Stochasticity of gene products from transcriptional pulsing

Srividya Iyer-Biswas,[1,*] F. Hayot,[2] and C. Jayaprakash[1]
[1]*Department of Physics, Ohio State University, Woodruff Avenue, Columbus, Ohio 43210, USA*
[2]*Department of Neurology, Mount Sinai School of Medicine, Levy Place, New York, New York 10029, USA*
(Received 28 October 2008; published 23 March 2009)

Transcriptional pulsing has been observed in both prokaryotes and eukaryotes and plays a crucial role in cell-to-cell variability of protein and mRNA numbers. An important issue is how the time constants associated with episodes of transcriptional bursting and mRNA and protein degradation rates lead to different cellular mRNA and protein distributions, starting from the transient regime leading to the steady state. We address this by deriving and then investigating the exact time-dependent solution of the master equation for a transcriptional pulsing model of mRNA distributions. We find a plethora of results. We show that, among others, bimodal and long-tailed (power-law) distributions occur in the steady state as the rate constants are varied over biologically significant time scales. Since steady state may not be reached experimentally we present results for the time evolution of the distributions. Because cellular behavior is determined by proteins, we also investigate the effect of the different mRNA distributions on the corresponding protein distributions using numerical simulations.

DOI: 10.1103/PhysRevE.79.031911 PACS number(s): 87.10.Mn

I. INTRODUCTION

Cell-to-cell variability in mRNA and protein numbers is now recognized as a major aspect of cellular response to stimuli, a variability which is hidden in cell population studies. The most egregious example of the latter is provided in cases where a graded average response hides the all-or-nothing behavior of single cells [1–3]. Variability of cellular response can have many origins, which are generally classified as extrinsic and intrinsic noise or fluctuations [4]. Many studies, both experimental and theoretical from bacteria to eukaryotes, have been undertaken to characterize intrinsic and extrinsic fluctuations [4–11]. Extrinsic fluctuations can have multiple origins, such as variations from cell to cell in the number of regulatory molecules or signaling cascade components. The source of intrinsic fluctuations is the random occurrence of reactions that can lead to variability for genetically identical cells in identical fixed environments. When the number of molecules is small, intrinsic noise can play a significant role in determining the behavior of individual cells. The difficulty in determining rate constants experimentally and the possibility of experiments being either in the steady state or in the transient regime make exact time-dependent solutions to models especially useful in the interpreting observed behavior of intrinsic noise. In this paper we present the exact time-dependent solution to a widely used model for transcriptional noise and discuss its implications.

II. EXPERIMENTAL AND THEORETICAL BACKGROUNDS

Intrinsic fluctuations arise from either noisy transcription, noisy translation, or both, the effects of which can be measured in single cell mRNA and protein experiments. The simplest model of mRNA and protein number distributions is to consider both transcription and translation as Poisson processes [7]. Recent experimental studies of mRNA distributions have shown strong evidence for transcriptional noise beyond what can be described by a simple Poisson process, i.e., distributions with variances significantly larger than the mean have been observed. In particular, transcriptional pulsing, where bursts of transcription alternate with quiescent periods, has been observed in both prokaryotes and eukaryotes. Raser and O'Shea [5], who studied intrinsic and extrinsic noise in *Saccharomyces cerevisiae*, showed that the noise associated with a particular promoter could be explained in a transcriptional pulsing model and confirmed it by mutational analysis. Transcriptional bursts were recorded in *E. coli* [12] by following mRNA production in time and their statistics computed. Evidence for a pulsing model of transcription, obtained from fluorescent microscopy, has also been presented for the expression of the discoidin Ia gene of *Dictyostelium* [13]. Transcriptional bursts have as well been detected in Chinese hamster ovary cells [10]. In these experiments the production of mRNA occurs in a sequence of bursts of transcriptional activity separated by quiescent periods. Transcriptional bursting, an intrinsically random phenomenon, thus becomes an important element to consider when evaluating cell-to-cell variability. One can predict that in many cases it will be a significant part of overall noise and most certainly of intrinsic noise.

Our study focuses on the consequences of transcriptional bursting in a simplified model of transcription that has been the subject of many studies and is believed to encapsulate the key features of bursting [2,5,10,12,14–16]. Related models that include feedback have been studied theoretically [17–19]. The complex phenomena that can occur in transcription (chromatin remodeling, enhanceosome formation, preinitiation assembly, etc.) are modeled through positing two states of gene activity: an inactive state, where no transcription occurs, and an active one, in which transcription occurs according to a Poisson process. The production of mRNA is thus pulsatile: temporally there are periods of in-

*srividya@mps.ohio-state.edu

activity interspersed with periods or bursts of transcriptional activity. Qualitative features of this model were presented in Ref. [2] and aspects of it relating to bursts explored and discussed in Ref. [12]. Raj *et al.* [10] provided a steady-state solution to the master equation of the transcriptional model considered here and analyzed it for some ranges of the rate constants relevant to their experiments. A related model with transcription in the presence of feedback without translation was solved exactly in the steady state by Hornos *et al.* [17].

In this paper, we provide a comprehensive analysis of a transcriptional pulsing model with an exact *transient* solution to the master equation for mRNA production. Given the range of time scales that can occurs in transcriptional processes in different organisms, it is important to highlight the most significant behaviors that can arise in this model and investigate how these depend on the many time scales. While the steady-state distribution was known [10] we have explored the entire range of behaviors that can arise: in particular we find a bimodal distribution that, surprisingly, exhibits power-law behavior between the peaks. A virtue of the model studied in this paper is that it is amenable to an analytic determination of the probability distribution of mRNA number as a function of time. To the best of our knowledge, the only previous case where a time-dependent analytical solution of a transcriptional model was known is that of the simple birth-death process. The results from our solution are directly relevant to experimental studies since the mRNA distribution may well be in the transient state and it is therefore useful to determine the time evolution of the distributions and characterize the time scales over which steady state is attained. Our time-dependent analytic solution allows us to address these issues in detail, revealing in particular how the mRNA lifetime shapes the time-dependent mRNA distribution.

Cellular behavior is however typically determined by proteins and not the corresponding mRNA. Therefore, we extend the model to one in which proteins are produced according to a birth-death process from mRNA produced from the transcriptional bursting model. This model has been used to interpret protein Fano factors measured experimentally in the transient and steady states [20]. An important issue is to what extent the protein distributions follow the mRNA distributions. It has been found experimentally [10] that when the protein decays faster than the parent mRNA, the two distributions are similar, while when the protein lifetime is longer, the distributions are different; this was explained by numerical simulations using the Gillespie algorithm [21] for experimentally relevant numbers for the pulsing model. We have performed numerical simulations of the model using the Gillespie algorithm to obtain the steady-state protein probability distributions for all the different behaviors of the underlying mRNA distribution that we have uncovered. That the two distributions are similar when the protein lifetime is shorter than the mRNA lifetime is not surprising. On the other hand, when the protein lifetime is longer the two distributions are not necessarily dissimilar; we have identified cases, depending on the pulsing rate, for which the two distributions can be similar or different. Our comprehensive study of the transcriptional pulsing model highlights how the steady-state shapes of mRNA and protein distributions depend on the ratios of rate constants and determine the time evolution to steady state. We thus provide an overview of possible behaviors which yields a framework for interpreting experimental results on transcriptional bursting across prokaryotes and eukaryotes.

III. MODEL AND FORMALISM

We study a model of transcriptional pulsing described by the following reactions where D and D^* denote the gene in the inactive and active states, respectively:

$$D \underset{c_b}{\overset{c_f}{\rightleftharpoons}} D^*, \tag{1}$$

$$D^* \xrightarrow{k_b} D^* + M, \tag{2}$$

$$M \xrightarrow{k_d} \varnothing . \tag{3}$$

The first equation describes the switching "on" and "off" of the gene at the rates c_f and c_b, respectively. The second and third equations describe transcription of the mRNA at a constant rate k_b in the active state and the subsequent degradation of the mRNA at a rate k_d. We present results for $P(m,t)$, the probability that the cell contains m mRNA molecules at a time t which describes cell-to-cell variability of mRNA copy number as a function of time.

It is convenient to define $P_0(m,t)$ and $P_1(m,t)$ to be the probability that at time t the cell has m mRNA molecules and the gene is in the inactive and active states, respectively. It is straightforward to write down the master equation for the two probabilities,

$$\begin{aligned}\frac{dP_0(m,t)}{dt} &= -c_f P_0(m,t) + c_b P_1(m,t) \\ &\quad + k_d[(m+1)P_0(m+1,t) - mP_0(m,t)] \\ &\quad \times \frac{dP_1(m,t)}{dt} = c_f P_0(m,t) - c_b P_1(m,t) \\ &\quad + k_d[(m+1)P_1(m+1,t) - mP_1(m,t)] \\ &\quad + k_b[P_1(m-1,t) - P_1(m,t)].\end{aligned} \tag{4}$$

We define the generating functions

$$G_\alpha(z,t) \equiv \sum_{m=0}^{\infty} z^m P_\alpha(m,t)$$

for $\alpha=0$ and 1. If we can evaluate $G(z,t)=G_0(z,t)+G_1(z,t)$ exactly, then the probability of having m mRNA transcripts at time t can be obtained by extracting the coefficient of the z^m term. We derive the equations obeyed by G_0 and G_1 and solve them exactly for the initial condition with zero mRNA, i.e., $P(m,0)=\delta_{m,0}$. The details are relegated to the Appendix. We find

$$\begin{aligned}G(z,t) &= F_s(t)\Phi[c_f, c_f + c_b; -k_b(1-z)] \\ &\quad + F_{ns}(t)\Phi[1-c_b, 2-c_f-c_b; -k_b(1-z)],\end{aligned} \tag{5}$$

where Φ is the (Kummer) confluent hypergeometric function

[22,23]. All the rate constants are measured in units of the decay rate k_d and time in units of k_d^{-1}. The coefficients $F_s(t)$ and $F_{ns}(t)$ are given by

$$F_s(t) = \Phi[-c_f, 1 - c_f - c_b; k_b e^{-t}(1-z)] \tag{6}$$

and

$$F_{ns}(t) = -\frac{c_f k_b (1-z)}{(c_f + c_b)(1 - c_f - c_b)} e^{-(c_f + c_b)t} \times \Phi[c_b, 1 + c_f + c_b; k_b e^{-t}(1-z)]. \tag{7}$$

In the steady-state limit $F_s \to 1$ and $F_{ns} \to 0$. For a general initial condition where the initial distribution $P(m,0)$ corresponds to the generating function $Q_0(z)$ the result can be obtained by dividing the left-hand side of Eq. (5) by $Q_0[1 - e^{-t}(1-z)]$.

A. Time scales

The importance of the time scales of different reactions in determining the behavior of models such as the one considered here has been discussed earlier [2,10,17–19]. We begin by briefly summarizing the different time scales that determine both the steady-state and temporal behaviors of the mRNA distributions. The model has four rate constants: the forward and backward rates for the gene to switch between the active and inactive states and the transcription and degradation rates that govern mRNA numbers, leading to three independent dimensionless ratios. The equation obeyed by the probability for the DNA to be in the excited state, denoted by $Q_1(t)$ can be obtained directly from Eq. (1): $dQ_1/dt = c_f - (c_f + c_b)Q_1$. This shows that the effective DNA relaxation rate to the steady state is governed by $c_f + c_b$. The mean mRNA number obeys the equation $d\langle m(t)\rangle/dt = -k_d\langle m\rangle + k_b Q_1(t)$. It is thus clear that the temporal behavior of the mean mRNA number is determined by the rates k_d and $c_f + c_b$, the latter entering since it determines the dynamics of the transcriptionally active state. As long as $k_d < c_f + c_b$, the mRNA decay rate sets the time scale over which relaxation to the steady state occurs. We find that the results of our exact solution can be interpreted in the most natural and transparent way when we measure time in units of k_d^{-1}, i.e., in terms of the mRNA lifetime. Thus we will use the three dimensionless ratios k_b/k_d, c_f/k_d, and c_b/k_d to organize our results. The mean number of mRNA in the steady state is given by the product of $c_f/(c_f + c_b)$, the fraction of the time the gene is in the activated state, and k_b/k_d, the mean value of mRNA if the gene is always "on." The ratio k_b/k_d clearly sets the scale for the number of mRNA and increasing it extends the range over which $P(m)$ is appreciable without a significant change of shape. The remaining ratios c_f/k_d and c_b/k_d determine the shape of the distribution.

B. Superposition of Poisson distributions

If the gene is always "on" mRNA kinetics follows a birth-death process. If no mRNAs are present initially, this implies that the mRNA distribution follows a Poisson distribution with the Poisson parameter, λ, given by the mean, which in steady state is equal to k_b/k_d, the ratio of transcription and degradation rates. Since the gene flips between the on and off states with the rates determined by c_f and c_b, the mRNA distribution is determined by a stochastic transcription rate $k_b\zeta(t)$, where $\zeta(t)$ is a dichotomous noise, i.e., it assumes values 0 or 1 corresponding to the gene being in the inactive or active state, respectively. The dynamics of the random variable ζ are determined by the stochastic chemical reaction described by Eq. (1). Thus the distribution of the mRNA number is described by a Poisson process in which the parameter λ itself is stochastic, a process known as a doubly stochastic Poisson process [24]. This provides an intuitively appealing way to view our exact result for $P(m,t)$: it can be written as a superposition of Poisson distributions with the Poisson mean itself distributed (see the Appendix). A formal proof that such a representation exists for this model and is unique, for the general time-dependent case, will be provided elsewhere.

We write

$$P(m,t) = \int d\lambda \rho(\lambda, t) e^{-\lambda} \frac{\lambda^m}{m!}, \tag{8}$$

where $\rho(\lambda,t)$ is the probability density of the random variable λ. Thus $\rho(\lambda,t)$ contains the same information as the probability distribution. The density in the steady sate $\rho(\lambda)$ can be calculated from the exact generating function and can be shown to be given by the scaled β distribution

$$\rho_{ss}(\lambda) = \left(\frac{k_b}{k_d}\right)^{1-(c_f+c_b)/k_d} \frac{\Gamma(c_f/k_d + c_b/k_d)}{\Gamma(c_f/k_d)\Gamma(c_b/k_d)} \times \lambda^{-1+c_f/k_d} \left(\frac{k_b}{k_d} - \lambda\right)^{-1+c_b/k_d} \tag{9}$$

for $0 < \lambda < k_b/k_d$ and zero otherwise. The maximum allowed value of λ corresponds to the gene being always on. We find it mathematically convenient to use this representation to derive the some of the results presented later; in addition it provides a simple way of visualizing the actual distribution in terms of Poisson distributions. Such superpositions have been studied in the context of stochastic processes, for example, in [25].

C. Steady-state distributions

Before discussing the full time-dependent distributions, in order to place our results in the context of the distributions that are asymptotically realized, we now describe the variety of steady-state distributions that occur in different regions of parameter space. These have been obtained by evaluating the exact steady-state distribution. While the steady-state distribution for this model [10] is in the literature, a complete characterization of the forms of the distribution that occur for different time scales is not available. Following the earlier discussion of time scales we classify the distributions by plotting c_b/k_d and c_f/k_d, respectively, along the x and y axes. We fix $k_b/k_d = 100$, a value chosen to make the range of m values over which prototypical behavior obtains broad. Changing k_b/k_d only alters the extent of the distribution without appreciably changing its shape. The results displayed

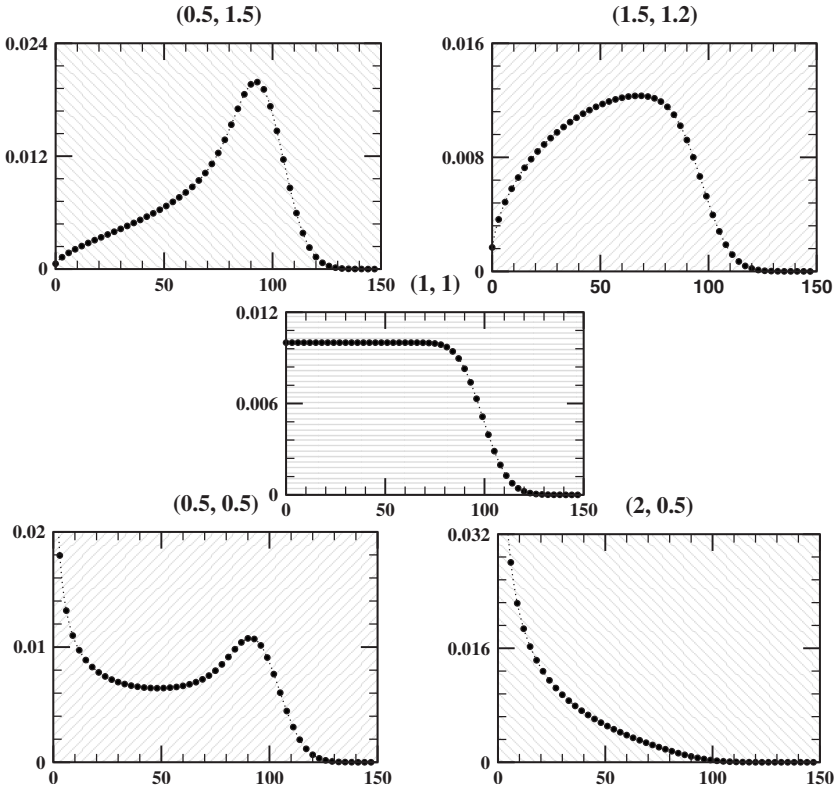

FIG. 1. Steady-state mRNA distributions $P(m)$ vs m, labeled by (c_b, c_f) in units of k_d obtained from the exact solution. The mRNA transcription rate k_b is $100k_d$ for all the distributions. Figure shows prototype distributions for the four major regimes c_f, $c_b \lessgtr k_d$ in parameter space corresponding to the four quadrants. We show the case where all the regimes overlap with the distribution for $c_f=c_b=k_d$ which is flat over roughly two decades for the values we have chosen. Bimodals with power-law decay in between are obtained in the lower left panel for c_f, $c_b<k_d$ while a power law occurs in the lower right panel. These distributions were plotted using the exact steady-state result obtained from Eqs. (5)–(7).

in Fig. 1 provide a bird's eye view of the strikingly different mRNA distributions that arise in different regions of parameter space. The experiments are performed on a variety of organisms both prokaryotic and eukaryotic; while rate constants are not known, quite different time scales occur. Therefore, we investigate values of c_f/k_d and c_b/k_d that encompass different biologically significant cases: for example, our choices include the vastly different rate constant values in the experiments of Raser and O'Shea [5] and Raj *et al.* [10].

We start with the interesting case displayed in the bottom left figure in Fig. 1, when the mRNA half life is shorter than the time scales over which the gene turns on or off, i.e., c_f, $c_b<k_d$. In the steady state at any given time, the gene is off in some cells. Since the mean duration of the pulse is larger than the mean lifetime of the mRNA, the mRNA produced in the previous occurrence of the on state would probably have decayed and so the number of transcripts will usually be small in these cells. This causes a peak in the mRNA distribution near $m=0$. In those cells in which the gene is on at the time of observation the number of transcripts can display a broad range of values depending on how long the gene was active as compared to the mRNA lifetime. Thus, we expect to observe a bimodal distribution, as was qualitatively argued in [2]. The result is shown in the lower left quadrant of Fig. 1. One finds a peak at $m=0$ and another peak at large $(\sim k_b/k_d)$ m values. If the two peaks are well separated, i.e., $k_b \gg k_d$, much of the intermediate region displays a power-law behavior with a power $-1+\frac{c_f}{k_d}$, a new result that follows from our analysis. The representation of $P(m)$ in terms of $\rho(\lambda)$ discussed earlier allows us to do a saddle-point approximation and deduce the exponent. We have verified this by plotting the exact distribution using MATHEMATICA and fitting it. The broad distribution of mRNA values reflects the broad range of times for which the gene has been active in different cells in the steady state. It is useful to remark that bimodal distributions have been obtained in models with feedback [8,26]. In contrast, in the transcriptional pulsing model, bimodality is obtained *without* the presence of a feedback loop.

Now imagine that we keep c_f fixed and vary c_b so that it is larger than the mRNA decay rate. This leads to a power-law behavior with the same exponent as in the bimodal region for mRNA numbers less than approximately $\frac{k_b}{c_b}$ which for appropriate choices of the rate constants can correspond to a significant range of mRNA values. This monotonic power-law decay, obtained in the case $c_f<k_d$ and $c_b>k_d$, is illustrated in the lower right quadrant of Fig. 1. This case has been treated analytically in a continuum approximation in [10,14].

When both the activation and inactivation rates are rapid, i.e., c_f, $c_b \gg k_d$, eliminating the fast reactions naively yields a simple birth-death process for the mRNA with an effective transcription rate $k_b c_f/(c_f+c_b)$. This would lead one to expect a Poisson distribution for the mRNA number. However, in this "quadrant," i.e., for c_f, $c_b>k_d$, the observed distribution has a broad single-humped shape as displayed in the upper right quadrant of Fig. 1, much broader than a Poisson distribution. This broadening occurs because the parameter λ itself is stochastic. When $c_f>k_d>c_b$ the gene is on most of the time. In the upper left quadrant of Fig. 1 the distribution is Poisson to a very good approximation. For the case where the four regions overlap when c_f, $c_b \sim k_d$, the distribution interpolates between these different possibilities. When c_f

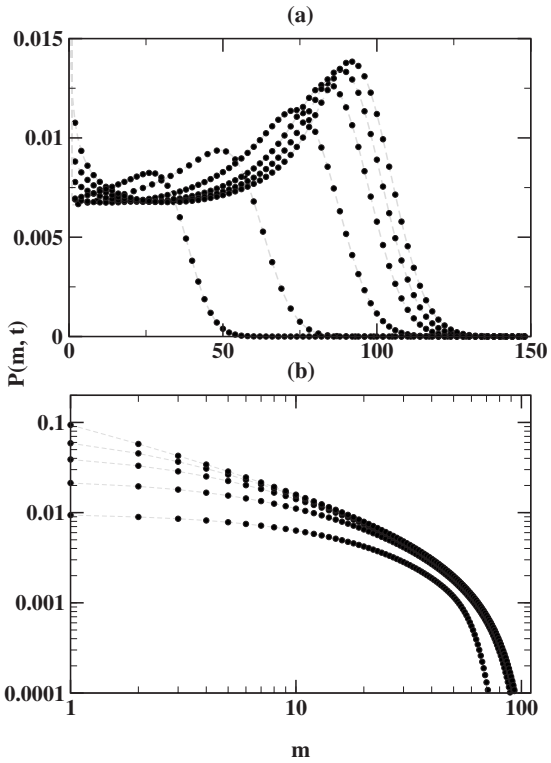

FIG. 2. Time evolution of $P(m,t)$ toward steady state as a function of m (from the exact expression in the Appendix) for the initial condition $P(0,0)=1$. The weights at $m=0$ are not displayed. (a) The time evolution of the mRNA distribution in the bimodal regime for $c_f=0.75k_d$, $c_b=0.5k_d$, and $k_b=100k_d$ at times $t=0.5,1,2,3,4$, and ∞ in units of k_d^{-1} is shown. As time increases, the weight at $m=0$ decreases while the mode at higher values acquires more weight and occurs at larger values of m. The weights at $m=0$ at these times are 0.69, 0.48, 0.23, 0.11, 0.06, and 0.016, respectively. (b) Log-log plot of $P(m,t)$ in the power-law regime is shown for $c_f=0.25k_d$, $c_b=2.5k_d$, and $k_b=100k_d$ at times $t=1,2,3,4$, and ∞ in units of k_d^{-1}. As time increases, the weight at $m=0$ decreases while the slope of the curve in the power-law region increases. The weights at $m=0$ at these times are, respectively, 0.79, 0.63, 0.52, 0.45, and 0.38.

$=c_b$ we find a flat region in m with a cutoff of roughly k_b/k_d (see Fig. 1, center).

D. Time evolution of probability distributions

Because of the range of different time scales, it can happen that at the time when measurements are made, the biological system has not attained steady state: we present results for how the distributions evolve to a steady state from an initial state with no mRNA and the gene in its inactive state. Using the time-dependent result for the distribution [Eqs. (5)–(7)], we evaluate the evolution using MATHEMATICA and plot the complete probability distribution as a function of time. Consider the case c_f, $c_b<k_d$ (bottom left in Fig. 1), where the mRNA distribution displays bimodality. Here the mRNA decay rate sets the scale for approach to the steady state. Figure 2(a) shows the evolution of the bimodal distribution as a function of time. For the given initial condition the second peak away from zero develops after a period of roughly twice the mRNA half life. Steady-state behavior sets in at about 4–5 times k_d^{-1}. It is clearly possible that, depending on the relative values of the cell cycle time and the mRNA half life, steady state and, therefore, full bimodality may not be observable.

Consider now the time evolution of the distribution that exhibits a range of power-law behavior in the steady state featured in the bottom right of Fig. 1. In Fig. 2(b) we plot $P(m,t)$ vs m on a double-logarithmic plot. We have chosen $c_f=0.25k_d$ and $c_b=2.5k_d$ to illustrate this case. Larger values of the transcription rate lead to a larger range of mRNA number over which power-law behavior obtains. It is clear that the exponent of the power law increases in magnitude with time and saturates at the steady-state value for t greater than about $4k_d^{-1}$. Thus the shape of the distribution depends crucially on the time (measured in units of the decay time) when experimental measurements are made.

E. Remarks on experiments

From the examples given in Fig. 1 it is clear that the complete probability distribution of mRNA number is required to characterize the behavior of the transcriptional pulsing model. Some of the experimental investigations, however, have focused on the variance. In the following we make remarks on attempts to represent a mRNA distribution by its mean and variance only.

There is danger in characterizing distributions solely by their mean and variance which can be calculated easily. An important result in Ref. [5] is the decrease in the noise strength (Fano factor, η, defined as the ratio of the variance σ^2 to the mean) with increase in the mean for genes with different activation rates. Here we show that a wide variety of distributions can underlie this correlation between the noise strength and the mean. The increase in the mean can be obtained in the model through an increase in the activating rate, c_f, and experimentally through mutations of an appropriate promoter [5]. Even though a smooth curve is obtained for the decrease of noise strength with the mean, the full mRNA distribution can differ significantly for different points along the curve. For specificity, we choose parameter values $k_b=200k_d$ and $c_b=k_d$ and vary the forward rate c_f for gene activation which changes the mean value. The result is shown in Fig. 3(a) and is similar to that obtained experimentally. The full probability distribution at three values of c_f, namely, $c_f=0.1k_d$, k_d, and $10k_d$, corresponding to mean values of 18, 100, and 181 respectively, are shown in Fig. 3(b): the distribution ranges from power-law decay of $P(m)$ for $c_f=0.1k_d$ to a broadened Poisson distribution for $c_f=10k_d$. As emphasized earlier, the value of mRNA degradation rate plays an important role in determining the type of mRNA distribution.

There are two popular measures of noise in terms of the first two moments of a probability distribution: the coefficient of noise, ξ, defined as the ratio of the standard deviation σ to the mean μ, and the noise strength or Fano factor, η, defined as the ratio of the variance σ^2 to the mean. The latter has the value of unity for a Poisson distribution and is therefore convenient for describing deviations from Poisson be-

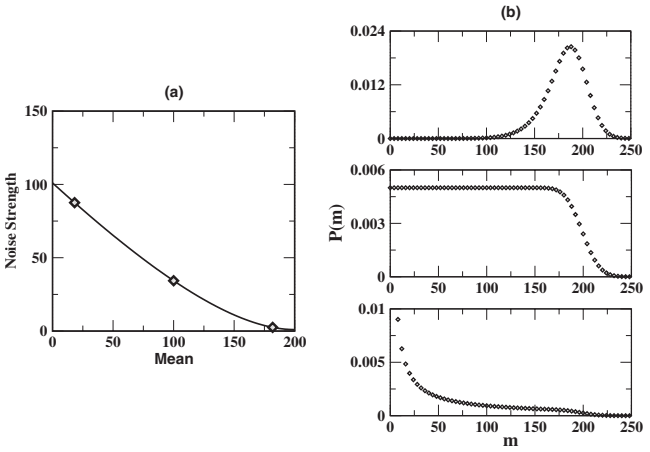

FIG. 3. Smoothly varying noise strengths as the activation rate c_f is varied can correspond to very different probability distributions. (a) Variation of noise strength (mRNA Fano factor) with activation rate for $k_b=200k_d$ and $c_b=k_d$. (b) The steady-state distributions $P(m)$ corresponding to the (diamond-shaped) points marked in (a). The points in (a) going from right to left and the corresponding figures from top to bottom in (b) are for $c_f/k_d=10$, 1 and 0.1, respectively. (b) Illustrates how different points on the same curve (a) can be associated with dramatically different mRNA distributions. These distributions were plotted using the exact results in Eqs. (5)–(7).

havior. In Fig. 4 we display constant η contours as a function of c_f/k_d and c_b/k_d for a fixed value of k_b/k_d on a logarithmic scale to encompass a broad range of parameter variation. When $c_f<k_d$ and $c_b>k_d$, the steady-state distribution $P(m)$ is monotonically decreasing and has a power-law region. In this region, a first approximation η is independent of c_f [and $\approx 1+k_b/(k_d+c_b)$] and the contours are roughly parallel to the c_f axis. This emphasizes the possibility that σ^2/μ is a constant for systems with power-law behavior in which c_f varies over a broad range of values. Since as we show later, the protein distribution can reflect the behavior of the corresponding mRNA distribution, the protein distribution can show a similar constancy of the Fano factor. Such a behavior has been observed experimentally in [20] where a pulsing model was discussed. In the region where c_f, $c_b<k_d$, then $\eta \approx 1+(k_b/k_d)/(1+c_f/c_b)$ and thus depends only on c_f/c_b. This is consistent with the contours in this region being straight lines with slope 1. For the region with c_f, $c_b>k_d$, there is rapid switching between on and off states and the Fano factor depends weakly on the rates c_f and c_b.

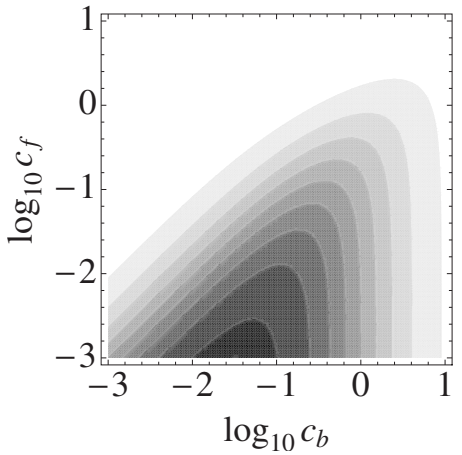

FIG. 4. Contour plot of the (exact) mRNA Fano factor η as c_b and c_f (in units of k_d) are varied, for $k_b=1000k_d$, in the steady state; c_b and c_f are varied on a $\log_{10}$ scale over 5 decades. Nine contours for different values of η are placed at intervals of 100, from 1 to 1001 with η increasing from light to dark values.

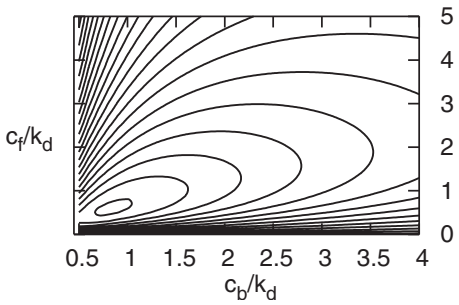

FIG. 5. Contour plot of the Shannon entropy of $P(m)$ defined in Eq. (10) for $k_b=100k_d$; separation between contours is 0.05 and the Shannon entropy decreases outward from the central contour.

A measure of noise that is sensitive to the shape of the entire distribution is the Shannon entropy. Given the exact solution we can directly evaluate the Shannon entropy associated with the steady-state distribution $P(m)$ defined by

$$\mathcal{S} = -\sum_i P(m)\log_2 P(m). \tag{10}$$

In Fig. 5 we display contours of constant entropy as a function of the forward and backward rates c_f and c_b. Not surprisingly, the values c_f/k_d, $c_b/k_d \simeq 1$ yield the largest entropy; $P(m)$ is then $\approx k_d/k_b$ for m from 1 to k_b/k_d and the entropy is well approximated by $\log(k_b/k_d)$. The power-law distribution also provides a range of values for the rates in which larger values of entropy can be obtained indicating that greater information content than the other distributions. Since the Shannon entropy is a measure of the amount of information required to describe the random variable on average it may be helpful in the interpretation of data.

We comment on the relevance of long-tailed distributions in biological systems. Consider dendritic cells that form a key component of the innate immune system and act as first responders to infection. Upon infection a series of complex steps occur culminating in the maturation of the dendritic cells that present the antigen to T cells after migration to the lymphoid organs, thereby connecting the innate and adaptive immune systems. The key first step in this process, in response to a virus for example, is the production of interferon β, a cytokine that is secreted into the intercellular medium and primes other dendritic cells. There are clearly several constraints on this step. It should be tightly controlled so that production occurs only in response to a pathogen; excess production of cytokines known as a cytokine storm that can have adverse consequences must be avoided and at the same time sufficient number must be produced by at least a few of the cells so that adequate response is initiated. The last two

requirements can be met by a broad distribution of mRNA. In addition, viruses such as those influenza mutate resulting in greater ability to evade the antagonistic action of the immune system. One may speculate that a broad distribution of the mRNA and the corresponding protein may confer greater ability to the dendritic cell to overcome viral mutations. For example, if the virus were to increase the rate of degradation of the mRNA or protein resulting in a narrowing of the distribution then some of the dendritic cells would still produce enough of the cytokines to respond to the pathogen. It has been recently observed that the interferon-β mRNA distribution in human dendritic cells is indeed broad [11].

F. Protein distributions

Given the variety of mRNA distributions that can result from genes undergoing transcriptional pulsing, it is important to understand how this affects the probability distributions for the corresponding protein. While a careful answer to this question would require detailed modeling of mRNA translocation and translation, we address this issue by extending the transcriptional pulsing model discussed thus far to include the following additional reactions that model production and degradation of the protein [7,10]:

$$M \xrightarrow{p_b} M + P, \tag{11}$$

$$P \xrightarrow{p_d} \varnothing . \tag{12}$$

The effective protein degradation rate would include contributions from dimerization and other gene-specific processes involving the loss of proteins.

The results from the numerical simulations we have performed using the Gillespie algorithm are consistent with the expectation that the steady-state protein distribution will mirror the mRNA distribution when the protein lifetime is shorter than the mRNA time scale. As pointed out earlier the time scale of mRNA dynamics is determined by the smaller of k_d and c_f+c_b. Even though for biologically relevant scenarios the protein lifetime is typically longer than the mRNA lifetime, protein distributions that mimic the mRNA distribution may thus be obtained for $c_f+c_b<p_d<k_d$. Conversely, when the protein lifetime is longer than the mRNA's, then the protein distribution may be qualitatively different from the mRNA's.

A neat argument has been proposed recently [27] for a model in which the gene is always on and the protein production occurs in instantaneous bursts separated by intervals that are exponentially distributed with an exponentially distributed number of proteins produced in each burst. They conclude that if the effective protein degradation is very slow compared to that of the mRNA, the protein distribution can be approximated by a gamma distribution. Similar arguments have previously been made in [28]. Their discussion is valid only when the protein lifetime is much larger than the mRNA lifetime and tacitly assumes that the number of mRNA is small $O(1)$. We find that in our model when the number of mRNAs display a power-law distribution the gamma distribution nevertheless provides a reasonable approximation for the proteins when $c_b<k_b$ and $c_f/(c_f+c_b)<k_b/k_d$. However, their general claim that the gamma distribution obtains even in the case of models such as ours is not true without the restrictions we have pointed out. Another case in point, for parameter values leading to a bimodal distribution for the mRNA, where c_f+c_b determines the mRNA time scale, the gamma distribution is not a good approximation whether or not the protein distributions are bimodal.

IV. DISCUSSION

In this paper, we have presented results for the time-dependent and steady-state probability distributions for mRNA based on an exact time-dependent solution to the master equation for a transcriptional pulsing model. A variety of mRNA distributions occur in different regimes of rate constants. Our aim is to provide a guide for the interpretation of data on cell-to-cell variability that could arise from transcriptional pulsing, both in the transient and the steady-state regimes. Transcriptional pulsing, entailed by the dynamics of chromatin remodeling, reinitiation, and similar processes, appears as a straightforward mechanism leading to bimodality and also to mRNA distributions with long tails. Having an exact analytic time-dependent solution for this model is especially useful given that long-tailed distributions are found to occur for a large range of biologically relevant parameters. Long-tailed distributions of mRNA have been seen in a variety of systems: the experiments of Raser and O'Shea [5] showed evidence for long tails which they attribute to transcriptional pulsing. In experiments on the gene ActB in cells from mouse pancreatic islets, the distribution of mRNA number m was found to be consistent with a log-normal distribution that would correspond to $P(m)\sim m^{-1}$ over some range of m [29]. More recently, a long-tailed distribution of the Interferon-β gene transcripts has been found in human dendritic cells [11]. For the latter two experiments our results should help clarify whether the origin of the mRNA behavior lies in transcriptional pulsing. Clearly, an important issue is to determine whether the distribution of the protein coded by the mRNA follows the corresponding mRNA distribution. We have used numerical simulations to determine when the two distributions are similar and when they are different from those of the mRNA. We presented results for the time dependence of the protein Fano factor. Our results on the range of cell-to-cell variability of mRNA and protein responses due to transcriptional pulsing should provide significant help in interpreting experiments.

ACKNOWLEDGMENTS

This work was supported by the National Institute of Allergy and Infectious Diseases through Contract No. HHSN266200500021C. We are grateful to Stuart Sealfon, James Wetmur, and Jianzhong Hu for discussions that stimulated this work. We thank Stuart Sealfon for a careful reading of the paper and comments. One of us (S.I.B.) thanks Rudro Rana Biswas for productive discussions. We thank the referee for a careful reading of the paper and suggestions.

APPENDIX: TIME DEPENDENT SOLUTION OF THE MASTER EQUATION

For the set of reactions described by Eqs. (1)–(3) in the text, we define $P_0(m,t)$ and $P_1(m,t)$ to be the probability that at time t the cell has m mRNA molecules and the gene is in the inactive and active states, respectively. It is straightforward to write down the master equation for the two probabilities

$$\frac{dP_0(m,t)}{dt} = -c_f P_0(m,t) + c_b P_1(m,t) + k_d[(m+1)P_0(m+1,t) - mP_0(m,t)],$$

$$\frac{dP_1(m,t)}{dt} = c_f P_0(m,t) - c_b P_1(m,t) + k_d[(m+1)P_1(m+1,t) - mP_1(m,t)] + k_b[P_1(m-1,t) - P_1(m,t)]. \quad \text{(A1)}$$

We define the generating functions

$$G_\alpha(z,t) \equiv \sum_{m=0}^{\infty} z^m P_\alpha(m,t)$$

for $\alpha=0$ and 1. The mRNA distribution (independent of the state of the gene) is determined by the sum $G \equiv G_0+G_1$. It is easy to deduce the equations obeyed by the generating functions from the master equations (with time rescaled by k_d),

$$\partial_t G_0(z,t) = -c_f G_0(z,t) + c_b G_1(z,t) + (1-z)\partial_z G_0(z,t), \quad \text{(A2)}$$

$$\partial_t G_1(z,t) = c_f G_0(z,t) - c_b G_1(z,t) + (1-z)\partial_z G_1(z,t) - k_b(1-z)G_1(z,t). \quad \text{(A3)}$$

All the rate constants are measured in units of k_d.

We simplify the above equations using an analog of the Galilean transformation by making the change of variables $v \equiv k_b(1-z)$ and $w \equiv ve^{-t} = k_b(1-z)e^{-t}$. In terms of the transformed variables, we have

$$v\partial_v G_0 = -c_f G_0 + c_b G_1, \quad \text{(A4)}$$

$$v\partial_v G_1 = c_f G_0 - c_b G_1 - vG_1. \quad \text{(A5)}$$

Adding the two equations we have the useful relation

$$\partial_v(G_0+G_1) = -G_1. \quad \text{(A6)}$$

Note that $G_0(z,t)$ and $G_1(z,t)$ (and hence, their sum) are functions of v only and independent of $w=k_b(1-z)e^{-t}$; the dependence on w is determined by the initial conditions.

It is convenient to derive a second-order differential equation for G. Therefore we differentiate the equations for G_0 and G_1 and obtain [using Eq. (A6)]

$$v\partial_v^2 G_0 + (1+c_f+c_b+v)\partial_v G_0 + c_f G_0 = 0,$$

$$v\partial_v^2 G_1 + (1+c_f+c_b+v)\partial_v G_1 + (1+c_f)G_1 = 0. \quad \text{(A7)}$$

We add the two equations and use Eq. (A6) to obtain

$$v\partial_v^2 G + (c_f+c_b+v)\partial_v G + c_f G = 0.$$

The substitution $G(v)=e^{-v}F(v)$ shows that $F(v)$ satisfies the confluent hypergeometric equation in the canonical form. The solution is given by

$$F = A(w)\Phi(c_b, c_f+c_b; v) + B_0(w)v^{1-c_f-c_b}\Phi(1-c_f, 2-c_f-c_b; v). \quad \text{(A8)}$$

Upon using the Kummer transformation, $e^{-v}\Phi(\alpha,\gamma;v) = \Phi(\gamma-\alpha,\gamma;-v)$, we obtain

$$G = A(w)\Phi(c_f, c_f+c_b; -v) + B_0(w)v^{1-c_f-c_b}\Phi(1-c_b, 2-c_f-c_b; -v). \quad \text{(A9)}$$

In order to obtain a well-defined power series in $v = k_b(1-z)$ for the generating function we must impose

$$B_0(w) = w^{c_f+c_b}B(w) = v^{c_f+c_b}e^{-(c_f+c_b)t}B(w).$$

This yields the form

$$G = A(w)\Phi(c_f, c_f+c_b; -v) + B(w)e^{-(c_f+c_b)t}v\Phi(1-c_b, 2-c_f-c_b; -v). \quad \text{(A10)}$$

We impose the initial conditions at $t=0$ which corresponds to $w=v$. The initial condition $P(m,t=0)=\delta_{m,0}$ leads to

$$G(w=v,v) = 1. \quad \text{(A11)}$$

For an arbitrary initial state described by the generating function $Q_0(z)$ the right-hand side is $Q_0(1-k_b^{-1}v)$ where we have used $v=k_b(1-z)$. We assume that the gene is initially in the inactive state and thus $G_1(z,t=0)=0$. The additional condition that arises from Eq. (A6) implies

$$\partial_v G(w,v)|_{w=v} = 0. \quad \text{(A12)}$$

Imposing these conditions we determine the unknown functions A and B_0. This involves judicious use of the Wronskian-type identity

$$\Phi(\alpha-\gamma+1, 1-\gamma; z)\Phi(\alpha,\gamma;z) - \frac{\alpha}{\gamma(1-\gamma)}z\Phi(\alpha-\gamma+1, 2-\gamma; z)\Phi(\alpha+1, \gamma+1, z) = e^z$$

that follows from results in Ref. [23] and other identities found there. The final result is

$$G(z,t) = F_s(t)\Phi[c_f, c_f+c_b; -k_b(1-z)] + F_{ns}(t)\Phi[1-c_b, 2-c_f-c_b; -k_b(1-z)],$$

where

$$F_s(t) = \Phi[-c_f, 1-c_f-c_b; k_b e^{-t}(1-z)]$$

and

$$F_{ns}(t) = -\frac{c_f k_b(1-z)}{(c_f+c_b)(1-c_f-c_b)}e^{-(c_f+c_b)t} \times \Phi[c_b, 1+c_f+c_b; k_b e^{-t}(1-z)].$$

This yields Eqs. (5)–(7). In the limit $t\to\infty$, $F_s(t)\to 1$ (since the argument of the confluent hypergeometric function van-

ishes) and $F_{ns}(t)\to 0$ exponentially. At $t=0$ the right-hand side of the above expression yields 1 since we have imposed Eq. (A11). For the general case since this initial condition is altered as indicated above, the general result is obtained by dividing the left-hand side by $Q_0[1-e^{-t}(1-z)]$.

We describe briefly the Poisson representation of the probability distribution function [Eq. (8) in the text] that is related to the generating function. From the definition of the generating function $G(z,t)$ for $P(m,t)$, we can obtain $P(m,t)$ by multiplying by z^{-m-1} and performing a contour integral around the unit circle

$$P(m,t) = \oint G(z,t) z^{-m-1} \frac{dz}{2\pi i}. \tag{A13}$$

Now using the Laplace transform of the generating function in the form

$$G(z,t) \equiv \int d\lambda \rho(\lambda,t) e^{-\lambda(1-z)}$$

and performing the z integral using Cauchy's formula we find the representation given in Eq. (8). For an intuitive interpretation it is crucial that the density be non-negative as it is in our case.

The steady-state functional form of $\rho(\lambda)$ given in Eq. (9) in the text can be obtained as follows. One of the integral representations of the confluent hypergeometric function is given by [30]

$$\int_0^u dx e^{\beta x} x^{\nu-1} (u-x)^{\mu-1} = \frac{\Gamma(\mu)\Gamma(\nu)}{\Gamma(\mu+\nu)} u^{\mu+\nu-1} \Phi(\nu, \mu+\nu; \beta u). \tag{A14}$$

Identifying $u \equiv k_b$, $\beta = -(1-z)$, $\nu = c_f$, and $\mu = c_b$ we can read off the steady state $\rho(\lambda)$ given in Eq. (9) of the text.

[1] S. Fiering, J. P. Northrop, G. P. Nolan, P. S. Mattila, G. R. Crabtree, and L. A. Herzenberg, Genes Dev. **4**, 1823 (1990).
[2] D. A. Hume, Blood **96**, 2323 (2000).
[3] M. S. Ko, BioEssays **14**, 341 (1992).
[4] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **99**, 12795 (2002).
[5] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).
[6] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, Nature (London) **422**, 633 (2003).
[7] M. Thattai and A. van Oudenaarden, Proc. Natl. Acad. Sci. U.S.A. **98**, 8614 (2001).
[8] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden, Nat. Genet. **31**, 69 (2002).
[9] J. Paulsson, Nature (London) **427**, 415 (2004).
[10] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, PLoS Biol. **4**, e309 (2006).
[11] J. Hu, S. C. Sealfon, F. Hayot, C. Jayaprakash, M. Kumar, A. C. Pendleton, A. Ganee, A. Fernandez-Sesma, T. M. Moran, and J. G. Wetmur, Nucleic Acids Res. **35**, 5232 (2007).
[12] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).
[13] J. Chubb, T. Trcek, S. Shenoy, and R. Singer, Curr. Biol. **16**, 1018 (2006).
[14] R. Karmakar and I. Bose, Phys. Biol. **1**, 197 (2004).
[15] T. B. Kepler and T. C. Elston, Biophys. J. **81**, 3116 (2001).
[16] L. Mariani, M. Lohning, A. Radbruch, and T. Hofer, Prog. Biophys. Mol. Biol. **86**, 45 (2004).
[17] 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).
[18] Y. Okabe, Y. Yagi, and M. Sasai, J. Chem. Phys. **127**, 105107 (2007).
[19] A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **102**, 18926 (2005).
[20] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, Y. P. E. O'Shea, and N. Barkai, Nat. Genet. **38**, 636 (2006).
[21] D. T. Gillespie, J. Chem. Phys. **115**, 1716 (2001).
[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1965).
[23] L. J. Slater, *Confluent Hypergeometric Functions* (Cambridge University Press, Cambridge, England, 1960).
[24] D. R. Cox and V. Isham, *Point Processes* (Chapman and Hall, London, 1980).
[25] C. W. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, 2nd ed. (Springer-Verlag, Berlin, 1994).
[26] J. E. Ferrell, Curr. Opin. Cell Biol. **14**, 140 (2002).
[27] N. Friedman, L. Cai, and X. S. Xie, Phys. Rev. Lett. **97**, 168302 (2006).
[28] J. Paulsson and M. Ehrenberg, Phys. Rev. Lett. **84**, 5447 (2000).
[29] M. Bengtsson, A. Stahlberg, P. Rorsman, and M. Kubista, Genome Res. **15**, 1388 (2005).
[30] I. S. Gradshteyn, I. M. Ryzhik, and A. Jeffrey, *Table of Integrals, Series, and Products* (Academic, New York, 1980).