Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression

Nir Friedman, Long Cai, and X. Sunney Xie

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge, Massachusetts 02138, USA

(Received 9 June 2006; published 19 October 2006)

We present an analytical framework describing the steady-state distribution of protein concentration in live cells, considering that protein production occurs in random bursts with an exponentially distributed number of molecules. We extend this framework for cases of transcription autoregulation and noise propagation in a simple genetic network. This model allows for the extraction of kinetic parameters of gene expression from steady-state distributions of protein concentration in a cell population, which are available from single cell data obtained by flow cytometry or fluorescence microscopy.

DOI: 10.1103/PhysRevLett.97.168302 PACS numbers: 82.39.Rt, 02.50.−r, 87.17.Aa

Gene expression, the flow of information from nucleic acids to proteins, consists of a set of biochemical reactions that occur stochastically inside living cells due to the small copy numbers of molecules involved. Consequently, the abundance of a given protein varies within a population of genetically identical cells growing in the same environment [1–3]. Stochasticity in gene expression has been studied theoretically since 1970s [4]. Recent advances in experimental methods allow direct observations of both real-time fluctuation in gene expression levels in individual live cells and their steady-state variations across a cell population [5–11]. It is conceivable that the two measurements are related, analogous to the ergodic principle in statistical physics. Here, we present a theoretical model that reconciles the time-resolved and population measurements based on the underlying mechanism of protein production.

We start with the simplest model shown in the kinetic scheme in Fig. 1(a). The protein is produced in bursts, in which an mRNA molecule is translated into a few protein molecules before its degradation. Assuming the lifetime of mRNAs is short compared to the lifetime of the protein, as in the case of bacteria or yeast, fluctuations in the mRNA level can be integrated out and proteins can be considered to be produced in random uncorrelated events, where each event results in an exponentially distributed number of proteins, as observed experimentally [10,11]. Protein production is then characterized by two parameters: the mean number of bursts per cell cycle, which we define as $a = k_1/\gamma_2$, and the mean number of protein molecules produced per burst, $b = k_2/\gamma_1$. Here, k_1 is the rate of transcription of a gene into mRNA, k_2 is the rate of translation of mRNA into protein, and $\gamma_{1,2}$ are the rates of degradation of the mRNA and protein, respectively, [Fig. 1(a)]. In addition, we consider the concentration of a given protein in a cell, $x(t)$, to be a continuous random variable that changes abruptly in time when a production burst occurs. Given this kinetic scheme of protein production, the distribution of a protein in a population of cells, $p(x)$, satisfies a continuous master equation [12]

$$\frac{\partial p(x)}{\partial t} = \frac{\partial}{\partial x}[\gamma_2 x p(x)] + k_1 \int_0^x dx' w(x, x') p(x'). \quad (1)$$

Here, $x = n/V$ is the concentration of a protein of interest in a cell, where n is the number of molecules of that protein inside a cell and V is the cell volume. The first term on the right-hand side of Eq. (1) describes decrease in the concentration of protein molecules, either due to dilution

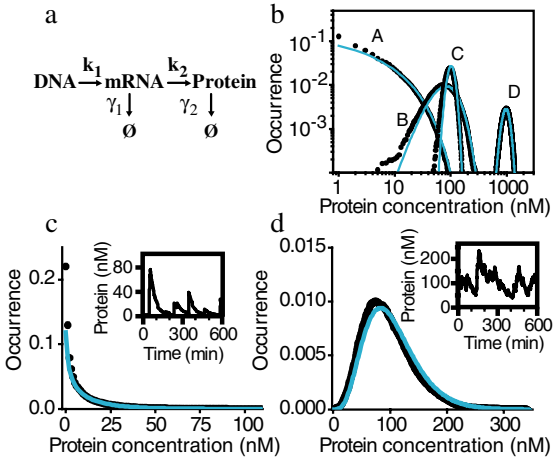

FIG. 1 (color online). Analytical solution to stochasticity in gene expression. (a) Kinetic scheme for describing noise in gene expression. (b) Distribution of protein concentration in a cell population. Simulation parameters are biologically relevant for bacteria: $\gamma_1 = 0.01\ s^{-1}$, $\gamma_2 = 4 \times 10^{-4}\ s^{-1}$, $V_0 = 1.7$ fl A: $a = 0.5$, $b = 30$ ln2; B: $a = 5$, $b = 30$ ln2; C: $a = 50$, $b = 3$ ln2; D: $a = 50$, $b = 30$ ln2. Simulations include random (binomial) partitioning of protein molecules during cell division and exponential growth in cell volume. Gamma distribution [lines, Eq. (5)] matches stochastic stimulation results (dots) well over a large range of parameter values. (c),(d) Simulation and analytic results for $a < 1$ and $a > 1$, corresponding to curves A, B of (b), respectively. In the first case, $p(x)$ peaks at $x = 0$, regardless of the value of b. In the second case, $p(x)$ peaks at $x > 0$. Insets show typical simulated trajectories.

during cell growth and division, or due to protein degradation. In general, $\gamma_2 = (\ln 2/T + \ln 2/T_1)$, with T the protein half-life time and T_1 the cell cycle time, in which the cell volume doubles. The second term describes protein production in bursts. $w(x, x') = w(x|x') - \delta(x - x')$, with $w(x|x')$ the conditional probability that the protein concentration in a cell will be x after a protein production burst, given that the concentration was x' before the burst. The δ function conserves the total probability density by accounting for the loss of density at existing protein concentration as a result of a burst away from x. We next assume that the burst size $(x - x')$ is independent of the current protein concentration, x', and that it follows some characteristic distribution, $\nu(x - x')$. Under these assumptions, $w(x, x') = w(x - x') = \nu(x - x') - \delta(x - x')$. At steady state, the concentration of a given protein per cell reflects the balance between production of new proteins and dilution or degradation of existing ones: $\partial p(x)/\partial t = 0$. Noting that with the above form of $w(x, x')$ the creation term in Eq. (1) becomes a convolution integral, we can write

$$-\frac{\partial}{\partial x}[xp(x)] = aw * p(x), \tag{2}$$

with the convolution integral denoted by: $w * p = \int_0^x w(x - x')p(x')dx'$. The Laplace transform of Eq. (2) yields a first order ordinary differential equation:

$$s\frac{\partial \hat{p}}{\partial s} = a\hat{w}\,\hat{p}\,. \tag{3}$$

Here, s is the Laplace space variable, while $\hat{w}$, $\hat{p}$ denote the Laplace transforms of $w(x)$, $p(x)$, respectively. Next, we note that the burst size distribution is given by an exponential distribution: $\nu(x) = (1/b)\exp(-x/b)$, with an average burst size b. This burst size distribution was recently measured experimentally in *E. coli* cells [10,11]. The corresponding Laplace transform of $w(x)$ is $\hat{w}(s) = -s/[s + (1/b)]$. With this, Eq. (3) can be solved to give:

$$\hat{p}(s) = [s + (1/b)]^{-a}. \tag{4}$$

Transforming back to real space gives the solution for the steady-state distribution:

$$p(x) = \frac{1}{b^a\Gamma(a)}x^{a-1}e^{-x/b}. \tag{5}$$

This is the Gamma distribution, with Γ denoting the Gamma function. We note that since the above derivation takes into account explicitly the exponentially distributed burst size without making the Gaussian noise approximation, the resulting Gamma distribution captures the asymmetry of experimentally observed distributions [7,13,14]. This functional form is the continuous equivalent of the Negative-Bionomial distribution derived by Paulsson *et al.* [15] using a discrete master equation.

The Gamma distribution is defined by the two parameters a and b, establishing a direct correspondence between the steady-state distribution and the kinetic parameters that describe protein expression. The two parameters are related to the mean, $\langle x\rangle$ and the standard deviation, σ, of $p(x)$: the Fano factor, $\sigma^2/\langle x\rangle = b$ is an estimator for average burst size relating to translation, and $\sigma^2/\langle x\rangle^2 = 1/a$ estimates the average time between bursts, relating to transcription. We note that, due to the continuous variable approximation, this result deviates from $\sigma^2/\langle x\rangle = b + 1$ as derived using a discrete master equation [15,16]. However, for $b > 1$, the deviation is small and majority of genes falls into this category [17].

We numerically simulated the reactions shown in Fig. 1(a) using Gillespie's algorithm [18] for various values of the rate constants. In those stochastic simulations, we assume a stable protein ($T \gg T_1$), and include random partitioning of proteins between daughter cells after cell division according to Binomial distribution, as well as an exponential growth in cell volume, which is doubled during a time T_1. The protein concentration is given by $x(t) = n(t)/V_0\exp[(\ln 2)(t/T_1)]$ with V_0 the cell volume just after cell division. Figure 1(b) shows that there exists an excellent agreement between the simulated protein concentration distribution and the corresponding Gamma distribution, with $a = k_1/\gamma_2$ and $b = (k_2/\gamma_1)/\langle V(t)\rangle = (k_2/\gamma_1)\ln 2$, where the average of $V(t)$ is taken over a cell cycle. We observe that, under almost all circumstances, the discreteness of the number of molecules does not lead to a significant difference between simulations and analytical results for $n \geq 1$. From this comparison, we conclude that the parameters a, b, which characterize the process of protein production, can be extracted from a protein concentration distribution with good accuracy, despite the simplifications used in the derivation. These simplifications are useful in gaining an intuitive view of the process and allow for extension of the model to more complex cases, as discussed below.

The parameter a determines two different regimes of Eq. (5). For $a < 1$, corresponding to less than one burst event per cell cycle on average, $p(x)$ peaks at zero. In this case, a large fraction of cells do not contain even a single copy of the protein of interest, regardless of the burst size b. For $a > 1$, $p(x)$ peaks at a nonzero x and most cells contain the protein at some level. These two cases are illustrated in Fig. 1(c) and 1(d), respectively.

In addition to the burstlike production of proteins, transcriptional bursts have also been observed recently [9]. The number of mRNA molecules produced per burst has been observed to be exponentially distributed. In this case, it can be shown that the number of proteins produced per transcriptional burst remains exponentially distributed. Thus, the steady-state protein distribution is still described by a Gamma distribution, even with mRNA produced in bursts. In a more general setting, transcription occurring in a time window with exponential dwell time, such as the on-off models of chromatin remodeling or transcriptional reini-

tiation [6,8], also produces distribution that can be well described by the Gamma distribution. In these cases, the a and b values extracted from the distribution can be generalized to denote the number of "on" windows per cell cycle and the number of proteins produced during each window, respectively.

An advantage of this analytical framework is that it can be extended to analytically describe the steady-state protein distribution for various cases of biological interest, as we demonstrate with the examples below. Many transcription factors—proteins that regulate the expression level of other genes—also regulate their own transcription [19]. Assuming that rates of transcription factor binding to ($k_{\rm on}$) and unbinding from ($k_{\rm off}$) the promoter are fast, autoregulation can be added to the model by allowing the burst rate to depend on the current level of x, by multiplying a with a response function $c(x)$. At steady state, the master equation is now modified:

$$-\frac{\partial}{\partial x}[xp(x)] = a\int_0^x dx' w(x-x')c(x')p(x'). \quad (6)$$

The Laplace transform of Eq. (6), assuming the same form of $w(x-x')$ as above, is

$$\frac{\partial \hat{p}}{\partial s} = -\frac{a}{s+(1/b)}(\hat{c} * \hat{p}). \quad (7)$$

Multiplying both sides by $[s+(1/b)]$ and inverse Laplace transform yield $\partial(xp)/\partial x + xp/b = acp$, which can be solved to give $p(x)$ in the presence of autoregulation:

$$p(x) = Ax^{-1}e^{-x/b}e^{a\int dx c(x)/x} \quad (8)$$

with A being a normalization factor. For a response in the form of a Hill function $c(x) = k^H/(k^H + x^H) + \varepsilon$, $p(x)$ can be integrated:

$$p(x) = Ax^{a(1+\varepsilon)-1}e^{-x/b}[1+(x/k)^H]^{-a/H}. \quad (9)$$

Here, $k = k_{\rm off}/k_{\rm on}$ is the equilibrium binding constant of the transcription factor to its DNA binding site, and $\varepsilon = k_\varepsilon/k_1$ is a small leakiness of the promoter, with k_ε the leakage transcription rate (See Fig. 2 for the corresponding kinetic scheme used in the stochastic simulations). A Hill coefficient $H > 0$ corresponds to negative feedback, and $H < 0$ to positive feedback.

Figure 2(a) shows that the calculated steady-state distribution of protein concentrations with negative feedback is squeezed compared to that without feedback. The calculated distribution with positive feedback may give rise to bistability as shown in Fig. 2(b). Both results are consistent with experimental observations [20]. As with the simpler case of Fig. 1, the agreement between the stochastic numerical simulations and the analytical model is good, again with no fitting parameters. We note that a bimodal population occurs in a narrow range of the parameter k. Outside this range, the population is unimodally distributed either at the low or high state, leading to a sharp switch, as shown in Fig. 2(c). Such sharp switches may exist in nature, where the binding constant of an activator to DNA changes in response to a signal, e.g., the concentration of a specific small molecule substrate.

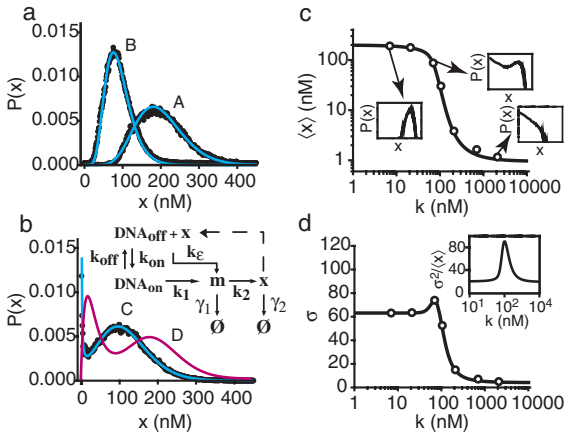

FIG. 2 (color online). Autoregulation. (a) Dots show simulated distribution of protein concentration without (curve A) and with (B) negative feedback. Lines are analytical solutions [Eq. (9)]. Parameters: $a = 10$, $b = 20$, $k = 70$ nM, $\varepsilon = 0.05$, $H = 1$. (b) Same as (a), for positive feedback. Inset: kinetic scheme. X: the autoregulated transcription factor, m: corresponding mRNA. Parameters: as in (a) except for : $H = -1$ (C, cyan) and $H = -4$, $\varepsilon = 0.2$ (D, magenta). In this case, $a\varepsilon > 1$ resulting in low peak at $x > 0$. (c) Mean of x as a function of k calculated from Eq. (9), for positive autoregulation (black curve). Parameters are as in [(b), curve C)]. Circles: values calculated from stochastic simulations. Representative distributions $p(x)$ are shown in insets. (d) Standard deviation of x vs k [(parameters as in (c)]. Inset: the Fano factor, $\sigma^2/\langle x\rangle$, peaking at the transition region.

From the analytical form of Eq. (9) it is seen that bistability can occur not only for $H < -1$, but also for the case $H = -1$ [Fig. 2(b)], which is not predicted by the deterministic rate equation without noise [21]. The analytical form of Eq. (9) facilitates calculations of moments of $p(x)$ as a function of a changing parameter. This is demonstrated in Fig. 2(c) and 2(d), where $\langle x\rangle$, σ and the Fano factor are plotted as a function of the binding constant k, for positive autoregulation.

We now extend the model to describe noise propagation in a simple genetic network—a repressor, R, that regulates gene x. We calculate the joint distribution $p(R, x)$ in a cell population, accounting for fluctuations in R and their effect on fluctuations in x. Here, the burst rate for producing x is dependent on the concentration of the repressor R and equals $a_x c(R)$. In the limit where fluctuations in R are fast compared with the rate of transcription of x, those can be averaged out, and the joint probability density of R, x is given by,

$$p(R,x) = p(R)p(x) = Bp(R)x^{a_x\langle c(R)\rangle - 1}e^{-x/b_x}. \quad (10)$$

Here, $p(R)$ is given by Eq. (5), a_i, b_i ($i = x, R$), are burst

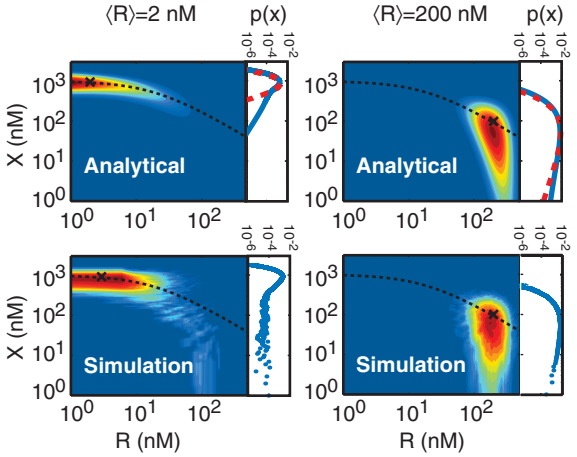

FIG. 3 (color). Joint distribution of a repressor, R, and a regulated gene, x. Contour plots of steady-state distributions obtained by analytic solution [Eq. (11), top row] and by numerical stochastic simulation (bottom row) for a case when fluctuation in the repressor is slow. Distributions are shown for 2 average values of repressor concentration. The black dashed line represents the deterministic steady-state solution for $x(R)$. Averages of the distributions are marked (X). The distribution $p(x)$ is shown on the right (blue line), compared to the case of fast repressor fluctuations [dashed-red line, Eq. (10): negligible extrinsic noise]. Parameters: γ_{1x}, γ_{2x}, V: as in Fig. 1(b); $k = 20$ nM, $k_{1x} = 0.01\ \mathrm{s}^{-1}$, $k_{2x} = 0.4\ \mathrm{s}^{-1}$, $(a_x = 25,\ b_x = 40)$; $\gamma_{1R} = 0.02\ \mathrm{s}^{-1}$, $\gamma_{2R} = \gamma_{2x}$, $k_{2R} = k_{2x}$, $(b_R = 20)$; left: $k_{1R} = 4 \times 10^{-5}\ \mathrm{s}^{-1}$ $(a_R = 0.1)$, right: $k_{1R} = 4 \times 10^{-3}\ \mathrm{s}^{-1}$ $(a_R = 10)$.

parameters describing production of the two proteins x and R, respectively, and B is a normalization factor. If fluctuations in R are slow, the joint distribution is given by Bayes rule:

$$p(R, x) = p(R)p(x|R) = Bp(R)x^{a_x c(R)-1}e^{-x/b_x}. \quad (11)$$

The numerical simulation and analytical solution for this scenario are illustrated in Fig. 3. Again, no fitting parameters are used, and the analytical form captures the features of the numerical result. This is an example of extrinsic noise, where fluctuations in an upstream regulator R contribute to the noise measured in x. Note that the analytical form shows that the effect of extrinsic noise in this case is asymmetric, increasing $p(x)$ mainly for small values of x [see $p(x)$ in top-left panel of Fig. 3]. Similar joint distributions were recently measured in live cells [22].

We have recently shown experimentally that dynamic information about the protein production process is preserved in, and can be extracted from the steady-state distribution [10]. Hence, this analytical framework can facilitate comparison of experimental data with suggested microscopic models. Because of the availability of whole-genome fluorescent protein fusion libraries and ease of scaling up flow cytometry studies [13,14], this analytical model provides a framework for extracting quantitative information on transcription and translation processes for genes in the whole genome from measured protein concentration distributions. It may be possible to tabulate a and b values from such data sets and identify common genetic regulatory elements in genes with similar noise behaviors. In addition, the ability to calculate analytically the shape of the protein distribution facilitates the understanding of stochasticity in biological decision making processes, such as developmental switches.

We thank Sam Kou, Hong Qian, and Johan Elf for helpful discussions and careful reading of the manuscript. We thank Eedo Mizrahi and Joel Stavans for help with the stochastic simulations code. This work was funded by the Department of Energy, Office of Biological and Environmental Research, Genomics: GTL Program, and in part by a National Institute of Health (NIH) Grant. L. C. is supported by the National Science Foundation, and N. F. by the Human Frontiers Science Program.

[1] J. Paulsson, Nature (London) **427**, 415 (2004).
[2] H. H. McAdams and A. Arkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 814 (1997).
[3] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, Nature Rev. Genet. **6**, 451 (2005).
[4] O. G. Berg, J. Theor. Biol. **71**, 587 (1978).
[5] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).
[6] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins, Nature (London) **422**, 633 (2003).
[7] E. M. Ozbudak *et al.*, Nat. Genet. **31**, 69 (2002).
[8] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).
[9] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).
[10] L. Cai, N. Friedman, and X. S. Xie, Nature (London) **440**, 358 (2006).
[11] J. Yu *et al.*, Science **311**, 1600 (2006).
[12] N. G. van Kampen, *Stochastic Process in Physics and Chemistry* (North-Holland, Amsterdam, 1992).
[13] J. R. S. Newman *et al.*, Nature (London) **441**, 840 (2006).
[14] A. Bar-Even *et al.*, Nat. Genet. **38**, 636 (2006).
[15] J. Paulsson and M. Ehrenberg, Phys. Rev. Lett. **84**, 5447 (2000).
[16] M. Thattai and A. van Oudenaarden, Proc. Natl. Acad. Sci. U.S.A. **98**, 8614 (2001).
[17] S. Ghaemmaghami *et al.*, Nature (London) **425**, 737 (2003).
[18] D. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[19] N. Rosenfeld, M. B. Elowitz, and U. Alon, J. Mol. Biol. **323**, 785 (2002).
[20] A. Becskei and L. Serrano, Nature (London) **405**, 590 (2000); F. J. Isaacs, J. Hasty, C. R. Cantor, and J. J. Collins, Proc. Natl. Acad. Sci. U.S.A. **100**, 7714 (2003); E. M. Ozbudak *et al.*, Nature (London) **427**, 737 (2004).
[21] T. B. Kepler and T. C. Elston, Biophys. J. **81**, 3116 (2001).
[22] N. Rosenfeld *et al.*, Science **307**, 1962 (2005); J. Pedraza and A. van Oudenaarden, Science **307**, 1965 (2005).