for the states with $n<2$:

$$\alpha(1) + \beta(1) - \alpha(0) - \frac{d}{dz}\alpha(0) - \beta(0) - \frac{d}{dz}\beta(0) + \alpha_1 + \beta_1 + \alpha_0 = 1. \tag{16}$$

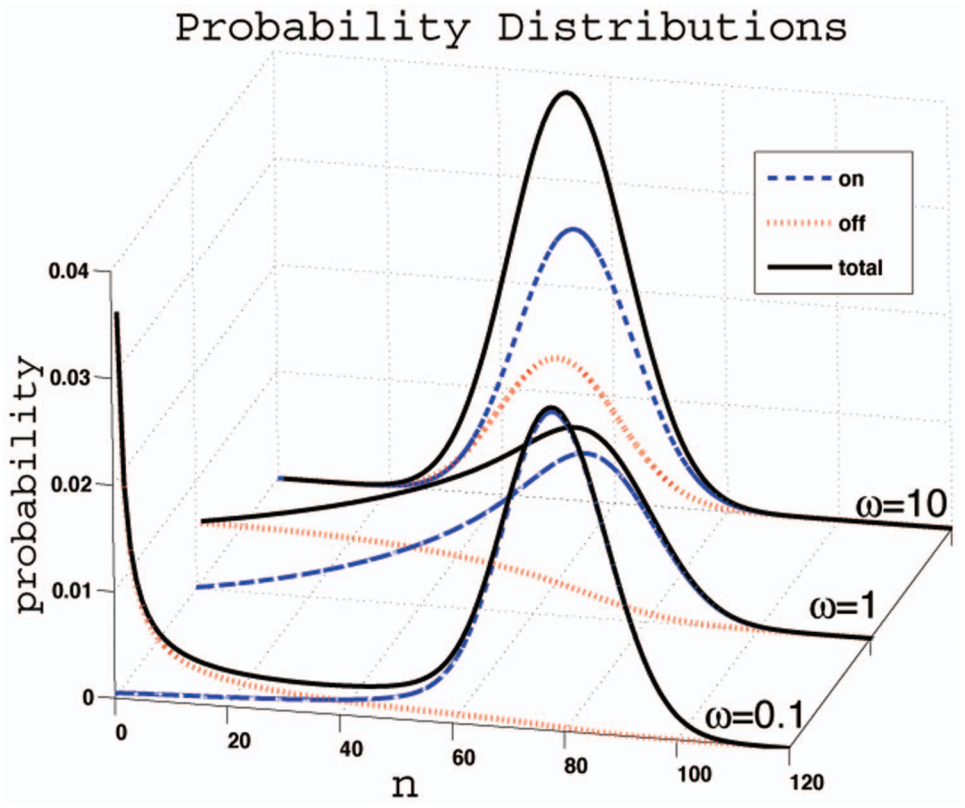

FIG. 1. (Color) The probabilities of the gene expression as a function of the number of proteins, n, for the on state, the off state, and the total. There are two peaks for small ω, but they converge to a single peak in the adiabatic regime of large ω. $X^{eq}=100$ and $X^{ad}=40$.

COMPARISON TO THE DETERMINISTIC MODEL

With these analytical solutions in hand, we are now in position to compare this exactly solved model with the commonly used deterministic mass-action approximation introduced by Ackers *et al.* [15]. To simplify the discussion we introduce the following parameters: $\omega=f/k$, $X^{eq}=f/h$, and $X^{ad}=(g_\alpha+g_\beta)/(2k)$. The parameter ω measures how rapidly the DNA state can equilibrate in its proteomic cloud in comparison to the characteristic time for protein degradation, which measures how fast the cloud itself fluctuates. X^{eq} is the

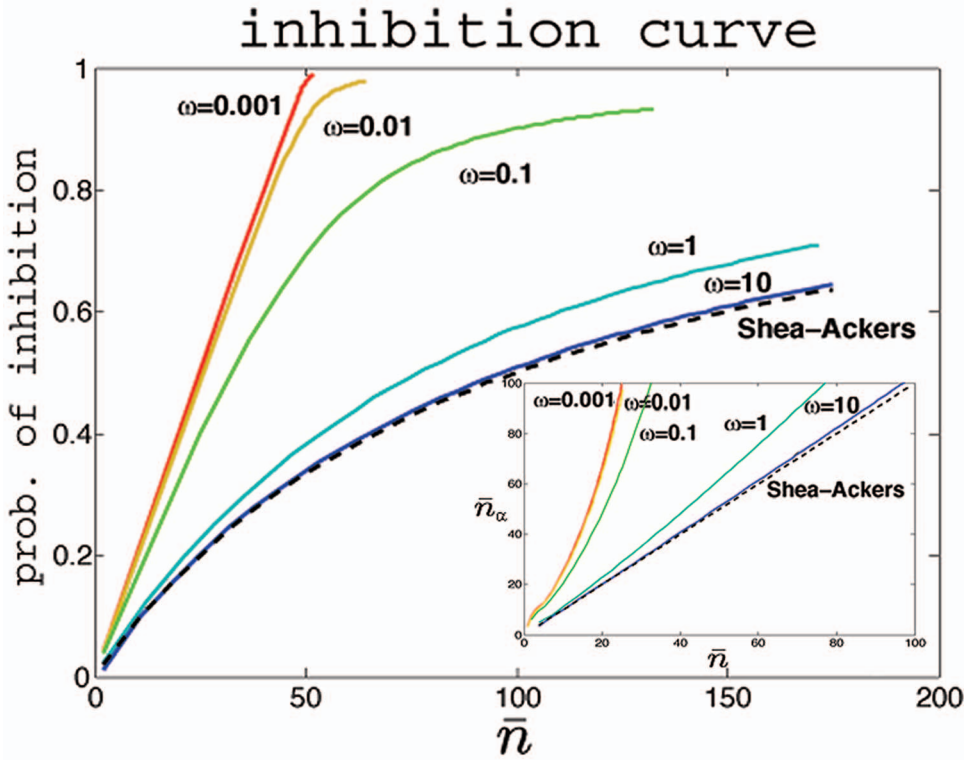

FIG. 2. (Color) Total probability of the DNA being found in the *off* state as a function of the average number of proteins, $\bar{n}$. In the adiabatic limit (large ω) we approach the behavior given by the equilibrium mass action law as in the treatment of Shea and Ackers, where $P_\beta=\bar{n}/(\bar{n}+X^{eq})$. $X^{eq}=100$. In our model we find $P_\beta=\overline{n_\alpha}/(\bar{n}_\alpha+X^{eq})$ exactly. The average number of proteins present when DNA is in the on state $\bar{n}_\alpha$ is different from $\bar{n}$, which includes the average number of proteins when the gene is off (inset).

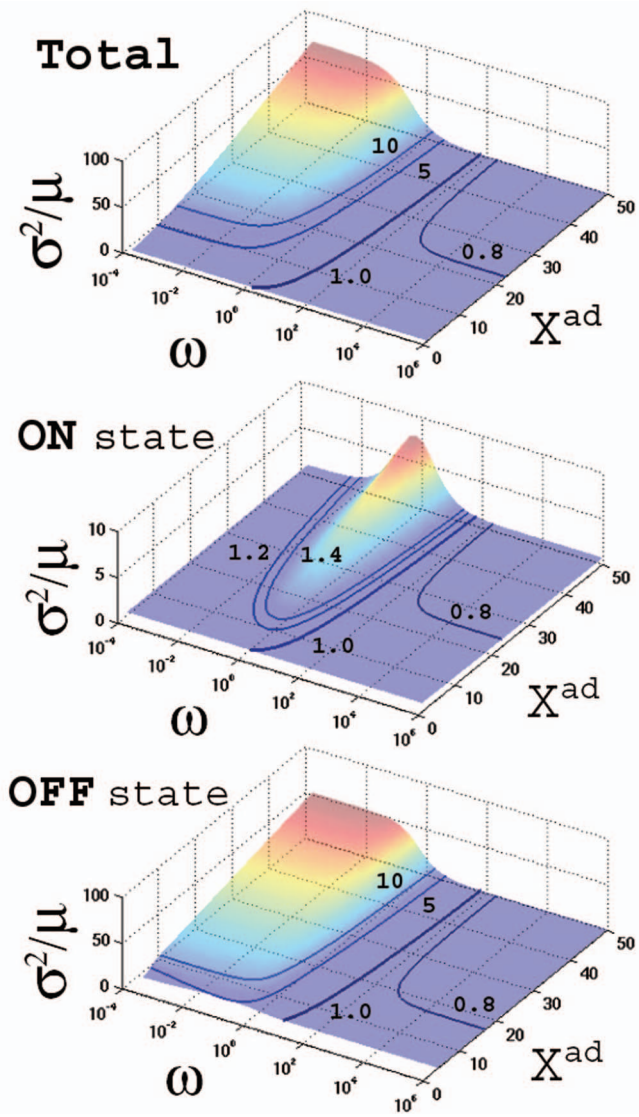

FIG. 3. (Color) Fano factor $F=\sigma^2/\mu$. Along the curve $F=1$, all distributions are Poisson like. The total distribution is independent of DNA state while the on state has the DNA free and the off state has protein bound to DNA. In the limit of large ω the adiabatic regime is reached, with an almost Poisson behavior. This regime should be equivalent to the Ackers *et al.* model. For intermediate ω's the overall fluctuations are large and therefore strongly deviate from Poisson behavior. In the *on* state, the distribution tends to Poisson behavior for ω very small, since the system behaves almost like a birth-death process. $X^{eq}=50$, $g_\beta=0$.

equilibrium constant of the binding and unbinding process. X^{ad} is a measure of the protein concentration, indicating the number of proteins when the system is half-inhibited.

The probability distributions for the protein number given the gene state (the total distribution $\alpha_n+\beta_n$, α_n for the on state and β_n for the off state) are shown in Fig. 1. The values of the switch characteristics used for the figure are $X^{eq}=100$ and $X^{ad}=40$ and $g_\beta=0$. These are typical values of the equilibrium switching threshold and mean protein copy number found in a small cell like *E. Coli*. For small values of ω the total probability distribution exhibits a two-peak structure, at g_β/k and g_α/k, corresponding to repressed protein production, when the DNA has protein bound at small n, and to the higher production from the free DNA at large n. In this limit, the *on* state behaves almost like an independent birth and death process since the binding and unbinding process becomes the slowest process in the system. Increasing the value of ω shifts both peaks to intermediate values, until there is only one peak at large ω. In the large-ω limit, the protein binding and unbinding process becomes extremely fast. This "adiabatic" regime should be equivalent to the Ackers *et al.* model in which the gene itself is taken to have an equilibrated average probability of being on or off. Most of the characterized genes are known to have high values of the adiabaticity parameter (e.g., when calculated from the transcription initiation rate obtained from [26]). Some systems exist, however, where ω is of order 1 (e.g., Cro protein in the λ-phage, parameters obtained from [27]). Also, the nonadiabatic regime may be important *in vivo*. For example, several *in vivo* mechanisms suggest that some proteins may be slow binders.

A more detailed understanding of the deviations from the Ackers *et al.* approximation can be made by noting that in the Ackers *et al.* model the probability of inhibition (P_β) is given by the equilibrium law of mass action as a function of the concentration of repressors. This concentration can be calculated using the first moments of the distribution $(d/dz)\alpha(z)$ and $(d/dz)\beta(z)$ at $z=1$, again with the corrections from the terms with $n<2$. Figure 2 shows how the exact solution for the master equation finally converges to the equilibrium approximation used by Ackers *et al.* $[P_\beta=\bar{n}/(\bar{n}+X^{eq})]$ in the limit of large ω.

To directly probe the effect of fluctuations, Fig. 3 shows the probability distributions compared to those that would arise from Poisson statistics: (a) independent of DNA state, (b) when the DNA is free (α_n), and (c) when the DNA is protein bound (β_n). The Fano factor $F=\sigma^2/\mu$ is plotted as a function of ω and X^{ad}, where μ and σ are the mean and standard deviations of the probability distributions. This factor would be 1 if the processes were purely Poisson processes. Notice that for very small ω, the Fano factor does limit to 1 when the DNA is in the on state. As discussed above, this is expected since, in this limit, the on state behaves almost like an independent birth and death process. The overall fluctuations are, however, quite large for intermediate ω's and therefore their contributions cannot be ignored in the overall mechanism. Indeed the Fano factor remains large even at ω values large enough for the probability of inhibition to agree with the equilibrium behavior. This shows DNA binding noise cannot be neglected.

In the large-ω regime (tending to the adiabatic limit), the Fano factor for the three distributions tends to values slightly smaller than 1. This indicates an almost Poisson behavior as one would expect for near-macroscopic kinetics.

DISCUSSION

The exact solution presented here for the self-regulated gene in a stationary regime establishes the basis for more complex problems yet to be solved. It provides an important analytical tool to understand the underlying mechanism gov-

erning these genetic networks. Already for this simple system, we notice that fluctuations become important for a large region of the parameter space. Figure 3 makes it clear that fluctuations cannot be ignored unless protein binding and unbinding are much faster than any other relevant time scale in the problem. Noise from binding and unbinding events dominates the shot noise of protein synthesis and degradation up to quite high values of the adiabaticity parameter. Figures 1 and 2 also demonstrate the effects of fluctuations. For small ω, binding is slow and therefore the stationary solution for the gene probabilities shown in Fig. 1 has two well-defined peaks. One peak corresponds to the repressed protein production (DNA with protein bound) and the other to the higher protein production (free DNA). As protein binding and unbinding become faster, these two peaks converge towards each other. Figure 2 shows how in this nonequilibrium system the probability of DNA being found in the protein-bound state deviates from the equilibrium mass action result. The self-repressing gene can become strongly anticooperative owing to nonadiabatic effects normally neglected in theories of gene regulation.

Some features of the genetic switch such as mRNA fluctuation and the time delays resulting from transcription and translation are not explicitly captured by this model. Although they might be essential in some cases, they may not always dominate the process of regulation. In prokaryotes, where there is no nucleus, transcription and translation occur within the same compartment, and mRNA is almost immediately translated [1]. Also many cases are being discovered where the regulation is performed by the RNA itself [28]. In cases like these, the approximation of having the synthesis of the transcription factors as one stochastic process seems plausible. This formulation of the problem of genetic regulation and its analytical solution will help the study of the specific cases where mRNA fluctuations and time delays play a determinant role.

While an otherwise isolated noninteracting self-regulating gene is a biological rarity, it would be straightforward to construct in the laboratory. The exact solution presented here would then make such an experiment a beautiful simplified system for understanding the importance of fluctuations that govern gene networks.

ACKNOWLEDGMENTS

This work was supported by the Center for Theoretical Biological Physics through National Science Foundation Grants Nos. PHY0216576 and PHY0225630. J.E.M.H. is supported by the Brazilian Agency FAPESP.

[1] M. Ptashne, *A Genetic Switch*, 2nd ed. (Cell Press and Blackwell Science, Cambridge, MA, 1992).
[2] M. Sasai and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 23742379 (2003).
[3] R. Metzler and P. G. Wolynes, Chem. Phys. **284**, 469 (2002).
[4] W. Bialek, Neural Comput. **13**, 2409 (2001).
[5] P. S. Swain, M. B. Elowitz, and E. D. Siggia, Proc. Natl. Acad. Sci. U.S.A. **99**, 1279512800 (2002).
[6] M. Thattai and A. van Oudenaarden, Proc. Natl. Acad. Sci. U.S.A. **98**, 86148619 (2001).
[7] N. E. Buchler, U. Gerland, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **100**, 51365141 (2003).
[8] A. Becskei, B. Seraphin, and L. Serrano, EMBO J. **20**, 2528 (2001).
[9] M. L. Simpson, C. D. Cox, and G. S. Sayler, J. Theor. Biol. **229**, 383 (2004).
[10] J. R. Pirone and T. C. Elston, J. Theor. Biol. **226**, 111 (2004).
[11] I. Bose, B. Ghosh, and R. Karmakar, Physica A **346**, 49 (2005).
[12] J. Paulsson, Nature (London) **427**, 415 (2004).
[13] P. J. Swain, J. Mol. Biol. **344**, 965 (2004).
[14] N. Rosenfeld, M. Elowitz, and U. Alon, J. Mol. Biol. **323**, 785 (2002).
[15] G. K. Ackers, A. D. Johnson, and M. A. Shea, Proc. Natl. Acad. Sci. U.S.A. **79**, 1129 (1982).
[16] J. Hasty, D. McMillen, F. Issacs, and J. J. Collins, Nat. Rev. Genet. **2**, 268279 (2001).
[17] P. J. Darling, J. M. Holt, and G. K. Ackers, J. Mol. Biol. **302**, 625638 (2000).
[18] A. Arkin, J. Ross, and H. H. McAdams, Genetics **149**, 16331648 (1998).
[19] E. Aurell, S. Brown, J. Johanson, and K. Sneppen, Phys. Rev. E **65**, 051914 (2002).
[20] T. B. Kepler and T. C. Elston, Biophys. J. **81**, 31163136 (2001).
[21] H. H. McAdams and A. Arkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 814819 (1997).
[22] J. Paulsson, O. G. Berg, and M. Ehrenberg, Proc. Natl. Acad. Sci. U.S.A. **97**, 71487153 (2000).
[23] B. E. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 11831186 (2002).
[24] P. Ao, J. Phys. A **37**, L25 (2004).
[25] *Handbook of Mathematical Functions*. Nat. Bur. Stand. Appl. Math. Series No. 55, edited by M. Abramowitz and I. A. Stegun (U.S. GPO, Washington, D.C., 1972).
[26] D. K. Hawley and W. R. McLure, J. Mol. Biol. **157**, 493 (1982).
[27] E. Aurell and K. Sneppen, Phys. Rev. Lett. **88**, 048101 (2002).
[28] J. Majewski and J. Ott, Genome Res. **12**, 1827 (2002).