Quantifying Extrinsic Noise in Gene Expression Using the Maximum Entropy Framework

Purushottam D. Dixit*
Biosciences Department, Brookhaven National Laboratory, Upton, New York

ABSTRACT We present a maximum entropy framework to separate intrinsic and extrinsic contributions to noisy gene expression solely from the profile of expression. We express the experimentally accessible probability distribution of the copy number of the gene product (mRNA or protein) by accounting for possible variations in extrinsic factors. The distribution of extrinsic factors is estimated using the maximum entropy principle. Our results show that extrinsic factors qualitatively and quantitatively affect the probability distribution of the gene product. We work out, in detail, the transcription of mRNA from a constitutively expressed promoter in *Escherichia coli*. We suggest that the variation in extrinsic factors may account for the observed wider-than-Poisson distribution of mRNA copy numbers. We successfully test our framework on a numerical simulation of a simple gene expression scheme that accounts for the variation in extrinsic factors. We also make falsifiable predictions, some of which are tested on previous experiments in *E. coli* whereas others need verification. Application of the presented framework to more complex situations is also discussed.

INTRODUCTION

Recent experiments show that the life cycle of a gene product inside the cell is stochastic. For any gene, there exists great cell-to-cell variation in the expression level of both the protein and the mRNA (1–10) and changing this variation has phenotypical and fitness effects (11–14). Recently, it was also shown that coregulated proteins have correlated variability (15). This variation arises from

1. The intrinsic statistical mechanical fluctuations in diffusion and binding of the molecules involved in gene expression; and
2. The variation in extrinsic factors that determine the state of the cell. Examples of extrinsic factors include the external environment (16,17), the epigenetic state of the cell (18,19), the time from last cell division, and levels of molecular machines such as RNA polymerase, ribosome, proteases, and RNases (3,4,20).

In a given population of cells, the total noise (coefficient of variation)

$$\eta_T = \frac{\langle m^2\rangle - \langle m\rangle^2}{\langle m\rangle^2} \quad (1)$$

serves as a useful experimental quantification of the variability in gene expression where $\langle m\rangle$ is the mean level of the gene product m (mRNA or protein) and $\langle m^2\rangle$–$\langle m\rangle^2$ is the variance.

For a constitutively expressing promoter, under simplifying conditions, the contribution to η_T associated with extrinsic factors, the extrinsic noise η_E, can be experimentally measured separately from the intrinsic noise η_I (3,6,15,20). The decomposition experiment usually involves expression of two identical copies of a single gene inside cells. Variation in local effects, e.g., binding and unbinding of transcription factors, affects the expression of the two genes in an uncorrelated manner. On the other hand, variation in global factors such as RNAP/ribosome/RNase levels affects them in a correlated manner. After comparing the statistics of the joint-expression system with that of a single gene expression system, the correlation between the two genes is identified as the extrinsic noise. It is now known that the extrinsic noise is the dominant contributor to gene expression (3,15) and can change the profile of gene expression in a nontrivial manner (21). Evidently, an important step toward the conceptual understanding of the noisy gene expression is to quantitatively account for the effect of variations in extrinsic factors on gene expression.

The major technical hurdle in building a comprehensive theory for extrinsic variation originates in the multitude of factors that contribute to it. Consequently, theoretical exploration of noisy gene expression has concentrated on intrinsic noise. Here, one generally employs the master equation framework (9,10,22–24). Briefly, we define a set of reactions $\mathcal{R}$ involving species $\mathcal{G}$ (protein, mRNA, etc.). A transition matrix for evolution of the probability distribution of $\mathcal{G}$ is constructed. The transition matrix contains information about the chemistry (rates, allosteric binding, etc.) and the topology (feedback, loops, etc.) of the reactions. The probability distribution $P(\mathcal{G}|t,\mathcal{K})$ is then sought in terms of the rate constants $\mathcal{K} = \{k_1,k_2,\ldots\}$ of all reactions and time t. Because closed form solutions for the master equation exist only for a few simple systems, much theoretical development explores efficient ways of simplifying the solution of the master equation (10,23,25).

Submitted February 1, 2013, and accepted for publication May 3, 2013.

*Correspondence: pdixit@bnl.gov

Editor: Sean Sun.

0006-3495/13/06/2743/8 $2.00

http://dx.doi.org/10.1016/j.bpj.2013.05.010

The chemical reactions are carried out by molecular machines such as RNA polymerase, ribosomes, and enzymes, among others. Moreover, these chemical reactions also depend on the chemical state of the cell, such as, for example, the time from cell division, the chromatin structure of DNA, the presence of DNA binding proteins, RNA degradation by small RNAs, and the presence of RNA binding proteins. All these variables differ from cell to cell and as a function of time. Hence, the rate constants $\mathcal{K}$ depend on the state of the cell and are themselves stochastic variables. This makes gene expression a doubly stochastic process (26,27). In the theoretical analysis, we interpret the variability in $\mathcal{K}$—which represents the variability in global factors—as the extrinsic variability. The theoretical decomposition will faithfully represent the experimentally quantified one if:

1. The underlying model of intrinsic noise is an accurate description of gene expression; and
2. If the effect of within-cell variation in the parameters on gene expression is negligible at timescales relevant for gene expression.

Due to the very large number of affectors, it is impossible to model the extrinsic variability from first principles. Consequently, the theoretical treatment has either assumed a small extrinsic contribution resulting in a linear susceptibility-like analysis (20) or assumed an *ad hoc* structure for the distribution of extrinsic factors (21,28). Here, instead of accounting for all the extrinsic contributors ab initio, we develop a maximum entropy framework to estimate $P(\mathcal{K})$, from limited information about the gene expression profile. We successfully test our results on a simplified numerical scheme for mRNA production that explicitly incorporates the variability in molecular machinery. Most importantly, we show that extrinsic factors can qualitatively and quantitatively affect the experimentally observed histogram of the gene expression product (protein or mRNA).

THEORY

For concreteness, consider a constitutively expressing promoter in a bacterial setting (see Fig. 1). Later, we will substantially simplify this example. Here, an inactive gene is converted to an active gene with rate constant k_1 and vice versa (rate constant k_{-1}). An mRNA molecule is transcribed from the active gene at a rate constant k_2. A protein is translated from the mRNA at a rate k_4. The mRNA and the protein are degraded at rates k_3 and k_5, respectively. The number of activated genes g, the number of mRNA molecules m, and the number of protein molecules p represent $\mathcal{G}$. The time from last division is itself a stochastic variable for a heterogeneous population (29) and can be included as a parameter with the reaction rate constants. We assume that the conditional distribution $P(\mathcal{G}|\mathcal{K})$ is known. Here, $\mathcal{G} = \{g,m,p\}$ and $\mathcal{K} = \{k_1,k_{-1},k_2,k_3,k_4,k_5\}$.

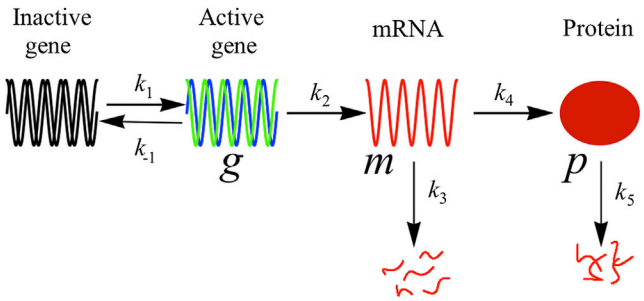

FIGURE 1 The most general case of a constitutively expressing promoter. An inactive gene (*black*) is turned into an active gene (and vice versa). The active gene (*blue* and *green*) is transcribed into an mRNA (*red*), which is then translated to a protein (*red ellipse*). The mRNA and the protein are also degraded. Various rate constants $\mathcal{K}$ govern the time evolution of $P(g,m,p|\mathcal{K})$, the joint probability distribution of g (number of activated genes), results for m (number of mRNA molecules), and the parameterization of p (number of protein molecules).

The maximum entropy framework

We now estimate the distribution of $\mathcal{K}$ using the maximum entropy (ME) framework (30). A brief introduction to ME can be found in the Supporting Material. Note that each point in the multidimensional $\mathcal{K}$-space represents a probability distribution in the $\mathcal{G}$-space. Consequently, the distribution whose entropy should be maximized is not $P(\mathcal{K})$ but the joint distribution $P(\mathcal{G},\mathcal{K})$ of species and rates (26,31).

The entropy $S[P(\mathcal{G},\mathcal{K})]$ of the joint distribution $P(\mathcal{G},\mathcal{K})$ is given by

$$S[P(\mathcal{G},\mathcal{K})] = -\sum_{\mathcal{G},\mathcal{K}} P(\mathcal{G},\mathcal{K}) \log P(\mathcal{G},\mathcal{K}) \tag{2}$$

$$= S[P(\mathcal{K})] + \sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K})P(\mathcal{K}). \tag{3}$$

Here,

$$P(\mathcal{K}) = \sum_{\mathcal{G}} P(\mathcal{G},\mathcal{K}), \tag{4}$$

$$S[P(\mathcal{K})] = -\sum_{\mathcal{K}} P(\mathcal{K}) \log P(\mathcal{K}), \tag{5}$$

and

$$S(\mathcal{G}|\mathcal{K}) = -\sum_{\mathcal{G}} P(\mathcal{G}|\mathcal{K}) \log P(\mathcal{G}|\mathcal{K}) \tag{6}$$

is the entropy of the conditional distribution $P(\mathcal{G}|\mathcal{K})$.

If we constrain the mean values of the rate constants $\langle k_1\rangle$, $\langle k_2\rangle$,..., the ME framework predicts that the joint distribution maximizes the entropy $S[P(\mathcal{G},\mathcal{K})]$ subject to the constraints. To find the distribution, we introduce Lagrange multipliers α_1, α_2,... corresponding to rate constants k_1, k_2,... and γ for normalization. The modified objective function is

$$S[P(\mathcal{G},\mathcal{K})] - \sum_j \alpha_j \left(\sum_{\mathcal{G},\mathcal{K}} P(\mathcal{G},\mathcal{K}) k_j - \langle k_j \rangle \right) + \gamma \left(\sum_{\mathcal{G},\mathcal{K}} P(\mathcal{G},\mathcal{K}) - 1 \right) \tag{7}$$

$$= S[P(\mathcal{K})] + \sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K}) P(\mathcal{K}) - \sum_j \alpha_j \left(\sum_{\mathcal{K}} P(\mathcal{K}) k_j - \langle k_j \rangle \right) + \gamma \left(\sum_{\mathcal{K}} P(\mathcal{K}) - 1 \right). \tag{8}$$

Note that the mean values of the rate constants are not directly observable from experiments. Employing them as constraints is a departure from the canonical understanding of the ME framework wherein probability distributions are predicted from moments calculated from experimental data. Yet, the ME framework can also be seen as an inference tool (26,31,32): ME predicts the logically consistent probability distribution if mean values of certain important parameters of an experiment are fixed.

Because we know the functional form of $P(\mathcal{G}|\mathcal{K})$, in Eq. 8 we have summed over all possible values of $\mathcal{G}$ at a given value of $\mathcal{K}$. Setting the derivative of Eq. 8 with respect to $P(\mathcal{K})$ equal to zero and solving, we get

$$P(\mathcal{K}) \propto \exp\left(S(\mathcal{G}|\mathcal{K}) - \sum_j \alpha_j k_j \right). \tag{9}$$

Equation 9 is the maximum entropy estimate of the distribution of $\mathcal{K}$ if we constrain only the mean values of the rate constants. Note that in addition to the usual exponentials (see the Supporting Material), the distribution also depends on the entropy $\mathcal{S}(\mathcal{G}|\mathcal{K})$ of the conditional distribution $P(\mathcal{G}|\mathcal{K})$.

Estimating $P(\mathcal{K})$ in an *N*-reporter experiment

Experimental advances allow us to construct more than one identical reporter for a gene inside a single cell (3,15). Mathematically, instead of generating samples of $\mathcal{G}$ from the distribution $P(\mathcal{G}|\mathcal{K})$ for a fixed value of $\mathcal{K}$, we can conceive an experiment where we can sample N identical experiments of the same species $\mathcal{G}$ from the joint distribution $P(\mathcal{G}_1,\mathcal{G}_2,\ldots,\mathcal{G}_N|\mathcal{K})$ at a fixed value of $\mathcal{K}$. Note that the variability in the extrinsic factors respecting the distribution $P(\mathcal{K})$ bears no relation to the number of reporters employed in a particular experiment. Consequently, we require the ME framework-predicted $P(\mathcal{K})$ to be independent of N (31).

If we assume that the N experiments are sampled independently of each other—this is a crucial assumption in N-reporter experiments (3,15)—we can write

$$P(\mathcal{G}_1,\mathcal{G}_2,\ldots,\mathcal{G}_N|\mathcal{K}) = \prod_{n=1}^{N} P(\mathcal{G}_n|\mathcal{K}). \tag{10}$$

Similar to the considerations above, to estimate $P(\mathcal{K})$ from this N-reporter experiment, we maximize the entropy of the joint distribution $P(\mathcal{G}_1,\mathcal{G}_2,\ldots,\mathcal{G}_N,\mathcal{K})$ constraining the mean values of the rate constants $\langle k_1\rangle,\langle k_2\rangle,\ldots$. The entropy of the joint distribution can be simplified using the independence in Eq. 10 as

$$S[P(\mathcal{G}_1,\mathcal{G}_2,\ldots,\mathcal{G}_N,\mathcal{K})] = S[P(\mathcal{K})] + N \sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K}) P(\mathcal{K}). \tag{11}$$

The modified objective function is given by (see Eq. 8)

$$S[P(\mathcal{K})] + N \sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K}) P(\mathcal{K}) - \sum_j \alpha_j \left(\sum_{\mathcal{K}} P(\mathcal{K}) k_j - \langle k_j \rangle \right) + \gamma \left(\sum_{\mathcal{K}} P(\mathcal{K}) - 1 \right). \tag{12}$$

Consequently, the ME framework estimates the distribution $P(\mathcal{K})$ as

$$P(\mathcal{K}) \propto \exp\left(N S(\mathcal{G}|\mathcal{K}) - \sum_j \alpha_j k_j \right). \tag{13}$$

Interestingly, the estimate of the variability $P(\mathcal{K})$ depends on the number of reporters (see Eq. 9 and Eq. 13) used in the experiment. This problem will be alleviated if we introduce the average entropy of a given experiment $\langle S(\mathcal{G}|\mathcal{K})\rangle$ as an additional constraint. This additional constraint is not an experimentally observable constraint but merely a requirement of consistency in the prediction over multiple experiments (26,31,33). Introducing the additional constraint $\langle S(\mathcal{G}|\mathcal{K})\rangle$ in the objective function by introducing a Lagrange multiplier μ_N, we write the modified objective function as

$$S[P(\mathcal{K})] + N \sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K}) P(\mathcal{K}) - \sum_j \alpha_j \left(\sum_{\mathcal{K}} P(\mathcal{K}) k_j - \langle k_j \rangle \right) + \gamma \left(\sum_{\mathcal{K}} P(\mathcal{K}) - 1 \right) + \mu_N \left(\sum_{\mathcal{K}} S(\mathcal{G}|\mathcal{K}) P(\mathcal{K}) - \langle S(\mathcal{G}|\mathcal{K}) \rangle \right). \tag{14}$$

Writing $N + \mu_N = \mu$ and maximizing with respect to $P(\mathcal{K})$, we get

$$P(\mathcal{K}) \propto \exp\left(\mu S(\mathcal{G}|\mathcal{K}) - \sum_j \alpha_j k_j \right). \tag{15}$$

Equation 15 is the main theoretical result of this work. Briefly, if we know that the rate constants $\mathcal{K}$ vary from cell to cell and as a function of time, and if, rather than precisely knowing them, we constrain only their mean values, the ME framework predicts the distribution $P(\mathcal{K})$ as Eq. 15. Note that in addition to the usual exponentials, the distribution also depends on the conditional entropy $S(\mathcal{G}|\mathcal{K})$. Similar results have been obtained for thermodynamic systems (26,33) and in estimating prior distributions in Bayesian inference (31).

Experimentally observed distribution of chemical species

The experimentally observable distribution $P(\mathcal{G})$ is obtained by summing over all possible variations in $\mathcal{K}$. We get

$$P(\mathcal{G};\mu,\alpha_1,\alpha_2,\ldots) \propto \sum_{\mathcal{K}} P(\mathcal{G}|\mathcal{K})\cdot\exp \times\left(\mu S(\mathcal{G}|\mathcal{K}) - \sum_j \alpha_j k_j\right). \tag{16}$$

Note that the distribution in Eq. 16 is parameterized by μ and $\alpha_1, \alpha_2,\ldots$. Each α_i corresponds to one rate constant k_i whereas μ governs the extrinsic variability. In short, the ME framework predicts extrinsic variability only with one additional parameter μ. Note that Eq. 16 provides a functional form for the $P(\mathcal{G})$ distribution. The parameters μ and $\{\alpha_i\}$ can be fit to suitable experimental measurements such as the moments of the distribution. Below, we will work out in detail the noise in the production of mRNA molecules from a constitutive promoter.

The distribution of mRNA copy numbers

Consider the simplified reaction scheme

$$\text{DNA} \xrightarrow{\gamma} \text{mRNA} \xrightarrow{\delta} \phi \tag{17}$$

of transcription and degradation of mRNA molecules of a particular gene. The value γ is the rate of transcription and δ is the rate of degradation.

In Eq. 17, we have neglected the activation states of the DNA molecule e.g., promoter fluctuations (4,5,10). Promoter fluctuations are thought to occur from (among other things) chromatin remodeling and binding and unbinding of transcription factors (11,18,19). The chromosome of the DNA of a bacteria like *E. coli* is structured in ~100–500 nucleoids (34). It is very likely that the chromatin structure extends locally to 10–50 genes around the gene studied and affects the transcription of all genes in a local region. Consequently, in a hypothetical dual-reporter experiment to study noise in mRNA production similar to Elowitz et al. (3), promoter fluctuations due to chromatin remodeling are likely to affect the expression of all genes localized in a given region on the DNA in a correlated fashion and will contribute to the extrinsic noise. On the other hand, promoter fluctuations arising due to noisy binding of transcription factors will act in an uncorrelated fashion in a hypothetical dual-reporter experiment. The contribution to mRNA noise due to noisy transcription factor binding will contribute to the intrinsic noise. Noisy transcription factor binding will result in a non-Poissonian process of mRNA production and will result in mRNA distributions that are wider than the Poisson statistics (18,19). In what follows, we neglect the contribution of noisy transcription factor binding to promoter fluctuations and effectively treat them as one of the local albeit extrinsic contributor to the variation in the effective rate of synthesis for the given gene. Below, we briefly discuss how to further parse the variability in the effective rate of synthesis into a contribution from promoter fluctuations and a contribution from other global extrinsic factors.

The solution of the reaction scheme at any time t and at steady state is a Poisson distribution

$$P(m|k) = \frac{e^{-k}k^m}{m!} \tag{18}$$

of mRNA copy number m with effective synthesis rate $k = \gamma/\delta(1-e^{-\delta t})$ (24).

The effective synthesis rate k depends, in a complicated manner, on various factors including chromatin remodeling (11,18,19), the states of many molecules in the cell including the components of RNA polymerase, the dynamics of assembly of the RNA polymerase holoenzyme, various RNase molecules, and other competing genes (3,20). Consequently, it varies from cell to cell and as a function of time from the start of the cell cycle. Thus, while studying gene expression in a population, instead of fixing a particular value of the effective synthesis rate k, we need to consider $P(k)$ the probability distribution of k. $P(k)$ quantifies the extrinsic contribution noisy gene expression.

For a given gene, experimentally assessing the variability in k is nontrivial—$P(k)$ has to be inferred from limited experimental information in respect to mean expression level, variation in gene expression level, etc. From Eq. 15, we see that the distribution $P(k)$ is given by

$$P(k) \propto \exp[(\mu\alpha - 1)S(k) - \alpha k]. \tag{19}$$

Here, $S(k)$ is the entropy of the conditional distribution $P(m|k)$, a Poisson distribution. Unfortunately, $S(k)$ does not have a closed form but $S(k) \sim \log k$. Thus,

$$P(k;\mu,\alpha) \propto k^{\mu\alpha-1}e^{-\alpha k}. \tag{20}$$

In Eq. 20, μ is the mean expression level and $\alpha = \eta_1/\eta_E$ is the ratio of the intrinsic and the extrinsic noise. The joint distribution $P(m,k)$ is then given by

$$P(m,k) = P(m|k)P(k) \propto \frac{e^{-\alpha k}k^{m+\mu\alpha-1}}{m!}. \tag{21}$$

The experimentally accessible histogram $P(m)$ is obtained by summing over all variations in k, i.e., summing over the variation in extrinsic factors,

$$P(m) = \sum_k P(m,k) \propto \sum_k \frac{e^{-\alpha k} k^{m+\mu\alpha-1}}{m!}. \quad (22)$$

We estimate $P(m)$ to be the negative binomial distribution (the discrete version of the γ-distribution),

$$P(m) \propto \frac{1}{(1+\alpha)^m} \times \frac{\Gamma[m+\alpha\mu]}{m!}. \quad (23)$$

Noise decomposition of experimental data

We estimate the total noise η_T from Eq. 23 (see the Supporting Material for details) as

$$\eta_T = \frac{1}{\mu}\left(1+\frac{1}{\alpha}\right) = \frac{1}{\mu}\left(1+\frac{\eta_E}{\eta_I}\right) \geq \frac{1}{\mu}$$

and

$$\begin{aligned} \eta_I &= \frac{1}{\mu}, \\ \eta_E &= \eta_T - \frac{1}{\mu}. \end{aligned} \quad (24)$$

The greater-than-Poisson relationship between η_T and the mean mRNA copy number μ (see Eq. 24) is sometimes attributed to nonPoissonian dynamics such as promoter fluctuations, chromatin remodeling, and mRNA synthesis bursts, among other causes (4,5,7,10,18,19). These effects themselves are thought to arise from cell-to-cell and dynamic variability in chromatin state and the state of DNA binding molecules (11,18,19). Additionally, we suggest that the cell-to-cell variation in other extrinsic factors (3,20) also contributes to the greater-than-Poisson relationship.

The ME framework predicts that Eqs. 23 and 24 completely determine the histogram of mRNA copy numbers from experimentally measured mean expression level μ and total noise η_T. Moreover, η_T is always >1 and η_I and η_E can be estimated from the histogram alone. Importantly, the framework estimates the hitherto elusive effect of extrinsic factors on gene expression regarding the distribution $P(k)$ of the effective synthesis rate k.

The joint distribution Eq. 21 also allows us to estimate potentially interesting moments; for example, we predict that the Pearson correlation coefficient

$$\rho_{mk} = \frac{1}{\sqrt{1+\alpha}} = \sqrt{\frac{\eta_E}{\eta_T}} \quad (25)$$

between effective mRNA synthesis rate and the mRNA copy number is the square-root of the ratio of extrinsic and total noise. These are some of the falsifiable predictions of the development presented here.

RESULTS AND DISCUSSION

Numerical validation of the ME-predicted distribution

We analyze a simple numerical scheme for the synthesis of rGene, the mRNA of a constitutively expressed gene. In the scheme, the variability in the effective synthesis rate k arises from the stochasticity in the production and degradation of the machinery (RNAP and RNase). We show that the ME-predicted distribution (Eq. 23) describes very accurately the numerically predicted distribution of mRNA copy number for different strengths of extrinsic noise (see Fig. 2 for a cartoon and the Supporting Material for details).

Let [X] denote the concentration of species X. In the model, the rate of synthesis $\gamma = \gamma_0$[RNAP] and the rate of degradation $\delta = \delta_0$[RNAse] of rGene, the mRNA of the gene under consideration, both depend on the concentration of the cellular proteins that carry out those reactions for [RNAP] (a proxy for the RNA polymerase complex) and [RNase] (a proxy for RNase), respectively. Both the proteins are themselves are stochastically synthesized and degraded. The variation in the proxies mimics the cell-to-cell variations in extrinsic factors. The effective synthesis rate k is directly proportional to the ratio [RNAP]/[RNase]. We implement the Gillespie algorithm (35) to estimate the steady-state distribution of [rGene], the mRNA copy number. The correlated dynamics of production of rGene, RNAP, and RNase play an important part in determining the dynamics of the variability in [rGene]. The steady-state joint distribution [RNAP] and [RNase] completely determines the steady-state distribution of [rGene] *if* the dynamics of synthesis and degradation of RNAP and RNase are not too fast compared to that of rGene. The parameters chosen for the simulation make sure that the timescale of synthesis and degradation of RNAP and RNase is of the same order as that of rGene. We only sample the distribution of mRNA copy numbers at long times ensuring that the steady state has been reached (see the Supporting Material for details). To clearly elucidate the effect of extrinsic factors on gene expression profile, in Fig. 3, we show the histogram of mRNA copy numbers for three different levels of noise, quantified by

$$\eta_k \equiv \frac{\langle k^2\rangle - \langle k\rangle^2}{\langle k\rangle^2} = \eta_E, \quad (26)$$

the coefficient of variation in k, keeping the mean expression constant. The equality $\eta_k = \eta_E$ is a consequence of

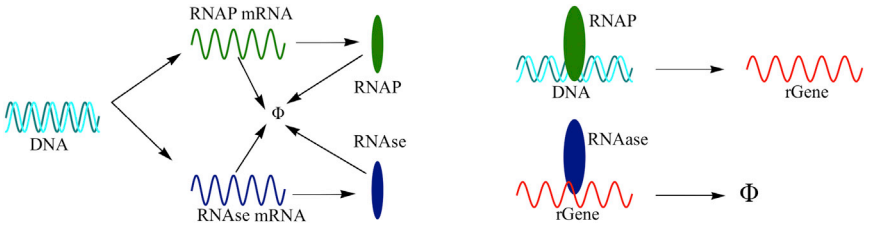

FIGURE 2 A cartoon of the simplified scheme of mRNA production that takes into account extrinsic factors in gene expression levels (see the Supporting Material for details). In the scheme, RNAP serves as the proxy for the RNA polymerase holoenzyme complex and RNase is the proxy for RNA degradation machinery. The rate of synthesis of rGene, the RNA of a given gene, is directly proportional to the concentration [RNAP] of the protein product of the RNAP gene. Similarly, the rate of degradation of rGene is directly proportional to the concentration [RNase], the protein product of RNase gene. RNAP and RNase themselves are synthesized and degraded stochastically.

the underlying single-step process and will not hold true for other cases.

In the left panel of Fig. 3, we show the histogram of mRNA copy numbers when the coefficient of variation η_k is low ($\eta_k \approx 5 \times 10^{-5}$). Observe that the histogram of mRNA copy numbers (*red circles*) is well described by a Poisson distribution (*black dashes*), as is expected. If we increase the variation in k ($\eta_k \approx 2.5$ in the *middle panel* and $\eta_k \approx 3.8$ in the *right panel*), the histogram of mRNA copy numbers gets broader and is best described by $P(m)$ (Eq. 23, *solid blue*) rather than Poisson distribution (*black dashes*). Thus, even though the mRNA synthesis and degradation is governed by a Poisson process with an effective synthesis rate k, the variation in the rate itself makes gene expression a doubly stochastic process (26,27) and leads to a histogram of mRNA copy numbers that is not Poisson-distributed and is best described by a Gamma-like distribution.

Interpreting experiments

Fig. 4 shows the best fit to the histogram of mRNA copy numbers for the *E. coli* gene TufA (7). The Poisson distribution does not capture the mRNA histogram whereas Eq. 23 describes it well (for a comparison with numerical simulations, see the *right panel* of Fig. 3). Also, recently, So et al. (18) showed that the distribution of mRNA copy numbers in *E. coli* is well described by a negative binomial distribution.

In Fig. 5, we show the measured total noise and the predicted log-binned average trends in the decomposition of the total noise into its intrinsic and extrinsic components. The components are estimated from Eq. 24 for ~130 genes, as reported in Taniguchi et al. (7). The noise decreases as mean expression level increases and both intrinsic and extrinsic components contribute significantly to the total noise. The total noise and the extrinsic noise saturate at high expression levels that are sometimes referred to as the "extrinsic limit" (4,7,8,15). Importantly, our framework also allows us to directly estimate the variation $P(k)$ of the effective synthesis rate k.

Incorporating promoter fluctuations explicitly

The mRNA histogram from a slightly involved model that captures the activation state of the DNA molecule (10,18,19) results in a distribution identical to Eq. 23. In that model, the deviation from Poisson distribution is ascribed entirely to promoter fluctuations. As mentioned above, promoter fluctuations arise, among other things, from chromatin remodeling (11,19) and are likely to affect the local region around the given gene (34). Within our framework, the variation in mRNA synthesis rate due to promoter fluctuations is treated as extrinsic and is automatically incorporated in the distribution of the effective synthesis rate.

We can further separate the variability in k due to promoter fluctuations from the variability due to other extrinsic

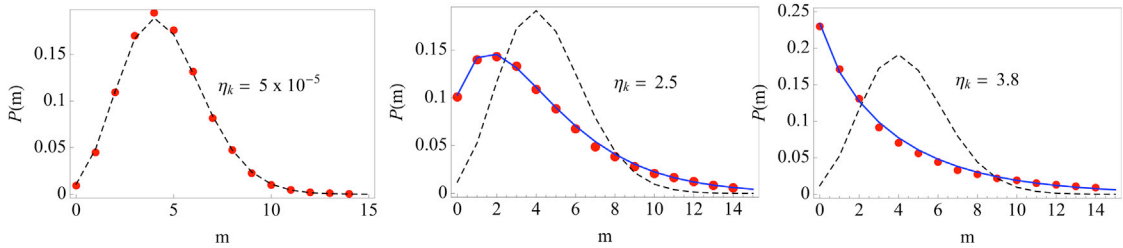

FIGURE 3 The histogram of mRNA copy numbers (*red dots*), the Poisson distribution fit (*dashed black lines*), and the marginal distribution fit (*solid blue*, see Eq. 23) for three different scenarios in the numerical simulation. The mean mRNA copy number $\mu \approx 4.4$ for all three cases. (*Left*) Small variations in extrinsic factors ($\eta_k \approx 5 \times 10^{-5}$) results in a histogram of mRNA copy numbers that is well described by a Poisson distribution. (*Middle*) Higher variation in extrinsic factors ($\eta_k \approx 2.5$) broadens the histogram of mRNA copy numbers. The marginal distribution $P(m)$ (see Eq. 23) fits the data well. (*Right*) High variation in extrinsic factors ($\eta_k \approx 3.8$). Again, note that the histogram of mRNA copy numbers is wider than a Poisson distribution and the marginal distribution $P(m)$ fits the simulation well.

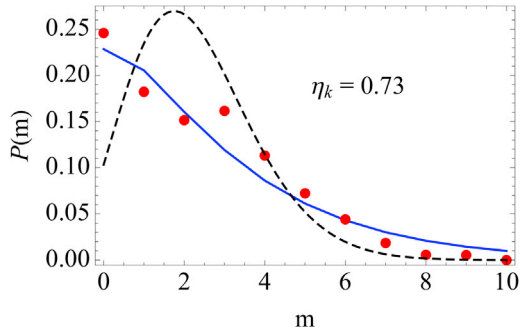

FIGURE 4 (*Dashed black lines*). Our results predict that the experimentally measured mRNA copy number histogram is described by Eq. 23 (*solid blue*). $\eta_k \approx 0.7$ is the estimated coefficient of variation of the effective synthesis rate k.

factors. The presence of other extrinsic factors can be tested in a number of ways. For example, if promoter fluctuations are the major contributor to the variation of effective synthesis rate, it can be shown that the experimentally estimated skewness

$$\gamma_1 = \frac{\langle m^3\rangle - 3\langle m\rangle\langle m^2\rangle + 2\langle m\rangle^3}{\left(\langle m^2\rangle - \langle m\rangle^2\right)^{3/2}} \qquad (27)$$

of the distribution of mRNA numbers will be roughly equal to twice the square-root of the total noise η_T. In the presence of other extrinsic noise, this relationship is somewhat modified (see the Supporting Material for details).

If promoter fluctuations are explicitly modeled, the distribution of mRNA copy numbers is characterized by at least two parameters (10,18). The development presented here will add one additional parameter to characterize the extrinsic variability beyond promoter fluctuations. Thus, the resulting distribution will be characterized by three parameters. Analyzing the reported experimental measurements of total noise to predict extrinsic noise beyond promoter fluctuations will consequently be an overfit. Yet, we note that if experimental measurements reliably estimate the third moment of the mRNA distribution, the presented framework will be able to parse the total noise into its extrinsic and intrinsic (which will include promoter fluctuations) contributions without the assistance of a two-color experiment (see the Supporting Material for details).

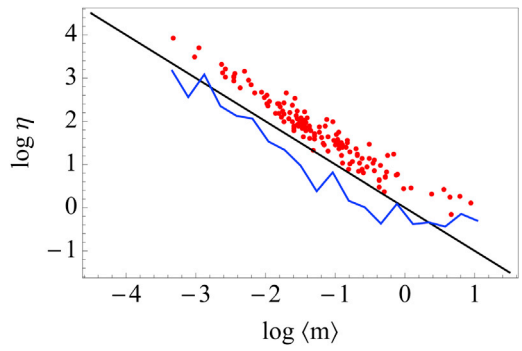

FIGURE 5 The experimentally measured total noise η_T (*red dots*) is always higher than what is expected from a Poisson distribution (*black line*, see Eq. 24). Our framework also allows us to predict the extrinsic noise η_E and the variation in the effective synthesis rate η_k. (*Blue line*) Log-binned average of η_E (also equal to η_k). Note that as opposed to proteins, for most mRNAs, intrinsic noise dominates the total noise for mRNAs. At higher mRNA numbers, the η_E dominates η_T. Within the ME framework, we can explicitly estimate the hitherto inaccessible variation in the effective synthesis rate as well.

CONCLUDING REMARKS

Measurements of the cell-to-cell variation in protein numbers show that the extrinsic contributions play a dominant role (3). Yet, much of the theoretical development in understanding noise in gene expression has focused on the effect of intrinsic contributors on statistical mechanical fluctuations in binding and diffusion of molecules. The limited treatment extrinsic noise has received (7,20,28) employs the linear fluctuation-dissipation like susceptibility analysis (20) or ad hoc assumptions about the nature of variation in extrinsic parameters (7,28).

To the best of our knowledge, for the first time, we have presented a framework that systematically estimates the static variation in the rate parameters of gene expression. In the context of the model, the extrinsic noise in gene expression arises solely because of the variation in the parameters, allowing us to separate the intrinsic and the extrinsic contributors to noisy gene expression from limited information about the gene expression profile. Consequently, a weakness of the presented framework is that the decomposition of the total noise in its intrinsic and extrinsic contributions depends on the accuracy of the gene expression model. We conclude that extrinsic factors can change the experimentally accessible histogram of mRNA copy numbers quantitatively and qualitatively. More importantly, the framework allows us to directly estimate the hitherto elusive variation in global extrinsic factors.

Specifically, we show that even if mRNA synthesis and degradation is described by a simple Poisson process, owing to the variation in the effective synthesis rate k, the experimentally accessible histogram of mRNA copy numbers is broader and we estimate it to be the negative binomial distribution (see Eq. 23). Consequently, we find that variation in the effective synthesis rate k contributes to the greater-than-Poisson relationship between noise η_T and the mean mRNA copy number $\langle m\rangle$ (see Eq. 24). We also predict that, in contrast to proteins (3), intrinsic and extrinsic factor variations both contribute significantly to the noisy expression of mRNA. Moreover, we directly probe the variation in effective mRNA synthesis rate k and show that the coefficient of variation η_k saturates at high expression levels (see Fig. 5, *bottom*).

Arguably, biologically interesting situations where noise is important are not limited to production of mRNA

molecules. One would like to know how noise affects the regulation of internal circuits, response to external stimuli, and fitness and evolution. It is clear that once the distribution of $\mathcal{G}$ is known as a function of $\mathcal{K}$, the application of the presented framework is, in principle, straightforward. Unfortunately, the conditional distribution $P(\mathcal{G}|\mathcal{K})$ is known for very few simple cases (similar to the one discussed in this work). We propose the following algorithm to overcome this difficulty.

Even though the entire distribution $P(\mathcal{G}|\mathcal{K})$ is almost always analytically inaccessible, the first two moments $\{\langle\mathcal{G}_i\rangle_{\mathcal{K}}\}$ and $\{\langle\mathcal{G}_i\mathcal{G}_j\rangle_{\mathcal{K}}\}$ can be estimated very accurately as analytical functions of $\mathcal{K}$ for a number of complicated situation using the well-known Ω expansion (9). Moreover, under the assumption of linear noise, the entropy $S(\mathcal{G}|\mathcal{K})$ can itself be approximated as $S(\mathcal{G}|\mathcal{K}) \sim \log\det\Sigma$, where $\sum_{ij} = \langle\mathcal{G}_i\mathcal{G}_j\rangle - \langle\mathcal{G}_i\rangle\langle\mathcal{G}_j\rangle$ is the covariance matrix. From here onwards, it is a straightforward exercise to compute $P(\mathcal{K})$ using Eq. 15. The intrinsic and extrinsic components can then be separated out analytically. We will implement the proposed program for protein synthesis and networks in the future.

SUPPORTING MATERIAL

Supporting analysis including equations and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(13)00560-2.

I thank Dr. Sergei Maslov and Dr. Adam de Graff for a critical reading and constructive suggestions, and Prof. Ken Dill, Prof. Dilip Asthagiri, and Ms. Shreya Saxena for stimulating conversations and suggestions about the manuscript. I also thank the reviewers for their critical reading of the manuscript and important suggestions in improving it considerably.

This work was supported by grant No.PM-031 from the Office of Biological Research of the U.S. Department of Energy.

REFERENCES

1. Bar-Even, A., J. Paulsson, …, N. Barkai. 2006. Noise in protein expression scales with natural protein abundance. *Nat. Genet.* 38:636–643.
2. Cai, L., N. Friedman, and X. S. Xie. 2006. Stochastic protein expression in individual cells at the single molecule level. *Nature.* 440:358–362.
3. Elowitz, M. B., A. J. Levine, …, P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science.* 297:1183–1186.
4. Kaufmann, B. B., and A. van Oudenaarden. 2007. Stochastic gene expression: from single molecules to the proteome. *Curr. Opin. Genet. Dev.* 17:107–112.
5. Raj, A., and A. van Oudenaarden. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell.* 135:216–226.
6. Rosenfeld, N., J. W. Young, …, M. B. Elowitz. 2005. Gene regulation at the single-cell level. *Science.* 307:1962–1965.
7. Taniguchi, Y., P. J. Choi, …, X. S. Xie. 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science.* 329:533–538.
8. Newman, J. R. S., S. Ghaemmaghami, …, J. S. Weissman. 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature.* 441:840–846.
9. Paulsson, J. 2005. Models of stochastic gene expression. *Phys. Life Rev.* 2:157–175.
10. Raj, A., C. S. Peskin, …, S. Tyagi. 2006. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309.
11. Kaern, M., T. C. Elston, …, J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.* 6:451–464.
12. Maheshri, N., and E. K. O'Shea. 2007. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* 36:413–434.
13. Maamar, H., A. Raj, and D. Dubnau. 2007. Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science.* 317:526–529.
14. Fraser, H. B., A. E. Hirsh, …, M. B. Eisen. 2004. Noise minimization in eukaryotic gene expression. *PLoS Biol.* 2:e137.
15. Stewart-Ornstein, J., J. S. Weissman, and H. El-Samad. 2012. Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. *Mol. Cell.* 45:483–493.
16. Chubb, J. R., T. Trcek, …, R. H. Singer. 2006. Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16:1018–1025.
17. Golding, I., and E. C. Cox. 2006. Eukaryotic transcription: what does it mean for a gene to be 'on'? *Curr. Biol.* 16:R371–R373.
18. So, L.-H., A. Ghosh, …, I. Golding. 2011. General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* 43:554–560.
19. Golding, I., J. Paulsson, …, E. C. Cox. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell.* 123:1025–1036.
20. Swain, P. S., M. B. Elowitz, and E. D. Siggia. 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA.* 99:12795–12800.
21. Shahrezaei, V., J. Ollivier, and P. Swain. 2008. Colored extrinsic fluctuations and stochastic gene expression. *Mol. Sys. Biol.* 4:196.
22. Thattai, M., and A. van Oudenaarden. 2001. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA.* 98:8614–8619.
23. Sánchez, A., and J. Kondev. 2008. Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. USA.* 105:5081–5086.
24. Hemberg, M., and M. Barahona. 2007. Perfect sampling of the master equation for gene regulatory networks. *Biophys. J.* 93:401–410.
25. Friedman, N., L. Cai, and X. S. Xie. 2006. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* 97:168302.
26. Dixit, P. D. 2012. A maximum entropy thermodynamics for small systems. *J. Chem. Phys.* 138:184111.
27. Tjostheim, D. 1986. Some doubly stochastic time series models. *J. Time Ser. Anal.* 7:51–72.
28. Scott, M., B. Ingalls, and M. Kaern. 2006. Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. *Chaos.* 16:026107.
29. Harley, C. B., and S. Goldstein. 1978. Cultured human fibroblasts: distribution of cell generations and a critical limit. *J. Cell. Physiol.* 97:509–516.
30. Jaynes, E. T. 1957. Information theory and statistical mechanics. I. *Phys. Rev.* 106:620–630.
31. Caticha, A., and R. Preuss. 2004. Maximum entropy and Bayesian data analysis: entropic prior distributions. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 70:046127.
32. Shore, J., and R. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inf. Theory.* 26:26–37.
33. Crooks, G. E. 2007. Beyond Boltzmann-Gibbs statistics: maximum entropy hyperensembles out of equilibrium. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 75:041119.
34. Reyes-Lamothe, R., X. Wang, and D. Sherratt. 2008. *Escherichia coli* and its chromosome. *Trends Microbiol.* 16:238–245.
35. Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361.