Intrinsic and extrinsic contributions to stochasticity in gene expression

Peter S. Swain*†‡, Michael B. Elowitz*§, and Eric D. Siggia*

*Center for Studies in Physics and Biology and §Laboratory for Cancer Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021

Edited by Robert H. Austin, Princeton University, Princeton, NJ, and approved June 17, 2002 (received for review January 23, 2002)

Gene expression is a stochastic, or "noisy," process. This noise comes about in two ways. The inherent stochasticity of biochemical processes such as transcription and translation generates "intrinsic" noise. In addition, fluctuations in the amounts or states of other cellular components lead indirectly to variation in the expression of a particular gene and thus represent "extrinsic" noise. Here, we show how the total variation in the level of expression of a given gene can be decomposed into its intrinsic and extrinsic components. We demonstrate theoretically that simultaneous measurement of two identical genes per cell enables discrimination of these two types of noise. Analytic expressions for intrinsic noise are given for a model that involves all the major steps in transcription and translation. These expressions give the sensitivity to various parameters, quantify the deviation from Poisson statistics, and provide a way of fitting experiment. Transcription dominates the intrinsic noise when the average number of proteins made per mRNA transcript is greater than ≃2. Below this number, translational effects also become important. Gene replication and cell division, included in the model, cause protein numbers to tend to a limit cycle. We calculate a general form for the extrinsic noise and illustrate it with the particular case of a single fluctuating extrinsic variable—a repressor protein, which acts on the gene of interest. All results are confirmed by stochastic simulation using plausible parameters for *Escherichia coli.*

Molecules are discrete entities. When present in large numbers, addition or removal of any single molecule typically has little effect on the properties of a system. However, stochastic fluctuations can become significant in smaller systems. In living cells, many components are present at very low copy numbers, [e.g., of order one for DNA loci and of order tens for transcription factors (1)]. Therefore, stochastic effects are thought to be particularly important for gene expression and have been invoked to explain cell–cell variations in clonal populations (2–4). Indeed, cellular components interact with one another in complex regulatory networks. Thus, fluctuations in even a single component may potentially affect the performance of the entire system.

Consider a particular gene of interest. The amount of protein it produces will vary from cell to cell in a population and over time in a single cell. These fluctuations originate in two ways: First, even if all cells were in precisely the same state, the reaction events leading to transcription and translation of the gene would still occur at different times, and in different orders, in different cells. Such stochastic effects are set locally by the gene sequence and the properties of the protein it encodes and will be referred to as "intrinsic" noise.

In addition, one must consider that other molecular species in the cell, e.g., RNA polymerase (RNAP), are themselves gene products and therefore will also vary over time and from cell to cell. This variation causes additional, and corresponding, fluctuations in the expression of the gene of interest that will be referred to as "extrinsic" noise. Thus, extrinsic sources of noise arise independently of the gene but act on it. Examples of extrinsic variables are numerous. They include the number of RNAPs or ribosomes, the stage in the cell cycle, the quantity of the protein, and mRNA degradation machinery, and the cell environment. In general, the total variation in gene expression will have both intrinsic and extrinsic sources. A particular cellular component will suffer intrinsic fluctuations in its own concentration and, at the same time, will be a source of extrinsic noise for other components with which it interacts.

Although the stochastic nature of gene expression has long been postulated (2), previous theoretical research (5–11) has concentrated on intrinsic noise. Excepting studies of plasmid copy number control (12), extrinsic effects have only been added in a post hoc manner (13). It is not known which molecular properties influence noise or even how a clear measurement of intrinsic noise could be obtained *in vivo*.

This paper seeks to address several problems. First, we distinguish between intrinsic and extrinsic sources of noise and integrate both within a single framework. Second, we model intrinsic noise at a level that allows direct connection with biochemical parameters, including those related to cell growth. Third, we suggest an experimental method that can be used to discriminate and quantify the two components of noise in living cells. Our approach, by integration of intrinsic and extrinsic effects, is general enough to allow comparison with experimental data (14).

Definitions

Fluctuations in the rates of transcription and translation of a particular gene result in corresponding fluctuations in the amount of its protein product. A natural and biologically relevant measure of the magnitude of gene expression noise is thus the size of protein fluctuations compared to their mean concentration. If $P(t)$ is the protein concentration at time t, then the protein noise, $\eta(t)$, is given by

$$\eta^2(t) = \frac{\langle P(t)^2\rangle - \langle P(t)\rangle^2}{\langle P(t)\rangle^2}, \qquad [1]$$

where the angled brackets denote an average over the probability distribution of P at time t. We will similarly use standard deviation divided by mean, or coefficient of variation, as a measure of noise in other distributions.

To examine the noise for a particular gene across a cell population, let the intrinsic and extrinsic variables (including time—cells are typically desynchronized) for that gene be given by vectors **I** and **E**, each of whose components represent a different source of noise. The expression level of the gene in one cell, as measured experimentally, is denoted P_k (with k a cell label). From a snapshot of N genetically identical cells, the P_ks can be averaged to find the moments of the protein distribution. This averaging process (where $m = 1$ and $m = 2$ for the mean and variance, respectively) is equivalent to

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: RNAP, RNA polymerase.

†To whom reprint requests should be addressed. E-mail: swain@cnd.mcgill.ca.

‡Present address: Department of Physiology, McGill University, 3655 Promenade Sir William Osler, Montréal, QC, Canada H3G 1Y6.

BIOPHYSICS

$$\frac{1}{N}\sum_{k=1}^{N} P_k^m \approx \int d\mathbf{E}\, d\mathbf{I}\, P^m(\mathbf{E}, \mathbf{I}) p(\mathbf{E}\mathbf{I}). \quad [2]$$

Here $p(\mathbf{EI})$ is the probability density function for the intrinsic and extrinsic variables, and $P(\mathbf{E}, \mathbf{I})$ is the measured expression level for particular values of $\mathbf{E}$ and $\mathbf{I}$. Using the product rule of probabilities, this becomes

$$\frac{1}{N}\sum_{k=1}^{N} P_k^m \approx \int d\mathbf{E}\, p(\mathbf{E}) \int d\mathbf{I}\, P^m(\mathbf{E}, \mathbf{I}) p(\mathbf{I}|\mathbf{E}). \quad [3]$$

The second integral is an average over the intrinsic variables with the extrinsic variables held fixed and shall be denoted by angled brackets:

$$\langle P^m(\mathbf{E})\rangle \equiv \int d\mathbf{I}\, P^m(\mathbf{E}, \mathbf{I}) p(\mathbf{I}|\mathbf{E}). \quad [4]$$

Averages over the extrinsic variables will be indicated with an overbar, so that Eq. **3** becomes

$$\frac{1}{N}\sum_{k=1}^{N} P_k^m = \overline{\langle P^m\rangle}, \quad [5]$$

that is, an average over both intrinsic and extrinsic noise sources.

Hence, the measured noise, η_{tot}, defined empirically by

$$\eta_{\text{tot}}^2 = \frac{\frac{1}{N}\sum_k P_k^2 - \left(\frac{1}{N}\sum_k P_k\right)^2}{\left(\frac{1}{N}\sum_k P_k\right)^2}, \quad [6]$$

is equivalent to

$$\eta_{\text{tot}}^2 = \frac{\overline{\langle P^2\rangle} - \left(\overline{\langle P\rangle}\right)^2}{\left(\overline{\langle P\rangle}\right)^2}. \quad [7]$$

This can be written as

$$\eta_{\text{tot}}^2 = \frac{\overline{\langle P^2\rangle - \langle P\rangle^2}}{\left(\overline{\langle P\rangle}\right)^2} + \frac{\overline{\langle P\rangle^2} - \left(\overline{\langle P\rangle}\right)^2}{\left(\overline{\langle P\rangle}\right)^2} \equiv \eta_{\text{int}}^2 + \eta_{\text{ext}}^2. \quad [8]$$

In other words, the square of the experimentally measurable noise is a direct sum of the intrinsic, η_{int}, and extrinsic, η_{ext}, contributions. The intrinsic noise, η_{int}, is proportional to the variance of the intrinsic distribution, calculated for a particular value of the extrinsic variables and then averaged over all possible values of these variables. The extrinsic noise, η_{ext}, vanishes as the extrinsic distributions become more and more spiked.

Finally, we need to address experimentally how both intrinsic and extrinsic contributions can be discriminated from the total noise, given by Eq. **6**. From Eq. **8**, this requires a measurement of the quantity $\overline{\langle P\rangle^2}$. Consider what would happen if two identical copies of the gene were present in the same (kth) cell, and their protein products, labeled $P_k^{(1)}$ and $P_k^{(2)}$, were measured simultaneously. These will have different values of the intrinsic variables, but, because both are present in a single cell, they will be exposed to the same intracellular environment and so have the same value of the extrinsic variables. Therefore, by summing their product, we obtain

$$\frac{1}{N}\sum_{k=1}^{N} P_k^{(1)} P_k^{(2)} \approx \int d\mathbf{E}\, d\mathbf{I}_1\, d\mathbf{I}_2\, P(\mathbf{E}, \mathbf{I}_1) P(\mathbf{E}, \mathbf{I}_2) p(\mathbf{E}\mathbf{I}_1\mathbf{I}_2) = \int d\mathbf{E}\, p(\mathbf{E}) \left[\int d\mathbf{I}\, P(\mathbf{E}, \mathbf{I}) p(\mathbf{I}|\mathbf{E})\right]^2 = \overline{\langle P\rangle^2}, \quad [9]$$

precisely the average needed.

Experimentally, two distinguishable variants of green fluorescent protein, corresponding to $P^{(1)}$ and $P^{(2)}$, would allow estimation of Eq. **9** (14). By considering quantities such as $\Sigma_k (P_k^{(1)} - P_k^{(2)})^2$, the intrinsic noise could be measured and η_{ext} extracted from the total noise by using Eq. **8**.

Expressions for Intrinsic Noise

To understand the sources of intrinsic fluctuations, consider the simplified model of gene expression shown in Fig. 1. All extrinsic variables (excepting time) are set to constant values; thus, the binding of RNAP, ribosomes, and degradosomes, for example, become first-order processes as their respective concentrations are held fixed. We model the bacterial cell cycle by allowing the gene copy number, n, to double at some (fixed) time t_d into each cycle and to halve at cell division (time T). Non-DNA molecular species are randomly distributed at cell division between the two daughter cells.

Fig. 1 *Inset* shows a simplified version of the model that can be solved analytically. Two effective rate constants, marked with primes, have been introduced and can be related to those of the full system. The mRNA half-life is given by the set of differential equations describing $\langle mR^U\rangle$ and $\langle mC^2\rangle$: $d_0' \simeq \log 2 \times (\ell_1 - \ell_0)/2$

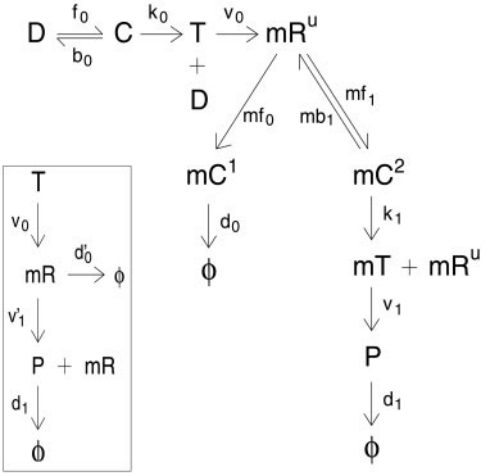

Fig. 1. Reaction scheme detailing constitutive expression of a protein *P*. All molecular species shown are intrinsic variables. Transcription is modeled (15) as reversible binding of RNAP to promoter, D (rates f_0 and b_0). Isomerization from closed to open complex and initiation of transcription are approximated as a first-order process (rate k_0). Only the leader region of the mRNA, mR^U, is followed. It is made by transcribing polymerase, T, at rate ν_0. mRNA is degraded by the binding of the degradosome (rate mf_0) to form complex mC^1, which decays in a first-order manner. Following ref. 6, ribosomes compete with degradosomes for the leader region of the mRNA and bind reversibly (rates mf_1 and mb_1). Start of translation is from the mC^2 state with rate k_1, which frees mR^U for further binding. Protein is translated (rate ν_1) in the mT state and decays with rate d_1. *Inset* shows a simplified model of translation, with mR now designating an entire mRNA molecule, degrading at rate d_0', and is translated with rate ν_1'.

with $\ell_1 = k_1 + mb_1 + mf_0 + mf_1$ and $\ell_0^2 = \ell_1^2 - 4\,mf_0(k_1 + mb_1)$. The number of proteins made from a particular mRNA is distributed geometrically (similar to ref. 6), and so the mean number of proteins produced per transcript, b, can be shown to be

$$b = \frac{1}{mf_0} \cdot \frac{k_1 mf_1}{k_1 + mb_1} \qquad [10]$$

and the overall translation rate is $\nu_1' = bd_0'$, by definition.

Simulation Results

Stochastic simulation (16, 17) was used to model the scheme of Fig. 1 by using parameter values published as supporting information on the PNAS web site, www.pnas.org. The probability of a given reaction occurring is equal to the product of the rate constant for that reaction and the number of potential reactants present. Time steps between reactions obey a Markov process and take account, for binary reactions, of the growing cell volume (17). The latter increases linearly (18) from an initial value until cell division (at time T), when it halves. Gene doubling and binomial partitioning of non-DNA molecules are included, and one daughter cell is followed at each division.

After a sufficient number of divisions, the protein number and concentration tend not to a steady-state but rather to the limit cycle (whose period is set by cell division), shown in Fig. 2. The slight kink in the protein number curve is due to the increased rate of protein production as the number of genes doubles (chosen arbitrarily to be at time $t_d = 0.4T$ into the cell cycle). The protein concentration is approximately the same before and after cell division once the limit cycle state has been reached. It falls initially (before gene replication) as protein is produced at a rate slower than that of cell growth.

The time scales associated with transcription and mRNA degradation are much shorter than the protein decay rate or the cell cycle time (see supporting information). Therefore, mRNA levels alternate between two approximately steady states, with a short transient in between. Despite the noise, this effect can be discerned in Fig. 2 (compare numbers for $t < t_d$ with those for $t > t_d$).

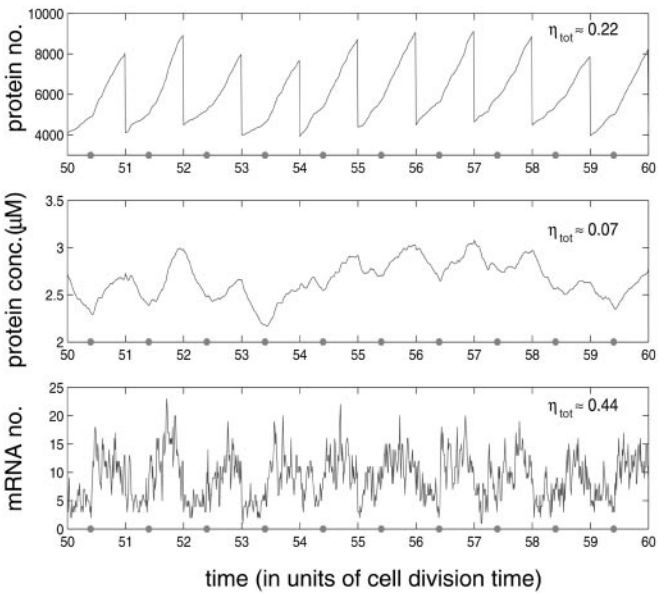

Fig. 2. Simulation results for protein and mRNA number using the model of Fig. 1 *Inset*. A strong promoter is used ($k_0 = 0.1\,s^{-1}$), and there is just one copy of the gene on the chromosome. On average, 15 proteins are synthesized per mRNA transcript, and the mRNA half-life is 1 min. Gene replication occurs every $t_d = 0.4T$ into the cell cycle and is marked with a small dot on the time axis. All other parameters are given in the supporting information. The total noise η_{tot} is defined in Eqs. **6** and **7**, with the overbar, in this simple case, just denoting a time (cell cycle) average. This is given for each species in the upper right-hand corner.

Analytical Results

Assuming that the molecules involved in transcription are in one of two steady states (depending on the gene copy number n), and that all other time dependence is absorbed into the protein distribution, then the simpler model, shown in Fig. 1 *Inset*, can be substituted for the full scheme and solved analytically (see supporting information).

The mean mRNA number satisfies

$$\langle mR \rangle \simeq \frac{f_0 k_0 n}{d_0' \ell} \qquad [11]$$

before replication, $t < t_d$ and twice this result for $t > t_d$. Here, $\ell = f_0 + b_0 + k_0$.

The mean protein number, which changes with time, obeys

$$\langle P(t) \rangle = \frac{\nu_1'}{d_1} \langle mR \rangle \phi_0(t) \qquad [12]$$

with ϕ_0 a continuous function of t,

$$\phi_0(t) = \begin{cases} 1 - \dfrac{e^{-d_1(T - t_d + t)}}{2 - e^{-d_1 T}} & \text{for } 0 \le t \le t_d \\ 2\left[1 - \dfrac{e^{-d_1(t - t_d)}}{2 - e^{-d_1 T}}\right] & \text{for } t_d \le t \le T \end{cases}. \qquad [13]$$

Note that the factor of two arising from gene replication is absorbed into the function ϕ_0, and so $\langle mR \rangle$ in Eq. **12** is given by Eq. **11** regardless of t being greater or less than t_d.

Eqs. **12** and **13** can be understood as being the solution of

$$\frac{d\langle P \rangle}{dt} = \begin{cases} \nu_1' \langle mR \rangle - d_1 \langle P \rangle & \text{for } 0 \le t < t_d \\ 2\nu_1' \langle mR \rangle - d_1 \langle P \rangle & \text{for } t_d \le t \le T \end{cases} \qquad [14]$$

with continuity at $t = t_d$ and the limit cycle boundary condition

$$\langle P(T) \rangle = 2\langle P(0) \rangle, \qquad [15]$$

which arises because of the partition at each cell division (see Fig. 2 *Upper*), that is, a simple birth-and-death process with the birth rate doubling after gene replication.

The noise in mRNA number is

$$\eta_{mR}^2 = \frac{\langle mR^2 \rangle - \langle mR \rangle^2}{\langle mR \rangle^2} = \frac{1}{\langle mR \rangle} - \frac{d_0' \nu_0 (d_0' + \ell + \nu_0)}{n(d_0' + \ell)(\ell + \nu_0)(d_0' + \nu_0)}. \qquad [16]$$

Eq. **16** is less than the Poisson value, $1/\langle mR \rangle$, because conservation of DNA species limits the maximum amount of C that can be present. This in turn gives an upper bound to the rate at which T is created, narrowing the distribution of T. A narrower T distribution leads to a narrower mRNA distribution and so to the negative term in Eq. **16**. Typically, however, this correction is rarely large.

The intrinsic noise in protein number (denoted $\hat{\eta}_{int}$, rather than η_{int}, as the extrinsic variables have not yet been averaged away) satisfies in the limit small d_1/d_0' (see supporting information for the full expression),

$$\hat{\eta}_{int}^2(t) = \frac{1}{\langle P(t) \rangle} + \frac{1}{\langle mR \rangle}\left(1 - \frac{f_0 k_0}{\ell^2}\right)\frac{d_1}{d_0'}\,\Phi_1(t) \qquad [17]$$

with

$$\Phi_1(t) = \frac{2 - e^{-d_1 T}}{2 + e^{-d_1 T}} \times \begin{cases} \dfrac{4 - e^{-2d_1 T} - 2e^{-2d_1 t} - e^{-2d_1(T+t-t_d)}}{[2 - e^{-d_1 T} - e^{-d_1(T+t-t_d)}]^2} & \text{for } 0 \le t \le t_d \\ \dfrac{4 - e^{-2d_1 T} - e^{-2d_1 t} - 2e^{-2d_1(t-t_d)}}{2[2 - e^{-d_1 T} - e^{-d_1(t-t_d)}]^2} & \text{for } t_d \le t \le T \end{cases} \quad [18]$$

for each traversal of the limit cycle. Note that $1/\langle P \rangle$ is also of order d_1/d_0' from Eq. **12** as $\nu_1' = bd_0'$. Eq. **17** contains a Poisson term, the mean $\langle P(t) \rangle$, and a non-Poisson term, which is a measure of the stochastic effects present in transcription. The limit $d_1/d_0' \ll 1$ taken is expected to be valid for many genes in *E. coli*, because protein lifetimes are typically hours, whereas those of mRNA are only minutes (see supporting information).

Fig. 3 shows the good agreement between theory and simulation. Because of the difference in gene copy number, the number of mRNAs for $t > t_d$ is approximately twice that for $t < t_d$ (see Eq. **11**). If cell cycle effects are temporarily set aside, the protein noise, dominated by mRNA number, will therefore be higher for a gene copy number of n than for a gene copy number of $2n$. However, because of the cell cycle, immediately after cell division the protein noise is low (being still determined by the previous $2n$ gene state) and will grow for $0 < t < t_d$ as it tends toward the higher value set by the cell being in a n gene state. Immediately after gene replication, the noise is high (from the cell having just left the n gene regime) and so for $t_d < t < T$ will fall as it tends toward the lower value prescribed by the cell's $2n$ state. Consequently, the intrinsic protein noise goes through a maximum at $t = t_d$.

The intrinsic noise, via Eq. **12** and using $\nu_1' = bd_0'$, can also be written as

$$\hat{\eta}^2_{\text{int}}(t) = \frac{d_1}{d_0'} \frac{1}{\langle mR \rangle} \left[\frac{1}{b} \Phi_0(t) + \left(1 - \frac{f_0 k_0}{\ell^2} \right) \Phi_1(t) \right], \quad [19]$$

with $\Phi_0(t) = 1/\phi_0(t)$ given in Eq. **13**. The parameter dependence and form of Eq. **19** can in fact be shown to hold for the full model of Fig. 1 when $d_1/d_0' \ll 1$. By inspecting Eq. **19**, only the first term depends on the parameters controlling translation rates. Therefore, if the second term dominates the first, transcription rather than translation will determine protein intrinsic noise. Expanding in $d_1 T$, this condition is fulfilled if

$$b \equiv \frac{1}{mf_0} \cdot \frac{k_1 mf_1}{k_1 + mb_1} \gg \frac{3}{2\left(1 - \dfrac{f_0 k_0}{\ell^2} \right)} \quad [20]$$

for $d_1 T \ll 1$. As $1 - f_0 k_0/\ell^2 \ge 3/4$, Eq. **20** will certainly be satisfied, and so transcription will dominate over translation if the number of proteins per transcript (or burst size in the terminology of refs. 6 and 10) is much greater than two, i.e., $b \gg 2$. In fact, the number of proteins per transcript may be of order tens (19) (although individual translation rates vary widely). In such cases, we conclude that transcription is the chief source of intrinsic noise. From Eq. **20**, noise at translation becomes important only if mf_1, the rate of ribosome binding, and k_1, the rate of commitment of a bound ribosome to carrying out translation, are low and if mf_0, the rate of mRNA degradation, is high.

Previous work claimed that translation controls protein noise (10). This conclusion was reached by using an alternative noise definition (variance, rather than standard deviation, over the mean), which divides out all dependence on transcriptional parameters. In contrast, Eq. **19** treats transcription and translation on an equal footing and keeps all parameter dependence explicit. (For example, if the gene copy number, n, is increased 100-fold, the noise is reduced as expected intuitively.) The conclusion that translation is a minor contributor to noise (for $b > 2$) is thus transparent. This conclusion is also confirmed by recent independent simulations of LacZ expression (7).

Eq. **19** makes explicit the dependence of the intrinsic noise on the parameters shown in Fig. 1. A high ratio of protein to mRNA lifetime (low d_1/d_0') reduces noise. Given Eq. **11**, the importance of the rates controlling the initiation of transcription can be seen; both terms in Eq. **19** decrease for high f_0, the "on" rate of the RNAP, a large isomerization rate, k_0, and to a lesser extent a low RNAP "off" rate, b_0. The square of the noise is independent of ν_0 and ν_1 and varies inversely with the gene copy number, n. Thus, for example, fast-growing bacteria, which undergo multiple rounds of initiation of DNA replication before division, are therefore expected to be intrinsically less noisy because of their higher n values (depending on gene distance from the origin of replication). The additional parameter dependence when translation also becomes important, i.e., when $b \ll 2$ and $\eta^2 \sim 1/b$, is given by Eq. **10**.

To illustrate the role of the cell cycle and the significance of the gene replication time, t_d, let us, for the moment, ignore all other effects and assume that time is the only extrinsic variable. Then, the extrinsic average (the overbar) in Eq. **8** is just a cell cycle average and

$$\eta^2_{\text{int}} = \frac{\displaystyle\int_0^T \frac{dt}{T} [\langle P(t)^2 \rangle - \langle P(t) \rangle^2]}{\left[\displaystyle\int_0^T \frac{dt}{T} \langle P(t) \rangle \right]^2}, \quad [21]$$

for example. This experimentally accessible value is an approximation to the "true" average value of the intrinsic noise, $\int dt\, \hat{\eta}^2_{\text{int}}(t)$, which involves only one integral and not two. Table 1 demonstrates the validity of Eq. **21** as an estimate of the true noise and the excellent agreement between theory and simulation.

In the limit where $d_1 T \to 0$ (appropriate for GFP in *E. coli*), protein does not decay but is diluted only because of partition at each cell division. Eq. **21** is then simple to evaluate and satisfies (with $\tau_d = t_d/T$)

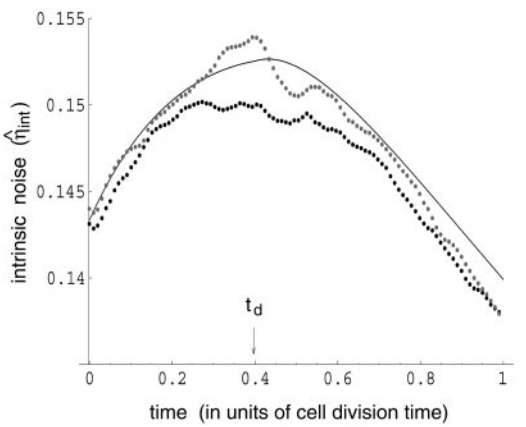

Fig. 3. Comparison of analytic solution and stochastic simulation. The noise in protein level (averaged over 5,000 runs) for the three different models is plotted as a function of time (in units of the cell cycle). Upper light dotted curve is the result of a simulation of the full model of protein expression, Fig. 1, with $k_0 = 0.01\ s^{-1}$. Dotted curve is a simulation of Fig. 1 *Inset*, whereas the full curve is a plot of Eq. **17**. At the beginning of the cell cycle, $\hat{\eta}_{\text{int}}$ is slightly greater than that at the end because of the random partitioning of proteins and mRNA into daughter cells on division.

$$\eta_{\rm int}^2 = \frac{1}{Td_0'\langle mR\rangle}\left\{\frac{2}{b(\tau_d^2-4\tau_d+6)}+\left(1-\frac{f_0k_0}{\ell^2}\right)\frac{4(3\tau_d^2-8\tau_d+10)}{3(\tau_d^2-4\tau_d+6)^2}\right\} \quad [22]$$

Both terms in Eq. **22** can be seen to increase monotonically with $\tau_d = t_d/T$; a low t_d implies that the cell spends most time in the high gene copy number state, $2n$, and so protein and mRNA levels increase, reducing noise. As t_d/T increases, Eq. **22** varies by a factor of around 0.35 to 0.7. Thus, a steady-state approximation, in which time dependence is ignored ($\Phi_0 = \Phi_1 = 1$), could overestimate intrinsic noise by as much as 65%.

Expressions for Extrinsic Noise

Time is, of course, not the only extrinsic variable; the sources of extrinsic noise are multiple, ranging from fluctuations in the bacterial growth rate to changes in the degree of DNA supercoiling and are often poorly understood. The simplest way to model these effects is to let each extrinsic variable, E_k, have mean μ_k and standard deviation, σ_k, and an independent, normal distribution. Using asymptotic expansion methods (20), performing an extrinsic average of a function $f(\mathbf{E})$ then gives

$$\int d\mathbf{E}\, p(\mathbf{E})f(\mathbf{E}) \simeq f(\mu) + \frac{1}{2}\sum_k \eta_k^2\partial_k^2 f|_\mu \quad [23]$$

in the limit of small extrinsic noise, $\eta_k = \sigma_k/\mu_k$. Here we write ∂_k for a partial derivative with respect to the variable $x_k = E_k/\mu_k$.

The extrinsic average arises in Eq. **8** as, experimentally, averages are taken over a population of cells, each cell having a different set of values of the extrinsic variables. The theoretical intrinsic noise, Eq. **19**, is calculated with all extrinsic variables fixed and therefore needs a correction term before it can be properly compared with experiment. Rather than also average over time explicitly, we will, for the sake of clarity, treat it just as a parameter in this section. The intrinsic noise, $\eta_{\rm int}^2$, then satisfies, via Eq. **23**,

$$\eta_{\rm int}^2 = \frac{\overline{\langle P^2\rangle - \langle P\rangle^2}}{\left(\overline{\langle P\rangle}\right)^2} \simeq \hat{\eta}_{\rm int}^2\left[1+\sum_k\frac{1}{2}\eta_k^2\left(\frac{\partial_k^2(\langle P^2\rangle-\langle P\rangle^2)}{\langle P^2\rangle-\langle P\rangle^2}-2\frac{\partial_k^2\langle P\rangle}{\langle P\rangle}\right)\right], \quad [24]$$

with $\hat{\eta}_{\rm int}$ given by Eq. **17** and all the extrinsic variables set to their mean values, μ_k. Experimentally, it is possible to continuously vary the rate of transcription from certain inducible promoters (14, 21). Eq. **24** implies that there is a correction to the more naive expectation, from Eq. **19**, that the intrinsic noise should vary as the inverse square of induction level (assuming that the latter is proportional to the number of mRNAs).

The extrinsic noise, $\eta_{\rm ext}$, obeys

$$\eta_{\rm ext}^2 = \frac{\overline{\langle P\rangle^2} - \left(\overline{\langle P\rangle}\right)^2}{\left(\overline{\langle P\rangle}\right)^2} \simeq \sum_k \chi_k^2\eta_k^2, \quad [25]$$

with all extrinsic variables set to their mean and $\chi_k = (\partial_k \log\langle P\rangle)^2$ defined, in analogy with statistical physics, as a noise "susceptibility." The individual χ_k measure the contribution of a particular extrinsic process to the total noise strength.

Table 1. Comparison of theory and simulation for a constitutively expressed gene

	Protein nos.		Protein concentration	
η	Simulation	Theory	Simulation	Theory
$\eta_{\rm tot}$	0.26 ± 0.003	0.26	0.15 ± 0.005	0.15
$\eta_{\rm int}$	0.15 ± 0.007	0.15	0.15 + 0.007	0.15
$\eta_{\rm ext}$	0.21 ± 0.004	0.21	0.05 ± 0.01	0.04

Time is the only extrinsic value, and there is consequently almost no extrinsic noise in the protein concentration (because this varies little during the cell cycle; see Fig. 2). The intrinsic noise in both cases, calculated by Eq. **21** (with an appropriate expression for cell volume when needed), is a very good approximation to $\int dt\,\hat{\eta}_{\rm int}^2$, which is found to be 0.15 ± 0.008 from simulation and 0.15 from integrating Eq. **17**. Parameter values are published as supporting information on the PNAS web site, except $k_0 = 0.01\ s^{-1}$. Values stated are mean results from 100 simulations and errors are ±1 SD.

An Example: Repression

To illustrate the effects of a fluctuating extrinsic variable, we consider a gene of interest that is repressed by another extrinsic protein. The repressor has a noise given by Eq. **17** with the appropriate parameter values for its own expression. To find the intrinsic noise, from Eq. **9**, we assume that two identical copies of our gene are present, both acted on by the same repressor.

Repression is modeled by the repressor, R, binding to the promoter and preventing access to it by RNAP (ref. 22). This repressor–DNA complex forms and decomposes with rates f_1 and b_1, respectively, implying that the mean mRNA number of the (repressed) protein obeys

$$\langle mR\rangle = \frac{f_0k_0n}{d_0'(\ell+\beta K)}, \quad [26]$$

where $\beta = b_0 + k_0$ and $K = f_1/b_1$.

In Eq. **26**, to make the extrinsic repressor concentration explicit, the "on" rate (which is really a second-order process) should be written as $f_1 = \hat{f}_1\mu_{\rm rep}$, where $\hat{f}_1$ is the binding rate of a single repressor to DNA, and $\mu_{\rm rep}$, the mean repressor number. Applying the definition of χ_k to Eq. **12**, with $\langle mR\rangle$ given by Eq. **26**, the repressor noise susceptibility can be found, and Eq. **8** becomes

$$\eta_{\rm tot}^2 \simeq \hat{\eta}_{\rm int}^2 + \frac{K^2\beta^2}{(\ell+\beta K)^2}\,\eta_{\rm rep}^2, \quad [27]$$

where $\hat{\eta}_{\rm int}^2$ is given by

$$\hat{\eta}_{\rm int}^2(t) = \frac{1}{\langle P(t)\rangle} + \frac{1}{\langle mR\rangle}\left(1-\frac{f_0k_0}{b_1}\cdot\frac{b_1(1+K)-\beta K}{(\ell+\beta K)^2}\right)\frac{d_1}{d_0'}\,\Phi_1(t), \quad [28]$$

with Eq. **26** again.

The total, intrinsic, and extrinsic noises found by simulation are compared with the corresponding theoretical values (with time again included properly as an extrinsic variable) in Table 2. The good agreement, as well as validating Eqs. **27** and **28**, also demonstrates the suitability of the Gaussian approximation implicit in Eq. **23** as the repressor is expressed in the simulation using the full scheme of Fig. 1 and not added in an ad hoc manner.

The repressor noise may dominate the extrinsic noise to such an extent that Eq. **27** is still valid when all other extrinsic variables fluctuate. For example, in ref. 14, the transcription rate of two distinguishable alleles of *gfp*, both having the same regulatory sequences, is controlled by a repressor. When the repressor concentration is systematically varied from low to high values (by adding different amounts of inducer), the extrinsic noise goes

Table 2. Comparison of theory and simulation for a gene acted on by a repressor

η	Simulation	Theory
η_{tot}	0.51 ± 0.02	0.51
η_{int}	0.45 ± 0.02	0.44
η_{ext}	0.24 ± 0.03	0.26

There are two extrinsic variables: time and the repressor. Noise is calculated for protein numbers, not concentrations. To find the intrinsic noise, two copies of the (repressed) gene were simulated (see text). All proteins (including the repressor) were created according to the full scheme of Fig. 1 (with, for simplicity, the same rate constants, although different t_ds). Parameter values are given in the supporting information, except $k_0 = 0.01\,s^{-1}$, $\hat{f}_1 = 5 \times 10^7\,M^{-1}{\cdot}s^{-1}$, $b_1 = 0.33\,s^{-1}$, and $t_d = 0.7T$ for the repressor gene. η_{rep} is calculated to be 0.17 by integrating Eq. **17** over one cell cycle. Because of the repressor, expression is reduced to about 10% of its constitutive level (the expression level of Table 1). Values given are mean results from 100 simulations, and errors are ±1 SD.

through a maximum (14). This phenomenon can be understood by realizing that the extrinsic noise is dominated by the repressor. Although χ_{rep} increases with increasing repressor concentration, η_{rep}, the noise in repressor number, decreases. As a result of these two opposing behaviors, η_{ext} (and therefore η_{tot}) exhibits a maximum as a function of repressor concentration. This behavior illustrates clearly the importance of noise susceptibilities in setting cell–cell variation.

Conclusion

We have presented a theoretical framework that enables interpretation of experimental measurements of stochasticity in gene expression. Cell–cell variation in expression of a single gene (η_{tot}) is not a direct measure of intrinsic noise. Rather, it contains both intrinsic and extrinsic contributions. In particular, extrinsic noise, a consequence of the different local environments of the gene in the different cells, must be considered.

Only the intrinsic variables (given in Fig. 1) vary from gene to gene, as well as moment to moment, within a particular cell. By changing the parameters that influence these variables, the cell can locally affect the noise in expression of a given gene. On the other hand, alterations in the extrinsic variables can potentially affect all genes within the cell (although the magnitude of these effects for one gene may be very different from those for another). Eqs. **17** and **19** are analytical expressions for the major component of the intrinsic noise. There is a Poisson term, expected for a birth-and-death process, determined by the protein mean, and an additional contribution coming from the noise generated during transcription (essentially a time average of the noise in mRNA level). Two noise regimes exist: if the translation efficiency, or burst size (6), b, is high (more than two proteins per transcript), as is believed to be typical in *E. coli*, then transcription dominates intrinsic noise. Otherwise, translational effects must also be considered.

All the major steps in transcription and translation are accounted for, and the complete parameter dependence of the noise is given by Eq. **19** with Eq. **10**. Intrinsic noise (except possibly for very short lived proteins) is unaffected by ν_0 and ν_1, the rates of transcription by RNAP and translation by a ribosome, respectively. As transcription usually dominates, f_0 and b_0, the "on" and "off" rates of RNAP as well as the isomerization rate, k_0, strongly influence noise. Longer-lived proteins (compared to mRNA lifetimes) and genes with high copy number are less stochastic. The chromosomal position of the gene also controls intrinsic noise—genes replicated early being less noisy.

The cell cycle drives protein numbers and intrinsic noise to a limit cycle. Protein numbers can be significantly different from the steady-state approximations used in the literature. The intrinsic noise itself does not change appreciably during the course of the cell cycle, but the cell cycle is crucial in determining its absolute magnitude.

The extrinsic noise is expected to be a linear sum of the noise in each of the extrinsic variables (see Eq. **25**), where the coefficients play the role of noise "susceptibilities." These susceptibilities determine the relative importance of each term in the total extrinsic noise and allow exploration of how the environment in which a gene is expressed influences its expression level. By simulating a repressed gene, where the repressor number is the only fluctuating extrinsic variable, we have verified our analytical expressions are quantitatively correct. Experimentally, the intrinsic and extrinsic noise can often be of similar magnitude (14). For a given gene, however, the quantity of interest is usually the intrinsic noise, which we have shown here can be measured by monitoring expression from two identical copies of the same gene integrated into each cell (see also ref. 14).

Our theoretical framework should provide support for experimental research (14) aimed at discovering whether noise is detrimental to the cell, whether it can be "regulated away" with higher-level circuitry (23), and to what extent it might confer evolutionary advantages on a clonal population.

We are grateful for conversations with S. Bekiranov, A. J. Levine, J. Paulsson, N. Rajewsky, B. Shraiman, N. Socci, and M. Zapotocky. P.S.S., M.B.E., and E.D.S. acknowledge support from the National Institutes of Health (GM59018), the Seaver Institute and Burroughs-Wellcome Fund, and the National Science Foundation (DMR0129848), respectively.

1. Guptasarma, P. (1995) *BioEssays* **17,** 987–997.
2. Spudich, J. L. & Koshland, D. E., Jr. (1976) *Nature (London)* **262,** 467–471.
3. McAdams, H. H. & Arkin, A. (1999) *Trends Genet.* **15,** 65–69.
4. Arkin, A., Ross, J. & McAdams, H. H. (1998) *Genetics* **149,** 1633–1648.
5. Ko, M. S. H. (1991) *J. Theor. Biol.* **153,** 181–194.
6. McAdams, H. H. & Arkin, A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 814–819.
7. Kierzek, A. M., Zaim, J. & Zielenkiewicz, P. (2001) *J. Biol. Chem.* **276,** 8165–8172.
8. Berg, O. G. (1978) *J. Theor. Biol.* **71,** 587–603.
9. Peccoud, J. & Ycart, B. (1995) *Theor. Popul. Biol.* **48,** 222–234.
10. Thattai, M. & Van Oudenaarden, A. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8614–8619.
11. Kepler, T. B. & Elston, T. C. (2001) *Biophys. J.* **81,** 3116–3136.
12. Paulsson, J. & Ehrenberg, M. (2001) *Q. Rev. Biophys.* **34,** 1–59.
13. Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 2075–2080.
14. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. (2002) *Science* **297,** 1183–1186.
15. Record, T. M., Reznikoff, W. S., Craig, M. L., McQuade, K. L. & Schlax, P. J. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), pp. 792–821.
16. Gillespie, D. T. (1977) *J. Phys. Chem.* **81,** 2340–2361.
17. Gibson, M. A. & Bruck, J. (2000) *J. Phys. Chem.* **104,** 1876–1889.
18. Kubitschek, H. E. (1990) *J. Bacteriol.* **172,** 94–101.
19. Bremer, H. & Dennis, P. P. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), pp. 1553–1569.
20. Bender, C. M. & Orszag, S. A. (1978) *Advanced Mathematical Methods for Scientists and Engineers* (McGraw–Hill, New York).
21. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A. (2002) *Nat. Genet.* **31,** 69–73.
22. Ptashne, M. (1992) *A Genetic Switch* (Cell and Blackwell, Cambridge, MA).
23. Becskei, A. & Serrano, L. (2000) *Nature (London)* **405,** 590–593.

Derivation of Analytical Expressions

Moments of Distributions Without Cell Division. To derive an approximate analytical solution to the model shown in Fig. 1 *Inset*, we at first neglect cell cycle effects. Taking advantage of the constraint on the number of DNA molecules

$$\langle D\rangle + \langle C\rangle = n, \qquad [1]$$

where n is the copy number of the gene of interest on the chromosome, we can write down the master equation for the system as a function of four variables. Let the separate species in Fig. 1, $\{D, C, T, mR, P\}$ be labeled $\{0, 1, 2, 3, 4\}$, respectively. Furthermore, let $p(n_1, n_2, n_3, n_4, t)$ be the probability that there exist, at time t, n_1 molecules of C, n_2 molecules of T, n_3 molecules of mRNA, and n_4 protein molecules, then

$$\begin{aligned}\frac{\partial}{\partial t}p(n_1, n_2, n_3, n_4, t) &= f_0\Big[(n - n_1 + 1)p(n_1 - 1, n_2, n_3, n_4, t) \\ &\quad -(n - n_1)p(n_1, n_2, n_3, n_4, t)\Big] + \cdots, \end{aligned} \qquad [2]$$

where the dots denote similar terms, one for each rate constant. Rather than write down this long equation, we transform it straight away to an expression for the generating function

$$F(z_1, z_2, z_3, z_4, t) = \sum_{n_1,n_2,n_3,n_4} z_1^{n_1} z_2^{n_2} z_3^{n_3} z_4^{n_4} p(n_1, n_2, n_3, n_4, t), \qquad [3]$$

which can be thought of as a kind of discrete Laplace transform. Defining

$$\begin{aligned} w = z_1 - 1 \quad ; \quad x = z_2 - 1 \\ y = z_3 - 1 \quad ; \quad z = z_4 - 1 \end{aligned} \qquad [4]$$

we arrive at

$$\begin{aligned}\frac{\partial F}{\partial t} &= f_0 nwF - \Big[f_0 w(1 + w) + b_0 w - k_0(x - w)\Big]\frac{\partial F}{\partial w} + v_0(y - x)\frac{\partial F}{\partial x} \\ &\quad +\Big[v_1' z(1 + y) - d_0' y\Big]\frac{\partial F}{\partial y} - d_1 z\frac{\partial F}{\partial z}. \end{aligned} \qquad [5]$$

Clearly, finding a full solution of Eq. **5** is very difficult. However, from Eq. **3** several properties of the solution are transparent. If all the z_i are set to unity ($w = x = y = z = 0$), then normalization implies that $F =$

1. Differentiating F with respect to a z_i, and then setting all z_i to unity, gives $\langle n_i \rangle$, whereas performing the same operation after two derivatives gives $\langle n_i(n_i-1)\rangle$. Because we intend to only calculate the intrinsic noise in protein levels, that is, the variance in n_4, an expansion of F around $z_i = 1$ (for small w, x, y, and z) should be suitable [this is equivalent to the method of compounding moments (1) and is exact in our case as each moment depends only on moments of lower or equal order].

Because the protein degradation rate is much smaller than all others in Fig. 1 *Inset*,

$$d_1 \ll \{f_0, b_0, k_0, v_0, d_0', v_1'\}, \qquad \textbf{[6]}$$

we assume that all the time dependence in F comes purely from the protein terms. Levels of C, T, and mR are assumed to be at their steady-state values. In this case, we can write

$$\begin{aligned} F(w,x,t,z,t) \;\simeq\; & 1 + wX_1 + xX_2 + yX_3 + zX_4(t) + \frac{1}{2}\Big[X_{11}w^2 + X_{22}x^2 \\ & + X_{33}y^2 + X_{44}(t)z^2 + 2X_{12}wx + 2X_{13}wy + 2X_{23}xy \\ & + 2X_{14}(t)wz + 2X_{24}(t) + 2X_{34}(t)yz\Big], \end{aligned} \qquad \textbf{[7]}$$

where, for example, $X_3 = \langle mR \rangle$, $X_{11} = \langle C^2 \rangle - \langle C \rangle$, and $X_{34}(t) = \langle mR\ P \rangle$.

Putting Eq. **7** into Eq. **5** and comparing coefficients gives,

$$X_1 = \tfrac{f_0 n}{\ell} \quad ; \quad X_2 = \tfrac{f_0 k_0 n}{v_0 \ell} \quad ; \quad X_3 = \tfrac{f_0 k_0 n}{d_0' \ell} \qquad \textbf{[8]}$$

and

$$\dot{X}_4 = v_1' X_3 - d_1 X_4. \qquad \textbf{[9]}$$

Similarly, it is possible to solve for all the X_{ij}. These obey

$$\begin{aligned} d_0' X_{33} &= v_0 X_{23} \\ v_0 X_{22} &= k_0 X_{12} \\ \ell X_{11} &= f_0(n-1)X_1 \\ (\ell + v_0)X_{12} - f_0 n X_2 &= k_0 X_{11} \\ (d_0' + \ell)X_{13} - f_0 n X_3 &= v_0 X_{12} \\ (d_0' + v_0)X_{23} - v_0 X_{22} &= k_0 X_{13}, \end{aligned} \qquad \textbf{[10]}$$

solution of which leads to

$$\eta_{11}^2 = \frac{\langle C^2 \rangle - \langle C \rangle^2}{\langle C \rangle^2} = \frac{1}{\langle C \rangle} - \frac{1}{n} \qquad \textbf{[11]}$$

$$\eta_{22}^2 = \frac{\langle T^2\rangle - \langle T\rangle^2}{\langle T\rangle^2} = \frac{1}{\langle T\rangle} - \frac{v_0}{n(\ell + v_0)} \qquad [12]$$

and

$$\eta_{13}^2 = \frac{\langle C\ mR\rangle - \langle C\rangle\langle mR\rangle}{\langle C\rangle\langle mR\rangle} = -\frac{d_0' v_0}{n(d_0' + \ell)(\ell + v_0)} \qquad [13]$$

$$\eta_{23}^2 = \frac{\langle T\ mR\ \rangle - \langle T\rangle\langle mR\rangle}{\langle T\rangle\langle mR\rangle} = -\frac{d_0' v_0 (d_0' + \ell + v_0)}{n(d_0' + \ell)(\ell + v_0)(d_0' + v_0)}, \qquad [14]$$

as well as the result for the mRNA noise given in the main paper. The cross-correlation functions are negative because of the constraint, **1**, which leads to $\eta_{01}^2 = -1/n$. For example, when $n = 1$, every time C is made, D vanishes and vice versa, giving exact negative correlation. This negative correlation propagates along the chain of different species in Fig. 1 *Inset*, leading to mRNA being made in a pulse-like manner with the number of C molecules increasing, then falling, resulting in a growth in T that falls to produce mR.

For the time-dependent cross correlations, we find

$$\begin{aligned}
\dot{X}_{14} &= v_1' X_{13} + f_0 n X_4 - (d_1 + \ell) X_{14} \\
\dot{X}_{24} &= v_1' X_{23} + k_0 X_{14} - (d_1 + v_0) X_{24} \\
\dot{X}_{34} &= v_1' (X_3 + X_{33}) + v_0 X_{24} - (d_0' + d_1) X_{34} \\
\dot{X}_{44} &= 2 v_1' X_{34} - 2 d_1 X_{44}, \qquad [15]
\end{aligned}$$

where the over dots denote differentiation with respect to time. Eq. **9** can be simply integrated (remembering $X_4 = \langle P\rangle$)

$$\langle P(t)\rangle = \frac{v_1' X_3}{d_1}\left(1 - \mathrm{e}^{-d_1 t}\right) + m\,\mathrm{e}^{-d_1 t}, \qquad [16]$$

where $\langle P(0)\rangle = m$. This result is the starting point for the solution of Eq. **15**. In keeping with approximation **6**, we assume sufficient time has passed that the only exponentials that need be considered in the solution (the others are very small) are those in $d_1 t$. In this case, for example,

$$X_{14}(t) = f_0 n \left(\frac{v_1' X_3}{d_1(d_1 + \ell)} + \frac{d_1 m - v_1' X_3}{d_1 \ell}\mathrm{e}^{-d_1 t}\right) + \frac{v_1' X_{13}}{d_1 + \ell}, \qquad [17]$$

with X_{24}, X_{34}, and X_{44} being given by similar, though more complicated, expressions. Upon simplification and using the definition

$$\lambda = \frac{v_1'}{d_1}\langle mR\rangle = \frac{v_1' f_0 k_0 n}{d_0' d_1 \ell} \qquad [18]$$

the equation for X_{44} gives

$$\begin{aligned}\hat{\sigma}^2_{\rm int}(t) &= \langle P(t)^2\rangle - \langle P(t)\rangle^2 \\ &= \left(1-{\rm e}^{-d_1 t}\right)\left(m\,{\rm e}^{-d_1 t} + \lambda\left[1+\lambda\Omega\left(1+{\rm e}^{-d_1 t}\right)\right]\right), \qquad [19]\end{aligned}$$

with Ω a measure of the fluctuations in mRNA,

$$\Omega = \frac{d_1}{d_0'+d_1}\left[\eta_{33}^2 + \frac{d_0'}{d_1+v_0}\left(\eta_{23}^2 + \frac{v_0}{d_1+\ell}\eta_{13}^2\right)\right]. \qquad [20]$$

Eq. **19** gives the intrinsic variance in protein number (with all extrinsic variables held fixed) as a function of time given that at $t=0$, $\langle P(0)\rangle = m$ and $\langle P(0)^2\rangle = m^2$. Ideally, m should not be a constant but should be determined by the cell cycle. To facilitate this, let us write down a generating function for just the protein that gives Eq. **16** and Eq. **19** on expansion. Formally, this generating function is

$$Q_m(z,t) = \sum_n q_{n|m}(t) z^n, \qquad [21]$$

where $q_{n|m}(t)$ is the probability of having n proteins at time t, given that there were m proteins at time $t=0$. Expanding around $z=1$,

$$Q_m(z,t) \simeq 1 + (z-1)\langle P(t)\rangle + \frac{1}{2}(z-1)^2\Big[\langle P^2(t)\rangle - \langle P(t)\rangle\Big] + \cdots \qquad [22]$$

and so this function is determined, from Eqs. **16** and **19**, up to order $(z-1)^3$.

As

$$\begin{aligned}\langle P(t)\rangle &= \langle P_0(t)\rangle + m\,{\rm e}^{-d_1 t} \\ \langle P^2(t)\rangle &= \langle P_0^2(t)\rangle + {\rm e}^{-d_1 t} m\Big(1+2\langle P_0(t)\rangle\Big) + m(m-1){\rm e}^{-2d_1 t}, \qquad [23]\end{aligned}$$

where the subscript zero denotes evaluation at $m=0$, one can write

$$Q_m(z,t) = Q_0(z,t)\left[1-{\rm e}^{-d_1 t} + z{\rm e}^{-d_1 t}\right]^m, \qquad [24]$$

which also has the desired property, Eq. **22**. This formulation will prove very useful.

In fact, as the gene encoding protein, P, is replicated at $t=t_d$, two generating functions need to be considered, $Q_m^{(1)}(z,t)$, which is valid when

the gene copy number is n and $Q_m^{(2)}(z,t)$, which holds after replication when the copy number is $2n$. Defining

$$Y = 1 - \mathrm{e}^{-d_1 t} \qquad [25]$$

then

$$\begin{aligned} Q_m^{(i)}(z,t) &= Q_0^{(i)}(z,t)\Big[Y + z(1-Y)\Big]^m \\ &= \sum_n z^n q_{n|m}^{(i)}(t), \end{aligned} \qquad [26]$$

where $q_{n|m}^{(i)}$ is now the probability of having n proteins at time t given m at $t = 0$ in (copy number) state i.

Including Cell Division. The number of proteins in the cell will be partly controlled by the cell cycle; dilution due to partition into daughter cells at the end of cell division can play a significant role in keeping protein numbers low. To incorporate this effect into our analysis, let $P_i(n)$ be the probability of finding n proteins at the start of the ith division cycle. Then $P_{i+1}(n)$ is related to $P_i(n)$ via a transfer probability $U(n|n')$,

$$P_{i+1}(n) = \sum_{n'} U(n|n') P_i(n'). \qquad [27]$$

In our calculation, just one daughter cell is followed, and we assume that each protein has a 50% probability of being kept in this cell (and so a 50% chance of being discarded into the one not followed). Given m proteins just before cell division, the probability of having n immediately after is just binomial

$$\binom{m}{n} 2^{-m}. \qquad [28]$$

For a cell cycle of length T, the transfer probability U is given by

$$U(n|n') = \sum_{m,m'} \binom{m}{n} 2^{-m} q_{m|m'}^{(2)}(T - t_d) q_{m'|n'}^{(1)}(t_d), \qquad [29]$$

where gene replication at time t_d is included, and the definitions of the $q^{(i)}$ have been used (see end of previous section).

After many divisions, the protein number, rather than tending to a steady-state, tends to a limit cycle. Mathematically, as the limit cycle is approached, $P_i(n)$ is expected to tend to $P^*(n)$, which obeys (see Eq. **27**),

$$P^*(n) = \sum_{n'} U(n|n')P^*(n'). \qquad [30]$$

To solve Eq. **30** for P^*, we again turn to generating functions. Defining

$$F^*(z) = \sum_{n=0}^{\infty} z^n P^*(n) \qquad [31]$$

multiplying Eq. **30** by z^n and summing over all n, gives

$$\begin{aligned} F^*(z) &= \sum_{m,m',n'} \sum_{n=0}^{m} z^n \binom{m}{n} 2^{-m} q^{(2)}_{m|m'}(T-t_d) q^{(1)}_{m'|n'}(t_d) P^*(n') \\ &= \sum_{m,m',n'} \left(\frac{1+z}{2}\right)^m q^{(2)}_{m|m'}(T-t_d) q^{(1)}_{m'|n'}(t_d) P^*(n'), \end{aligned} \qquad [32]$$

where Eq. **29** has been used. From definition **26**, this can be written as

$$\begin{aligned} F^*(z) &= \sum_{m',n'} Q^{(2)}_0\Big((1+z)/2, T-t_d\Big) \Big[Y_1 + \frac{1}{2}(1+z)(1-Y_1)\Big]^{m'} \\ &\quad \times q^{(1)}_{m'|n'}(t_d) P^*(n'), \end{aligned} \qquad [33]$$

with

$$Y_1 = 1 - \mathrm{e}^{-d_1(T-t_d)}. \qquad [34]$$

The power of writing $Q^{(i)}_m(z,t)$ in the form **26** should now be apparent; it also allows the sum over m' to be evaluated similarly,

$$\begin{aligned} F^*(z) &= Q^{(2)}_0\Big((1+z)/2, T-t_d\Big) Q^{(1)}_0\Big(Y_1 + (1+z)(1-Y_1)/2, t_d\Big) \\ &\quad \times \sum_{n'} \Big[Y_2 + (Y_1 + (1+z)(1-Y_1)/2)(1-Y_2)\Big]^{n'} P^*(n'), \end{aligned} \qquad [35]$$

where Y_2 is

$$Y_2 = 1 - \mathrm{e}^{-d_1 t_d}. \qquad [36]$$

Finally, Eq. **31** allows the last summation to be carried out

$$\begin{aligned} F^*(z) &= Q^{(2)}_0\Big((1+z)/2, T-t_d\Big) Q^{(1)}_0\Big(Y_1 + (1+z)(1-Y_1)/2, t_d\Big) \\ &\quad \times F^*\Big(Y_2 + (Y_1 + (1+z)(1-Y_1)/2)(1-Y_2)\Big). \end{aligned} \qquad [37]$$

The solution of Eq. **37** will give the generating function related to $P^*(n)$, the probability of finding n proteins at the beginning of the cell cycle given that the bacterium has divided enough times to have reached a limit cycle state. Because we are only interested in calculating the variance in protein number, only the first two moments of $F^*(z)$ are required. Writing

$$F^*(z) = 1 + (z-1)f_1^* + \frac{1}{2}(z-1)^2 f_2^* + \cdots \qquad [38]$$

and then comparing coefficients of $z-1$ in Eq. **37** allows f_1^* and f_2^* to be determined: for example,

$$f_1^* = \frac{2\langle P_0(T-t_d)\rangle + (1-Y_1)\langle P_0(t_d)\rangle}{1+Y_1+Y_2-Y_1Y_2}, \qquad [39]$$

with a similar expression for f_2^*.

Because we now know the probability distribution $P^*(n)$, we can write the generating function for the protein number during the two stages of the cell cycle. Given that the protein number has reached a limit cycle state, and defining $t=0$ to be at the beginning of this cycle, i.e., immediately after cell division, then for

$0 \le t \le t_d$:

$$\begin{aligned} F^{(1)}(z,t) &= \sum_{n,m} z^n q^{(1)}_{n|m}(t) P^*(m) \\ &= \sum_m Q_0^{(1)}(z,t)\Big[Y + z(1-Y)\Big]^m P^*(m) \\ &= Q_0^{(1)}(z,t) F^*\Big(Y + z(1-Y)\Big), \end{aligned} \qquad [40]$$

with the summations evaluated by using Eq. **26** and Eq. **31** again.

$t_d \le t \le T$: In this case, because of gene replication, the expression is a little more complicated. Defining

$$Y' = 1 - \mathrm{e}^{-d_1(t-t_d)}, \qquad [41]$$

one has

$$F^{(2)}(z,t) = \sum_{n,m,m'} z^n q^{(2)}_{n|m}(t-t_d) q^{(1)}_{m|m'}(t_d) P^*(m')$$

$$
\begin{aligned}
&= \sum_{m,m'} Q_0^{(2)}(z, t-t_d)\Big[Y' + z(1-Y')\Big]^m q^{(1)}_{m|m'}(t_d) P^*(m') \\
&= Q_0^{(2)}(z, t-t_d) Q_0^{(1)}\Big(Y' + z(1-Y'), t_d\Big) \\
&\quad \times \sum_{m'} \Big[Y_2 + \Big(Y' + z(1-Y')\Big)(1-Y_2)\Big]^{m'} P^*(m') \\
&= Q_0^{(2)}(z, t-t_d) Q_0^{(1)}\Big(Y' + z(1-Y'), t_d\Big) \\
&\quad \times F^*\Big(Y_2 + \Big(Y' + z(1-Y')\Big)(1-Y_2)\Big). \qquad \textbf{[42]}
\end{aligned}
$$

Differentiation of these two generating functions with respect to z will give the mean and variance of the intrinsic protein number distribution. The protein mean is given in the main paper, and the noise satisfies

$$\hat{\eta}^2_{\rm int}(t) = \frac{1}{\langle P(t)\rangle} + \Omega \Phi_1(t), \qquad \textbf{[43]}$$

with $\Phi_1(t)$ given in the main paper and Ω by Eq. **20**. Using expressions **13**, **14** and that for the mRNA noise, Eq. **20** simplifies to

$$\Omega \simeq \frac{d_1}{d_0'}\left(1 - \frac{f_0 k_0}{\ell^2}\right) \cdot \frac{1}{\langle mR\rangle} \qquad \textbf{[44]}$$

in the limit of $d_1/d_0' \ll 1$. Eqs. **43** and **44** comprise the expression for the intrinsic protein noise, $\hat{\eta}_{\rm int}$, given in the main paper.

Parameters Used in Simulations

All parameter values are given in Table 3.

Process	Parameters
RNAP binding to DNA	Free RNAP concentration = 30 nM (2) Binding rate 1.4×10^7 $M^{-1}s^{-1}$ for λP_L (3) $\Rightarrow f_0 = 0.42$ s^{-1} $b_0 = 0.1$ s^{-1} [chosen to give an equilibrium constant of 1.4×10^8 M^{-1} (2)]
Transcription initiation rate	k_0; ranges from 0.001 s^{-1} to 0.1 s^{-1} (4) (closed to open complex isomerization)
Formation and degradation of RBS on mRNA	$v_0 = 0.3$ s^{-1} [RNAP moving at 40 nt s^{-1} (5)] $mf_0 = 0.114$ s^{-1} (chosen so that the average number of proteins per transcript = 15) $d_0 = 0.1$ s^{-1}
Binding of ribosome	Free ribosome concentration = 400 nM (order of magnitude larger than RNAP) binding rate 1×10^7 $M^{-1}s^{-1}$ $\Rightarrow$ $mf_1 = 4.0$ s^{-1} $mb_1 = 0.4$ s^{-1} [chosen to given an equilibrium constant of 2.5×10^7 M^{-1} (6)]
Translation	$k_1 = 0.3$ s^{-1} (6) $v_1 = 0.048$ s^{-1} [given a 1000 nt protein and a translation rate of 48 nt s^{-1} (7)]
Protein degradation	$d_1 = 6.42 \times 10^{-5}$ s^{-1} ($t_{\frac{1}{2}} \simeq 3$ hours)
Cell cycle time Gene replication time	$T = 60$ min (chosen for at most two chromosomes per cell) $t_d = 0.4\,T$
Cell volume and growth	Linear growth (8) $V(t) = V_0(1 + t/T)$ for $0 \leq t \leq T$ and $V_0 = 2.5 \times 10^{-15}\ell$

Table 3. Parameters suitable for constitutive gene expression in *Escherichia coli*. Abbreviations: RNA polymerase (RNAP), ribosome binding site (RBS), nucleotide (nt).

References

1. Van Kampen, N. G. (1990) *Stochastic Processes in Physics and Chemistry* (Elsevier, New York).

2. McClure, W. R. (1983) in *Biochemistry of Metabolic Processes*, eds. Lennon, D. L. F., Stratman, F. W. & Zahlten, R. N. (Elsevier, New York).

3. Lanzer, M. & Bujard, H. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8973–8977.

4. Hawley, D. K. & McClure, W. R. (1982) *J. Mol. Biol.* **157**, 493–525.

5. Manor, H., Goodman, D. & Stent, G. S. (1969) *J. Mol. Biol.* **39**, 1–29.

6. Draper, D. E. (1996) in *Escherichia coli and Salmonella : Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, D.C.).

7. Bremer, H. & Dennis, P. P. (1996) in *Escherichia coli and Salmonella : Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, D.C.).

8. Kubitschek, H. E. (1990) *J. Bacteriol.* **172**, 94–101.