Stochastic branching-diffusion models for gene expression

David Cottrell[a], Peter S. Swain[b,1], and Paul F. Tupper[c,1]

[a]Department of Mathematics and Statistics, McGill University, Montreal, Canada; [b]SynthSys–Synthetic and Systems Biology, University of Edinburgh, Edinburgh, United Kingdom; and [c]Department of Mathematics, Simon Fraser University, Burnaby, Canada

Edited by Charles S. Peskin, New York University, Manhattan, NY, and approved April 3, 2012 (received for review January 20, 2012)

A challenge to both understanding and modeling biochemical networks is integrating the effects of diffusion and stochasticity. Here, we use the theory of branching processes to give exact analytical expressions for the mean and variance of protein numbers as a function of time and position in a spatial version of an established model of gene expression. We show that both the mean and the magnitude of fluctuations are determined by the protein's Kuramoto length—the typical distance a protein diffuses over its lifetime—and find that the covariance between local concentrations of proteins often increases if there are substantial bursts of synthesis during translation. Using high-throughput data, we estimate that the Kuramoto length of cytoplasmic proteins in budding yeast to be an order of magnitude larger than the cell diameter, implying that many such proteins should have an approximately uniform concentration. For constitutively expressed proteins that live substantially longer than their mRNA, we give an exact expression for the deviation of their local fluctuations from Poisson fluctuations. If the Kuramoto length of mRNA is sufficiently small, we predict that such local fluctuations become approximately Poisson in bacteria in much of the cell, unless translational bursting is exceptionally strong. Our results therefore demonstrate that diffusion can act to both increase and decrease the complexity of fluctuations in biochemical networks.

A challenge in systems biology is to understand how the spatial structure of cells influences signal transduction and information processing (1). Diffusion is not only fundamental to many developmental processes (2), but also to responses in differentiated cells: it is, for example, necessary for nanoclusters of signaling proteins to form temporarily at the cell membrane (3) and to allow some cells to polarize (4). Yet, most modelers assume that a cell is "well stirred"—that the effects of diffusion are negligible and that any location in the cell is identical to any other. Modeling space and diffusion in biochemical networks is particularly challenging because these networks are now recognized to often be substantially stochastic (5–7). Consequently, the standard approach is to use numerical simulation (8), but many simulations are required to build intuition on a system of interest. Further, multiple different methodologies for such simulations exist because of the difficulties of the underlying theory of stochastic reaction-diffusion systems (9).

Here, we present analytical solutions to a reaction-diffusion version of a well-known model of gene expression (10). We give expressions for the spatial correlations in the system and determine the limits under which the well-stirred approximation holds and when diffusion dominates, generating fluctuations that, unlike the well-stirred case, are locally approximately Poissonian. As well as building intuition, our analytical results should provide useful tests to validate algorithms for stochastic spatial simulations.

The key to our approach is to consider gene expression as a stochastic branching process. When solving stochastic systems, the master equation is usually mapped to a partial differential equation describing the evolution of the generating function of the system. With space and diffusion, however, this equation becomes a partial differential equation for a functional (11), which is difficult to solve. In our approach, spatial diffusion generates a system of partial differential equations for the evolution of a set of generating functions. These equations are considerably more tractable.

To begin, we illustrate our method by first considering gene expression without diffusion.

Gene Expression Without Diffusion

We consider a model of constitutive gene expression: Transcription of the gene can always occur. Fig. 1*A* shows the processes involved. The probability of having N_2 mRNAs and N_3 proteins obeys a master equation:

$$\begin{aligned}\frac{\partial P_{N_2,N_3}}{\partial t} &= v_2(P_{N_2-1,N_3} - P_{N_2,N_3}) + v_3 N_2(P_{N_2,N_3-1} - P_{N_2,N_3}) \\ &\quad + d_2[(N_2+1)P_{N_2+1,N_3} - N_2 P_{N_2,N_3}] \\ &\quad + d_3[(N_3+1)P_{N_2,N_3+1} - N_3 P_{N_2,N_3}]. \end{aligned} \quad [1]$$

This model of gene expression is a branching process because each molecule behaves independently and does not interact with any other molecules: there are no binary reactions. This independence means that a system with initially N molecules either of one or of several different species is statistically identical to the sum of N independent systems, each of which has a single initial molecule. We can consequently describe the evolution of the system by three different generating functions, all having one initial molecule, which can be either a DNA, an mRNA, or a protein.

We will first consider the standard generating function and its evolution (12). Letting N_i be the number of molecules of species i, where $i = 1$ for DNA, $i = 2$ for mRNA, and $i = 3$ for protein, then the generating function, defined as $\sum_n s^n P_n$ for a one dimensional system, can be written as

$$g(t, s_1, s_2, s_3) = \mathbb{E}\left[\prod_{i=1}^{3} s_i^{N_i(t)}\right], \quad [2]$$

where the s_i are auxiliary variables with $0 \le s_i \le 1$. We do not explicitly write the dependence on the initial condition. The expectation is taken over $P_{N_1,N_2,N_3}(t)$ and is shown as a time dependence in the exponent of Eq. **2**. Rescaling time by the degradation rate of protein, d_3, so that $\tau = d_3 t$ and substituting Eq. **2** into Eq. **1** gives

$$\frac{\partial g}{\partial \tau} = \left[a(s_2 - 1) + \gamma(b s_2(s_3 - 1) - (s_2 - 1))\frac{\partial}{\partial s_2} - (s_3 - 1)\frac{\partial}{\partial s_3}\right] g, \quad [3]$$

where $a = v_2/d_3$ is the number of mRNAs transcribed during a typical lifetime of a protein and $b = v_3/d_2$ is the burst size or the

Author contributions: D.C., P.S.S., and P.F.T. performed research; D.C., P.S.S., and P.F.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: peter.swain@ed.ac.uk or pft3@sfu.ca.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201103109/-/DCSupplemental.

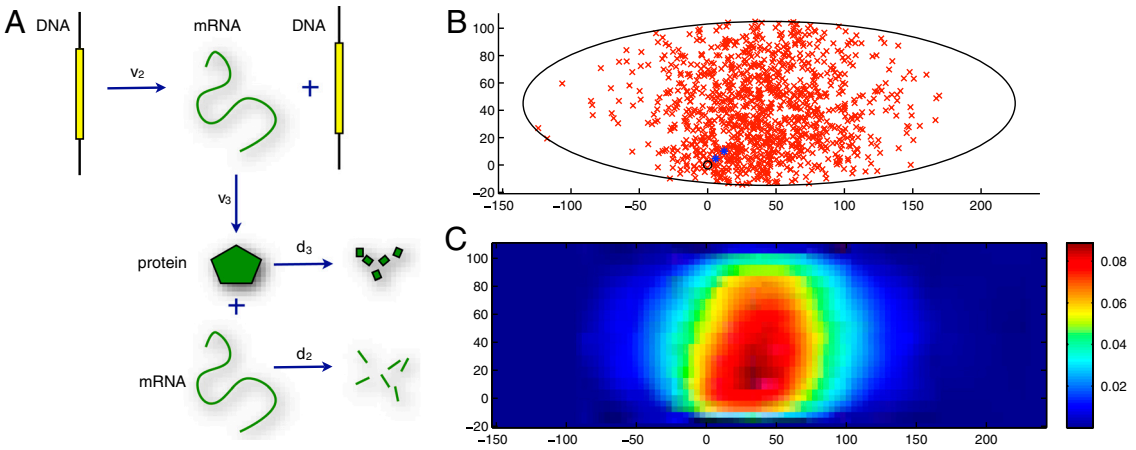

Fig. 1. (*A*) A model of constitutive gene expression (10). DNA is transcribed into mRNA with a probability per unit time of v_2; mRNA is translated into protein with a probability per unit time of v_3. Both mRNA and protein are degraded: mRNA with a probability per unit time of d_2 and protein with a probability per unit time of d_3. (*B*) Protein usually diffuses further from the DNA than mRNA because mRNA has both a lower diffusion coefficient and degrades faster. A snapshot of simulation results for an elliptical cell. Red crosses denote protein, blue asterisks denote mRNA, and the black circle at the origin denotes the DNA. We set $d_2 = 0.006\ \text{s}^{-1}$ (a half-life of 2 min), $d_3 = 0.0002\ \text{s}^{-1}$ (a half-life of 1 h), $v_2 = 0.01\ \text{s}^{-1}$, and $v_3 = 0.12\ \text{s}^{-1}$ (10). Consequently, $a \simeq 52$, $b = 20$, and $\gamma = 30$. The diffusion coefficients are $D_2 = 0.5\ \mu\text{m}^2\,\text{s}^{-1}$ and $D_3 = 5\ \mu\text{m}^2\,\text{s}^{-1}$ giving Kuramoto lengths of $\kappa_2 = 9.3\ \mu\text{m}$ and $\kappa_3 = 160\ \mu\text{m}$. Axis labels are in micrometers. (*C*) A corresponding heat map of the steady-state number density of molecules of protein.

typical number of proteins synthesized from a single mRNA during the mRNA's lifetime (13). The parameter γ is the ratio of the lifetime of the protein to the mRNA—$\gamma = d_2/d_3$—and is typically larger than one (5). Eq. **3** has the initial condition $g(0, s_1, s_2, s_3) = \prod_{i=1}^{3} s_i^{N_i(0)}$ when there is initially $N_i(0)$ molecules of species i. It can be solved approximately for $\gamma \gg 1$ giving a negative binomial distribution for protein numbers at steady state (5).

We consider an alternative description of this system using techniques from the theory of branching processes (14, 15). The number of molecules of species i at time t is a sum of the number of species of type i generated by each of the initial molecules. Writing $N_{i|j}^{(k)}(t)$ to be the number of molecules of species i at time t that are generated from the kth initial molecule of species j, then the number of molecules of species i is $N_i(t) = \sum_{j=1}^{3}\sum_{k=1}^{N_j(0)} N_{i|j}^{(k)}(t)$. The $N_{i|j}^{(k)}$ for $k = 1, \ldots, N_j(0)$ are identically distributed and independent. From Eq. **2**, we can then write the generating function as

$$g = \mathbb{E}\left[\prod_{i=1}^{3}\prod_{j=1}^{3}\prod_{k=1}^{N_j(0)} s_i^{N_{i|j}^{(k)}(t)}\right] = \prod_{j=1}^{3}\prod_{k=1}^{N_j(0)} \mathbb{E}\left[\prod_{i=1}^{3} s_i^{N_{i|j}^{(k)}(t)}\right], \qquad [4]$$

but, again from Eq. **2**, this last expectation is the generating function for a gene expression process that starts with a single molecule of species j. We will define u_j to be such a generating function:

$$u_j(t, s_1, s_2, s_3) = g(t, s_1, s_2, s_3 | 1 \text{ molecule of species } j). \qquad [5]$$

Hence,

$$g = \prod_{j=1}^{3}\prod_{k=1}^{N_j(0)} u_j = \prod_{j=1}^{3} u_j^{N_j(0)}, \qquad [6]$$

and so the three u_j collectively contain the same information as the generating function g.

From the chemical reactions occurring during gene expression, we can derive how the u_j evolve over time. For example, consider the generating function corresponding to a single initial DNA molecule, u_1, at time $t + dt$ for a small time interval dt. From Eq. **2**, this generating function obeys

$$u_1(t + dt) = \mathbb{E}\left[\prod_{i=1}^{3} s_i^{N_i(t+dt)} \middle| \mathbf{N}(0) = (1, 0, 0)\right] \qquad [7]$$

by definition. If we consider the time interval dt to be at the start of the dynamics of the system when only a single DNA molecule is present and to be small enough that only one reaction can possibly have occurred then this reaction can only be the synthesis of an mRNA. Consequently, remembering that v_2 is the probability of synthesis of an mRNA per unit time, we can write

$$\begin{aligned} u_1(t + dt) &= v_2 dt \mathbb{E}\left[\prod_{i=1}^{3} s_i^{N_i(t+dt)} \middle| \mathbf{N}(dt) = (1, 1, 0)\right] \\ &+ (1 - v_2 dt)\mathbb{E}\left[\prod_{i=1}^{3} s_i^{N_i(t+dt)} \middle| \mathbf{N}(dt) = (1, 0, 0)\right] \\ &= v_2 dt u_1 u_2 + (1 - v_2 dt)u_1, \end{aligned} \qquad [8]$$

and so that

$$\frac{\partial}{\partial t} u_1 = v_2 u_1 (u_2 - 1). \qquad [9]$$

Similarly, and if we rescale time to $\tau = d_3 t$, the three generating functions satisfy

$$\frac{\partial}{\partial \tau}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} a u_1 (u_2 - 1) \\ \gamma(b u_2 (u_3 - 1) - (u_2 - 1)) \\ -(u_3 - 1) \end{bmatrix}, \qquad [10]$$

with the initial condition $u_j(0, s_1, s_2, s_3) = s_j$ because at $t = 0$ only one molecule is present.

Gene Expression with Diffusion

We now include diffusion within Γ, a region of $\mathbb{R}^d$ (16, 17) (Fig. 1 *B* and *C*). Γ may either be all of $\mathbb{R}^d$ or a subregion with a boundary reflecting diffusing molecules. Let mRNA molecules have a diffusion coefficient of D_2 and protein molecules have a diffusion coefficient of D_3. We will consider DNA molecules that neither diffuse nor decay but are fixed at a point ξ_0 in Γ. If space has d dimensions then ξ_0 is a d-dimensional vector. We will let $x_i^{(k)}$ be the spatial location of the k'th molecule of species i. The generating function now depends on s_i that are functions of space, $s_i = s_i(\xi)$, with ξ in Γ, but has a similar form to Eq. **2**:

$$g(t, s_1, s_2, s_3) = \mathbb{E}\left[\prod_{i=1}^{3}\prod_{k=1}^{N_i(t)} s_i(x_i^{(k)}(t))\right]. \qquad [11]$$

The moments of the distribution for the number of molecules are derived analogously to the well-stirred model but now by functional differentiation of $g(t, s_1, s_2, s_3)$ by the variables $s_i(\xi)$ and then evaluating at $s_i = 1$. As before, we define generating functions for systems that initially have only one molecule of species j at location ξ_0

$$u_j(t, s_1, s_2, s_3|\xi_0) = g(t, s_1, s_2, s_3|1 \text{ molecule of species } j \text{ at } \xi_0), \quad [12]$$

and then the generating function can be expressed in terms of these new u_j

$$g(t, s_1, s_2, s_3) = \prod_{j=1}^{3} \prod_{k=1}^{N_j(0)} u_j(t, s_1, s_2, s_3|x_j^{(k)}(0)). \quad [13]$$

The evolution equations for the u_i are also similar although they contain the Laplacian operator

$$\left(\frac{\partial}{\partial \tau} - \mathbf{D}\nabla^2_{\xi_0}\right)\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} au_1(u_2 - 1) \\ \gamma(bu_2(u_3 - 1) - (u_2 - 1)) \\ -(u_3 - 1) \end{bmatrix}, \quad [14]$$

with the initial conditions $u_i(0, s_1, s_2, s_3|\xi_0) = s_i(\xi_0)$ and where the matrix $\mathbf{D}$ is the diffusion matrix: $\mathbf{D} = d_3^{-1}\text{diag}(0, D_2, D_3)$. Eq. **14** holds for all ξ_0 in Γ, with the normal derivatives of the u_i being zero on the boundary of Γ, if applicable. By using branching processes, we have a Laplacian evaluated at the coordinates of the initial condition. Consequently, the evolution equations for the moments of the process are simpler than those derived directly from g.

Eq. **10** can be recovered from Eq. **14** if the initial conditions are constant in space: $u_i(0, s_1, s_2, s_3|\xi_0) = \text{constant}$ because then the term $\nabla^2_{\xi_0}\mathbf{u}$ vanishes.

First- and Second-Order Statistics

Including space, we must now consider fields for the molecular species rather than numbers of molecules. Let the field for species i be $X_i(t, \xi)$. It is defined as a distribution determined by the locations of the molecules of species i: $X_i(t, \xi) = \sum_{k=1}^{N_i(t)} \delta(x_i^{(k)}(t) - \xi)$. The field for DNA is, however, deterministic and constant with respect to time because we have one DNA molecule fixed at ξ_0.

We can obtain the moments of the process at a single point in time by differentiating the u_i with respect to $s_i(\xi)$. The first-order statistics are mean density fields; the second-order statistics are distributions on $\mathbb{R}^d \times \mathbb{R}^d$. Let $M_{i|j}$ be the mean density of species i given one initial molecule of species j at ξ_0. Then

$$M_{i|j}(\tau, \xi|\xi_0) = \mathbb{E}[X_i(\tau, \xi)|X_k(0, \xi) = \delta_{jk}\delta(\xi - \xi_0)] = \left.\frac{\delta u_j}{\delta s_i(\xi)}\right|_{s=1}. \quad [15]$$

For example, with this definition, the mean protein field is

$$\left.\frac{\delta u_j}{\delta s_3(\xi)}\right|_{s=1} = \left.\mathbb{E}\left[\frac{\delta}{\delta s_3(\xi)}\prod_{i=1}^{3}\prod_{k=1}^{N_i(t)} s_i(x_i^{(k)}(t))\right]\right|_{s=1}$$
$$= \mathbb{E}\left[\sum_{k=1}^{N_i(t)} \delta(\xi - x_3^{(k)}(t))\right] = \mathbb{E}[X_3(t, \xi)], \quad [16]$$

where we have not explicitly written the dependence on one initial molecule of species j at ξ_0. Integrating Eq. **15** over a volume gives the expected number of molecules of species j in the volume. Similarly, let $C_{ii'|j}$ denote the covariance densities between molecules of species i in one spatial location and molecules of species i' at another, given one initial molecule of species j at ξ_0, then

$$C_{ii'|j}(\tau, \xi, \xi'|\xi_0) = \mathbb{E}[X_i(\tau, \xi)X_{i'}(\tau, \xi')] - M_{i|j}(\tau, \xi|\xi_0)M_{i'|j}(\tau, \xi'|\xi_0)$$
$$= \left.\frac{\delta^2 u_j}{\delta s_i(\xi)\delta s_{i'}(\xi')}\right|_{s=1} + M_{i|j}(\tau, \xi|\xi_0)\delta_{ii'}\delta(\xi - \xi') - M_{i|j}(\tau, \xi|\xi_0)M_{i'|j}(\tau, \xi'|\xi_0), \quad [17]$$

where again we have not written explicitly the dependence of the expectation on one initial molecule of species j at ξ_0. Integrating Eq. **17** with respect to ξ over one volume and with respect to ξ' over another volume gives the expected covariance between the number of molecules of species i in the first volume with the number of molecules of species i' in the second volume.

Analytical Results

Reaction-diffusion systems are often characterized by the Kuramoto length (18): the distance a molecule typically diffuses over its lifetime. We will define a Kuramoto length for both mRNA and protein:

$$\kappa_2 = \sqrt{\frac{D_2}{d_2}}; \qquad \kappa_3 = \sqrt{\frac{D_3}{d_3}}. \quad [18]$$

We will consider either a d-dimensional space or specialize to three dimensions.

In principle, our main results hold for arbitrary geometries and are presented in terms of a general probability density describing diffusion of single molecules. Usually, and for simplicity, we will give results for free space, assuming that gene expression occurs sufficiently far from a confining membrane. Then the diffusion density, $f(t, \xi|\xi_0)$, is a Gaussian function: $f(t, \xi|\xi_0) = (4\pi t)^{-\frac{d}{2}}\exp[-\frac{|\xi - \xi_0|^2}{4t}]$. Alternatively, for Brownian diffusion in a region Γ with molecules reflecting off the boundary, we must solve the diffusion equation for $f(t, \xi|\xi_0)$ with the boundary condition that the spatial derivative of f normal to the boundary is zero. For general domains Γ, this system does not have a closed-form solution, but a series solution exists in terms of the eigenvalues and eigenmodes of the Laplacian operator on the region (19).

Mean and Covariance. To begin, we calculate the first and second moments of the distribution as integrals over the diffusion density. We functionally differentiate Eq. **14** with respect to $s_i(\xi)$ and evaluate at $s_i = 1$ to obtain a system of inhomogeneous linear reaction-diffusion equations. These equations can be solved using Green's function techniques (*SI Text*).

The mean density of mRNA is then

$$M_{2|1}(\tau, \xi|\xi_0) = a\int_0^\tau e^{-\gamma\tau_1} f(\kappa_2^2\gamma\tau_1, \xi|\xi_0)\mathrm{d}\tau_1, \quad [19]$$

and the mean density of protein is

$$M_{3|1}(\tau, \xi|\xi_0) = ab\gamma\int_0^\tau\int_0^{\tau_1} e^{-\gamma(\tau_1 - \tau_2) - \tau_2} \times f(\kappa_2^2\gamma(\tau_1 - \tau_2) + \kappa_3^2\tau_2, \xi|\xi_0)\mathrm{d}\tau_2\mathrm{d}\tau_1. \quad [20]$$

The single-time covariance matrix, $\mathbf{C}_{|1}$, has components $C_{ii'|1}$ and obeys

$$\mathbf{C}_{|1} = \delta(\xi - \xi')\begin{bmatrix} 0 & 0 & 0 \\ 0 & M_{2|1} & 0 \\ 0 & 0 & M_{3|1} \end{bmatrix} + ab\gamma \int_0^\tau \int_0^{\tau_1} e^{-\gamma(\tau_1-\tau_2)} \times \int_{\mathbb{R}^d} f(\kappa_2^2\gamma(\tau_1-\tau_2), \zeta|\xi_0)\mathbf{F}(\tau_2, \xi, \xi'|\zeta)\mathrm{d}\zeta\mathrm{d}\tau_2\mathrm{d}\tau_1, \quad [21]$$

where the matrix $\mathbf{F}$ is

$$\mathbf{F}(\tau, \xi, \xi'|\xi_0) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & M_{2|2}M'_{3|3} \\ 0 & M_{3|3}M'_{2|2} & M_{3|2}M'_{3|2} + M_{3|2}M'_{3|3} \end{bmatrix}. \quad [22]$$

We use a prime to denote evaluation at ξ' so that $M'_{j|i} = M_{j|i}(\tau, \xi'|\xi_0)$ and $M_{j|i} = M_{j|i}(\tau, \xi|\xi_0)$. The mean densities with different initial conditions are

$$\begin{aligned} M_{2|2}(\tau, \xi|\xi_0) &= e^{-\gamma\tau} f(\kappa_2^2\gamma\tau, \xi|\xi_0) \\ M_{3|2}(\tau, \xi|\xi_0) &= b\gamma \int_0^\tau e^{-\gamma(\tau-\tau_1)-\tau_1} \\ &\quad \times f(\kappa_2^2\gamma(\tau-\tau_1) + \kappa_3^2\tau_1, \xi|\xi_0)\mathrm{d}\tau_1 \\ M_{3|3}(\tau, \xi|\xi_0) &= e^{-\tau} f(\kappa_3^2\tau, \xi|\xi_0). \end{aligned} \quad [23]$$

We will use these statistics to find radial correlation functions, which can be calculated explicitly.

Specializing to free space, our results when integrated over all space agree, as expected, with the nonspatial case. Setting $\xi_0 = 0$ and letting the overline denote integration over all spatial variables, the average total number of mRNAs in all of $\mathbb{R}^d$ is

$$\begin{aligned} \overline{M_{2|1}}(\tau) &= \int_{\mathbb{R}^d} M_{2|1}(\tau, \xi|0)\mathrm{d}\xi \\ &= a\int_0^\tau e^{-\gamma\tau_1} \int_{\mathbb{R}^d} f(\kappa_2^2\gamma\tau_1, \xi|0)\mathrm{d}\xi\mathrm{d}\tau_1 = \frac{a}{\gamma}(1 - e^{-\gamma\tau}), \end{aligned} \quad [24]$$

agreeing with earlier work (5). Similarly, the average total number of protein molecules is

$$\overline{M_{3|1}}(\tau) = ab\left(1 - \frac{\gamma}{\gamma - 1}e^{-\tau} + e^{-\gamma\tau}\right), \quad [25]$$

as expected (5, 10).

For the covariance functions, we need to integrate over ξ and ξ' to compare with the nonspatial system. We find that the protein-protein covariance in the limit of $\tau \to \infty$ is

$$\overline{C_{33|1}}(\tau) \overset{\tau\to\infty}{\to} ab\left(1 + \frac{b\gamma}{1+\gamma}\right), \quad [26]$$

again as expected (5, 10). For small τ, we find $\overline{C_{33|1}}(\tau) = ab(1 + \frac{1}{3}b\gamma^2\tau^3) + o(\tau^4)$, a behavior similar to Poisson fluctuations.

The Steady-State Limit: Means. We use Laplace transforms to calculate the steady-state limits (*SI Text*). Assuming free space and that the DNA is fixed at the origin ($\xi_0 = 0$), the mean mRNA and protein densities at steady-state are

$$M_{2|1}(\xi) = \frac{a}{\gamma}\frac{e^{-|\xi|/\kappa_2}}{4\pi|\xi|\kappa_2^2}; \qquad M_{3|1}(\xi) = ab\frac{e^{-|\xi|/\kappa_3} - e^{-|\xi|/\kappa_2}}{4\pi|\xi|(\kappa_3^2 - \kappa_2^2)}. \quad [27]$$

We can derive the steady-state mean squared distance of mRNA and protein from the origin from these equations, by integrating the product of $|\xi|^2$ and either $M_{2|1}$ or $M_{3|1}$ over all space:

$$\langle|\xi_2|^2\rangle = \frac{6a\kappa_2^2}{\gamma}; \qquad \langle|\xi_3|^2\rangle = 6ab(\kappa_2^2 + \kappa_3^2),$$

and we see that the mean squared distance of protein is given by the sum of the squares of the Kuramoto lengths of mRNA and protein because protein is synthesized from a diffusing source (of mRNA).

The Steady-State Limit: Radial Correlation Functions. Even assuming free space, the covariance densities do not appear to yield simple analytic forms in the limit as $\tau \to \infty$. Instead, we can calculate explicit expression for the radial correlation functions where the radius r is the distance between the two fields whose covariation we are studying. For species i and j with fields X_i and X_j at time τ, we define $\Gamma_{ij}(\tau, r)$ to be the radial correlation function. It satisfies

$$r^{d-1}A_d\Gamma_{ij}(\tau, r) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \delta(r - |\xi'|)C_{ij|1}(\tau, \xi, \xi + \xi')\mathrm{d}\xi\mathrm{d}\xi', \quad [28]$$

where A_d is the area of the unit sphere in d dimensions, and ξ and ξ' are vectors measured from ξ_0, the location of the DNA molecule. The radial correlation function is still a density and determines how on average the covariance between a molecule of species i and a molecule of species j depends on the distance between these two molecules. It is calculated by averaging over all possible positions of these molecules relative to the DNA molecule. With this definition, we find that the radial correlation functions can be written as integrals (over time), but become explicit functions when $\tau \to \infty$. Considering only $d = 3$, free space, and writing $\kappa^2 = \frac{\gamma\kappa_2^2 + \kappa_3^2}{\gamma + 1}$, then

$$\Gamma_{22}(r) = \frac{\delta(r)}{4\pi r^2}\frac{a}{\gamma}; \qquad \Gamma_{23}(r) = \frac{ab}{\gamma\kappa_2^2 + \kappa_3^2}\frac{e^{-r/\kappa}}{4\pi r} \quad [29]$$

and

$$\Gamma_{33}(r) = \frac{\delta(r)}{4\pi r^2}ab + \frac{ab^2}{\kappa_3^2 - \kappa_2^2}\frac{e^{-r/\kappa_3} - e^{-r/\kappa}}{4\pi r}. \quad [30]$$

The $\delta(r)$ is interpreted in the right-handed sense: $\int_0^\epsilon \delta(z)g(z)\mathrm{d}z = g(0)$ for all $\epsilon > 0$. Eqs. **29** and **30** should be integrated over r to be interpreted as covariances of numbers of molecules. Integrating Eq. **29** over any region that does not include $r = 0$ shows that there is no covariance between spatially separated mRNAs, and integrating Eq. **30** over all r (multiplying by $4\pi r^2$ because we are using spherical coordinates) recovers Eq. **26**.

We further verified Eqs. **27**, **29**, and **30** by comparing integrals of these densities with Monte Carlo simulations (*SI Text*).

Limiting Cases

To more easily interpret the expressions for the mean of the protein field, Eq. **20**, and its covariance, Eq. **17**, we consider two limiting cases: the limit of rapid degradation of mRNA compared to degradation of protein ($\gamma \to \infty$) and the limit of fast diffusion of proteins ($\kappa_3 \to \infty$).

The Limit of Fast Degradation of mRNA Compared to Protein. Many proteins have substantially longer lifetimes than their corresponding mRNAs (5). In the limit of $\gamma \gg 1$, with a and b remaining finite, mRNA diffuses relatively little over its lifetime compared to protein, and it appears that all the protein is synthesized at the DNA at the length scale associated with the diffusion of the protein, κ_3. Taking $\gamma \to \infty$ and $\xi_0 = 0$, we find at steady-state and in free space that

$$M_{3|1}(\xi) = \frac{ab}{4\pi|\xi|\kappa_3^2}\exp(-|\xi|/\kappa_3)$$

$$C_{33|1}(\xi,\xi') = \delta(\xi-\xi')M_{3|1}(\xi) + \frac{ab^2}{2\pi^3\kappa_3^6}\frac{K_2(\sqrt{2\theta})}{\theta}, \quad [31]$$

where $\theta = (|\xi|^2 + |\xi'|^2)/\kappa_3^2$. The function K_2 is the order 2 modified Bessel function of the second kind. In this $\gamma \gg 1$ limit, Eq. **31** implies that the correlation between the protein field at two different points depends only on the distance of each of the points to the DNA, and not on the position of the points relative to each other. These correlations arise because of the "burstiness" of the synthesis of protein at the origin: A single fixed Poisson source of protein would lead to a protein density field that is uncorrelated at distinct points.

One application of Eq. **31** is to estimate how far the distribution of proteins differs from being Poisson. For a reaction-diffusion system at equilibrium, detailed balance holds and the distribution of numbers of each chemical species is Poisson (11). Most biological systems are, however, far from equilibrium, but fluctuations averaged over a small volume should be dominated by diffusion and can be approximately Poisson (11). For a small region of volume, we can find the mean and variance of the number of molecules in that volume by integrating $M_{3|1}$ and $C_{33|1}$ over the region. At steady-state, we would obtain $C_{33|1}(\xi,\xi') = \delta(\xi-\xi')M_{3|1}(\xi)$ for a Poisson field, and the mean and variance would be equal. We can therefore measure the deviation from being Poisson in a small region by computing the Fano factor (the variance divided by the mean). At steady-state, when $\gamma \to \infty$, and for a small volume $\Delta\xi^3$ at a distance ξ from the origin, the result is (*SI Text*)

$$\text{Fano factor} = 1 + b\left(\frac{\Delta\xi}{\kappa_3}\right)^3 \mathcal{G}(\xi/\kappa_3), \quad [32]$$

where b is the burst size or the typical number of proteins synthesized per mRNA and $\mathcal{G}(z) = K_2(2z)\exp(z)/(2\pi^2 z)$ (*SI Text*).

We see from Eq. **32** that for large γ the deviation from being Poisson becomes small from either the burst size b being small, the volume of the region measured in units of the protein's Kuramoto length being small, or else the region being far from the DNA source (Fig. 2). This requirement for $\gamma \gg 1$ (the ratio of protein to mRNA lifetimes being large) implies short mRNA lifetimes and so the Kuramoto length of mRNA is negligible. (The effective lifetime of a protein cannot be increased arbitrarily because proteins are lost at cell division as well as being degraded.) Nevertheless, Eq. **32** confirms the observations of Saunders and Howard (20).

Limit of Fast Protein Diffusion. Spatially extended chemical systems are often assumed to be spatially homogeneous, obviating the need for considering the location of molecules. One justification for this assumption is that if diffusion rates of relevant species are fast compared to reaction rates, the species behave as if they are uniformly distributed over the region in question. For example, Grima and Schnell argue that diffusion effects become important when the average intermolecular distance and the Kuramoto length are of the same order (8). In our model, we can examine the limit of large κ_3. In free space, this limit implies that the steady-state protein density will converge to zero everywhere because proteins will rapidly leave any neighborhood of the DNA. To obtain a nontrivial limit, we therefore include diffusion in confined geometries. Let Γ be a reference region of dimensionless unit diameter (such as a sphere or a cube). We consider our system confined to the region $L\Gamma$ where L has units of length, so that our region has volume $L^d\text{Vol}(\Gamma)$ with d being the dimension of space, and use a series solution for $f(t,\xi|\xi_0)$ (*SI Text*).

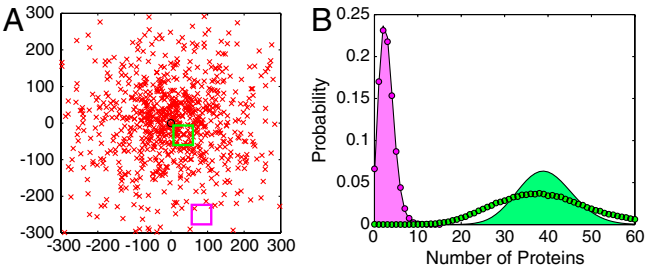

Fig. 2. Local fluctuations in protein numbers become approximately Poisson sufficiently far from the DNA. (*A*) A snapshot of a simulation of gene expression with diffusion in free space. Protein is denoted by red crosses and the DNA by the black circle. In this snapshot, there is by chance no mRNA. Parameter values are given in Fig. 1. Axis labels are in micrometers. (*B*) Histograms of counts of proteins sampled from the purple and green boxes in *A*. A total of 10^6 samples were taken at 10 s intervals. Each box has $\Delta\xi/\kappa_3 = 0.31$. The colored dots show data collected from the simulation. The filled curves show the best Poisson fits to the data. For the green box $|\xi|/\kappa_3 = 0.28$ and Eq. **32** predicts a deviation from being Poisson of about 0.8; for the purple box $|\xi|/\kappa_3 = 1.55$ with a predicted deviation from Poisson of about 0.005.

We find a uniform protein concentration approximately holds when the diameter of the region of interest is small compared to the protein's Kuramoto length. To be well mixed, the mean protein density needs to be constant over the region. We computed a series expansion of $M_{3|1}$ in powers of L/κ_3 at steady-state. To first-order, we find

$$M_{3|1}(\xi) = \frac{ab}{L^d\text{Vol}(\Gamma)}\left[1 + \left(\frac{L}{\kappa_3}\right)^2 \eta(\xi/L) + \cdots\right], \quad [33]$$

where η is a dimensionless function of position depending on the geometry of Γ. We see that $M_{3|1}$ approaches the well-stirred result of a concentration of $ab/L^d\text{Vol}(\Gamma)$ when $L \ll \kappa_3$.

Estimating the Kuramoto Lengths of Cytoplasmic Proteins

Our results show that the Kuramoto length determines both the size and the nature of local fluctuations, and we therefore estimated the Kuramoto length for cytoplasmic proteins in budding yeast. Although there are high-throughput measurements of protein lifetimes (21), there have been almost no measurements of diffusion coefficients in budding yeast. To estimate diffusion coefficients, we therefore used the measured diffusion coefficient for the kinase Fus3p (22) and rescaled this diffusion coefficient by the cubed root of the ratio of the molecular mass of Fus3p to the molecular mass of the protein of interest (*SI Text*). Although this approach assumes that proteins are uniformly dense spheres and so our results are only approximate, we believe that they are still informative.

All of the proteins we considered had Kuramoto lengths larger than a typical cell diameter (*SI Text*): Both the mean and the median Kuramoto lengths are well over 100 μm (160 and 128 μm, respectively), but the diameter of a cell is only approximately 4 μm (23). We expect this difference to also hold for some proteins in bacteria. For example, Green Fluorescent Protein (GFP) has a diffusion coefficient of 7.7 $\mu m^2\,s^{-1}$ in *Escherichia coli* (24) and typically decays only through dilution. With a cell-cycle time of 40 min, GFP then has a Kuramoto length of approximately 160 μm. An *E. coli* cell, however, has a length of around 2.5 μm (25).

Our results therefore indicate that cytoplasmic proteins are, perhaps typically, approximately uniformly distributed in *E. coli* and in budding yeast, at least assuming spherical proteins with constant diffusion coefficients and Brownian diffusion. Although we have not explicitly included the nucleus in our calculations, Eq. **33** is valid for the volume between two spheres providing this volume is not small compared to the volume of the outer sphere. We have also assumed constitutive expression, and Eq. **33** may

change with sufficiently large bursts of transcription. In contrast, for mRNAs, even neglecting nuclear export, we predict the Kuramoto length to be an order of magnitude smaller than the Kuramoto length of a protein and so of similar size to the cell diameter (assuming diffusion coefficients that are an order of magnitude less than those of protein and degradation rates that are an order of magnitude greater).

Discussion

To include spatial effects into models of biochemical networks is challenging. Cells have complex internal geometries, are intracellularly heterogeneous, and are packed with molecules, potentially generating substantial volume exclusion (26). Furthermore, combining spatial and stochastic effects even in simple geometries and homogeneous environments is mathematically challenging (11). Here, we have shown how techniques from the theory of branching processes can be used to derive analytical expressions for both the local mean and variance of proteins in an established model of gene expression, at least for point molecules and homogeneous, Brownian diffusion. Such Brownian diffusion is appropriate for proteins, at least when measured in *E. coli* (27), but anomalous diffusion (28) and active transport (29) has been reported for mRNA.

Our approach is extensible to other first-order biochemical networks in arbitrary geometries and with Markovian diffusion. For example, we need not only consider constitutively expressed genes but can also include regulated gene expression. Such expression is often modeled using a promoter with two states: one "off," with no expression, and the other "on," with a constant probability of expression per unit time (30). This model has been applied widely from bacteria (31) to human cells (32) and fits within our framework of branching processes (*SI Text*). To include the nucleus in our model, we should use a diffusion density for a confined region (such as that between two spheres) to describe the cytoplasm and consider the source of transcription in Fig. 1*A* not as a DNA molecule but as a nuclear pore complex stochastically exporting mRNA. Indeed, we can describe the export of mRNA from multiple nuclear pores diffusing on the outer membrane of the nucleus as a branching process.

Despite the complexity of the intracellular environment, our results indicate the local fluctuations of some proteins in bacteria can be well approximated by Poisson fluctuations, at least for those constitutively expressed and with small Kuramoto lengths for their mRNA. For example, assuming $\Delta\xi$ in Eq. **32** is 1/8 of the length of the cell, a conservative estimate of the Kuramoto length of 10 μm, and that the limit of large γ holds (and so negligible Kuramoto lengths for mRNA), then to have a correction to Poisson behavior of at least 10% in Eq. **32**, the burst size must be greater than 50 if the local volume is a distance of 0.75 μm from the DNA and greater than 600 if the local volume is at distance 1.75 μm. From measurements of around 1,000 genes in *E. coli*, however, over 92% of genes have proteins expressed with a b value less than 50 and over 99% have a b value of less than 600 (33). Almost all genes, though, even if the Kuramoto length for their mRNA is sufficiently small, would still have non-Poissonian fluctuations in protein numbers if the local volume is only 0.25 μm away from the DNA.

Our results imply that cytoplasmic proteins are often uniformly distributed in budding yeast and *E. coli* providing the levels of these proteins have had time to reach steady-state and the approximations we have made, particularly of spherical proteins and constitutive expression, are valid. Further, we have shown that diffusion can cause local fluctuations to be close to Poisson sufficiently far from, but not close to, the DNA. We expect non-Poisson behavior in cellular compartments that have diameters smaller than the Kuramoto length of protein, however, because of the reflecting boundary conditions imposed by the walls of the compartment (12). We also predict non-Poissonian behavior if translational bursting is sufficiently strong. Although our results are most applicable to bacteria because we have not explicitly included the nucleus, we expect that the intuition gained holds more generally (11, 12). With our analytical calculations, we have thus demonstrated that although space and diffusion are often a source of increased complexity when studying intracellular dynamics as a whole, they have, in contrast, the potential to simplify local behavior.

ACKNOWLEDGMENTS. We thank Ramon Grima for commenting on the manuscript. P.S.S. is supported by a Scottish Universities Life Sciences Alliance (SULSA) chair in Systems Biology. P.F.T. holds a Canada Research Chair in Applied Mathematics. P.F.T. and D.C. were supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

1. Kholodenko BN, Hancock JF, Kolch W (2010) Signalling ballet in space and time. *Nat Rev Mol Cell Biol* 11:414–426.
2. Turing A (1952) The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 237:37–72.
3. Tian T, et al. (2007) Plasma membrane nanoswitches generate high-fidelity Ras signal transduction. *Nat Cell Biol* 9:905–914.
4. Howell AS, et al. (2009) Singularity in polarization: Rewiring yeast cells to make two buds. *Cell* 139:731–743.
5. Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA* 105:17256–17261.
6. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* 135:216–226.
7. Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. *Nature* 467:167–173.
8. Grima R, Schnell S (2008) Modelling reaction kinetics inside cells. *Essays Biochem* 45:41–56.
9. Lemerle C, Di Ventura B, Serrano L (2005) Space as the final frontier in stochastic simulations of biological systems. *FEBS Lett* 579:1789–1794.
10. Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA* 98:8614–8619.
11. Gardiner CW (1990) *Handbook of Stochastic Methods* (Springer, Berlin).
12. Van Kampen NG (1981) *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam).
13. Friedman N, Cai L, Xie XS (2006) Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Phys Rev Lett* 97:168302.
14. Athreya K, Ney P (1972) *Branching Processes* (Springer, Berlin).
15. Haccou P, Jagers P, Vatutin VA, Dieckmann U (2007) *Branching Processes: Variation, Growth, and Extinction of Populations* (Cambridge Univ Press, Cambridge, UK).
16. Engländer J (2007) Branching diffusions, superdiffusions and random media. *Prob Surv* 4:303–364.
17. Etheridge AM (2000) *An Introduction to Superprocesses* (American Mathematical Society, Providence, Rhode Island).
18. Kuramoto Y (1974) Effects of diffusion on the fluctuations in open chemical systems. *Prog Theor Phys* 52:711–713.
19. Carslaw HS, Jaeger JC (1986) *Conduction of Heat in Solids* (Oxford Univ Press, Oxford).
20. Saunders TE, Howard M (2009) Morphogen profiles can be optimized to buffer against noise. *Phys Rev E Stat Nonlin Soft Matter Phys* 80:041902.
21. 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.
22. Maeder CI, et al. (2007) Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. *Nat Cell Biol* 9:1319–1326.
23. Jorgensen P, et al. (2007) The size of the nucleus increases as yeast cells grow. *Mol Biol Cell* 18:3523–3532.
24. Elowitz MB, Surette MG, Wolf PE, Stock JB, Leibler S (1999) Protein mobility in the cytoplasm of Escherichia coli. *J Bacteriol* 181:197–203.
25. Grossman N, Ron EZ, Woldringh CL (1982) Changes in cell dimensions during amino acid starvation of Escherichia coli. *J Bacteriol* 152:35–41.
26. Zhou HX, Rivas G, Minton AP (2008) Macromolecular crowding and confinement: Biochemical, biophysical, and potential physiological consequences. *Annu Rev Biophys* 37:375–397.
27. English BP, et al. (2011) Single-molecule investigations of the stringent response machinery in living bacterial cells. *Proc Natl Acad Sci USA* 108:E365–E373.
28. Golding I, Cox EC (2006) Physical nature of bacterial cytoplasm. *Phys Rev Lett* 96:098102.
29. Bassell G, Singer RH (1997) mRNA and cytoskeletal filaments. *Curr Opin Cell Biol* 9:109–115.
30. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet* 6:451–464.
31. Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036.
32. Sigal A, et al. (2006) Variability and memory of protein levels in human cells. *Nature* 444:643–646.
33. Taniguchi Y, et al. (2010) Quantifying E coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533–538.