Correction

APPLIED MATHEMATICS, BIOPHYSICS

Correction for "Analytical distributions for stochastic gene expression," by Vahid Shahrezaei and Peter S. Swain, which appeared in issue 45, November 11, 2008, of *Proc Natl Acad Sci USA* (105:17256–17261; first published November 6, 2008; 10.1073/pnas.0803850105).

The authors note that due to a printer's error, Eq. **13** appeared incorrectly in part. The corrected equation appears below.

$$F(z) = [f(z)]^a = (1 + b - bz)^{-a} \qquad \textbf{[13]}$$

www.pnas.org/cgi/doi/10.1073/pnas.0811545106

Analytical distributions for stochastic gene expression

Vahid Shahrezaei[1] and Peter S. Swain[2,3]

Centre for Non-linear Dynamics, Department of Physiology, McGill University, 3655 Promenade Sir William Osler, Montreal, QC, Canada, H3G 1Y6

Edited by Charles S. Peskin, New York University, New York, NY, and approved September 5, 2008 (received for review April 22, 2008)

Gene expression is significantly stochastic making modeling of genetic networks challenging. We present an approximation that allows the calculation of not only the mean and variance, but also the distribution of protein numbers. We assume that proteins decay substantially more slowly than their mRNA and confirm that many genes satisfy this relation by using high-throughput data from budding yeast. For a two-stage model of gene expression, with transcription and translation as first-order reactions, we calculate the protein distribution for all times greater than several mRNA lifetimes and thus qualitatively predict the distribution of times for protein levels to first cross an arbitrary threshold. If in addition the fluctuates between inactive and active states, we can find the steady-state protein distribution, which can be bimodal if fluctuations of the promoter are slow. We show that our assumptions imply that protein synthesis occurs in geometrically distributed bursts and allows mRNA to be eliminated from a master equation description. In general, we find that protein distributions are asymmetric and may be poorly characterized by their mean and variance. Through maximum likelihood methods, our expressions should therefore allow more quantitative comparisons with experimental data. More generally, we introduce a technique to derive a simpler, effective dynamics for a stochastic system by eliminating a fast variable.

intrinsic noise | bursts | master equation | adiabatic approximation

Gene expression in both prokaryotes and eukaryotes is inherently stochastic (1–4). This stochasticity is both controlled and exploited by cells and, as such, must be included in models of genetic networks (5, 6). Here we will focus on describing intrinsic fluctuations, those generated by the random timing of individual chemical reactions, but extrinsic fluctuations are equally important and arise from the interactions of the system of interest with other stochastic systems in the cell or its environment (7, 8). Typically, experimental data are compared with predictions of mean behaviors and sometimes with the predicted standard deviation around this mean, because protein distributions are often difficult to derive analytically, even for models with only intrinsic fluctuations.

We will propose a general, although approximate, method for solving the master equation for models of gene expression. Our approach exploits the difference in lifetimes of mRNA and protein and is valid when the protein lifetime is greater than the mRNA lifetime. Typically, proteins exist for at least several mRNA lifetimes, and protein fluctuations are determined by only time-averaged properties of mRNA fluctuations. Following others (7, 9–11), we will use this time-averaging to simplify the mathematical description of stochastic gene expression.

For many organisms, single-cell experiments have shown that gene expression can be described by a three-stage model (3, 4, 12–14). The promoter of the gene of interest can transition between two states (10, 15–17), one active and one inactive. Such transitions could be from changes in chromatin structure, from binding and unbinding of proteins involved in transcription (3, 4, 12), or from pausing by RNA polymerase (18). Transcription can only occur if the promoter region is active. Both transcription and translation, as well as the degradation of mRNAs and proteins, are usually modeled as first-order chemical reactions (5).

By taking the limit of a large ratio of protein to mRNA lifetimes, we will study the three-stage model and a simpler two-stage version where the promoter is always active. For this two-stage model, we will derive the protein distribution as a function of time. We will derive the steady-state protein distribution for the full, three-stage model. We also include expressions for the corresponding mRNA distributions (14, 16) in the supporting information (SI) *Appendix*.

A Two-Stage Model of Gene Expression

We will first consider the model of gene expression in Fig. 1*A* (9). This model assumes that the promoter is always active and so has two stochastic variables: the number of mRNAs and the number of proteins. The probability of having m mRNAs and n proteins at time t satisfies a master equation:

$$\begin{aligned}\frac{\partial P_{m,n}}{\partial t} &= v_0(P_{m-1,n} - P_{m,n}) + v_1 m(P_{m,n-1} - P_{m,n}) \\ &\quad + d_0[(m+1)P_{m+1,n} - mP_{m,n}] \\ &\quad + d_1[(n+1)P_{m,n+1} - nP_{m,n}] \end{aligned} \qquad [1]$$

with v_0 being the probability per unit time of transcription, v_1 being the probability per unit time of translation, d_0 being the probability per unit time of degradation of an mRNA, and d_1 being the probability per unit time of degradation of a protein. By defining the generating function, $F(z',z)$, by $F(z',z) = \sum_{m,n} z'^m z^n P_{m,n}$, we can convert Eq. **1** into a first-order partial differential equation:

$$\frac{\partial F}{\partial v} - \gamma\left[b(1+u) - \frac{u}{v}\right]\frac{\partial F}{\partial u} + \frac{1}{v}\frac{\partial F}{\partial \tau} = a\frac{u}{v}F, \qquad [2]$$

where we have rescaled (19), with $a = v_0/d_1, b = v_1/d_0, \gamma = d_0/d_1$, and $\tau = d_1 t$, and where $u = z' - 1$ and $v = z - 1$.

If the protein lifetime is much greater than the mRNA lifetime and $\gamma \gg 1$, Eq. **2** can be solved by using the method of characteristics. Let r measure the distance along a characteristic which starts at $\tau = 0$ with $u = u_0$ and $v = v_0$ for some constant u_0 and v_0, then Eq. **2** is equivalent to (20)

$$\begin{aligned} &\frac{dv}{dr} = 1; && \frac{d\tau}{dr} = \frac{1}{v} \\ &\gamma^{-1}\frac{du}{dr} = \frac{u}{v} - b(1+u); && \frac{dF}{dr} = \frac{au}{v}F. \end{aligned} \qquad [3]$$

Consequently direct integration implies $r = v = v_0 e^{\tau}$. For $\gamma \gg 1$, $u(v)$ obeys (*SI Appendix*)

$$u(v) \simeq \left(u_0 - \frac{bv_0}{1-bv_0}\right) e^{-\gamma b(v-v_0)} \left(\frac{v}{v_0}\right)^{\gamma} + \frac{bv}{1-bv} \qquad [4]$$

or

$$u(v) \simeq \frac{bv}{1-bv} \qquad [5]$$

Author contributions: V.S. and P.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Department of Mathematics, Imperial College, London SW7 2BZ, UK.

[2]Present address: Center for Systems Biology, University of Edinburgh, Edinburgh EH9 3JU, UK.

[3]To whom correspondence should be addressed. E-mail: swain@cnd.mcgill.ca.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803850105/DCSupplemental.

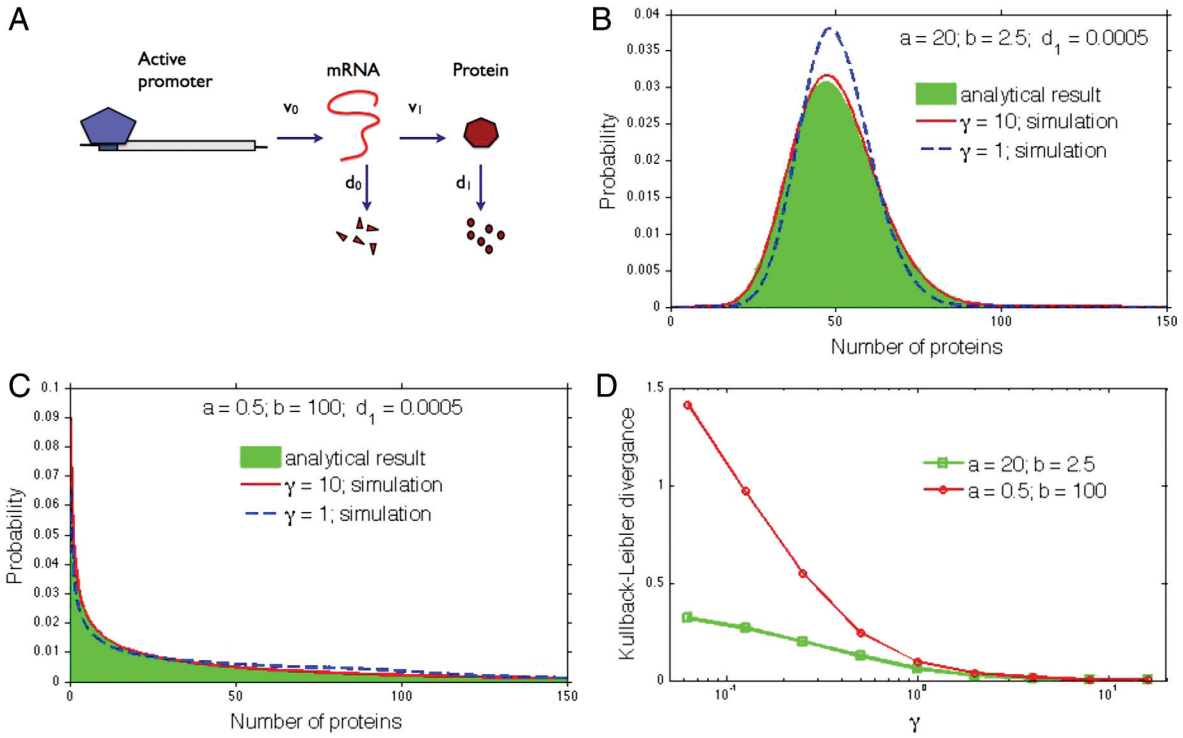

Fig. 1. Predictions and simulations for a two-stage model of gene expression. (*A*) Both transcription and translation are modeled as first-order processes: transcription occurs with a probability v_0 per unit time and translation with a probability of v_1 per unit time. Degradation of mRNA and protein are also both first-order processes: mRNA degrades with a probability d_0 per unit time and protein degrades with a probability d_1 per unit time. (*B* and *C*) A comparison of Eq. **9**, shown as the distribution in green, and stochastic simulations for large and small γ. Protein distributions can be either peaked or have a maximum at $n = 0$ (19). The mean number of mRNAs, a/γ, is either 2 or 20 in *B* and either 0.05 or 0.5 in *C*. *D* The accuracy of Eq. **9** improves with larger γ. The Kullback–Leibler divergence between the analytical and simulated protein distributions is plotted as a function of γ. For $\gamma \gg 1$, the distributions become almost indistinguishable.

as $v = v_0 e^{\tau} > v_0$ for $\tau > 0$. When $\gamma \gg 1$, u rapidly tends to a fixed function of v: for most of a protein's lifetime, the dynamics of mRNA is at steady state. The generating function then obeys

$$\frac{dF}{dv} \simeq \frac{ab}{1-bv}F. \qquad [6]$$

Intuitively, Eq. **6** arises from Eq. **2** because large γ causes the term in square brackets in Eq. **2** to tend to zero to keep $F(u,v)$ finite and well defined. Eq. **6** describes only the distribution for protein numbers: $F(u,v)$ is just a function of v. Terms of higher order in γ^{-1} will depend on u. Large γ implies that most of the mass of the joint probability distribution of mRNA and protein is peaked at $m = 0$: $P_{m,n} \simeq P_{0,n}$.

We can find the probability distribution for protein numbers as a function of time by integrating Eq. **6**. Integration gives

$$F(z,\tau) = \left[\frac{1-b(z-1)e^{-\tau}}{1+b-bz}\right]^a \qquad [7]$$

assuming that no proteins exist at $\tau = 0$. From the definition of a generating function, expanding $F(z)$ in z gives (*SI Appendix*)

$$P_n(\tau) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^n\left(\frac{1+be^{-\tau}}{1+b}\right)^a \times {}_2F_1\left(-n,-a,1-a-n;\frac{1+b}{e^{\tau}+b}\right) \qquad [8]$$

where $P_n(\tau) = P_{0,n}(\tau)$. Here, ${}_2F_1(a,b,c;z)$ is a hypergeometric function and Γ denotes the gamma function (21). Eq. **8** is valid when $\gamma \gg 1$, $\tau \gg \gamma^{-1}$ to allow the mRNA distribution to reach steady state, and a and b are finite. The mean, $\langle n\rangle = ab(1-e^{-\tau})$, and the variance, $\langle n^2\rangle - \langle n\rangle^2 = \langle n\rangle(1+b+be^{-\tau})$, of Eq. **8** agree with earlier results (9). At steady state $\tau \gg 1$ and

$$P_n = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^n\left(1-\frac{b}{1+b}\right)^a \qquad [9]$$

which is a negative binomial distribution. We verified Eq. **8** and Eq. **9** with stochastic simulations by using the Gibson–Bruck version (22) of the Gillespie algorithm (23) and the *Facile* network complier and stochastic simulator (24). If $\gamma \gg 1$, Eq. **9** accurately predicts the distribution described by Eq. **1** (Fig. 1*B* and *C*), but it fails as expected for smaller γ. This effect can be quantified by calculating the Kullback–Leibler divergence between the predicted and simulated distributions for different γ (Fig. 1*D*). Eq. **8** is illustrated in Fig. 2*A* and *B*. As well as $\gamma \gg 1$, times with $\tau > \gamma^{-1}$ are necessary for negligible Kullback–Leibler divergences (Fig. 2*C*).

Eq. **8** allows complete characterization of the Markov process underlying the two-stage model. The "propagator" probability, $P_{n|k}(\tau)$, which is the probability of having n proteins at time τ given k proteins initially, satisfies (*SI Appendix*)

$$P_{n|k}(\tau) = \sum_{r=0}^{k}\binom{k}{r}P_{n-r}(\tau)(1-e^{-\tau})^{k-r}e^{-r\tau} \qquad [10]$$

where $P_n(\tau) = 0$ if $n < 0$. With Eqs. **8** and **10**, two-stage gene expression is in principle completely characterized for $\gamma \gg 1$ and $\tau \gg \gamma^{-1}$. For example, we can calculate how the noise in protein numbers, η (their standard deviation divided by their mean), changes with time. If protein numbers initially have a distribution $P_k^{(0)}$, then at a time τ their distribution will be $\sum_k P_{n|k}(\tau)P_k^{(0)}$. The noise of this distribution can either increase, decrease, or behave nonmonotonically as time increases (Fig. 2*D*). We can

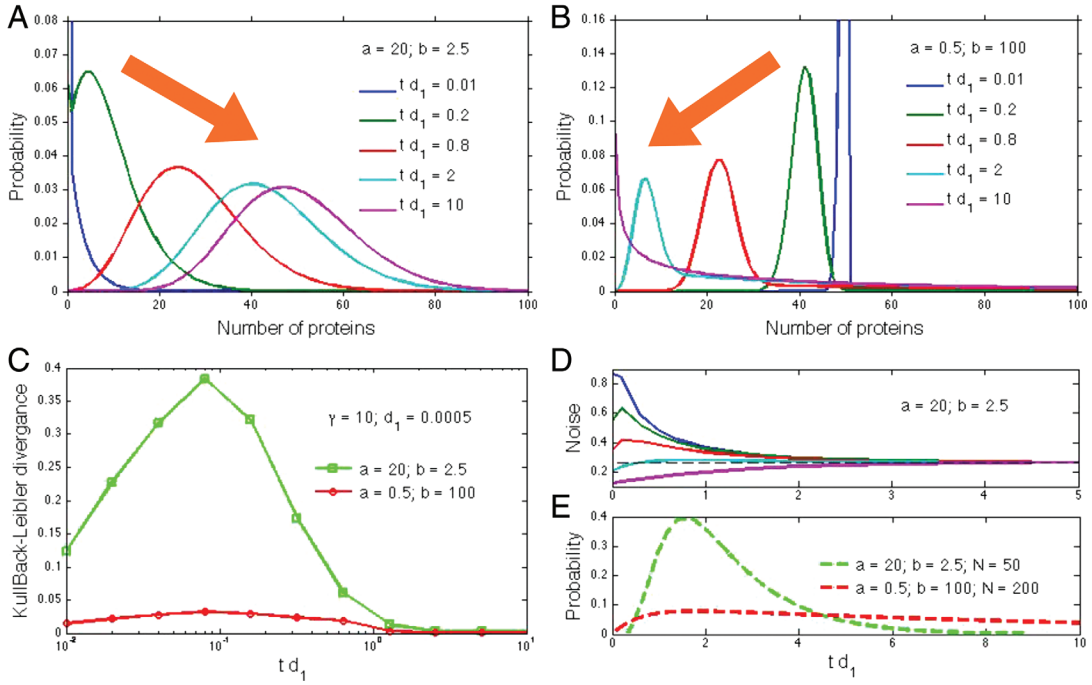

Fig. 2. Predictions for the time-dependent solution of the two-stage model of gene expression. (*A* and *B*) The distribution of protein numbers at different times with time increasing in the direction of the arrow. Parameters in *A* correspond to Fig. 1*B*. There are zero proteins initially. Parameters in *B* correspond to Fig. 1*C*. There are 50 proteins initially. *C* The Kullback–Leibler divergence for the distributions of *A* and *B*. The divergence decreases as $\tau = td_1$ grows above γ^{-1}. It is small for small times because both the simulations and the calculations start from the same initial distribution. (*D*) Noise in protein numbers as a function of time. Initially, proteins have a negative binomial distribution chosen with a particular magnitude of noise. The noise at steady state is shown by a dashed line. (*E*) The calculated distributions for the first-time protein levels reach a given threshold, *N*, if initially there are zero proteins. These distributions are qualitative with the probability typically underestimated for small td_1. They obey a renewal equation (26), which we solve numerically.

also calculate non-steady-state autocorrelation functions and first-passage time distributions for protein levels to first cross a threshold, N (with some standard numerics). In general, such distributions are only qualitative because contributions from times with $\tau < \gamma^{-1}$ are always relevant. Accuracy can be improved by having $\gamma \gg 10$ and a sufficiently high threshold (Fig. 2*E* and *SI Appendix*).

We can derive Eq. **9** more intuitively. An mRNA undergoes a competition between translation and degradation because ribosomes and degradosomes bind to it mutually exclusively (25). For each competition, the probability of a ribosome binding to the mRNA is $\frac{v_1}{v_1+d_0} = \frac{b}{1+b}$. If we assume that proteins have longer lifetimes than mRNAs ($\gamma \gg 1$), then each protein synthesized from a given mRNA will not on average be degraded before the mRNA is degraded. On protein timescales, all the proteins synthesized from an mRNA will appear to be synthesized simultaneously (*SI Appendix* Fig. S1). Consequently, the probability of r new proteins being produced by the synthesis and degradation of one mRNA is equal to the probability of an mRNA being translated r times. This probability is (25)

$$P_r = \left(\frac{b}{1+b}\right)^r \left(1 - \frac{b}{1+b}\right) \qquad [11]$$

which is a geometric, or "burst," distribution. Alternatively, we can consider the lifetime t' of each mRNA. This lifetime is stochastic and satisfies $P(t') = d_0 e^{-d_0 t'}$, the distribution expected for any first-order decay process (26). Protein synthesis is also first order, and the number of proteins, r, synthesized by an mRNA during its lifetime satisfies a Poisson process: $\frac{(v_1 t')^r}{r!} e^{-v_1 t'}$ (26). Consequently, the probable number of proteins synthesized from a particular mRNA is given by

$$P(r) = \int_0^\infty dt' d_0 e^{-d_0 t'} \frac{(v_1 t')^r}{r!} e^{-v_1 t'} \qquad [12]$$

which integrates to Eq. **11**. Eq. **11** is equivalent to an exponential distribution with a parameter λ, where $\lambda = -\log\left(1 - \frac{b}{1+b}\right)$ (27). If $b < 1$, then $\lambda \simeq b$. Exponential bursts of protein synthesis have been characterized experimentally (28, 29). We note that Eq. **11** has a generating function $f(z) = (1 + b - bz)^{-1}$.

Given that the synthesis and degradation of one mRNA generates a burst of r proteins, then the number of proteins at steady state is given by the typical number of mRNAs synthesized during a protein lifetime, $\frac{v_0}{d_1} = a$, and the r_i for each mRNA. The number of proteins n will be sum of these r_i. If we assume that there are sufficient ribosomes and charged tRNAs, then translation from each mRNA is independent. The generating function of a sum of independent variables is the product of their individual generating functions (26). Consequently, the generating function for P_n, $F(z)$, satisfies

$$[F(z)]^a = \prod_{i=1}^{a} [f(z)]^a = (1 + b - bz)^{-a} \qquad [13]$$

which is Eq. **7** when $\tau \gg 1$, and so derives Eq. **9**.

By assuming explicitly that protein synthesis occurs in bursts, we can derive an effective master equation for gene expression that considers only proteins, but implicitly includes mRNA fluctuations (19, 30). We will show that this master equation has Eq. **8** as its solution and so is equivalent to the large γ approximation to Eq. **1**, the master equation for both mRNA and protein. If we assume that each mRNA synthesized leaves behind a burst of r proteins then

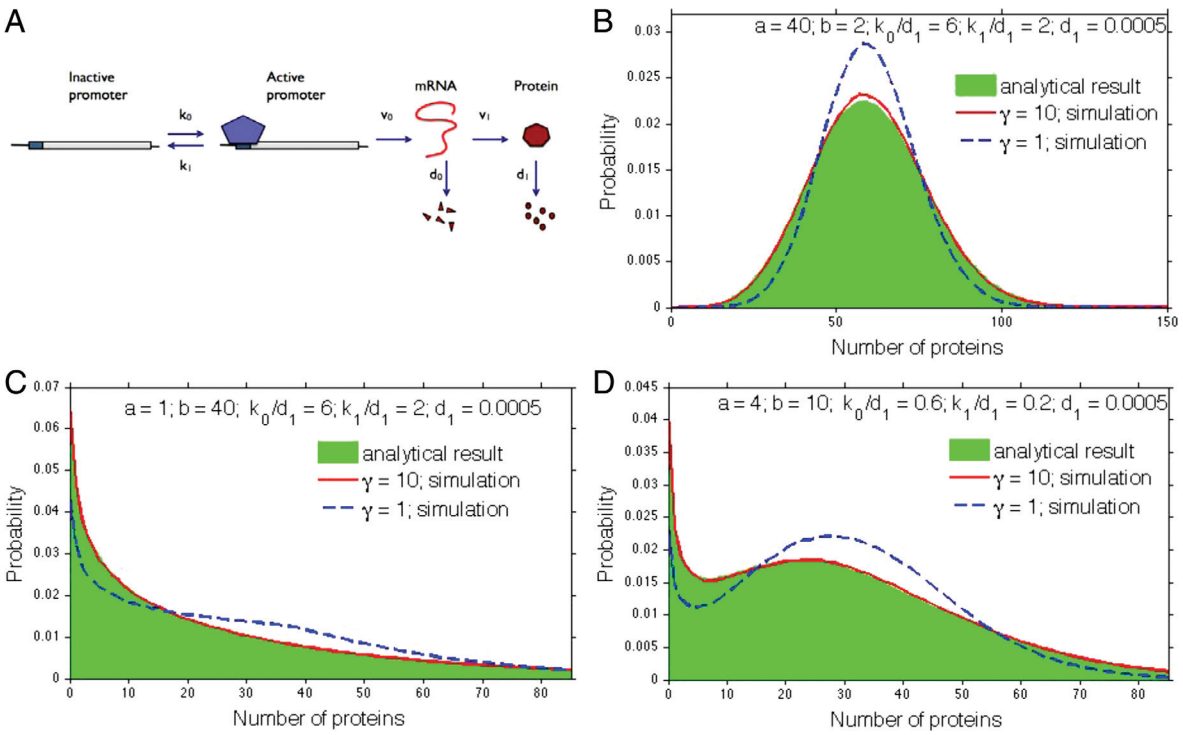

Fig. 3. Predictions and simulations for a three-stage model of gene expression. (*A*) The region of the DNA containing the promoter region transitions between inactive and active forms with probabilities per unit time of k_0 and k_1. As an example, we show the TATA-box binding protein driving the transition. (*B–D*) A comparison of Eq. **18**, shown as the distribution in green, and stochastic simulations for large and small γ. The mean number of mRNAs, $\frac{ak_0}{\gamma(k_0+k_1)}$, is either 3 or 30 in *B*, 0.075 or 0.75 in *C*, and either 0.3 or 3 in *D*.

$$\frac{\partial P_n}{\partial \tau} = a\left[\left(1-\frac{b}{1+b}\right)\sum_{r=0}^{n}\left(\frac{b}{1+b}\right)^r P_{n-r} - P_n\right] + (n+1)P_{n+1} - nP_n \qquad [14]$$

where the size of each burst has been determined by Eq. **11** (30). Eq. **14** has Eq. **7** as its generating function (*SI Appendix*). By introducing bursts of protein synthesis, mRNA fluctuations can be absorbed into a one-variable master equation provided $\gamma \gg 1$. Friedman *et al.* used a continuous version of this approach with an exponential burst distribution inspired by their experimental results (28, 29). They derived a gamma distribution for steady-state protein numbers (19). Eq. **9** tends to this distribution

$$P_n \to \frac{n^{a-1}e^{-n/b}}{b^a\Gamma(a)} \qquad [15]$$

for large n (*SI Appendix*). Friedman *et al.* (19) also demonstrated that the burst approximation remains valid when negative or positive feedback is included.

In summary, we have shown that exploiting the difference between protein and mRNA lifetimes through a large value of γ, but finite a and b, allows powerful mathematical simplifications. Large γ implies that mRNA is at steady state for most of the lifetime of a protein and that the probability mass of the joint distribution of protein and mRNA is peaked at zero mRNAs, although the mean number of mRNAs need not be zero (Fig. 1). The number of proteins translated from an mRNA obeys a geometric distribution in both the two-stage and three-stage models (25), but large γ implies that the proteins translated from an mRNA all appear, on protein timescales, simultaneously so that the synthesis and degradation of an mRNA leaves behind a geometric burst of proteins. If $\gamma < 1$, then proteins synthesized from a particular mRNA will be degraded as further proteins are synthesized, and the distribution describing the number of proteins remaining once the mRNA is degraded will no longer be geometric. Explicitly including geometric bursts accurately describes the effects of mRNA fluctuations on the distribution of protein numbers when $\gamma \gg 1$. It allows the model of Fig. 1*A* to be described by a one-variable master equation: Eq. **14**.

A Three-Stage Model of Gene Expression

We next consider the full three-stage model of gene expression (Fig. 3*A*). We find the protein distribution for this system by taking the large γ limit of the master equation. Let $P^{(0)}_{m,n}$ be the probability of having m mRNAs and n proteins when the DNA is inactive and $P^{(1)}_{m,n}$ be the probability of having m mRNAs and n proteins when the DNA is active. We then have two coupled equations:

$$\frac{\partial P^{(0)}_{n,m}}{\partial \tau} = \kappa_1 P^{(1)}_{m,n} - \kappa_0 P^{(0)}_{m,n} + (n+1)P^{(0)}_{m,n+1} - nP^{(0)}_{m,n} + \gamma\left[(m+1)P^{(0)}_{m+1,n} - mP^{(0)}_{m,n} + bm\left(P^{(0)}_{m,n-1} - P^{(0)}_{m,n}\right)\right] \qquad [16]$$

$$\frac{\partial P^{(1)}_{n,m}}{\partial \tau} = -\kappa_1 P^{(1)}_{m,n} + \kappa_0 P^{(0)}_{m,n} + (n+1)P^{(1)}_{m,n+1} - nP^{(1)}_{m,n} + a\left(P^{(1)}_{m-1,n} - P^{(1)}_{m,n}\right) + \gamma\left[(m+1)P^{(1)}_{m+1,n} - mP^{(1)}_{m,n} + bm\left(P^{(1)}_{m,n-1} - P^{(1)}_{m,n}\right)\right] \qquad [17]$$

where $\kappa_0 = k_0/d_1$ and $\kappa_1 = k_1/d_1$.

We solve Eqs. **16** and **17** at steady state by taking the large γ limit of the equivalent equations for their generating functions (a generating function is defined for each state of the promoter). Our

approach is a natural extension of the method used to solve the two-stage model (*SI Appendix*). We find that

$$P_n = \frac{\Gamma(\alpha+n)\Gamma(\beta+n)\Gamma(\kappa_0+\kappa_1)}{\Gamma(n+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\kappa_0+\kappa_1+n)} \times \left(\frac{b}{1+b}\right)^n \left(1-\frac{b}{1+b}\right)^\alpha \times {}_2F_1\left(\alpha+n, \kappa_0+\kappa_1-\beta, \kappa_0+\kappa_1+n; \frac{b}{1+b}\right) \quad [18]$$

where

$$\alpha = \frac{1}{2}(a+\kappa_0+\kappa_1+\phi) \quad [19]$$

$$\beta = \frac{1}{2}(a+\kappa_0+\kappa_1-\phi) \quad [20]$$

and $\phi^2 = (a+\kappa_0+\kappa_1)^2 - 4a\kappa_0$. Eq. **18** is valid when $\gamma \gg 1$ and a and b are finite. The mean of this distribution is $\langle n\rangle = \frac{abk_0}{k_0+k_1}$ and the protein noise, η, satisfies

$$\eta^2 = \frac{1}{\langle n\rangle} + \gamma^{-1}\frac{1}{\langle m\rangle} + \frac{d_1}{d_1+k_0+k_1}\eta_D^2 \quad [21]$$

where $\langle m\rangle$ is the mean number of mRNAs, and is inversely proportional to γ, and η_D is the noise in the active state of DNA: $\eta_D^2 = k_1/k_0$ (4). As well as a Poisson-like term expected for any birth-and-death process, protein noise has time-averaged contributions from fluctuations in the number of mRNAs and fluctuations in the state of DNA. We verify Eq. **18** by simulation in Fig. 3.

The protein distribution for the three-stage model can have similar behavior to the two-stage model of Fig. 1*A*, but it can also generate a bimodal distribution with a peak both at zero and nonzero numbers of molecules (Fig. 3*D*). This bimodality is not a reflection of an underlying bistability, but arises from slow transitions driving the DNA between active and inactive states (5, 17, 31, 32).

As expected, Eq. **18** recovers the negative binomial distribution under certain conditions. It tends to Eq. **9** when $\kappa_1 \to 0$: the DNA is then always active at steady state. When $\kappa_1 = 0$, Eqs. **19** and **20** imply that $\alpha = a$ and $\beta = \kappa_0$, and recall that ${}_2F_1(a,0,c;z) = 1$ for all a, c, and z. Similarly, when κ_0 and κ_1 are both large, but κ_0/κ_1 is fixed, then $\alpha \to \kappa_0+\kappa_1$ and $\beta \to \frac{\kappa_0 a}{\kappa_0+\kappa_1}$. Consequently,

$$P_n \to \frac{\Gamma(\beta+n)}{\Gamma(n+1)\Gamma(\beta)}\left(\frac{b}{1+b}\right)^n\left(1-\frac{b}{1+b}\right)^\beta \quad [22]$$

because ${}_2F_1(a,b,a;z) = (1-z)^{-b}$. With fast switching of the DNA between active and inactive states, Eq. **18** becomes Eq. **9**, but with a replaced by $\frac{\kappa_0 a}{\kappa_0+\kappa_1}$.

Discussion

We have shown we can calculate distributions for protein numbers by assuming that protein lifetimes are longer than mRNA lifetimes with a, the number of mRNAs transcribed during a protein remaining, and b, the number of proteins translated during a mRNA lifetime, remaining finite. Fig. 4*A* shows the ratio γ measured for almost 2,000 genes in budding yeast. Approximately 80% of the genes have $\gamma > 1$ and the median value is ≈3 (we include the dataset in *SI Appendix*). We therefore expect our predicted distributions to be widely applicable in budding yeast. In bacteria, too, γ is expected to be above 1 because mRNA lifetimes are usually minutes (they are typically tens of minutes in yeast) and protein lifetimes are often determined by the length of the cell cycle (typically 30 or more minutes) (33).

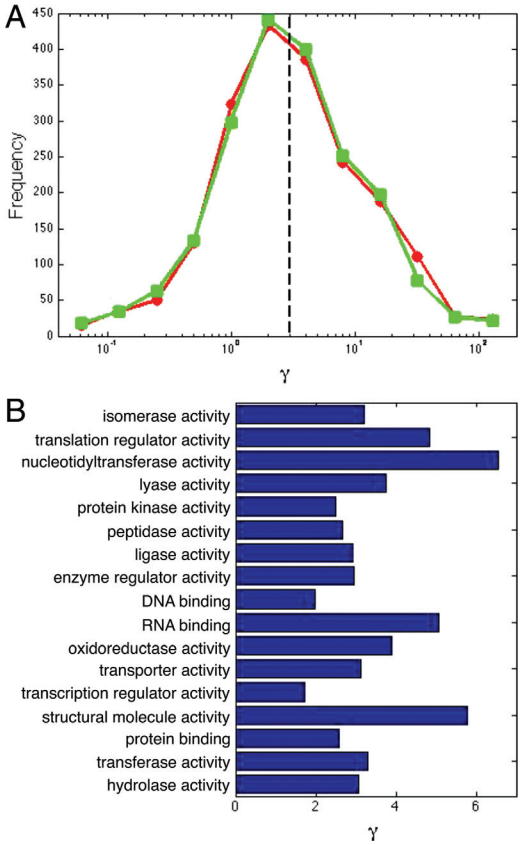

Fig. 4. The ratio of the protein to mRNA lifetime, γ, for 1,962 genes in budding yeast. (*A*) Most proteins have $\gamma > 1$. Protein lifetimes are from Belle *et al.* (41) and mRNA lifetimes are from Grigull *et al.* (circles) (42) or from Wang *et al.* (squares) (43). The median of γ is $\simeq 3$ (shown by a dashed line), whereas its mean is > 10 (although this value is probably erroneously high because of outliers). Overall, we found little correlation between mRNA and protein lifetimes. (*B*) The median value of γ for genes in different gene ontology classes. We plot the mean of the medians for the two datasets. Errors in the medians are ≈25% (using 1,000 bootstrap samples for each gene ontology class). Gene annotations are from the *Saccharomyces cerevisiae* genome database (www.yeastgenome.org).

Values of $\gamma > 1$ reduce protein fluctuations by allowing more averaging of the underlying mRNA fluctuations (Eq. **21**). We indeed observe a small, but statistically significant, negative correlation between total noise and γ by using the data of Newman *et al.* (34) (a rank correlation of $\simeq -0.2$ with a P value of 10^{-6}). In Fig. 4*B*, we have calculated the median γ for yeast genes in different gene ontology classes. All classes have a median $\gamma > 1$. Proteins involved in transferring nucleotidyl groups, which include RNA and DNA polymerases, have high median $\gamma > 5$, presumably because high stochasticity in these proteins can undermine many cellular processes. Similarly, proteins that contribute to the structural integrity of protein complexes have a median $\gamma > 5$. Large fluctuations can vastly reduce the efficiency of complex assembly by preventing complete complexes forming because of a shortage of one or more components (11, 35). Perhaps surprisingly transcription factors have a low median $\gamma > 1$. Although low γ does increase stochasticity, it can allow quick response times if the protein degradation rate is high. A high protein degradation rate may also keep numbers of transcription factors low to reduce deleterious nonspecific chromosomal binding.

We show that protein synthesis occurs in bursts in both the two- and the three-stage model when $\gamma \gg 1$. Such bursts of gene expression have been measured in bacteria and eukaryotes (12–14, 28, 29). They allow mRNA to be replaced in the master equation by a geometric distribution for protein synthesis for

all times greater than several mRNA lifetimes if their source is translation and the protein lifetime is substantially longer than the mRNA lifetime. Such an approach has already been proposed (19), but without determining its validity. Similarly, if mRNA fluctuations are negligible, the master equation reduces to one variable (protein), and describing the protein distribution becomes substantially easier (31, 36).

An important problem in systems biology is to determine which properties of biochemical networks and the intracellular environment must be modeled to make accurate, quantitative predictions. Besides obscuring the process driving the observed phenotype, models more complex than needed are harder to correctly parameterize and to simulate to generate predictions. Our results show that complexity, here two states of the promoter, can be modeled by effective parameters that under certain conditions will give accurate predictions of the entire distribution of protein numbers: Eq. **22**. Alternatively, they show that not only the mean and variance (37), but also the protein distribution may not have enough information to determine the biochemical mechanism generating gene expression from measurements of protein levels: Eqs. **9** and **22**. Such effects are likely to be compounded by non-steady-state dynamics (Fig. 2) and extrinsic fluctuations. Collecting data on the corresponding mRNA distribution may disfavor the two-stage over the three-stage model because mRNA distributions in the three-stage model can have two peaks even though the protein distribution has only one (6). In general, though, time series measurements, preferably with and without perturbations, may provide the most discriminative power (37).

Experimental measurements are best compared with the predicted distribution rather than its mean, standard deviation, or mode. Both the protein and mRNA distributions are typically not symmetric and may not be unimodal. Consequently, the mean and the mode can be significantly different, and the standard deviation can be a poor measure of the width of the distribution at half-maximum (38). Such distributions are poorly characterized by the commonly used coefficient of variation because they are not locally Gaussian around their mean (Figs. 1–3). In addition, fitting moments to find model parameters can be challenging. Moments, more so than distributions, are functions of combinations of parameters and can also be badly estimated without large amounts of data, particularly for asymmetric distributions. We therefore believe a Bayesian or maximum likelihood approach is most suitable where the experimental protocol is replicated by the fitting procedure and explicitly accounts for the shape of the distribution and the number of measurements. For example, irrespective of how many measurements are available, the likelihood of the data for a particular set of parameters can always be determined from the assumed distribution of protein numbers. Our analytical expressions will greatly speed up such approaches by avoiding large numbers of simulations and by aiding in deconvolving extrinsic fluctuations that can substantially change the shape of protein distributions (8).

Our results should also allow more general fluctuation analyzes of gene expression data. Such analyzes convert fluorescence measurements into absolute units (numbers of molecules) by exploiting that the magnitude of fluctuations is determined by the number of molecules independently of how those numbers are measured (39). Converting into absolute units is essential if information from different experiments is to be combined into a larger, predictive framework, a goal of systems biology.

More generally, our approach is an example of a technique to simplify the dynamics of a stochastic system by exploiting differences in timescales. We remove a fast stochastic variable through replacing a constant parameter (the parameter a) by a time-dependent parameter (the burst distribution) whose variation captures the effects of fluctuations in the fast variable on the dynamics of the slow one (40).

ACKNOWLEDGMENTS. We thank an anonymous referee for showing us Eq. **12**. P.S.S. is a recipient of a Tier II Canada Research Chair. V.S. and P.S.S. are supported by National Sciences and Engineering Research Council (Canada).

1. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31:69–73.
2. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186.
3. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422:633–637.
4. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304:1811–1814.
5. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* 6:451–464.
6. Shahrezaei V, Swain PS (2008) The stochastic nature of biochemical networks. *Curr Opin Biotechnol* 19:369–374.
7. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 99:12795–12800.
8. Shahrezaei V, Ollivier JF, Swain PS (2008) Colored extrinsic fluctuations and stochastic gene expression. *Mol Syst Biol* 4:196.
9. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA* 98:8614–8619.
10. Kepler TB, Elston TC (2001) Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys J* 81:3116–3136.
11. Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol* 344:965–976.
12. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036.
13. Chubb JR, Trcek T, Shenoy SM, Singer RH (2006) Transcriptional pulsing of a developmental gene. *Curr Biol* 16:1018–1025.
14. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4:e309.
15. Ko MS (1991) A stochastic model for gene induction. *J Theor Biol* 153:181–194.
16. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. *Theor Popul Biol* 48:222–234.
17. Karmakar R, Bose I (2004) Graded and binary responses in stochastic gene expression. *Phys Biol* 1:197–204.
18. Voliotis M, Cohen N, Molina-Paras C, Liverpool TB (2008) Fluctuations, pauses, and backtracking in DNA transcription. *Biophys J* 94:334–348.
19. Friedman N, Cai L, Xie XS (2006) Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys Rev Lett* 97:16830.
20. Zwillinger D (1989) *Handbook of Differential Equations* (Academic Press, New York).
21. Abramowitz M, Stegun IA (1984) *Pocketbook of Mathematical Functions* (Harri Deutsch Publishing, Frankfurt am Main, Germany).
22. Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *J Phys Chem A* 104:1876–1889.
23. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361.
24. Siso-Nadal F, Ollivier JF, Swain PS (2007) Facile: A command-line network compiler for systems biology. *BMC Syst Biol* 1:36.
25. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA* 94:814–819.
26. van Kampen NG (1990) *Stochastic Processes in Physics and Chemistry* (Elsevier, New York).
27. Prochaska BJ (1973) A note on the relationship between the geometric and exponential distributions. *Am Statistician* 27:27.
28. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* 440:358–362.
29. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. *Science* 311:1600–1603.
30. Paulsson J, Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Phys Rev Lett* 84:5447–5450.
31. Hornos JE, *et al.* (2005) Self-regulating gene: An exact solution. *Phys Rev E* 72:051907.
32. Pirone JR, Elston TC (2004) Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression. *J Theor Biol* 226:111–121.
33. Bremer H, Dennis PP (1996) Modulation of chemical composition and other parameters of the cell by growth rate. ed Neidhardt FC *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ASM Press, Washington, DC), pp 1553–1569.
34. Newman JR, *et al.* (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* 441:840–846.
35. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol* 2:834–838.
36. Walczak AM, Sasai M, Wolynes PG (2005) Self-consistent proteomic field theory of stochastic gene switches. *Biophys J* 88:828–850.
37. Pedraza JM, Paulsson J (2008) Effects of molecular memory and bursting on fluctuations in gene expression. *Science* 319:339–343.
38. Samoilov MS, Arkin AP (2006) Deviant effects in molecular reaction pathways. *Nat Biotechnol* 24:1235–1240.
39. Rosenfeld N, Perkins TJ, Alon U, Elowitz MB, Swain PS (2006) A fluctuation method to quantify in vivo fluorescence data. *Biophys J* 91:759–766.
40. Shibata T (2003) Fluctuating reaction rates and their application to problems of gene expression. *Phys Rev E* 67:061906.
41. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK (2006) Quantification of protein half-lives in the budding yeast proteome. *Proc Natl Acad Sci USA* 103: 13004–13009.
42. Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals post-transcriptional control of ribosome biogenesis factors. *Mol Cell Biol* 24:5534–5547.
43. Wang Y, *et al.* (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* 99:5860–5865.

Analytical distributions for stochastic gene expression: Supporting information

Vahid Shahrezaei and Peter S. Swain

Derivation of the protein distribution for a two-stage model of gene expression

From the master equation

The generating function for the master equation of the two-stage model satisfies (Eq. 1 in the main text)

$$\frac{1}{v}\frac{\partial F}{\partial \tau}+\frac{\partial F}{\partial v}-\gamma\left[b(1+u)-\frac{u}{v}\right]\frac{\partial F}{\partial u}=a\frac{u}{v}F \tag{23}$$

where $F(z',z)$ is defined as $\sum_{m,n}(z')^m z^n P_{m,n}$, and we have let $u=z'-1$ and $v=z-1$. If r measure the distance along a characteristic which starts at $\tau=0$ with $u=u_0$ and $v=v_0$ for some constant u_0 and v_0, then Eq. 23 becomes

$$\begin{array}{lll}\frac{dv}{dr}=1 & ; & \frac{d\tau}{dr}=\frac{1}{v}\\ \frac{du}{dr}=-\gamma\left[b(1+u)-\frac{u}{v}\right] & ; & \frac{dF}{dr}=\frac{au}{v}F.\end{array} \tag{24}$$

Consequently, $v=r$ and

$$\frac{du}{dv}=-\gamma\left[b(1+u)-\frac{u}{v}\right] \tag{25}$$

which has solution

$$u(v)=\mathrm{e}^{-\gamma bv}v^{\gamma}\left[C-b\gamma\int^{v}dv'\frac{\mathrm{e}^{\gamma bv'}}{v'^{\gamma}}\right] \tag{26}$$

for a constant C as can be verified by differentiation. By Taylor expanding $\mathrm{e}^{\gamma bv}$ so that $\mathrm{e}^{\gamma bv}=\sum_n\frac{(\gamma bv)^n}{n!}$, we can evaluate the integral in Eq. 26,

$$u(v)=\mathrm{e}^{-\gamma bv}\left[Cv^{\gamma}-\sum_{n=0}^{\infty}\frac{(\gamma bv)^{n+1}}{n!(n-\gamma+1)}\right]. \tag{27}$$

We can also carry out the sum in Eq. 27 in the limit of $\gamma\gg 1$ following Bender and Orzag [1]. By comparing the ratio of the $n-1$'th and the n'th term, we see that the elements of the sum have a maximum when $n\simeq\gamma bv$. For $\gamma\gg 1$, the sum will be dominated by terms with n near γbv. We therefore let $n=\gamma bv+s$ for some s, then $n!$ can be shown to be approximately [1]

$$n!\simeq(\gamma bv)^n\mathrm{e}^{-\gamma bv}\mathrm{e}^{\frac{s^2}{2\gamma bv}}\sqrt{2\pi\gamma bv} \tag{28}$$

using Stirling's approximation. Consequently, by approximating the sum as an integral and extending the range of the integral to $-\infty$,

$$\begin{aligned}
\sum_{n=0}^{\infty} \frac{(\gamma bv)^{n+1}}{n!(n-\gamma+1)} &\simeq \int_{-\infty}^{\infty} ds\, \frac{\mathrm{e}^{-\frac{s^2}{2\gamma bv}}}{\sqrt{2\pi\gamma bv}} \cdot \frac{\gamma bv \mathrm{e}^{\gamma bv}}{\gamma(bv-1)+s+1} \\
&= \int_{-\infty}^{\infty} ds\, \frac{\mathrm{e}^{-\frac{s^2}{2\gamma bv}}}{\sqrt{2\pi\gamma bv}} \cdot . \frac{bv\mathrm{e}^{\gamma bv}}{bv-1}\left[1+\gamma^{-1}\left(\frac{s+1}{bv-1}\right)\right]^{-1} \\
&= \frac{bv\mathrm{e}^{\gamma bv}}{bv-1}\int_{-\infty}^{\infty} ds\, \frac{\mathrm{e}^{-\frac{s^2}{2\gamma bv}}}{\sqrt{2\pi\gamma bv}} + O\left(\gamma^{-1}\right) \\
&\simeq \frac{bv\mathrm{e}^{\gamma bv}}{bv-1} \qquad (29)
\end{aligned}$$

to the lowest order in γ. From Eq. 27, u satisfies

$$u(v) \simeq C\mathrm{e}^{-\gamma bv} v^{\gamma} + \frac{bv}{1-bv} \qquad (30)$$

when $\gamma \gg 1$. We evaluate C using $u = u_0$ when $v = v_0$ giving

$$\begin{aligned}
u &\simeq \left(u_0 - \frac{bv_0}{1-bv_0}\right)\mathrm{e}^{-\gamma b(v-v_0)}\left(\frac{v}{v_0}\right)^{\gamma} + \frac{bv}{1-bv} \\
&\simeq \frac{bv}{1-bv}. \qquad (31)
\end{aligned}$$

when $\gamma \gg 1$ because $v = v_0\mathrm{e}^{\tau} > v_0$ from Eq. 24.

Finding the generating function

Using Eq. 31, Eq. 24 becomes

$$\frac{dF}{dv} = \frac{ab}{1-bv}F \qquad (32)$$

or, on integrating,

$$\log\frac{F(v)}{F(v_0)} = -a\log\left(\frac{1-bv}{1-bv_0}\right) \qquad (33)$$

because $F(v_0) = F(\tau = 0)$. If initially we have k proteins then

$$F(v_0) = \sum P_n(\tau=0)z^n = \sum \delta_{n,k} z^n = z^k = (1+v_0)^k. \qquad (34)$$

For our approximation, Eq. 31, to be valid, enough time must have passed for mRNA levels to have reached steady-state. Strictly, this initial condition is only valid for non-zero τ of the order of $d_1/d_0 = \gamma^{-1}$. Finally, inserting Eq. 34 into Eq. 33 gives

$$F(z,\tau) = \left[\frac{1-b(z-1)\mathrm{e}^{-\tau}}{1-bz+b}\right]^{a}\left[1+(z-1)\mathrm{e}^{-\tau}\right]^{k} \qquad (35)$$

because $v_0 = (z-1)\mathrm{e}^{-\tau}$. When $k = 0$, Eq. 35 becomes Eq. 7.

Deriving the probability distribution for proteins

We can find $P_n(\tau)$, the probability of having n proteins at time τ given initially zero proteins, by differentiating Eq. 35 when $k = 0$. By definition, P_n satisfies $P_n = \frac{1}{n!}\frac{\partial^n}{\partial z^n}F(z,\tau)\Big|_{z=0}$. By writing

$$F(z,\tau) = \left(\frac{1+b\mathrm{e}^{-\tau}}{1+b}\right)^a \cdot \frac{\left[1-\frac{b}{1+b}z\right]^{-a}}{\left[1-\frac{b}{\mathrm{e}^{\tau}+b}z\right]^{-a}}, \tag{36}$$

we can make use of the identities

$$\frac{\partial^n}{\partial z^n}[1-qz]^{-a}\Big|_{z=0} = \frac{\Gamma(a+n)}{\Gamma(a)}q^n \tag{37}$$

and

$$\frac{\partial^n}{\partial z^n}\frac{x(z)}{y(z)} = n!\sum_{k=0}^{n}\frac{\partial^{n-k}}{\partial z^{n-k}}x(z)\cdot\sum_{j=0}^{k}\frac{(-1)^j(k+1)y(z)^{-j-1}}{(j+1)!(n-k)!(k-j)!}\frac{\partial^k}{\partial z^k}y(z)^j \tag{38}$$

which is given at Wolfram Research (`functions.wolfram.com/GeneralIdentities/9`).

Interpreting $x(z)$ as the numerator of the quotient in Eq. 36 and $y(z)$ as its denominator, we find

$$\begin{aligned} P_n(\tau) &= \left(\frac{1+b\mathrm{e}^{-\tau}}{1+b}\right)^a \sum_{k=0}^{n}\frac{\Gamma(a+n-k)}{\Gamma(a)}\left(\frac{b}{1+b}\right)^{n-k} \\ &\quad\times\sum_{j=0}^{k}\frac{(-1)^j(k+1)}{(j+1)!(n-k)!(k-j)!}\cdot\frac{\Gamma(aj+k)}{\Gamma(aj)}\cdot\left(\frac{b}{\mathrm{e}^{\tau}+b}\right)^k \end{aligned} \tag{39}$$

where we can use

$$\sum_{j=1}^{k}\frac{(-1)^j\Gamma(aj+k)}{\Gamma(aj)(j+1)!(k-j)!} = \frac{(-1)^k\Gamma(a+1)}{\Gamma(a-k+1)(k+1)!} \tag{40}$$

to simplify further. Eq. 40 can be verified by directly expanding the sum. Consequently,

$$P_n(\tau) = \left(\frac{b}{1+b}\right)^n\left(\frac{1+b\mathrm{e}^{-\tau}}{1+b}\right)^a\sum_{k=0}^{n}\frac{(-1)^k}{k!}\frac{\Gamma(a-k+n)}{\Gamma(n-k+1)\Gamma(a-k+1)}\left(\frac{1+b}{\mathrm{e}^{\tau}+b}\right)^k. \tag{41}$$

The hypergeometric function ${}_2F_1(a,b,c;z)$ obeys

$${}_2F_1(-n,b,c;z) = \sum_{k=0}^{n}(-1)^k\frac{\Gamma(n+1)}{\Gamma(n-k+1)}\frac{(b)_k}{(c)_k}\frac{z^k}{k!} \tag{42}$$

when a is a negative integer and where $(b)_k$ and $(c)_k$ are Pochhammer symbols [2]. From their definition, $(a)_k = \Gamma(a+k)/\Gamma(a)$, the Pochhammer symbols satisfy

$$\Gamma(a+1) = (-1)^k(-a)_k\Gamma(a-k+1). \tag{43}$$

Writing $\Gamma(a-k+n) = \Gamma(a+n-1-k+1)$ and using Eq. 42 and Eq. 43, we find that

$$P_n(\tau) = \frac{1}{n!}\left(\frac{b}{1+b}\right)^n\left(\frac{1+b\mathrm{e}^{-\tau}}{1+b}\right)^a\frac{\Gamma(a+n)}{\Gamma(a)}\,{}_2F_1\left(-n,-a,1-a-n;\frac{1+b}{\mathrm{e}^{\tau}+b}\right) \tag{44}$$

which is valid for $\gamma \gg 1$, $\tau > \gamma^{-1}$, and a and b finite.

Deriving the 'propagator' probability

By differentiating Eq. 35 for non-zero k, we can express the 'propagator' probability, $P_{n|k}(\tau)$, in terms of Eq. 44. From the definition of $P_n(\tau)$, Eq. 35 can be written as

$$F(z,\tau) = \left[\sum_{n=0}^{\infty} P_n(\tau) z^n\right] \left[1 - \mathrm{e}^{-\tau} + z\mathrm{e}^{-\tau}\right]^k \tag{45}$$

or

$$F(z,\tau) = \sum_{n=0}^{\infty} P_n(\tau) z^n \sum_{r=0}^{k} \binom{k}{r} \left(1 - \mathrm{e}^{-\tau}\right)^{k-r} \left(z\mathrm{e}^{-\tau}\right)^r \tag{46}$$

using the binomial theorem. From the coefficients of the powers of z, we find

$$P_{n|k}(\tau) = \sum_{r=0}^{k} \binom{k}{r} P_{n-r}(\tau) \left(1 - \mathrm{e}^{-\tau}\right)^{k-r} \mathrm{e}^{-r\tau} \tag{47}$$

because $F(z,\tau) = \sum_n P_{n|k}(\tau) z^n$ and remembering that $P_n(\tau) = 0$ if $n < 0$.

Finding the probability distribution for the first passage time

With $P_n(\tau)$ and $P_{n|k}(\tau)$, we can find the distribution for the first time the number of proteins reaches a threshold N. We define this distribution to be $f_N(\tau)$. It obeys a renewal equation [3]

$$P_N(\tau) = \int_0^{\tau} d\tau' f_N(\tau') P_{N|N}(\tau - \tau'). \tag{48}$$

The probability of having N proteins at time τ is equal to the sum of the probability of first reaching N proteins at τ' and then returning to N proteins at a time $\tau - \tau'$ later for all times τ' less than τ. We have assumed that the initial number of proteins is zero, but this assumption is not necessary.

Eq. 48 is a Volterra integral equation of the first kind and can be straightforwardly solved numerically [4]. If $N > 0$ then $f_N(0) = 0$ and $P_{N|N}(0) = 1$ by definition. Consequently, by discretizing and letting $\tau_i = i\epsilon$ for integer i and small ϵ, we can write the integral in Eq. 48 as a trapezium rule:

$$\int_0^{\tau_i} d\tau' f_N(\tau') P_{N|N}(\tau_i - \tau') \simeq \epsilon \left[\frac{1}{2} f_N(\tau_i) + \sum_{j=1}^{i-1} P_{N|N}(\tau_i - \tau_j) f_N(\tau_j)\right]. \tag{49}$$

Inserting Eq. 49 into Eq. 48 gives a series of equations for $f_N(\tau_i)$ which we solve iteratively:

$$f_N(\tau_1) = \frac{2P_N(\tau_1)}{\epsilon} \tag{50}$$

$$f_N(\tau_i) = 2\left[\frac{P_N(\tau_i)}{\epsilon} - \sum_{j=1}^{i-1} P_{N|N}(\tau_{i-j}) f_N(\tau_j)\right]. \tag{51}$$

We implement Eqs. 50 and 51 in Matlab (The Mathworks, Natick, Massachusetts). Our code is available at `www.cnd.mcgill.ca/~swain`.

We use

$$\langle n(\tau_1) n(\tau_2)\rangle = \sum_{n,n'} n n' P_{n|n'}(\tau_2 - \tau_1) P_{n'}(\tau_1) \tag{52}$$

to find the auto-correlation function. We evaluate the sum in Eq. 52 numerically, cutting off the sums when n is many times the mean steady-state value: $\langle n\rangle = ab$.

High γ implies bursts of protein synthesis

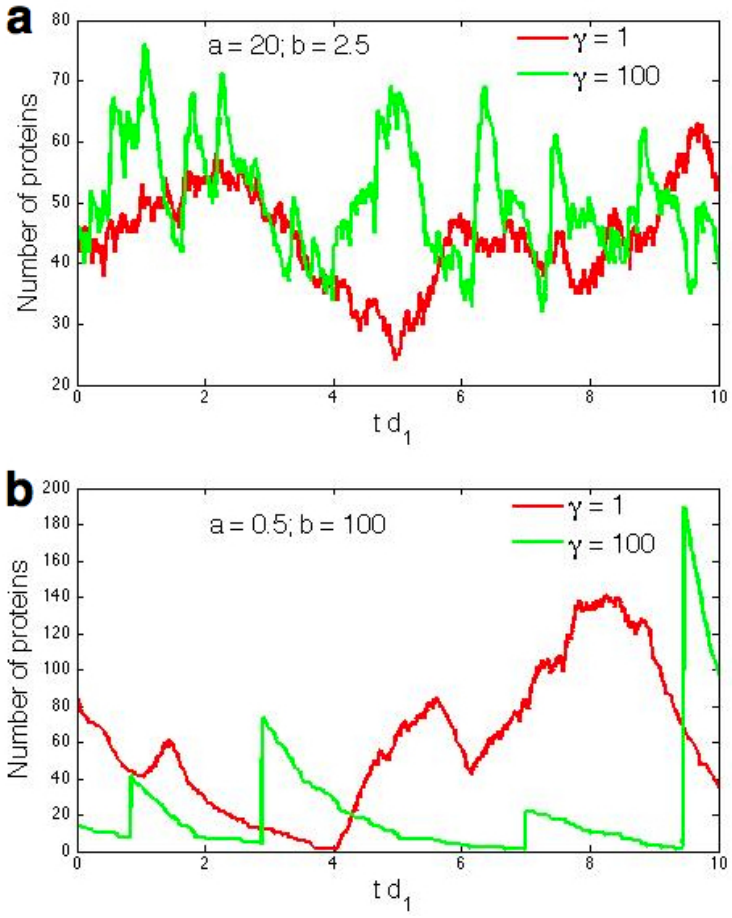

Figure 5: As γ increases, protein synthesis occurs in bursts. Time courses of protein numbers from simulations of the two-stage model of of Fig. 1. When γ is increased to 100 from 1, we see steep bursts of synthesis: short-lived mRNAs are only able to be occasionally translated before being degraded. The protein degradation rate is $d_1 = 0.0005\text{s}^{-1}$. **a** $a = 20$ and $b = 2.5$. **b** $a = 0.5$ and $b = 100$. Both examples have a mean protein number of 50.

Solving the master equation for bursts of protein synthesis

When $\gamma \gg 1$, the distribution for protein numbers can also be derived by only considering $P_n(\tau)$, the probability of having n proteins at time τ, if this probability obeys a master equation where proteins are synthesized in bursts. We let the size r of a burst obey a geometric distribution,

$$P(r) = \left(\frac{b}{1+b}\right)^r \left(1 - \frac{b}{1+b}\right). \tag{53}$$

The corresponding master equation is

$$\frac{\partial P_n}{\partial \tau} = a\left[\left(1 - \frac{b}{1+b}\right)\sum_{r=0}^{n}\left(\frac{b}{1+b}\right)^r P_{n-r} - P_n\right] + (n+1)P_{n+1} - nP_n \tag{54}$$

which can be converted into an equation for the generating function, $F(z) = \sum_n z^n P_n(\tau)$.

The generating function obeys

$$\frac{\partial F}{\partial \tau} = (1-z)\frac{\partial F}{\partial z} - aF + a\left(1 - \frac{b}{1+b}\right)\sum_{n=0}^{\infty}\sum_{r=0}^{n} z^n \left(\frac{b}{1+b}\right)^r P_{n-r} \tag{55}$$

where we need to evaluate the sums over n and r. Relabelling and resuming

$$\begin{aligned}
\sum_{n=0}^{\infty}\sum_{r=0}^{n} z^n \left(\frac{b}{1+b}\right)^r P_{n-r} &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} z^n \left(\frac{b}{1+b}\right)^{n-k} P_k \\
&= \sum_{k=0}^{\infty}\left(\frac{b}{1+b}\right)^{-k} P_k \sum_{n=k}^{\infty}\left(\frac{bz}{1+b}\right)^n \\
&= \sum_{k=0}^{\infty} \frac{P_k \left(\frac{bz}{1+b}\right)^k}{\left(1-\frac{bz}{1+b}\right)\left(\frac{b}{1+b}\right)^k} \\
&= \frac{F(z)}{1-\frac{bz}{1+b}}
\end{aligned} \tag{56}$$

where we use the definition of the generating function. Consequently, Eq. 55 becomes

$$\frac{\partial F}{\partial \tau} = (1-z)\frac{\partial F}{\partial z} + \left(\frac{1-\frac{b}{1+b}}{1-\frac{bz}{1+b}} - 1\right) aF \tag{57}$$

or

$$\frac{1}{v}\frac{\partial F}{\partial \tau} + \frac{\partial F}{\partial v} = \frac{ab}{1-bv}F \tag{58}$$

with $v = z - 1$. This partial differential equation is Eq. 23 when $\gamma \gg 1$ and Eq. 31 holds.

Derivation of the gamma distribution for protein numbers

We can derive the gamma distribution for protein numbers found by Friedman *et al.* [5] when n is large. If $P(n|a,b)$ is the negative binomial distribution and $\Gamma(n|a,b)$ is the gamma distribution, then

$$P(n|a,b) = \int_0^{\infty} d\lambda \, \frac{\mathrm{e}^{-\lambda}\lambda^n}{n!}\Gamma(\lambda|a,b) \tag{59}$$

which is a general relation between the negative binomial and gamma distributions. It can be verified by evaluating the integral using the definition of a gamma function [2]. If we approximate the Poisson distribution by a normal distribution and write $z = \lambda - n$, Eq. 59 becomes

$$\begin{aligned}
P(n|a,b) &\simeq \int_{-\infty}^{\infty} dz \, \frac{\mathrm{e}^{-\frac{z^2}{2(z+n)}}}{\sqrt{2\pi(z+n)}}\Gamma(z+n|a,b) \\
&= \int_{-\infty}^{\infty} dz \frac{\mathrm{e}^{-\frac{z^2}{2n}\left(1+\frac{z}{n}\right)^{-1}}}{\sqrt{2\pi n}} \cdot \left(1+\frac{z}{n}\right)^{-\frac{1}{2}} \Gamma\Big(n\Big[1+\frac{z}{n}\Big]\Big| a, b\Big).
\end{aligned} \tag{60}$$

We note that only values of z close to zero contribute to the integral when $n \gg 1$ because $z = 0$ is the minimum of the exponent in the integrand. Then $n \gg 1$ implies $z/n \ll 1$, and so

$$\begin{aligned} P(n|a,b) &\simeq \int_{-\infty}^{\infty} dz \, \frac{\mathrm{e}^{-\frac{z^2}{2n}}}{\sqrt{2\pi n}} \Gamma(n|a,b) \\ &= \Gamma(n|a,b) \end{aligned} \tag{61}$$

for large n, as expected [5].

Derivation of the protein distribution for a three-stage model of gene expression

We can use the same approximation of large γ to find the protein distribution for the three-stage model. Let $P^{(0)}_{m,n}$ be the probability of having m mRNAs and n proteins when the DNA is inactive and $P^{(1)}_{m,n}$ be the probability of having m mRNAs and n proteins when the DNA is active. The master equation consists of two coupled equations:

$$\begin{aligned} \frac{\partial P^{(0)}_{n,m}}{\partial \tau} &= \kappa_1 P^{(1)}_{m,n} - \kappa_0 P^{(0)}_{m,n} + (n+1) P^{(0)}_{m,n+1} - n P^{(0)}_{m,n} \\ &\quad + \gamma \left[(m+1) P^{(0)}_{m+1,n} - m P^{(0)}_{m,n} + bm \left(P^{(0)}_{m,n-1} - P^{(0)}_{m,n} \right) \right] \end{aligned} \tag{62}$$

$$\begin{aligned} \frac{\partial P^{(1)}_{n,m}}{\partial \tau} &= -\kappa_1 P^{(1)}_{m,n} + \kappa_0 P^{(0)}_{m,n} + (n+1) P^{(1)}_{m,n+1} - n P^{(1)}_{m,n} + a \left(P^{(1)}_{m-1,n} - P^{(1)}_{m,n} \right) \\ &\quad + \gamma \left[(m+1) P^{(1)}_{m+1,n} - m P^{(1)}_{m,n} + bm \left(P^{(1)}_{m,n-1} - P^{(1)}_{m,n} \right) \right] \end{aligned} \tag{63}$$

where $\kappa_0 = k_0/d_1$ and $\kappa_1 = k_1/d_1$. By defining two generating functions

$$f^{(0)}(z',z) = \textstyle\sum_{m,n} (z')^m z^n P^{(0)}_{m,n} \quad ; \quad f^{(1)}(z',z) = \textstyle\sum_{m,n} (z')^m z^n P^{(1)}_{m,n}, \tag{64}$$

these equations become

$$\frac{1}{v} \frac{\partial f^{(0)}}{\partial \tau} = \frac{1}{v} \left[\kappa_1 f^{(1)} - \kappa_0 f^{(0)} \right] - \frac{\partial f^{(0)}}{\partial v} + \gamma \left[b(1+u) - \frac{u}{v} \right] \frac{\partial f^{(0)}}{\partial u} \tag{65}$$

$$\frac{1}{v} \frac{\partial f^{(1)}}{\partial \tau} = \frac{1}{v} \left[-\kappa_1 f^{(1)} + \kappa_0 f^{(0)} \right] - \frac{\partial f^{(1)}}{\partial v} + a \frac{u}{v} f^{(1)} + \gamma \left[b(1+u) - \frac{u}{v} \right] \frac{\partial f^{(1)}}{\partial u} \tag{66}$$

with $u = z' - 1$ and $v = z - 1$.

At steady-state $\frac{\partial f^{(0)}}{\partial \tau} = \frac{\partial f^{(1)}}{\partial \tau} = 0$, and we find using the method of characteristics that

$$\begin{array}{lll} \frac{dv}{dr} = 1 & ; & \frac{du}{dr} = -\gamma \left[b(1+u) - \frac{u}{v} \right] \\ \frac{df^{(0)}}{dr} = \frac{1}{v} \left[\kappa_1 f^{(1)} - \kappa_0 f^{(0)} \right] & ; & \frac{df^{(1)}}{dr} = \frac{1}{v} \left[-\kappa_1 f^{(1)} + \kappa_0 f^{(0)} \right] + a \frac{u}{v} f^{(1)} \end{array} \tag{67}$$

where r measures the distance along a characteristic. Both u and v obey Eq. 24 again. Consequently, $v = r$ and $u \simeq \frac{bv}{1-bv}$ from Eq. 31 when $\gamma \gg 1$. From Eq. 67, we therefore obtain the two coupled differential equations:

$$v \frac{df^{(0)}}{\partial v} = \kappa_1 f^{(1)} - \kappa_0 f^{(0)} \tag{68}$$

$$v \frac{df^{(1)}}{\partial v} = -\kappa_1 f^{(1)} + \kappa_0 f^{(0)} + \frac{abv}{1-bv} f^{(1)}. \tag{69}$$

Following Hornos *et al.* [6], Eqs. 68 and 69 can be reduced to one differential equation for $f^{(0)}(v)$ by solving Eq. 68 for $f^{(1)}$ in terms of $f^{(0)}$ and its derivative, and inserting the result into Eq. 69. This equation becomes a second-order differential equation:

$$v(bv-1)\frac{df^{(0)}}{dv^2}+\Big[(\kappa_0+\kappa_1)(bv-1)+bv(1+a)-1\Big]\frac{df^{(0)}}{dv}+ab\kappa_0 f^{(0)}=0. \tag{70}$$

Eq. 70 has solution

$$f^{(0)}(v)=C\,{}_2F_1(\alpha,\beta,1-\kappa_0-\kappa_1;bv) \tag{71}$$

where ${}_2F_1(a,b,c;z)$ is a hypergeometric function,

$$\alpha = \frac{1}{2}\left(a+\kappa_0+\kappa_1+\sqrt{(a+\kappa_0+\kappa_1)^2-4a\kappa_0}\right) \tag{72}$$

$$\beta = \frac{1}{2}\left(a+\kappa_0+\kappa_1-\sqrt{(a+\kappa_0+\kappa_1)^2-4a\kappa_0}\right), \tag{73}$$

and C is a constant of integration.

We can find the generating function for protein numbers, $F(z)=f^{(0)}(z)+f^{(1)}(z)$, by using our solution for $f^{(0)}$ and Eq. 68 to find $f^{(1)}$. Determining the constant of integration C from $F(1)=1$ and using the relation $c(c+1)\,{}_2F_1(a,b,c;z)=c(c+1)\,{}_2F_1(a,b,c+1;z)+abz\,{}_2F_1(a+1,b+1,c+2;z)$, we find that

$$F(z)=\,{}_2F_1\Big(\alpha,\beta,\kappa_0+\kappa_1;b(z-1)\Big), \tag{74}$$

replacing v by $z-1$.

Expanding the generating function around $z=0$ determines the probabilities P_n. Using properties of the n-th derivatives with respect to z of the hypergeometric function, ${}_2F_1^{(n)}(a,b,c;z)$, we can write

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty}{}_2F_1^{(n)}\Big(\alpha,\beta,\kappa_0+\kappa_1;-b\Big)\frac{b^n}{n!}z^n \\ &= \sum_{n=0}^{\infty}\frac{\Gamma(\alpha+n)\Gamma(\beta+n)\Gamma(\kappa_0+\kappa_1)b^n}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\kappa_0+\kappa_1+n)n!}\,{}_2F_1\Big(\alpha+n,\beta+n,\kappa_0+\kappa_1+n;-b\Big)z^n \end{aligned} \tag{75}$$

and P_n can be found from the definition of $F(z)$: $F(z)=\sum_n P_n z^n$. With the linear transformation formulae for hypergeometric functions [2], we write P_n as

$$\begin{aligned} P_n &= \frac{\Gamma(\alpha+n)\Gamma(\beta+n)\Gamma(\kappa_0+\kappa_1)}{\Gamma(n+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\kappa_0+\kappa_1+n)}\left(\frac{b}{1+b}\right)^n\left(1-\frac{b}{1+b}\right)^{\alpha} \\ &\quad\times{}_2F_1\left(\alpha+n,\kappa_0+\kappa_1-\beta,\kappa_0+\kappa_1+n;\frac{b}{1+b}\right). \end{aligned} \tag{76}$$

The exact mRNA distributions

For completeness, we include the mRNA distributions for the two-stage and three-stage models. With initially zero mRNAs, the two-stage model has a Poisson distribution:

$$P_m(t)=\mathrm{e}^{-\langle m(t)\rangle}\frac{\langle m(t)\rangle^m}{m!} \tag{77}$$

where $\langle m(t)\rangle = m_s\left(1-\mathrm{e}^{-d_0 t}\right)$ and $m_s = v_0/d_0$ is the steady-state number of mRNAs. The propagator probability satisfies

$$P_{m|k}(t) = \sum_{r=0}^{k} \binom{k}{r} P_{m-r}(t)\left(1-\mathrm{e}^{-d_0 t}\right)^{k-r} \mathrm{e}^{-r d_0 t} \tag{78}$$

with $P_m(t) = 0$ if $m < 0$.

The steady-state distribution of mRNA for the three-stage model was first derived by Peccoud and Ycart, although they did not recognize it as such [7], and also by Raj *et al.* [8]. The exact probability of having m RNAs at steady-state is

$$P_m = \frac{m_s^m \mathrm{e}^{-m_s}}{m!} \cdot \frac{\Gamma(\zeta_0+m)\Gamma(\zeta_0+\zeta_1)}{\Gamma(\zeta_0+\zeta_1+m)\Gamma(\zeta_0)}\, {}_1F_1(\zeta_1, \zeta_0+\zeta_1+m; m_s) \tag{79}$$

where $m_s = v_0/d_0$, $\zeta_0 = k_0/d_0$, and $\zeta_1 = k_1/d_0$, and ${}_1F_1(a,b;z)$ is the confluent hypergeometric function of the first kind [2]. Eq. 79 like Eq. 18 can be bimodal. For $\zeta_1 = k_1/d_0 \gg 1$, Eq. 79 tends to a negative binomial distribution [8], because then mRNA synthesis is more burst-like. The distribution becomes Poisson when k_1 is zero, and the three-stage model reduces to the two-stage model.

References

[1] Bender, C. M. and Orzag, S. A. *Advanced Mathematical Methods for Scientists and Engineers.* Springer, New York, New York, (1999).

[2] Abramowitz, M. and Stegun, I. A. *Pocketbook of Mathematical Functions.* Harri Deutsch Publishing, Frankfurt am Main, (1984).

[3] van Kampen, N. G. *Stochastic Processes in Physics and Chemistry.* Elsevier, New York, New York, (1990).

[4] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical recipes in C++.* Cambridge University Press, New York, New York, (2002).

[5] Friedman, N., Cai, L., and Xie, X. S. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* **97**, 16830 (2006).

[6] Hornos, J. E., Schultz, D., Innocentini, G. C., Wang, J., Walczak, A. M., Onuchic, J. N., and Wolynes, P. G. Self-regulating gene: an exact solution. *Phys. Rev. E.* **72**, 051907 (2005).

[7] Peccoud, J. and Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).

[8] Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).