Stochastic single-gene autoregulation

Tomás Aquino,[1] Elsa Abranches,[2,3] and Ana Nunes[1,*]

[1]*Centro de Física da Matéria Condensada and Departamento de Física, Faculdade de Ciências da Universidade de Lisboa, P-1649-003 Lisboa, Portugal*

[2]*Instituto de Medicina Molecular and Instituto de Histologia e Biologia do Desenvolvimento, Faculdade de Medicina da Universidade de Lisboa, Avenida Professor Egas Moniz, P-1649-028 Lisboa, Portugal*

[3]*Champalimaud Neuroscience Programme, Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, P-2780-156 Oeiras, Portugal*

(Received 16 December 2011; revised manuscript received 7 April 2012; published 19 June 2012)

A detailed stochastic model of single-gene autoregulation is established and its solutions are explored when mRNA dynamics is fast compared with protein dynamics and in the opposite regime. The model includes all the sources of randomness that are intrinsic to the autoregulation process and it considers both transcriptional and post-transcriptional regulation. The time-scale separation allows the derivation of analytic expressions for the equilibrium distributions of protein and mRNA. These distributions are generally well described in the continuous approximation, which is used to discuss the qualitative features of the protein equilibrium distributions as a function of the biological parameters in the fast mRNA regime. The performance of the time-scale approximation is assessed by comparison with simulations of the full stochastic system, and a good quantitative agreement is found for a wide range of parameter values. We show that either unimodal or bimodal equilibrium protein distributions can arise, and we discuss the autoregulation mechanisms associated with bimodality.

DOI: 10.1103/PhysRevE.85.061913 PACS number(s): 87.10.Mn, 02.50.Ey, 05.40.Ca

I. INTRODUCTION

The role of stochasticity in cells and microorganisms has been discussed theoretically since the 1970s [1,2]. Because cellular processes often rely on chemical reactions, and, correspondingly, on chance encounters between molecules or molecular complexes, stochastic effects due to small numbers are ubiquitous in the cell. In particular, cellular decision processes, which are of of paramount importance, as they allow cells to react to the internal and external media, are based on gene activation and regulation, often depending on random association and dissociation events. While many works focus on the limits imposed by stochasticity and the evolution of noise-minimization strategies [1,3–5], there is a growing interest in possible functional roles of noise. Generically, the basic role of randomness in gene expression is to provide a natural means of generating phenotype variability across a population, enhancing its capacity to quickly adapt to fast-changing conditions.

The evolution of experimental molecular biology techniques has made single-cell measurements possible and brought numerous confirmations of the presence of stochastic effects in gene expression [6], prompting a renewed interest in the mechanisms underlying gene expression and regulation, in general, and, specifically, in the sources of randomness affecting them. The fact that genes coding for specific proteins are often present in single copies may introduce considerable noise. Furthermore, mRNAs are commonly present at low copy numbers, from a few to a few hundred molecules, and many proteins also exist at low numbers. Because transcription, translation, and degradation events are stochastic, finite-size fluctuations in mRNA and protein numbers become important. Stochastic effects may suffice to drive long excursions of a gene's expression to higher or lower values, producing well-defined pulses in single-cell protein abundances over time and/or multimodal protein expression distributions in a population. Fluctuations of the biological parameters of the system under consideration are another source of randomness. For example, we characterize an active gene by a constant effective transcription rate, while this rate may depend on the presence of transcription factors whose concentration fluctuations induce fluctuations of the effective rate. Examples of theoretical approaches to these ideas can be found in [7–9].

The recent development of single-molecule techniques led to the experimental identification of another, more specific source of variability in gene expression that accounts for the heavy-tailed distributions often found in measures of population distributions of protein and mRNA abundance: both transcription and translation have been found, in many cases, to occur in time-localized bursts resulting in a geometrically distributed number of molecules (see [10–13]).

As experimental evidence of these sources of randomness accumulates [14,15], the tools of statistical physics are being called upon for the development of a theoretical understanding of the underlying mechanics in noisy gene expression. Several models of the simplest elements of a gene regulatory network have been studied as stochastic processes that include a representation of some of these sources of randomness [7,16–19]. As expected, the stationary solutions of these models may differ significantly from what one would obtain by simply adding a noise term to the equations stemming from a deterministic description. Moreover, the analytic solutions that can be obtained under certain assumptions were found to be in agreement with a wide set of experimental data [13].

In this paper, we make use of these tools to study a bottom-up model for single-gene auto-regulation that includes all the sources of randomness that are intrinsic to the autoregulation process and is applicable in general to any protein species, autoregulated by means either of transcriptional, as it is commonly considered, or of post-transcriptional regulation.

*anunes@ptmat.fc.ul.pt

Analytic solutions of the general model obtained in two complementary approximations for the relative timescales of protein and mRNA dynamics are discussed in terms of the qualitative features of the equilibrium protein distributions. The conditions for these approximations to hold are studied in some detail, and in their expected region of validity we find good quantitative agreement with the results of stochastic simulations of the full system. We use the analytic solutions to discuss the conditions under which single-gene autoregulation gives rise to bimodal protein distributions. Although these distributions are often associated in the literature with the presence of more complex regulation mechanisms, we find that those conditions are quite general.

The paper is organized as follows. In Sec. II, we establish the stochastic model. In Sec. III we present the solutions of the model for the protein and mRNA equilibrium distributions in the time-scale separation approximations, and we discuss the qualitative features of the former. In Sec. IV, we study the validity of the approximations and compare the approximate analytic solutions with the results of simulations. We conclude in Sec. V. The six appendices contain technical details which are too cumbersome to include in the main text.

II. MODEL

We study the cell-level dynamics, and corresponding population distributions, of a single protein capable of autoregulation and its mRNA. Protein and mRNA concentrations are controlled by the balance between production and degradation events. In transcriptional regulation (see Fig. 1, left arrow), the regulatory feedback is mediated by binding of a molecule, whose concentration depends on that of the protein itself, to the promoter region in the DNA to alter the transcription rate of its mRNA. This is the most commonly studied mechanism of gene regulation, but other mechanisms have been reported in the recent literature that act post-transcription, at the mRNA rather than at the promoter level [20]. In this translational regulation scenario, the regulator molecule interacts with the mRNA to change its rate of protein production (see Fig. 1, right arrow). In what follows we derive the master equations that govern protein and mRNA abundances in both these scenarios, starting with transcriptional regulation.

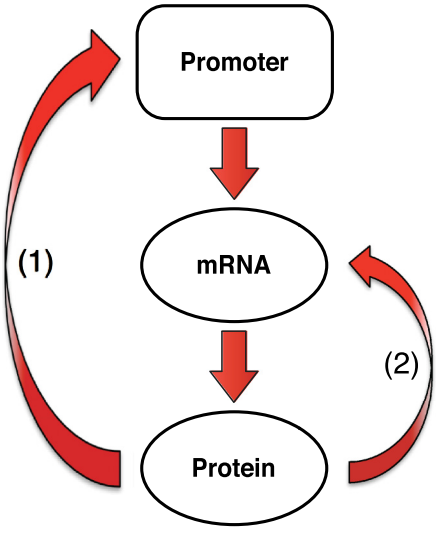

FIG. 1. (Color online) Basic structure of the dynamics of a single protein that autoregulates either (1) transcriptionally or (2) translationally. Arrows representing protein and mRNA degradation have been omitted.

For concreteness, we consider regulation to be effected by protein dimers, in agreement with experimental evidence for some particular proteins [21]. Other choices, such as monomer binding [16] and a general cooperative binding modeled by a Hill function [7], have been used in the literature. We assume the protein and mRNA populations to be noninteracting except for the fact that proteins dimerize prior to binding to the promoter. The time scale of promoter reactions (of the order of seconds) is assumed to be much shorter than that of the mRNA (of the order of minutes to hours) and protein (usually of the order of hours), in agreement with data for the typical time scales of these processes (see, e.g., [3] and [22]). Since regulation takes place at the DNA level, we need to model processes at three different levels: the promoter's (DNA), the mRNA's, and the protein's.

A. DNA level

For promoter dynamics, we essentially follow [23], adapted to a fully stochastic description. The promoter site is assumed to bind only one dimer molecule at a time. To avoid unnecessarily heavy notation, in what follows we assume a given value of protein copy number n in the cell; probabilities should accordingly be taken as conditional probabilities given n. Denote by P_f the free promoter state and by P_b the bound state of the promoter and a dimer. For each instant t, let $p(P_f,t)$ be the probability of the promoter's being free and $p(P_b,t)$ the probability of its being bound to a dimer. The evolution of the probability of the bound state is governed by the master equation

$$\dot{p}(P_b,t \mid j) = j\, k^+ p(P_f,t \mid j) - k^- p(P_b,t \mid j), \qquad (1)$$

where k^+ and k^- are the promoter site binding and unbinding rates, and j is the number of dimer molecules available for binding to the promoter. Since at all times the promoter is either free or bound to a dimer, we also have $p(P_f,t \mid j) + p(P_b,t \mid j) = 1$.

The number of dimers in the cell as a function of the protein copy number is given in a rate equation description by (see, e.g., [24] or Appendix A)

$$n_2(n) = \frac{n}{2} + a^2 - \sqrt{n + a^2}, \qquad (2)$$

where a is a dimensionless parameter defined by $a \equiv \sqrt{V/(8k_d)}$, V is the cell's volume, and k_d is the ratio of the dimerization and undimerization rates. If there are $n_2(n)$ dimers in the cell, the equilibrium probability distribution for the number j of dimers available through diffusion for binding to a promoter with characteristic volume V_P much smaller than the cell's volume V is given by [24]

$$\mathcal{P}_j(\lambda n_2(n)), \qquad (3)$$

where $\mathcal{P}_j(\theta)$ is the Poisson distribution of mean θ (evaluated at j), and $\lambda \equiv V_P/V \ll 1$. When writing down (3) we have taken into account that for typical values of protein (and protein dimer) diffusion coefficients and dimerization rates

(see [3], [25], and [26]), dimer formation and dissociation within the small volume V_P occur with a negligible probability compared to diffusion into and out of V_P. Note that λ is typically very small, since promoters have linear dimensions in the nanometer range and cells in the micrometer range. As discussed, for example, in [27], other transport mechanisms more efficient than three-dimensional diffusion must be at play that enable the promoter to gauge the actual number of molecules in the cell. Assuming that transport does not distinguish between dimers, and that the number of dimers does not influence the transport of a single dimer (essentially, that dimers are independent regarding transport, as is the case for diffusion), the distribution of dimers in V_P is binomial in general, with an "effective rate of volumes" parameter λ. In the relevant limit $\lambda \ll 1$ we regain (3).

We now explicitly take into account that the promoter time scale is much shorter than the protein time scale by assuming that the distributions $p(P_f,t)$, $p(P_b,t)$ have time to reach equilibrium for each fixed value of the number of proteins. Using the equilibrium dimer number distribution, we have

$$p^{\mathrm{eq}}(P) = \sum_{j\geqslant 0} p^{\mathrm{eq}}(P \mid j)\mathcal{P}_j(\lambda n_2(n)), \tag{4}$$

with $P \in \{P_f, P_b\}$. Solving Eq. (1) in equilibrium ($\dot{p} = 0$) and substituting in (4) leads to

$$p^{\mathrm{eq}}(P_f \mid n) = \sum_{j\geqslant 0} \frac{1}{1+kj}\mathcal{P}_j(\lambda n_2(n)), \tag{5a}$$

$$p^{\mathrm{eq}}(P_b \mid n) = \sum_{j\geqslant 0} \frac{kj}{1+kj}\mathcal{P}_j(\lambda n_2(n)), \tag{5b}$$

where we have introduced the dimensionless parameter $k \equiv k^+/k^-$ and we now emphasize the dependence on protein copy number n.

B. mRNA and protein levels

The production of mRNA and protein molecules in the cell has been found, in many cases, to occur in sharp geometrical bursts [10–13]. Although the concept of bursts and the mechanisms underlying them are still open to discussion (see, e.g., [28] and [29], a basic description stems from two simple ideas. First, if transcription and translation events are widely spaced compared to their duration, it is reasonable to speak of burst events. Second, the geometric distribution relates to the number of consecutive "heads" in the throwing of a (generally biased) coin; thus, if during a burst event there is a fixed probability that *another* molecule will be produced, a geometrically distributed number of molecules results. A major achievement of this burst description is that the resulting predicted form of unregulated protein expression distributions [7,30] is remarkably simple and fits an impressive number of experimental distributions measured for yeast populations [13]. We adopt here an approach in which bursts are formulated in a stochastic framework both for transcription and for translation.

Owing to the time-scale separation between promoter and mRNA/protein dynamics, for transcriptional regulation the latter are described in chemical reaction notation by

$$\begin{cases} \varnothing \xrightarrow{\beta_m f(n)} \mu_m m, \\ m \xrightarrow{\delta_m} \varnothing, \\ m \xrightarrow{\beta_p} m + \mu_p p, \\ p \xrightarrow{\delta_p} \varnothing. \end{cases} \tag{6}$$

Here m is the mRNA and p is the protein, while n stands for the protein copy number. f is the regulation function, such that

$$f(n) = \sum_{j\geqslant 0} \frac{1+\rho kj}{1+kj}\mathcal{P}_j(\lambda n_2(n)). \tag{7}$$

Thus, β_m is the transcription rate when the promoter is free, and $\rho\beta_m$ is the transcription rate when the promoter is bound to a dimer; the protein exhibits negative autoregulation (autoinhibition) if $\rho < 1$ and positive autoregulation (autoactivation) if $\rho > 1$; μ_m is the mean transcriptional burst size. With the burst scenario in mind, the transcription rates above are to be interpreted as the mean rates at which a transcription event takes place; this event is modeled as the instantaneous transcription of a certain number (drawn from a geometric distribution) of mRNA molecules. We assume here that regulation affects only the base transcription rate, and not the burst size. Finally, δ_m is the mRNA degradation rate. Similar definitions stand for the protein parameters (with β_p the translation rate, interpreted as the rate at which a single mRNA molecule initiates an instantaneous translational burst, μ_p the mean translational burst size, and δ_p the protein degradation rate).

It is interesting to see that the time-scale separation for promoter dynamics allows all details of regulation to be condensed in the regulation function. Different regulatory dynamics affecting only the transcription rate and obeying the same time-scale separation may be modeled in this framework simply by considering a different form of $f(n)$. Note also that a useful approximation to the regulation function as defined by (7) exists if $k \ll 1$. If $\lambda n_2(n)$ is small, the low j terms of the sum will dominate; Taylor expansion of the denominator to lowest order in kj (for $kj \ll 1$) and explicit calculation of the sum lead to

$$f(n) \approx \frac{1+\rho k\lambda n_2(n)}{1+k\lambda n_2(n)}. \tag{8}$$

If $\lambda n_2(n)$ is large, the large j terms dominate, and the approximation given by (8) remains valid because $(1+\rho k\alpha)/(1+k\alpha) \approx \rho$ for large α. Direct numerical calculation reveals that (8) is a good approximation overall, even for moderately large values of $k < 1$ (see Fig. 2).

Let $E_i(\theta) \equiv \frac{(\theta-1)^{i-1}}{\theta^i}$ be the geometric distribution of mean θ (evaluated at i), conditioned to nonzero values $i \geqslant 1$ because a burst of zero molecules has no physical meaning [33]. Let also $p_{j,n}(t)$ be the joint probability distribution of mRNA and protein copy numbers (evaluated at mRNA copy number j and protein copy number n) at time t. Then the master equation

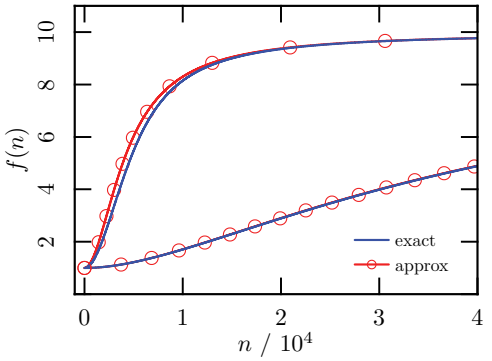

FIG. 2. (Color online) Approximation (8) for the regulation function. We fixed $\lambda = 10^{-2}$, $\rho = 10$, $a = 10^2$ for a typical example. Bottom curves, $k = 10^{-2}$; top curves, $k = 0.5$.

for process (6) reads

$$\dot{p}_{j,n}(t) = \Big[\beta_m f(n) \sum_{i\geqslant 1} E_i(\mu_m)\big(\mathbb{E}_m^{-i} - 1\big) + \delta_m(\mathbb{E}_m - 1)j + \beta_p j \sum_{i\geqslant 1} E_i(\mu_p)\big(\mathbb{E}_p^{-i} - 1\big) + \delta_p(\mathbb{E}_p - 1)n\Big] p_{j,n}(t), \tag{9}$$

where we have made use of the "step operators" $\mathbb{E}_m$, $\mathbb{E}_p$, defined by

$$\mathbb{E}_m^i g_{j,n}(t) = g_{j+i,n}(t), \quad \mathbb{E}_p^i g_{j,n}(t) = g_{j,n+i}(t), \tag{10}$$

for any function g depending on the mRNA copy number j, protein copy number n, and time t. With this notation, each term of the sum in the first term on the right-hand side of (9) stands for the creation of i mRNA molecules through a burst of size i, with bursts of arbitrary size occurring at rate $\beta_m f(n)$. The second term represents the degradation of one mRNA molecule, occurring at rate $\delta_m j$. Each term of the sum in the third term on the right-hand side stands for the creation of i protein molecules through a burst of size i, with bursts of arbitrary size occurring at rate $\beta_p j$. Finally, the last term in (9) represents the degradation of one protein molecule, occurring at rate $\delta_p n$.

C. Translational regulation

Consider now the case of translational regulation (Fig. 1, right arrow). In this case we assume that mRNA production proceeds through bursts without protein regulation and that the rate of production of protein bursts in translation is modulated by a regulation function $\tilde{f}(j,n)$ depending on mRNA and protein copy numbers and describing an interaction (direct or indirect) of the protein with its mRNA. Then mRNA and protein dynamics are described in chemical reaction notation by

$$\begin{cases} \varnothing \xrightarrow{\beta_m} \mu_m m, \\ m \xrightarrow{\delta_m} \varnothing, \\ m \xrightarrow{\beta_p \tilde{f}(j,n)} m + \mu_p p, \\ p \xrightarrow{\delta_p} \varnothing. \end{cases} \tag{11}$$

The master equation for process (11) then reads

$$\dot{p}_{j,n}(t) = \Big[\beta_m \sum_{i\geqslant 1} E_i(\mu_m)\big(\mathbb{E}_m^{-i} - 1\big) + \delta_m(\mathbb{E}_m - 1)j + \beta_p j \sum_{i\geqslant 1} E_i(\mu_p)\big(\mathbb{E}_p^{-i} - 1\big)\tilde{f}(j,n) + \delta_p(\mathbb{E}_p - 1)n\Big] p_{j,n}(t). \tag{12}$$

III. APPROXIMATE SOLUTIONS FOR THE EQUILIBRIUM DISTRIBUTIONS

In what follows, n and j always stand for protein and mRNA copy numbers, respectively. The coupling between mRNA and protein reactions leads to correlations between the random variables corresponding to n and j. As a result, the joint distribution $p_{j,n}$ does not factorize and separate master equations for n and j do not exist. Studying the solutions of the master equations (9) and (12) in general calls for direct numerical simulations of the dynamics or numerical integration techniques. However, further time-scale separations between mRNA and protein dynamics may be explored to simplify the problem. We say that mRNA is fast (compared to protein) if we can write the joint equilibrium distribution as

$$p_{j,n}^{\mathrm{eq}} = q_{j|n}^{\mathrm{eq}} p_n^{\mathrm{eq}}, \tag{13}$$

where $q_{j|n}^{\mathrm{eq}}$ is the equilibrium solution to the master equation for mRNA with fixed n. This means that mRNA dynamics are fast enough for a large number of mRNA-only reactions to take place before an n-changing reaction occurs, so that $q_{j|n}$ reaches equilibrium and the time spent out of equilibrium is negligible. Then, by substituting (13) in the appropriate general master equation and summing over j, we obtain an equation for p_n^{eq} independent of j. The physical idea is that for a certain n the mRNA will essentially sample the distribution $q_{j|n}^{\mathrm{eq}}$, and j-dependent quantities are correspondingly averaged over this distribution.

Similarly, we say that protein is fast if we may write

$$p_{j,n}^{\mathrm{eq}} = p_{n|j}^{\mathrm{eq}} q_j^{\mathrm{eq}}, \tag{14}$$

with analogous interpretations. In this case, for each j the $p_{n|j}^{\mathrm{eq}}$ distribution is sampled and n-dependent quantities are averaged over it.

We postpone to Sec. IV the analysis of the conditions under which such a separation holds as a good approximation and use it here to write down an equation for p_n^{eq} or q_n^{eq} from which approximate analytic expressions for the stationary solutions of the master equations (9) and (12) will be derived.

A. Transcriptional regulation under fast mRNA dynamics

In this section we consider transcriptional regulation for the case of fast mRNA compared to protein dynamics. We explore both the discrete scenario and a continuous approximation. It is convenient in this case to consider fixed n, since fast mRNA dynamics should allow the mRNA copy number to equilibrate for each fixed protein copy number. This means that we are considering the reactions

$$\begin{cases} \varnothing \xrightarrow{\beta_m f(n)} \mu_m m, \\ m \xrightarrow{\delta_m} \varnothing \end{cases} \tag{15}$$

at fixed n. Let $q_{j|n}(t)$ be the distribution of mRNA copy number (evaluated at j) at time t, given n. The master equation for this process has the simple form

$$\dot{q}_{j|n}(t) = \left[\beta_m f(n) \sum_{i\geqslant 1} E_i(\mu_m)\left(\mathbb{E}_m^{-i} - 1\right) + \delta_m(\mathbb{E}_m - 1)j \right] q_{j|n}(t). \tag{16}$$

Let $q^{\mathrm{eq}}_{|n}$ be the equilibrium distribution of mRNA copy number, for each protein copy number n. The mean value of mRNA corresponding to this distribution can be found to be (see Appendix B)

$$\langle id \rangle_{q^{\mathrm{eq}}_{|n}} = \mu_m \gamma_m f(n), \tag{17}$$

where $\gamma_m = \beta_m/\delta_m$, and id is the identity function.

In the protein time scale, we have the reactions

$$\begin{cases} m \xrightarrow{\beta_p} m + \mu_p p, \\ p \xrightarrow{\delta_p} \varnothing. \end{cases} \tag{18}$$

Let $p_n(t)$ be the distribution of the protein copy number (evaluated at n) at time t. According to (13), the master equation for this process reads

$$\begin{aligned} \dot{p}_n(t) &= \left[\sum_{\substack{j\geqslant 0,\\ i\geqslant 1}} \beta_p E_i(\mu_p)\left(\mathbb{E}_p^{-i} - 1\right) j q^{\mathrm{eq}}_{j|n} + \delta_p(\mathbb{E}_p - 1)n \right] \\ &\quad \times p_n(t) \qquad (19) \\ &= \left[r\delta_p \sum_{i\geqslant 1} E_i(\mu_p)\left(\mathbb{E}_p^{-i} - 1\right) f(n) + \delta_p(\mathbb{E}_p - 1)n \right] \\ &\quad \times p_n(t), \end{aligned}$$

where $r \equiv \mu_m\gamma_m\gamma_p$ and $\gamma_p \equiv \beta_p/\delta_p$. The parameter r is the prefactor of the average effective rate of translation burst events scaled by the degradation rate of the protein, $rf(n)$. We will see that, together with the average translational burst size μ_p, it determines the protein equilibrium distribution in this approximation.

As expected, when mRNA is fast, protein dynamics depends at each time only on the average mRNA corresponding to the available protein number n. Specifically, the translation rate becomes proportional to $\langle id \rangle_{q^{\mathrm{eq}}_{|n}}$, which is in turn proportional to $f(n)$. Through this mechanism, promoter-level regulation gauges the number of proteins present in the cell at a certain time. Note also that further details of mRNA dynamics, including burst-like production, are lost at the level of protein.

Let us consider as well a continuous approximation of the dynamics. For this we take $x \equiv \lambda n$ as an "approximately continuous" variable (recall that $\lambda \ll 1$). A continuous master equation for the distribution $p(x,t)$ of protein "concentration" x reads (see Appendix C)

$$\begin{aligned} \dot{p}(x,t) = r\delta_p \int_0^x f(y)[E(x-y,\tilde{\mu}_p) - \delta_D(x-y)]p(y,t)dy \\ + \delta_p\, \partial_x[xp(x,t)], \end{aligned} \tag{20}$$

where $E(x,\theta) \equiv (1/\theta)e^{-x/\theta}$ is the exponential probability distribution of mean θ evaluated at x and δ_D is the Dirac δ. For simplicity we have chosen to keep the symbol f, such that $f(x) = f(n)$ for $x = \lambda n$. The exponential distribution term accounts for the contribution to $p(x,t)$ due to bursts leading to concentration x, and the Dirac δ term accounts for bursts away from x; $\tilde{\mu}_p$ is the rescaled burst size, $\tilde{\mu}_p \equiv \lambda\mu_p$. The last term is due to protein degradation.

The equilibrium solution of (20) can be found to be (see Appendix D)

$$p^{\mathrm{eq}}(x) = A_c\, x^{-1} e^{-x/\tilde{\mu}_p} e^{r\int_c^x du f(u)/u}, \tag{21}$$

where the constant A_c depends on the arbitrary constant c and is determined by normalization.

If we solve Eq. (19) directly in the discrete setting (see Appendix E), we find the solution

$$p^{\mathrm{eq}}_n = \frac{rp^{\mathrm{eq}}_0}{n} \prod_{i=1}^{n-1} \left(r\frac{f(i)}{i} + \frac{\mu_p - 1}{\mu_p} \right), \tag{22}$$

for $n \geqslant 1$, with p^{eq}_0 determined by normalization.

Generically, the continuous approximations presented throughout this section are very accurate for burst sizes of order 10 and higher. It should be noted, however, that very sharp peaks (with a width of the order of a single molecule) that arise for zero protein or mRNA in some parameter ranges are not well captured by the continuous approximation.

The role of the biological parameters in the qualitative features of the protein distribution is particularly clear in the continuous setting. To study some of these features, consider the derivative of the probability distribution given by (21); concentrations $x > 0$, where probability peaks correspond to $\partial_x p^{\mathrm{eq}}(x) = 0$, leading to

$$r\tilde{\mu}_p f(x) = x + \tilde{\mu}_p. \tag{23}$$

Let us consider the regulation function as given by the approximation described by (8). In the continuous description we write

$$f(x) \approx \frac{1 + \rho k x_2(x)}{1 + k x_2(x)}, \tag{24}$$

with

$$x_2(x) \equiv \lambda n_2(n) = \frac{x}{2} + \tilde{a}^2 - \sqrt{x + \tilde{a}^2} \tag{25}$$

and $\tilde{a} = a/\sqrt{\lambda}$. By noting that Eq. (23) is equivalent to a quartic equation in $z = \sqrt{x + \tilde{a}^2}$, it is easy to prove that p^{eq} is at most bimodal (see Appendix F).

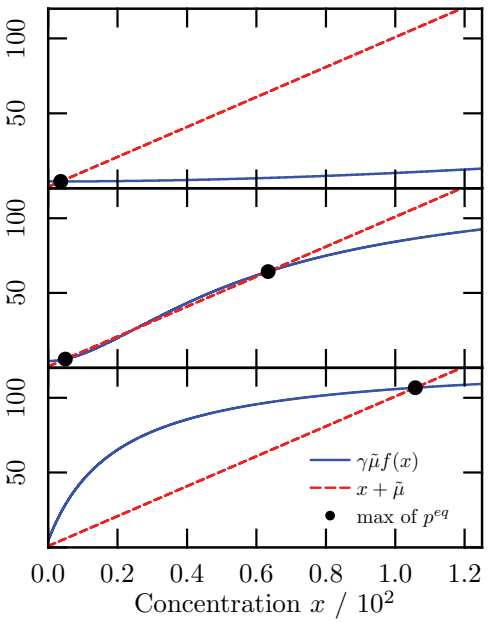

FIG. 3. (Color online) Illustration of the effect of varying the dimerization parameter $\tilde{a}$ when bimodality is possible. For low dimerization (top) there is only a low concentration equilibrium, and for high dimerization (bottom) there is only a high concentration equilibrium. Bimodality without a peak at 0 arises only for intermediate dimerization (middle). Example parameters are $r = 5$, $\tilde{\mu}_p = 0.9$, $\rho = 28$, $k = 10^{-1}$, and (top) $\tilde{a} = 50$, (middle) $\tilde{a} = 5$, and (bottom) $\tilde{a} = 0$.

In the case of negative autoregulation ($\rho < 1$), p^{eq} is always unimodal because the regulation function is monotonically decreasing. Positive autoregulation ($\rho > 1$) is necessary for more structured distributions, and bimodal distributions do in fact arise for some parameter sets. It is interesting to note that in the limit of weak dimerization (large $\tilde{a}$), p^{eq} is always unimodal, while in the limit of strong dimerization (small $\tilde{a}$), it is unimodal if $\gamma > 1$ and bimodal with a peak at 0 if $\gamma < 1$; bimodal distributions that do not peak at 0 are present only for intermediate dimerization (see Fig. 3). Near parameter regions allowing for bimodality, the shape of p^{eq} is also very sensitive to the promoter affinity k (see Fig. 4). Varying r and ρ affects bimodality as well, but the values of these parameters

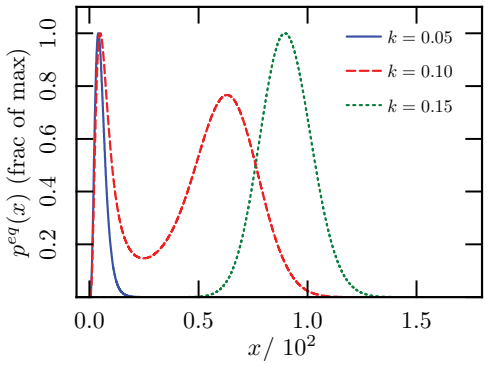

FIG. 4. (Color online) Illustration of the effect of varying the promoter affinity k when bimodality is possible. We fixed $r = 5$, $\tilde{\mu}_p = 0.9$, $\rho = 28$, and $\tilde{a} = 5$.

have a stronger effect on the peak positions. Finally, the burst size parameter $\tilde{\mu}_p$ also affects the position and relative size of peaks in p^{eq}, but its essential role is to produce the heavy-tailed distributions commonly observed experimentally.

It is now easy to obtain the distribution of mRNA expression. For the continuous approximation, taking into account master equation (16), we have, as in (20),

$$\dot{q}(z,t \mid x) = \delta_m \partial_z [z q(z,t \mid x)] + \beta_m f(x) \times \int_0^z [E(z-w,\tilde{\mu}_m) - \delta(z-w)] q(w,t \mid x) dw. \tag{26}$$

This is an evolution equation for the distribution of a "continuous" mRNA concentration variable $z \equiv \lambda j$, given a fixed protein concentration $x = \lambda n$ (with $\tilde{\mu}_m$ again a rescaled burst size). Since f depends on protein but not mRNA concentration, we find for the equilibrium distribution (see Appendix D) a Γ distribution:

$$q^{\mathrm{eq}}(z \mid x) = G(z, \gamma_m f(x), \tilde{\mu}_m). \tag{27}$$

To find the equilibrium distribution of mRNA, we take the integral over all values of protein concentration, weighted by the respective probabilities given by (21):

$$\begin{aligned} q^{\mathrm{eq}}(z) &= \int_0^\infty q^{\mathrm{eq}}(z \mid x) p^{\mathrm{eq}}(x)\, dx \\ &= \langle G(z, \gamma_m f, \tilde{\mu}_m) \rangle_{p^{\mathrm{eq}}}. \end{aligned} \tag{28}$$

Similarly, the solution for the discrete dynamics, corresponding to Eq. (16), is given by a negative binomial distribution (cf. Appendix E):

$$q^{\mathrm{eq}}_{j|n} = N_j \left(\frac{\mu_m}{\mu_m - 1} \gamma_m f(n), \frac{1}{\mu_m} \right). \tag{29}$$

The discrete equilibrium distribution for mRNA is found in this case by summing over all n, weighing with the discrete protein distribution given by (22)

$$\begin{aligned} q^{\mathrm{eq}}_j &= \sum_{n \geqslant 0} q^{\mathrm{eq}}_{j|n} p^{\mathrm{eq}}_n \\ &= \left\langle N_j \left(\frac{\mu_m}{\mu_m - 1} \gamma_m f, \frac{1}{\mu_m} \right) \right\rangle_{p^{\mathrm{eq}}}. \end{aligned} \tag{30}$$

The performance of the continuous approximation is similar for mRNA and for protein.

B. Transcriptional regulation under fast protein dynamics

It is now convenient to consider fixed j, since in this case protein dynamics is much faster and will equilibrate. Let $p_{n|j}(t)$ be the distribution of the protein copy number (evaluated at n) at time t, given j. We have again reactions (18), but in this case we write the master equation for a fixed mRNA copy number j:

$$\dot{p}_{n|j}(t) = \left[\beta_p j \sum_{i \geqslant 1} E_i(\mu_p) \left(\mathbb{E}_p^{-i} - 1 \right) + \delta_p (\mathbb{E}_p - 1) n \right] p_{n|j}(t). \tag{31}$$

In the continuous approximation, we find

$$\dot{p}(x,t \mid z) = \tilde{\gamma}_p \delta_p \int_0^x [E(x-y,\tilde{\mu}_p) - \delta_D(x-y)]p(y,t)dy + \delta_p \partial_x [x p(x,t)], \tag{32}$$

where $\tilde{\gamma}_p \equiv \gamma_p/\lambda$. This equation can be solved for the equilibrium distribution in exactly the same way as Eq. (26), yielding

$$p^{\mathrm{eq}}(x \mid z) = G(x, \tilde{\gamma}_p z, \tilde{\mu}_p). \tag{33}$$

Similarly, the discrete solution [Eq. (31)] is

$$p^{\mathrm{eq}}_{n|j} = N_n\left(\frac{\mu_p}{\mu_p - 1}\gamma_p j, \frac{1}{\mu_p}\right), \tag{34}$$

where $N_n(0,\cdot) \equiv \delta_{n,0}$, with $\delta_{n,0}$ a Kronecker δ symbol.

Following arguments similar to those leading to Eq. (19), the master equation for mRNA reads, in this case,

$$\dot{q}_j(t) = \left[\beta_m \sum_{i\geqslant 1} E_i(\mu_m)\left(\mathbb{E}_m^{-i} - 1\right)\langle f\rangle_{p^{\mathrm{eq}}_{|j}} + \delta_m(\mathbb{E}_m - 1)j\right] q_j(t), \tag{35}$$

and the corresponding continuous master equation is

$$\dot{q}(z,t) = \delta_m\, \partial_z[zq(z,t)] + \beta_m \int_0^z \langle f\rangle_{p^{\mathrm{eq}}(|w)} \times [E(z-w,\tilde{\mu}_m) - \delta_D(z-w)]q(w,t)dw. \tag{36}$$

The equilibrium solution of Eq. (36) can be found through the same method as the one used for Eq. (20), yielding

$$q^{\mathrm{eq}}(z) = A_c z^{-1} e^{-z/\tilde{\mu}_m} e^{\gamma_m \int_c^z du \langle f\rangle_{p^{\mathrm{eq}}(|u)}/u}, \tag{37}$$

where A_c is again a normalization constant. The discrete solution, for Eq. (35), is

$$q^{\mathrm{eq}}_j = \frac{\gamma_m q^{\mathrm{eq}}_0}{j} \prod_{i=1}^{j-1}\left(\gamma_m \frac{\langle f\rangle_{p^{\mathrm{eq}}_{|i}}}{i} + \frac{\mu_m - 1}{\mu_m}\right), \tag{38}$$

for $j \geqslant 1$, with q^{eq}_0 determined by normalization [note that $\langle f\rangle_{p^{\mathrm{eq}}_{|0}} = f(0) = 1$].

In the continuous approximation, the distribution of protein concentration follows immediately from the integration of the conditional distribution given by Eq. (33):

$$p^{\mathrm{eq}}(x) = \int_0^\infty p^{\mathrm{eq}}(x \mid z) q^{\mathrm{eq}}(z) dz = \langle G(x, \tilde{\gamma}_p\, id, \tilde{\mu}_p)\rangle_{q^{\mathrm{eq}}}. \tag{39}$$

The corresponding discrete distribution is

$$p^{\mathrm{eq}}_n = \sum_{j\geqslant 0} p^{\mathrm{eq}}_{n|j} q^{\mathrm{eq}}_j = \left\langle N_n\left(\frac{\mu_p}{\mu_p - 1}\gamma_p\, id, \frac{1}{\mu_p}\right)\right\rangle_{q^{\mathrm{eq}}}. \tag{40}$$

As expected, in this time-scale regime the role of the regulation function is confined to the level of mRNA. The protein distribution depends only on the mRNA distribution, plus the translation rate and protein burst size.

C. Translational regulation

In this scenario, mRNA production takes place through bursts without protein regulation and so mRNA reaches equilibrium independently of protein concentrations. Formally, mRNA dynamics decouples from the general master equation, (12), yielding, for the mRNA distribution $q_j(t)$, the master equation

$$\dot{q}_j(t) = \left[\beta_m \sum_{i\geqslant 1} E_i(\mu_m)\left(\mathbb{E}_m^{-i} - 1\right) + \delta_m(\mathbb{E}_m - 1)j\right] q_j(t). \tag{41}$$

The equilibrium solution for an unregulated process of this type (see Appendix E) is a negative binomial,

$$q^{\mathrm{eq}}_j = N_j\left(\frac{\mu_m}{\mu_m - 1}\gamma_m, \frac{1}{\mu_m}\right), \tag{42}$$

whose average is $\gamma_m \mu_m$.

In the fast mRNA dynamics approximation, the master equation for protein abundances reads

$$\dot{p}_n(t) = \left[\sum_{i\geqslant 1} \beta_p E_i(\mu_p)\left(\mathbb{E}_p^{-i} - 1\right)\langle id\, \tilde{f}(\cdot,n)\rangle_{q^{\mathrm{eq}}} + \delta_p(\mathbb{E}_p - 1)n\right] p_n(t), \tag{43}$$

which, with the simple regulation function $\tilde{f}(j,n) = f(n)$, reduces to

$$\dot{p}_n(t) = \left[r\delta_p \sum_{i\geqslant 1} E_i(\mu_p)\left(\mathbb{E}_p^{-i} - 1\right) f(n) + \delta_p(\mathbb{E}_p - 1)n\right] p^{\mathrm{eq}}_n. \tag{44}$$

For a general regulation function $\tilde{f}(j,n)$, (44) still holds, where now $f(n) \equiv \langle id\, \tilde{f}(\cdot,n)\rangle_{q^{\mathrm{eq}}}/(\gamma_m \mu_m)$.

Equation (44) is the same equation that describes the distribution of protein with transcriptional regulation under fast mRNA dynamics [compare to Eq. (19)]. We thus see that the protein equilibrium distribution is the same as found in Sec. III A, with the appropriate interpretation of the new regulation function f. Moreover, as we see in Sec. IV, this solution holds under less stringent conditions than that of Eq. (19).

Finally, let us consider the fast protein approximation in the translational regulation scenario. By the same arguments as in Sec. III B, the equilibrium protein distribution will be given by

$$p^{\mathrm{eq}}_n = \sum_{j\geqslant 0} p^{\mathrm{eq}}_{n|j} q^{\mathrm{eq}}_j, \tag{45}$$

where q^{eq}_j is the negative binomial, (42), and $p^{\mathrm{eq}}_{n|j}$ is the equilibrium distribution for a fixed mRNA copy number j.

The latter is the stationary solution to

$$\dot{p}_{n|j}(t) = \left[\sum_{i \geqslant 1} \beta_p E_i(\mu_p)(\mathbb{E}^{-i} - 1) j \tilde{f}(j,n) + \delta_p(\mathbb{E}_p - 1)n \right] p_{n|j}(t), \tag{46}$$

already found in Sec. III A [see (19) and (22)] to be given by

$$p^{\mathrm{eq}}_{n|j} = \frac{j\gamma_p}{n} p^{\mathrm{eq}}_0 \prod_{i=1}^{n-1} \left(j\gamma_p \frac{\tilde{f}(j,i)}{i} + \frac{\mu_p - 1}{\mu_p} \right) \tag{47}$$

or, in the continuous approximation for n [see (21)], by

$$p^{\mathrm{eq}}_{|j}(x) = A_c\, x^{-1} e^{-x/\tilde{\mu}_p} e^{j\gamma_p \int_c^x du\, \tilde{f}(j,u)/u}. \tag{48}$$

For the purpose of comparison of this approximation with simulation results, we use the analytic solution given by (45) with (42) and the continuous approximation for $p^{\mathrm{eq}}_{n|j}$ given by (48).

IV. VALIDITY OF THE TIME-SCALE SEPARATION APPROXIMATIONS

In this section we study the conditions under which the time-scale separation assumptions used in Sec. III should hold approximately. We illustrate the quality of the approximate analytic solutions for protein by comparing them with the results of simulations of the full stochastic process described by the master equations (9) and (12) using the Gillespie algorithm [31]. In order to illustrate the agreement with the analytic distributions, the simulation curves shown below were plotted for sampling sizes such that the error bars at each data point are smaller than the markers. We have checked that the structure of the curves obtained from these simulations is robust down to order 10^5 independent samples.

Let the subscripts f and s refer to the fast and slow species, respectively, and let α and σ be the mean and standard deviation of the equilibrium distribution, respectively. Consider also the average times T for a change of one molecule to occur. Two conditions must be met:

(1) The fast species must approach equilibrium quickly compared to T_s. If a change in the slow species produces a change in the absolute value $\Delta\alpha_f$ in the equilibrium average of the fast species, we must have

$$\frac{T^{\mathrm{up}}_f}{T_s} \Delta\alpha_f \ll 1, \tag{49}$$

where T^{up}_f is the average time for the production of one copy of the fast species, since it is straightforward to check for each case that re-equilibrating following a burst is the most demanding scenario.

(2) The fast species must accurately sample the equilibrium average within a time interval T_s. The relative standard error of the mean for N *uncorrelated* samples from a distribution of mean α and standard deviation σ is given by

$$\epsilon = \frac{\sigma}{\alpha\sqrt{N}}. \tag{50}$$

When the fast species dynamically samples the equilibrium distribution, two uncorrelated measurements will be spaced in time approximately by the correlation time $\tau_f = 1/\delta_f$. Thus, considering the relative error in an interval T_s, we find the condition

$$\frac{\sigma_f}{\alpha_f \sqrt{T_s \delta_f}} \ll 1. \tag{51}$$

We now study the constraints imposed by applying conditions 1. and 2 self-consistently in the hypothesis of fast mRNA and fast protein, with both transcriptional and translational regulation. For transcriptional regulation under fast mRNA we have

$$\begin{aligned} \alpha_f &= \gamma_m f(n) \mu_m, \\ \sigma_f &= \sqrt{\gamma_m f(n)} \mu_m. \end{aligned} \tag{52}$$

After a protein event leading to n we may write

$$\begin{aligned} T^{\mathrm{up}}_f &= [\mu_m \beta_m f(n)]^{-1}, \\ \Delta\alpha_f &= \mu_m \gamma_m \Delta f, \end{aligned} \tag{53}$$

where Δf is the absolute value of the change in the value of f associated with the protein event. Since production and degradation reactions must be balanced in equilibrium, we may estimate T_s for macroscopically occupied n as

$$T_s = [\mu_p \beta_p \mu_m \gamma_m f(n)]^{-1}. \tag{54}$$

Then, setting $\delta = \delta_p/\delta_m$, we find, for condition 1,

$$\delta \mu_p r \Delta f \ll 1 \tag{55}$$

and, for condition 2,

$$\sqrt{\delta \mu_p r/\gamma_m} \ll 1. \tag{56}$$

We may combine the two conditions and write

$$\sqrt{\delta r \mu_p}(\sqrt{\delta r \mu_p}\Delta f + 1/\sqrt{\gamma_m}) \ll 1. \tag{57}$$

Note that Δf is bounded by $|\rho - 1|$. It should be mentioned that in the interpretation of a protein translation burst as the protein production of a single mRNA molecule along its whole lifetime, $\beta_p = \delta_m(\mu_p - 1)/\mu_p$ (see [33]), $\delta r \mu_p = (\mu_p - 1)\mu_m\gamma_m$, and the conditions of the fast mRNA approximation may be met only for very low protein burst sizes, even in the absence of regulation.

In Fig. 5 we plot the equilibrium protein distribution obtained from simulations of the stochastic process (9), together with the analytic solution, (22). The parameter values were chosen taking into account condition (57). There is excellent quantitative agreement between approximate and exact solutions.

For transcriptional regulation under fast protein we have, after an mRNA event leading to j,

$$\begin{aligned} \alpha_f &= \mu_p \gamma_p j, \\ \sigma_f &= \sqrt{\gamma_p j} \mu_p, \\ T^{\mathrm{up}}_f &= [\mu_p \beta_p j]^{-1}, \\ T_s &= [\delta_m j]^{-1}. \end{aligned} \tag{58}$$

Since the variation in j due to a burst is of order μ_m, this leads to

$$\mu_m/\delta \ll 1 \tag{59}$$

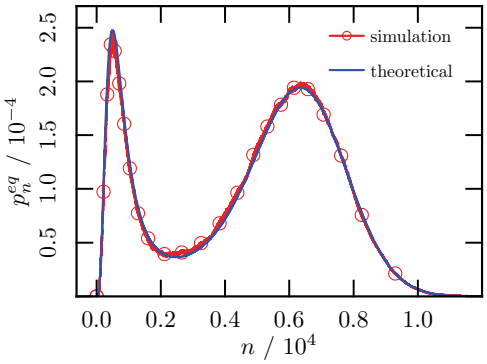

FIG. 5. (Color online) Illustration of the fast mRNA approximation with transcriptional regulation. Parameters are $r = 5$, $\mu_p = 90$, $\gamma_m = 2.25 \times 10^2$, $\mu_m = 2$, $\delta = 10^{-2}$, $\rho = 28$, $k = 0.1$, $a = 50$, and $\lambda = 10^{-2}$. Error bars are smaller than markers.

for condition 1, and for condition 2 we find

$$1/\sqrt{\delta\gamma_p} \ll 1. \tag{60}$$

The combined condition is

$$\frac{1}{\sqrt{\delta}}(\mu_m/\sqrt{\delta} + 1/\sqrt{\gamma_p}) \ll 1. \tag{61}$$

In Fig. 6 we illustrate the behavior of the analytic solution given by (40) versus simulations of the full stochastic process, (9). The parameter values were chosen taking into account condition (61), and once again, there is excellent quantitative agreement.

Consider now the case of translational regulation. The fast mRNA approximation may be treated in much the same way as the corresponding transcriptional regulation case. Note, however, that the mRNA-only reactions now decouple, and the equilibrium solution for the mRNA distribution does not depend on n. Thus $\Delta\alpha_f = 0$, and condition 1 imposes no constraints. The resulting constraint is due to condition 2 only and becomes

$$\sqrt{\frac{\delta r \mu_p f(n)}{\gamma_m}} \ll 1, \tag{62}$$

to be considered for macroscopically occupied values of n. Recall that $f(n)$ is bounded by $\max(1,\rho)$. Figure 7 again shows excellent quantitative agreement between the analytic solution derived in Sec. III C for fast mRNA dynamics and the equilibrium distributions obtained from simulations of (12).

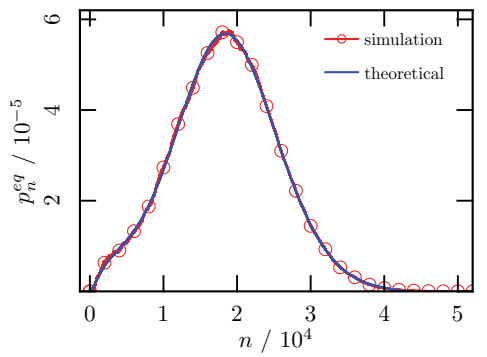

FIG. 6. (Color online) Illustration of the fast protein approximation with transcriptional regulation. Parameters are $\gamma_p = 3$, $\mu_p = 20$, $\gamma_m = 3$, $\mu_m = 20$, $\delta = 10^2$, $\rho = 7.5$, $k = 0.25$, $a = 200$, and $\lambda = 10^{-2}$. Error bars are smaller than markers.

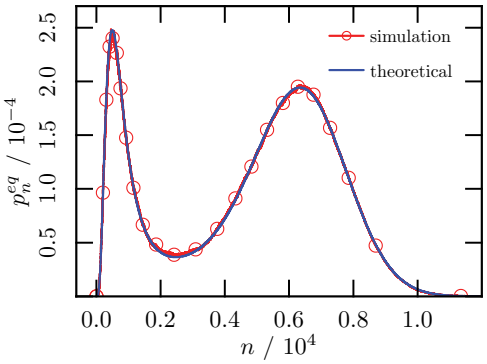

FIG. 7. (Color online) Illustration of the fast mRNA approximation with translational regulation. We took $\tilde{f}(j,n) = f(n)$, and the parameters are $r = 5$, $\mu_p = 90$, $\gamma_m = 6.3 \times 10^4$, $\mu_m = 2$, $\delta = 1$, $\rho = 28$, $k = 0.1$, $a = 50$, and $\lambda = 10^{-2}$. Error bars are smaller than markers.

Note the similarity between Fig. 7 and Fig. 5, which is due to the fact that we have considered for the translational regulation function $\tilde{f}(j,n) = f(n)$, with $f(n)$ the transcriptional regulation function used for the results in Fig. 5. Note also that the ratio δ between protein and mRNA decay rates is far higher in the case of Fig. 7, illustrating that, in terms of the relative stability of protein and mRNA, the fast mRNA approximation is much less demanding for translational regulation than for transcriptional regulation.

For the case of translational regulation in the fast protein approximation, the distribution of the fast species at fixed j is more structured and may be bimodal in general. Furthermore, the dependence of peak positions for the protein distribution on j is also more complicated. A calculation of the parameter constraints imposed by conditions 1 and 2 in this scenario would not lead to simple estimates such as the ones found in the previous three cases. However, the arguments and calculations above provide the intuition that the separation regime will be reached for a certain set of parameters if δ is made large enough, as shown in Fig. 8. This figure also illustrates the

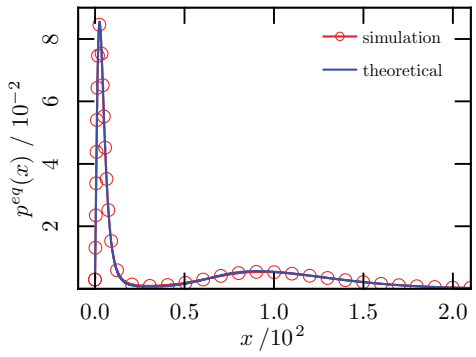

FIG. 8. (Color online) Illustration of the fast protein approximation with translational regulation. We took $\tilde{f}(j,n) = f(n)$, and the parameters are $\gamma_p = 0.25$, $\mu_p = 90$, $\gamma_m = 10$, $\mu_m = 2$, $\delta = 10^3$, $\rho = 28$, $k = 0.1$, $a = 50$, and $\lambda = 10^{-2}$. Error bars are smaller than markers.

good performance, despite the pronounced peak at low x, of the continuous approximation, which was used to plot the analytic curve (see Sec. III C). In the preceding figures the approximation for continuous n is also very good (data not shown).

The parameter values for the figures shown throughout this section were chosen so as to illustrate the accuracy of the predicted distributions when the conditions of validity of the time-scale separation approximation are met. The fulfillment within a large margin of these conditions was favored over biological realism. However, we have found, in a broader exploration of parameter space, that the theoretical distributions derived in Sec. III yield useful approximations for a large set of biologically plausible parameter combinations.

V. DISCUSSION AND CONCLUSIONS

In this paper we have established a detailed stochastic model of single-gene auto-regulation and explored its solutions when mRNA dynamics is fast compared with protein dynamics and in the opposite regime. The time-scale separation allows the derivation of analytic closed-form expressions for the equilibrium distributions of protein and mRNA. Except for very small numbers of molecules, these distributions are well described in the continuous approximation, which we discuss in detail. We typically find distributions that differ significantly from Gaussian distributions and exhibit heavy tails. This is the effect of an essential ingredient of the model, the transcriptional and translational bursts, which typically have a magnitude comparable to the system size. The continuous approximation is well suited to the description of the qualitative features of the protein equilibrium distributions as a function of the biological parameters for fast mRNA. In particular, we find that for positive autoregulation and intermediate values of the dimerization parameter a, the protein equilibrium distributions are bimodal, with two nonzero peaks in a significant range of the remaining parameters. In more general terms, our results show that a fully stochastic description of single-gene positive autoregulation generates structured protein distributions that otherwise can only be explained in the framework of more complex gene regulatory networks.

We have discussed the conditions under which the time-scale separations hold in good approximation and illustrated the performance of both regimes for transcriptional and for translational regulation in comparison with simulations. We found a broad range of parameter values where one of the two opposite time-scale regimes provides a very good approximation. In this range, statistical measures such as mean and variance commonly used in the biological literature to characterize experimental results on protein and mRNA abundances in ensembles of cells can be readily computed from the analytic equilibrium distributions derived in Sec. III. However, mean and variance fall short of characterizing distributions that can be unimodal or bimodal and nonuniformly heavy tailed. For the purpose of comparing the results of the model with real data, it is best to consider the full analytic distributions.

Evidence of translational regulation reported in the biological literature raises the question of understanding its role in the context of stochastic gene expression. Recent work has shown how different post-transcriptional regulation mechanisms modulate noise in protein distributions [32]. Here we have shown that the equilibrium protein distributions for translational regulation have the same form as those that arise under transcriptional regulation in the case of fast mRNA. In particular, the structured protein distributions produced by transcriptional autoregulation with fast mRNA are also produced by translational autoregulation under less demanding conditions in terms of protein and mRNA relative stability. On the other hand, we have shown that in the translational regulation scenario these structured protein distributions are often found as equilibrium solutions also for fast protein dynamics. These properties suggest an additional biological rationale for translational regulation: it allows for efficient autoregulation, circumventing mRNA stability. This idea concurs with the analysis in [22] based on experimental data for protein-mRNA lifetime pairs.

Transcriptional and translational bursts are an essential ingredient of the model whose possible underlying mechanisms and statistics are currently being discussed in the literature. Throughout the paper, we have assumed the simplest form for these bursts. Extending these results, in particular, the validity of the time-scale separation approximation, to the case of more complex mRNA and protein production statistics (see [29]) will be the subject of future work.

ACKNOWLEDGMENTS

Authors T.A. and A.N. wish to thank two anonymous referees and R. Travasso for their useful criticism of a previous version of this paper. The authors are also grateful to D. Henrique for many helpful discussions of the biological foundations and applications of the model. Financial support from the Portuguese funding agency Fundação para a Ciência e a Tecnologia (FCT) under Contract No. POCTI/ISFL/2/261 is gratefully acknowledged. T.A. and E.A. were also supported by the FCT under Grant Nos. PTDC-FIS-70973-2006 (T.A.) and SFRH/BPD/26854/2006 (E.A.).

APPENDIX A: DIMER DYNAMICS

Consider a cell of volume V where there are n copies of some molecular species that can be characterized by a dimerization rate k_d^+ (dimensions volume $\times$ time^{-1}) and an undimerization rate k_d^- (dimensions time^{-1}). Our goal here is to find the explicit form of $n_2(n)$, the number of dimers as a function of the (fixed) total copy number n. The equations governing dimerization dynamics of this species at a fixed total density $\phi \equiv n/V$ are

$$\dot{\phi}_1 = k_d^+ \phi_f^2 - k_d^- \phi_2, \tag{A1a}$$

$$\phi = \phi_f + 2\phi_2. \tag{A1b}$$

Equation (A1a) is the rate equation for temporal dynamics, and the conservation equation, (A1b), reflects that molecules are either free ($\phi_f \equiv n_f/V$) or bound in pairs as dimers ($\phi_2 \equiv n_2/V$).

Defining $k_d \equiv k_d^+/k_d^-$, Eq. (A1a) yields, in equilibrium,

$$\phi_2 = k_d \phi_f^2. \tag{A2}$$

Using Eq. (A1b) for ϕ_f leads, in terms of copy number, to the desired result,

$$n_2(n) = \frac{n}{2} + a^2 - a\sqrt{n + a^2}, \tag{A3}$$

where a is a dimensionless parameter defined by $a \equiv \sqrt{V/(8k_d)}$.

It is also interesting to note that there are two limits in which (A3) becomes very simple and intuitive. On the one hand, if $a^2 \ll n$, we find

$$n_2(n) \approx \frac{n}{2}. \tag{A4}$$

In physical terms, this can be understood as follows: for a certain density n/V, if k_d is high enough, most proteins will bind in dimers; conversely, for a certain k_d, if the density is high enough, most proteins will again be bound because of the increased collision probability. On the other hand, if $a^2 \gg n$, we are in the opposite limit, where most proteins will be free. Taylor expansion of the square root leads, in lowest order, to

$$n_2(n) \approx \frac{k_d}{V} n^2. \tag{A5}$$

This result can also be found by setting $\phi_f \approx \phi$ in (A2).

APPENDIX B: MEAN mRNA IN EQUILIBRIUM (FAST mRNA)

Consider the mRNA master equation, (16). Multiplying both sides by j and summing over j, we find an equation for the mean,

$$\partial_t \langle id \rangle_{q_{|n}(t)} = \left[\beta_m f(n) \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j\left(\mathbb{E}_m^{-i} - 1\right) + \delta_m \sum_{j \geqslant 0} j(\mathbb{E}_m - 1) j \right] q_{j|n}(t). \tag{B1}$$

Let us compute (omitting the arguments t, n for simplicity)

$$\begin{aligned} \sum_{\substack{j \geqslant 0, \\ i \geqslant 1}} E_i(\mu_m) j\left(\mathbb{E}_m^{-i} - 1\right) q_j &= \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j(q_{j-i} - q_j) \\ &= \sum_{i \geqslant 1} i E_i(\mu_m) \sum_{j \geqslant 0} q_j \\ &= \mu_m, \end{aligned} \tag{B2}$$

where we have made use of the fact that $q_j = 0$ whenever the copy number j is negative. Now let us look at

$$\begin{aligned} \sum_{j \geqslant 0} j(\mathbb{E}_m - 1) j q_j &= \sum_{j \geqslant 0} j(j+1) q_{j+1} - \sum_{j \geqslant 0} j^2 q_j \\ &= \sum_{j \geqslant 1} (j-1) j q_j - \sum_{j \geqslant 0} j^2 q_j \\ &= -\sum_{j \geqslant 0} j q_j \; = -\langle id \rangle_{q_{|n}(t)}. \end{aligned} \tag{B3}$$

Since we are looking for the equilibrium mean we now set the left-hand side of (B1) to 0, and using results (B2) and (B3) we find the desired result:

$$\langle id \rangle_{q_{|n}^{\mathrm{eq}}} = \mu_m \gamma_m f(n). \tag{B4}$$

APPENDIX C: CONTINUOUS APPROXIMATION

Here we study a continuous approximation for equations of the form

$$\dot{p}_n(t) = \left[r\delta \sum_{i \geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1) f(n) + \delta(\mathbb{E} - 1) n \right] p_n(t), \tag{C1}$$

where f is some function of (protein or mRNA) copy number n, $r \neq 0$ and $\delta \neq 0$ are constants, and the step operator $\mathbb{E}$ raises n. For some time t, let copy number n be fixed, and let $x = \lambda n$ be the corresponding concentration. In accordance with the main text, the convention $f(n) = f(x)$ is used. First, note that a reasonable definition for the continuous distribution obeys

$$\begin{aligned} p_n(t) &\equiv p(x,t)\lambda[(n + 1/2) - (n - 1/2)] \\ &\approx \int_{n-1/2}^{n+1/2} p(x,t)\, dx = \lambda p(x,t). \end{aligned} \tag{C2}$$

Now consider the conditioned geometric distribution. We have

$$E_n(\mu) = \frac{(\mu - 1)^{(n-1)}}{\mu^n} = \frac{1}{\mu - 1} e^{-n \log(1 - 1/\mu)}. \tag{C3}$$

If we take $\mu \gg 1$ (which is biologically common, especially for proteins; see, e.g., [11] and [13]) and expand $\log(1 - 1/\mu)$ around $1/\mu = 0$, we find to lowest order

$$E_n(\mu) \approx \frac{1}{\mu} e^{-n/\mu} = \lambda \frac{1}{\tilde{\mu}} e^{-x/\tilde{\mu}} = \lambda E(x, \tilde{\mu}), \tag{C4}$$

with $\tilde{\mu} = \lambda\mu$. Now note that, apart from the constant coefficients, the creation term in Eq. (C1) may be written

$$\begin{aligned} &\sum_{i \geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1) f(n) p_n(t) \\ &\quad = \sum_{i \geqslant 1} E_i(\mu) f(n - i) p_{n-i} - f(n) p_n(t) \\ &\quad = \sum_{i=0}^{n} (E_{n-i}(\mu) - \delta_{n,i}) f(i) p_i, \end{aligned} \tag{C5}$$

where $\delta_{n,i}$ is a Kronecker δ symbol. Note that the upper limit of the sum can be extended to infinity by taking $E_j(\mu) = 0$ for $j \leqslant 0$, and the lower limit can be extended to negative infinity since $p_i = 0$ for $i < 0$.

The Kronecker δ term reads

$$\begin{aligned} \sum_{i=0}^{n} \delta_{n,i} f(i) p_i &= f(n) p_n(t) \\ &= \lambda \int_0^x \delta_D(x - y) f(y) p(y,t)\, dy, \end{aligned} \tag{C6}$$

where δ_D is the Dirac δ. Note that, for a meaningful conversion to the continuous case, the lower limit of the integral must be strictly included (in order to encompass the contribution of the δ function). Thus, the upper and lower limits of the integral may be extended to infinity.

For the conditioned geometric distribution term in (C5), we may write

$$\begin{aligned}\sum_{i=0}^{n} & E_{n-i}(\mu) f(i) p_i \\ &\approx \sum_{i=0}^{n} \lambda E(\lambda(n-i),\tilde{\mu}) f(i) \lambda p(\lambda i,t) \\ &\approx \lambda \int_0^x E(x-y,\tilde{\mu}) f(y) p(y,t) dy, \end{aligned} \tag{C7}$$

where, again, the upper and lower limits of the integral may be extended to plus and minus infinity by considering, respectively, $E(y,\tilde{\mu}) = 0$ and $p(y,t) = 0$ for negative y. Here, the approximations $\mu \gg 1$ (approximating the conditioned geometric distribution with an exponential distribution) and $\lambda \ll 1$ (approximating the sum with an integral, i.e., considering x continuous) have been explicitly used.

Finally, the degradation term in Eq. (C1) reads, apart from a factor of δ,

$$\begin{aligned}(\mathbb{E}-1) n p_n(t) &= [(n+1)p_{n+1}(t) - n p_n(t)] \\ &= \frac{1}{\lambda}[(x+\lambda)\lambda p(x+\lambda)(t) - x\lambda p(x,t)] \\ &\approx \lambda \partial_x (x p(x,t)), \end{aligned} \tag{C8}$$

where we again make use of $\lambda \ll 1$ to approximate a finite difference with a derivative. Noting that $\dot{p}_n(t) = \lambda \dot{p}(x,t)$ and collecting terms, we find

$$\begin{aligned}\dot{p}(x,t) = r\delta \int_0^x f(y)[E(x-y,\tilde{\mu}) - \delta_D(x-y)] p(y,t)\, dy \\ + \delta\, \partial_x [x p(x,t)]. \end{aligned} \tag{C9}$$

APPENDIX D: CONTINUOUS EQUILIBRIUM DISTRIBUTIONS

Here we follow [7] to obtain an analytical solution to Eq. (C9). As discussed in Appendix C, the upper and lower integration limits may be extended to plus and minus infinity, respectively. Thus, defining

$$w(x,\tilde{\mu}) = E(x,\tilde{\mu}) - \delta_D(x), \tag{D1}$$

we may write

$$\dot{p}(x,t) = r\delta(w(\tilde{\mu}) * fp)(x,t) + \delta\, \partial_x [x p(x,t)], \tag{D2}$$

where $*$ is a convolution product. In equilibrium we have

$$-\partial_x [x p^{\mathrm{eq}}(x)] = r(w(\tilde{\mu}) * f p^{\mathrm{eq}})(x). \tag{D3}$$

Laplace transformation of this equation leads to

$$\begin{aligned} s\, \partial_s \hat{p}(s) &= r\hat{w}(s)\mathcal{L}(f p^{\mathrm{eq}})(s) \\ &= r\hat{w}(s)(\hat{f} * \hat{p})(s) \\ &= -r\frac{s}{s+1/\tilde{\mu}}(\hat{f} * \hat{p})(s). \end{aligned} \tag{D4}$$

Here, $\hat{g}(s) = \mathcal{L}(g)(s) = \int_0^{+\infty} e^{-sx} g(x)\, dx$ (integration limit 0 strictly included) is the Laplace transform of function g (evaluated at s), and $\hat{p} = \mathcal{L}(p^{\mathrm{eq}})$. Convolution theorems have been used in the first and second lines, and in the third line the explicit form of $\hat{w}(s)$ was substituted. Rearranging terms, we have

$$(s + 1/\tilde{\mu})\hat{p}(s) = -r(\hat{f} * \hat{p})(s), \tag{D5}$$

which inverse transforms to

$$\partial_x [x p^{\mathrm{eq}}(x)] = (r f(x)/x - 1/\tilde{\mu})\, x p^{\mathrm{eq}}(x). \tag{D6}$$

This equation can easily be solved, leading to

$$p^{\mathrm{eq}}(x) = A_c\, x^{-1} e^{-x/\tilde{\mu}} e^{r\int_c^x du f(u)/u}. \tag{D7}$$

The constant A_c is determined by normalization (depending on the arbitrary integration limit c).

Consider now the case $f(x) = 1$, for all x. Solving the integral in (21) and normalizing the probability distribution to integral unity, we find

$$\begin{aligned} p^{\mathrm{eq}}(x) &= \frac{x^{r-1} e^{-x/\tilde{\mu}}}{\tilde{\mu}^r \Gamma(r)} \\ &= G(x,r,\tilde{\mu}). \end{aligned} \tag{D8}$$

This is the Γ distribution of parameters r and $\tilde{\mu}$ (Γ is the Euler Γ function). With $r = \mu_m \gamma_m \gamma_p$ and $\tilde{\mu}$ the mean rescaled protein burst size (with definitions according to the main text), this is the equilibrium solution for unregulated protein dynamics with fast mRNA.

APPENDIX E: DISCRETE EQUILIBRIUM DISTRIBUTIONS

In this Appendix we analyze, directly in the discrete setting, Eq. (C1). Analogously to the continuous case, the discrete master equation may be written

$$\dot{p}_n(t) = r\delta(w(\mu) * fp)(n) + \delta[(n+1)p_{n+1}(t) - n p_n(t)], \tag{E1}$$

where $*$ is now the discrete convolution product, and $w(n,\mu) = E_n(\mu) - \delta_{n,0}$. We now follow the procedures in Appendix D using the Z transform instead of the Laplace transform, $\hat{g}(s) = \mathcal{Z}(g)(s) = \sum_{n=0}^{+\infty} s^{-n} g(n)$, $\mathcal{Z}(p^{\mathrm{eq}}) = \hat{p}$. The corresponding equation in "momentum space" is

$$s(s-1)\partial_s \hat{p}(s) + \frac{s}{\mu}\partial_s \hat{p}(s) = -r\left(\hat{f} * \hat{p}\right)(s). \tag{E2}$$

Inverse transforming, we get

$$(n+1)p_{n+1}^{\mathrm{eq}} + (1/\mu - 1) n p_n^{\mathrm{eq}} = r f(n) p_n^{\mathrm{eq}}, \tag{E3}$$

leading to the recurrence relation:

$$\begin{aligned} p_1^{\mathrm{eq}} &= r f(0) p_0^{\mathrm{eq}}, \\ (n+1)p_{n+1} &= \left(r\frac{f(n)}{n} + \frac{\mu-1}{\mu}\right) n p_n^{\mathrm{eq}}, \quad n \geqslant 1. \end{aligned} \tag{E4}$$

This is easily solved, yielding

$$p_n^{\mathrm{eq}} = \frac{r f(0) p_0^{\mathrm{eq}}}{n} \prod_{i=1}^{n-1} \left(r\frac{f(i)}{i} + \frac{\mu-1}{\mu}\right) \tag{E5}$$

for all $n \geqslant 1$, with $p_0^{\rm eq}$ determined by normalization (and the standard convention that the product equals 1 when the upper limit is smaller than the lower). Note that if f is a regulation function per the main text, we have $f(0) = 1$, since the promoter is necessarily free when no protein is present.

Consider now the case $f(n) = 1$ for all n. Write (E5) as

$$p_n^{\rm eq} = \frac{\mu}{\mu-1}\frac{rf(0)p_0^{\rm eq}}{n}\left(\frac{\mu-1}{\mu}\right)^n \prod_{i=1}^{n-1}\left(\frac{\mu}{\mu-1}r\frac{f(i)}{i}+1\right)$$
$$= \frac{r'f(0)p_0^{\rm eq}}{n}\left(\frac{\mu-1}{\mu}\right)^n \prod_{i=1}^{n-1}\left(r'\frac{f(i)}{i}+1\right), \qquad \text{(E6)}$$

with $r' = r\mu/(\mu-1)$. The product can be solved explicitly is terms of Γ functions, and normalizing to unit sum we find

$$p_n^{\rm eq} = \frac{1}{\mu^{r'}}\left(\frac{\mu-1}{\mu}\right)^n \frac{\Gamma(n+r')}{\Gamma(r')\Gamma(n+1)}$$
$$= N_n\left(r', \frac{1}{\mu}\right). \qquad \text{(E7)}$$

This is the negative binomial distribution of parameters γ' and $1/\mu$. The parameters are defined such that

$$N_n(k,p) = p^k(1-p)^n\binom{n+k-1}{k-1}. \qquad \text{(E8)}$$

As in the continuous case (Appendix D), with $r = \mu_m\gamma_m\gamma_p$ and μ the mean protein burst size μ_p (definitions according to the main text), this is the discrete solution for unregulated protein dynamics with fast mRNA (as reported, e.g., in [30]).

APPENDIX F: BIMODAL EQUILIBRIUM PROTEIN DISTRIBUTIONS

Consider the continuous equilibrium distribution for protein with fast mRNA, given by (21). The derivative of this probability distribution is given by

$$\partial_x p^{\rm eq}(x) = [r\tilde{\mu}f(x) - (x+\tilde{\mu})]\frac{p^{\rm eq}(x)}{\tilde{\mu}x}. \qquad \text{(F1)}$$

If $p^{\rm eq}$ peaks at 0 [i.e., if $\partial_x p^{\rm eq}(0) < 0$], the term in brackets in Eq. (F1) must be negative at 0. Because $p^{\rm eq}(x) > 0$ for all $x > 0$, other extrema of $p^{\rm eq}$ must satisfy

$$r\tilde{\mu}f(x) - (x+\tilde{\mu}) = 0. \qquad \text{(F2)}$$

Consider $f(x)$ as given by (8). A change of variables to $z = \sqrt{x+\tilde{a}^2}$ in Eq. (F2) leads to an equivalent quartic equation,

$$P_4(z) = -z^4 + 2\tilde{a}z^3 + \alpha_2 z^2 + \alpha_1 z + \alpha_0 = 0, \qquad \text{(F3)}$$

where the α_i are real constants determined by the biological parameters. The equation $P_4''(z) = 0$ is quadratic in z and has two solutions:

$$z = \frac{\tilde{a}}{2} \pm \sqrt{\left(\frac{\tilde{a}}{2}\right)^2 + \frac{\alpha_2}{6}}. \qquad \text{(F4)}$$

If they are real, one of these solutions necessarily obeys $z < \tilde{a}$. Therefore, $P_4''(z)$ has at most one root in $z > \tilde{a}$.

We now proceed to prove that $p^{\rm eq}$ is at most bimodal. Since zeros of P_4 correspond alternately to maxima and minima of $p^{\rm eq}$, the presence of more than two maxima requires at least four positive roots of $P_4(z)$ in $z > \tilde{a}$. But then $P_4''(z)$ would have at least two roots in $z > \tilde{a}$.

[1] H. Berg and E. Purcell, Biophys. J. **20**, 193 (1977).
[2] O. Berg, J. Theor. Biol. **71**, 587 (1978).
[3] W. Bialek and S. Setayeshgar, Proc. Natl. Acad. Sci. USA **102**, 10040 (2005).
[4] W. Bialek and S. Setayeshgar, Phys. Rev. Lett. **100**, 258101 (2008).
[5] H. Fraser, A. Hirsh, G. Giaever, J. Kumm, and M. Eisen, PLoS Biol. **2**, e137 (2004).
[6] M. Elowitz, A. Levine, E. Siggia, and P. Swain, Science **297**, 1183 (2002).
[7] N. Friedman, L. Cai, and X. S. Xie, Phys. Rev. Lett. **97**, 168302 (2006).
[8] T. Kalmar, C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. Martinez Arias, PLoS Biol. **7**, e1000149 (2009).
[9] P. Rué and J. Garcia-Ojalvo, Math. Biosci. **231**, 90 (2011).
[10] L. Cai, N. Friedman, and X. Xie, Nature **440**, 358 (2006).
[11] B. Kaufmann and A. van Oudenaarden, Curr. Opin. Genet. Dev. **17**, 107 (2007).
[12] E. Ozbudak, M. Thattai, I. Kurtser, A. Grossman, and A. van Oudenaarden, Nat. Genet. **31**, 69 (2002).
[13] Y. Taniguchi, P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, Science **329**, 533 (2010).
[14] D. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Science **322**, 472 (2011).
[15] D. Larson, D. Zenklusen, B. Wu, J. Chao, and R. Singer, Science **322**, 475 (2011).
[16] J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Phys. Rev. E **72**, 051907 (2005).
[17] A. Mugler, A. M. Walczak, and C. H. Wiggins, Phys. Rev. E **80**, 041921 (2009).
[18] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Phys. Rev. E **79**, 031911 (2009).
[19] M. Assaf, E. Roberts, and Z. Luthey-Schulten, Phys. Rev. Lett. **106**, 248102 (2011).
[20] L. Waters and G. Storz, Cell **136**, 615 (2009).
[21] J. Wang, D. Levasseur, and S. Orkin, Proc. Natl. Acad. Sci. USA **105**, 6326 (2008).
[22] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, Nature **473**, 337 (2011).
[23] R. Guantes and J. Poyatos, PLoS Comput. Biol. **4**, e1000235 (2008).
[24] N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992).
[25] M. Liu, P. Li, and J. Giddings, Protein Sci. **2**, 1520 (1993).
[26] D. Xu, C. Tsai, and R. Nussinov, Protein Sci. **7**, 533 (1998).

[27] O. Bénichou, Y. Kafri, M. Sheinman, and R. Voituriez, Phys. Rev. Lett. **103**, 138102 (2009).
[28] J. Paulsson, Phys. Life Rev. **2**, 157 (2005).
[29] V. Elgart, T. Jia, A. Fenley, and R. Kulkarni, Phys. Biol. **8**, 046001 (2011).
[30] V. Shahrezaei and P. Swain, Proc. Natl. Acad. Sci. USA **105**, 17256 (2008).
[31] D. Gillespie, J. Chem. Phys. **115**, 1716 (2001).
[32] T. Jia and R.V. Kulkarni, Phys. Rev. Lett. **105**, 018101 (2010).
[33] Note that, if $E^0(\theta)$ is the nonconditioned geometrical distribution of mean θ, the relation $E_i(\theta) = \frac{\theta}{\theta-1} E_i^0(\theta - 1)$ holds for all $i \geqslant 1$. Thus, "conditioned bursting" with frequency α and mean θ is equivalent to "nonconditioned bursting" with frequency $\frac{\theta}{\theta-1}\alpha$ and mean $\theta - 1$, and the difference becomes relevant only for $\theta \sim 1$.