RESEARCH ARTICLE

Transcriptional Bursting in Gene Expression: Analytical Results for General Stochastic Models

Niraj Kumar[1], Abhyudai Singh[2], Rahul V. Kulkarni[1]*

1 Department of Physics, University of Massachusetts Boston, Boston, Massachusetts, United States of America, **2** Department of Electrical and Computer Engineering, University of Delaware, Newark, Delaware, United States of America

* rahul.kulkarni@umb.edu

Abstract

Gene expression in individual cells is highly variable and sporadic, often resulting in the synthesis of mRNAs and proteins in bursts. Such bursting has important consequences for cell-fate decisions in diverse processes ranging from HIV-1 viral infections to stem-cell differentiation. It is generally assumed that bursts are geometrically distributed and that they arrive according to a Poisson process. On the other hand, recent single-cell experiments provide evidence for complex burst arrival processes, highlighting the need for analysis of more general stochastic models. To address this issue, we invoke a mapping between general stochastic models of gene expression and systems studied in queueing theory to derive exact analytical expressions for the moments associated with mRNA/protein steady-state distributions. These results are then used to derive noise signatures, i.e. explicit conditions based entirely on experimentally measurable quantities, that determine if the burst distributions deviate from the geometric distribution or if burst arrival deviates from a Poisson process. For non-Poisson arrivals, we develop approaches for accurate estimation of burst parameters. The proposed approaches can lead to new insights into transcriptional bursting based on measurements of steady-state mRNA/protein distributions.

Author Summary

One of the fundamental problems in biology is understanding how phenotypic variations arise among individuals in a population. Recent research has shown that phenotypic variations can arise due to probabilistic cell-fate decisions driven by inherent randomness (noise) in the process of gene expression. One of the manifestations of such stochasticity in gene expression is the production of mRNAs and proteins in bursts. Bursting in gene expression is known to impact cell-fate in diverse systems ranging from latency in HIV-1 viral infections to cellular differentiation. Recent single-cell experiments provide evidence for complex arrival processes leading to bursting, however an analytical framework connecting such burst arrival processes with the corresponding higher moments of mRNA/

protein distributions is currently lacking. We address this issue by invoking a mapping between general models of gene expression and systems studied in queueing theory. The framework developed and the results derived lead to new approaches for testing commonly used assumptions in modeling gene expression and for accurate estimation of burst parameters. Notably, the phenomenon of stochastic bursting has been observed in a wide range of disciplines ranging from neuroscience and finance to cell biology. The approaches developed and results obtained in this work will thus contribute towards quantitative characterization of burst processes in diverse systems of current interest.

Introduction

The cellular response to fluctuating environments requires adjustments to cellular phenotypes driven by underlying changes in gene expression. Given the inherent stochasticity of cellular reactions, biological circuits controlling gene expression have to operate in the presence of significant noise [1–15]. While noise reduction and filtering is essential for several cellular processes [16], cells can also amplify and utilize intrinsic noise to generate phenotypic diversity that enables survival under stressful conditions [17]. Recent studies have demonstrated the importance of such bet-hedging survival strategies in diverse processes ranging from viral infections to bacterial competence [17]. Quantifying the kinetic mechanisms of gene expression that drive variations in a population of cells will thus contribute towards a fundamental understanding of cellular functions with important applications to human health.

Recent experiments focusing on gene expression at the single-cell level have revealed striking differences from the corresponding population-averaged behavior. In particular, it has been demonstrated that transcription in single cells is sporadic, with mRNA synthesis often occurring in bursts followed by variable periods of inactivity [7, 18–28]. Such transcriptional bursting can give rise to high variability in gene expression products and to phenotypic variations in a population of genetically identical cells [29–32]. Furthermore, dynamical parameters that characterize transcriptional bursting of key genes can significantly influence cell-fate decisions in diverse processes ranging from HIV-1 viral infections to stem-cell differentiation [17]. Correspondingly, there is significant interest in developing approaches for quantifying parameters related to transcriptional bursting such as frequency and mean burst size.

In recent years, multiple studies have provided evidence for bursty synthesis of mRNAs [20–25, 33, 34] and proteins [35, 36]. Experimental approaches in such studies include both steady-state measurements and time-dependent measurements of the mean and variance of gene expression products at the single-cell level. While obtaining time-lapse measurements of bursts at the single-cell level can be challenging, steady-state measurements at the single-cell level are now carried out routinely. It would thus be desirable to develop approaches for making inferences about burst parameters in gene expression using steady-state measurements at the single-cell level.

As noted in [37], steady-state measurements of the mean and variance alone cannot be used for estimating burst parameters for general models of gene expression, e.g. when burst arrival is governed by complex promoter-based regulation [38]. Additional insights into processes leading to transcriptional bursting can potentially be obtained using measurements of higher moments. However, analytical results for higher moments of steady-state mRNA and protein distributions in general models of expression have not been obtained so far. The derivation of the corresponding analytical expressions will elucidate how measurement of higher moments can potentially lead to quantification of burst parameters. To address these issues, it is essential to develop and analyze a general class of stochastic models of gene expression.

A simple stochastic model that is widely used in analyzing bursting in gene expression is the random telegraph model that takes into account the switching of promoter between transcriptionally active (ON) and inactive (OFF) states [39–41]. This model has been used as the basis for several studies focusing on inferring gene expression parameters based on observations of the mean and variance of mRNA/protein distributions [13, 27, 42]. In this model, in the limit that we have transcriptional bursting, the arrival of bursts is a Poisson process. Correspondingly, the waiting-time distribution between arrival of mRNA bursts is assumed to be exponential. In general, this assumption is not valid as there are multiple kinetic steps involved in promoter activation [37, 43, 44]. Recent experiments on mammalian genes [7, 45, 46] have demonstrated that the waiting-time for arrival of bursts does not have an exponential distribution. In view of these experimental observations, it is natural to ask: Using steady-state measurements, can we infer if the burst arrival process is *not* a Poisson process? If so, how can we estimate the corresponding burst parameters?

Furthermore, in estimating burst size it is commonly assumed that mRNA/protein bursts are geometrically distributed. This assumption, which has been validated by experimental observations for some genes, is derived from the corresponding distribution of bursts in the random telegraph model. However, given the complexity and diversity of gene expression mechanisms, it is possible that several promoters involve multiple rate-limiting steps in the transition from the ON state to the OFF state. In such cases, the transcriptional burst size distribution will not be a geometric distribution. This observation leads to the following question: Can we use steady-state measurements of moments to determine if the burst distribution deviates from a geometric distribution?

The aim of this paper is to address the above questions by considering models with general arrival processes for mRNA creation. The paper is organized as follows. First, we introduce a class of gene expression models with general arrival processes leading to mRNA/protein bursts with arbitrary burst distribution. Then we review the mapping from gene expression models to systems studied in queuing theory [43, 47, 48] and use this mapping to derive steady-state moments for mRNA/protein distributions. The analytical expressions obtained for the steady-state moments are used to develop approaches for estimating burst parameters for general arrival processes. Finally, we use the results obtained to derive conditions relating experimentally measurable quantities that determine if the arrival of mRNA bursts deviates from a Poisson process and if the distribution of mRNA bursts deviates from a geometric distribution.

Results

Model and preliminaries

We consider a general model of gene expression [43] as outlined in Fig 1. In the model, mRNAs are produced in bursts, with $f(t)$ representing a general arrival time distribution for mRNA bursts. The mRNA burst distribution can be arbitrary. Each mRNA then produces proteins with rate k_p, and finally, both mRNAs and proteins decay with rates μ_m and μ_p, respectively. Note that the model also allows for post-transcriptional regulation since the protein burst distribution from each mRNA can be arbitrary; the only assumption is that each mRNA produces proteins independently.

In the limit $\mu_p \ll \mu_m$, we can use the bursty synthesis approximation [40] for analyzing protein dynamics. This approximation consists of two steps: 1) obtaining the distribution of proteins produced from each mRNA and 2) assuming that the proteins are produced in instantaneous bursts. The corresponding distribution for the number of proteins created is referred to as the protein burst distribution. A detailed justification of the validity of this approximation has been provided in previous work [40, 49].
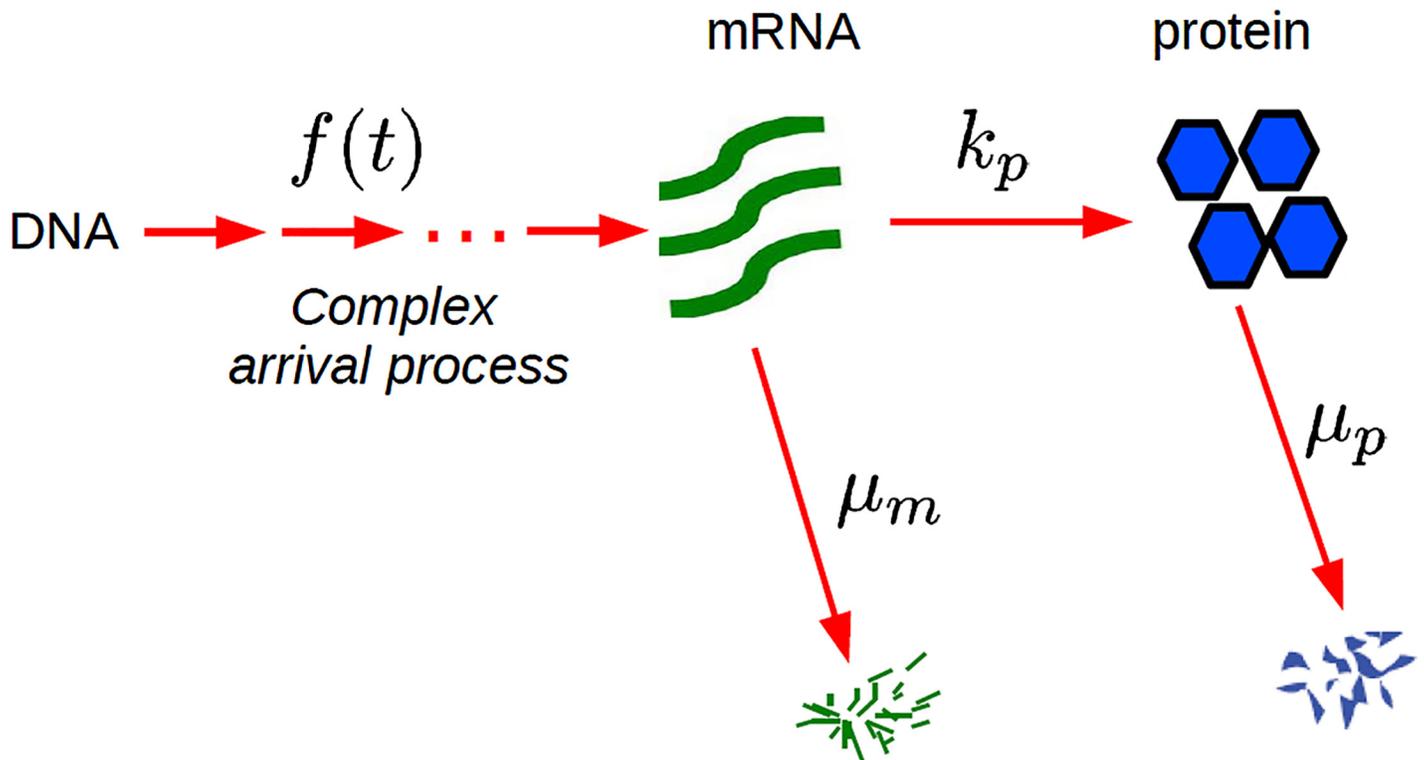
Fig 1. Kinetic scheme for the gene expression with general arrival time distributions. Bursts of mRNAs arrive with a general arrival time distributions f (t). Each mRNA produces proteins with rate k_p and mRNAs and proteins decay with rates μ_m and μ_p, respectively.

Let $a^p(z) = \sum_{n=0}^{\infty} z^n p(n)$ denote the generating function of the protein burst distribution p (n) produced by a *single* mRNA, and let $A^p(z) = \sum_{n=0}^{\infty} z^n P(n)$ denote the generating function of the protein burst distribution $P(n)$ produced by *all* the mRNAs in a burst. If we denote by $A^m(z)$ the generating function of the mRNA burst distribution, then we have the following relation between the generating functions

$$A^p(z) = A^m[a^p(z)]. \tag{1}$$

The above relation follows from the observation that the number of proteins produced in a burst is a compound random variable: the sum of m independent identical random variables, each of which corresponds to the number of proteins produced from a single mRNA in the burst and m itself is a random variable denoting the number of mRNAs produced in the burst.

While the analytical results that we derive are valid for general mRNA and protein burst distributions, we will primarily focus on a specific class of burst distributions. Simple kinetic models and the results from multiple experiments indicate that mRNA burst distributions are geometric [35]. Similarly, the burst distribution of proteins produced from a single mRNA is a geometric distribution with mean $\langle p_b \rangle = k_p/\mu_m$. For a geometric distribution with mean $\langle p_b \rangle$, the generating function is given by

$$a^p(z) = \frac{1}{[1 + \langle p_b \rangle (1 - z)]}.$$

If we condition the geometric distribution on the production of at least 1 mRNA, then the generating function for the corresponding conditional geometric distribution is given by

$$A^m(z) = \frac{z}{[1 + \langle m_b \rangle (1 - z)]}$$

with $(1 + \langle m_b \rangle)$ as the mean mRNA burst size. Note that in the limit $\langle m_b \rangle \to 0$, this distribution reduces to exactly 1 mRNA produced in each burst. Thus the conditional geometric distribution provides a unified representation of both Poisson arrival process for mRNAs ($\langle m_b \rangle \to 0$) and processes leading to transcriptional bursting ($\langle m_b \rangle > 0$).

Consider now the protein burst distribution produced by an underlying conditional geometric mRNA burst distribution with mean $(1 + \langle m_b \rangle)$. Using Eq (1), we see that the corresponding generating function of the protein burst distribution is given by

$$A^p(z) = \frac{1}{1 + \langle m_b \rangle \langle p_b \rangle (1 - z)}.$$

This is the generating function for a geometric distribution with mean $b = (1 + \langle m_b \rangle)\langle p_b \rangle)$, where $\langle p_b \rangle = k_p / \mu_p$ represents the mean protein burst size from a single mRNA.

Single-cell experiments have demonstrated that the protein burst mean b can be directly measured in some cases [35]. However, if the protein production rate k_p is not known, the preceding analysis implies that measurements of protein burst distributions (which determine b) cannot be used to determine the degree of transcriptional bursting $(1 + \langle m_b \rangle)$. Since the mean transcriptional burst size is an important parameter characterizing bursting, it is of interest to develop approaches for estimating it based on available experiments. Previous work [50] has argued that the mean transcriptional burst size cannot be determined using measurements of protein burst distributions alone or by using only protein steady-state distributions. It was suggested that combining such measurements can potentially provide a way of determining the mean transcriptional burst size. To explore this possibility, it is necessary to derive analytical results connecting moments of burst and steady-state distributions for general kinetic schemes.

Mapping to queueing theory: Results for moments

To obtain steady-state moments for the model outlined in Fig 1, we invoke the mapping of this gene expression model to systems studied in queueing theory [43, 48, 51, 52]. Broadly speaking, queueing theory is the mathematical theory of waiting lines formed by customers who, arriving according to some random protocol, stay in the system until they receive service from a group of servers. Such queues are typically characterized by specifying a) the stochastic process governing the arrival of customers, b) distribution of number of customers in each arrival, c) the stochastic process governing departure of customers, and d) the number of servers. When the gene expression model in Fig 1 is expressed in the language of queueing theory, individual mRNAs/proteins are the analogs of customers in queueing models. The production of mRNAs/proteins in bursts corresponds to the arrival of customers in batches. Just as the customers leave the queue after receiving service, mRNAs/proteins exit the system upon degradation. Thus the waiting-time distribution for mRNA/protein decay is the analog of service time distribution for customers in queueing models. For the model in Fig 1, their decay time distribution is the exponential distribution. Also, since mRNAs/proteins are degraded independently of each other, the corresponding number of servers in queueing models is ∞ (which ensures that presence of a customer in the system does not affect the service time of any other customer in the system).

Based on the above mapping, the queueing system corresponding to the model outlined in Fig 1 is the $GI^X/M/\infty$ system [43, 48]. In this model, the symbol G refers to a general waiting-

time distribution for the arrival process, I^X denotes customers arriving in batches of independently distributed random sizes X, M stands for Markovian (i.e. exponential) service-time distribution for customers and '∞' stands for infinite servers.

For the $GI^X/M/\infty$ model, exact results for iteratively obtaining the moments of the steady-state distribution of the number of customers have been derived [48]. Using these results, explicit expressions for the first four moments of the steady-state distribution are provided in the Supplementary S1 Text. Applying the mapping discussed above, these results can be translated into exact expressions for the moments of mRNA/protein steady-state distributions, as discussed below.

Let us first examine the expressions for steady-state means of mRNAs, $\langle m_s \rangle$, and proteins, $\langle p_s \rangle$, which are given by

$$\langle m_s \rangle = \frac{k_b}{\mu_m} \langle m_b \rangle, \quad \langle p_s \rangle = \frac{k_b}{\mu_p} b, \tag{2}$$

where k_b stands for the mean arrival rate of mRNA bursts and $b = \langle m_b \rangle \langle p_b \rangle$ is the mean of the protein burst distribution (including contributions from all the mRNAs). Although Eq (2) has been derived by assuming that the arrival of mRNAs/proteins is a renewal process, it is valid for arbitrary arrival processes. This is because Eq (2) is a direct consequence of Little's Law [47, 53] which is valid for general arrival processes.

The above equations, Eq (2), can be used to determine the mean transcriptional burst size, provided the protein burst distribution can be measured experimentally. To see this, we note that dividing the expressions for the mean mRNA and protein levels leads to

$$\frac{b}{\langle m_b \rangle} = \frac{\mu_p}{\mu_m} \frac{\langle p_s \rangle}{\langle m_s \rangle}. \tag{3}$$

Since the steady-state means $\langle m_s \rangle$ and $\langle p_s \rangle$ as well as the degradation rates μ_m and μ_p are parameters that can be measured experimentally, the above equation implies that the ratio $b/\langle m_b \rangle$ can be determined experimentally. Given $b/\langle m_b \rangle = k_p/\mu_m$, this implies that the mean protein production rate k_p can also be determined experimentally. This is an important result since it provides an approach for determining the mean protein production rate k_p that is valid for arbitrary arrival processes for mRNAs. Furthermore, the above equation implies that, if the mean of protein burst distribution b can be measured [36], then the mean transcriptional burst size $\langle m_b \rangle$ can also be determined. Thus, if we have measurements for mean mRNA and protein numbers and also the mean of protein burst distribution, then these measurements can be used to determine the degree of transcriptional bursting $\langle m_b \rangle$ as well as the parameters $\langle p_b \rangle$ and k_p. It is noteworthy that this procedure for estimating the burst parameters is valid for arbitrary stochastic processes corresponding to mRNA transcription.

We next turn to expressions for higher moments of mRNA and protein steady-state distributions. The noise in mRNA steady-state distributions is given by

$$\frac{\sigma_{m_s}^2}{\langle m_s \rangle^2} = \frac{1}{\langle m_s \rangle} + \frac{\mu_m}{k_b} + \frac{\mu_m}{2k_b} \left[K_g(\mu_m) - 1 + \frac{\sigma_{m_b}^2}{\langle m_b \rangle^2} - \left(1 + \frac{1}{\langle m_b \rangle} \right) \right], \tag{4}$$

where $\sigma_{m_b}^2$ is the variance of mRNA burst distribution and $K_g(\mu_m)$ is the so-called gestation factor,

$$K_g(\mu_m) = 1 + 2 \left[\frac{f_L(\mu_m)}{1 - f_L(\mu_m)} - \frac{k_b}{\mu_m} \right], \tag{5}$$

with $f_L(s)$ denoting the Laplace transform of arrival time distribution of mRNA bursts. The function $K_g(\mu_m)$ encodes information about the arrival process. Specifically, we note that for Poisson arrivals, we have $K_g(\mu_m) = 1$.

For proteins (in the burst limit $\mu_m \gg \mu_p$), we obtain [43]

$$\frac{\sigma_{p_s}^2}{\langle p_s \rangle^2} = \frac{1}{\langle p_s \rangle} + \frac{\mu_p}{k_b} + \frac{\mu_p}{2k_b}\left[K_g(\mu_p) - 1 + \frac{\sigma_{m_b}^2}{\langle m_b \rangle^2} - \left(1 - \frac{1}{\langle m_b \rangle}\right) + \left(\frac{\sigma_{p_b}^2}{\langle p_b \rangle^2} - \left(1 + \frac{1}{\langle p_b \rangle}\right)\right)\frac{1}{\langle m_b \rangle}\right] \quad (6)$$

where $K_g(\mu_p)$ is given by Eq (5) and $\sigma_{p_b}^2$ is the variance of protein burst distribution produced by a single mRNA. The expression for protein noise is composed of the noise term for the basic two-stage model of gene expression [40] and additive noise contributions due to: a) deviations from exponential waiting-time distribution for the arrival process, b) deviations from conditional geometric distributions for mRNA burst distributions and c) deviations from geometric distributions for protein burst distributions. For both mRNAs and proteins, the noise in steady-state distributions depends on all the moments of the burst arrival time distribution through the term K_g. Therefore, arrival processes corresponding to different kinetic schemes for transcription will make different contributions to the overall noise, even if they have identical means and variances for the the burst arrival time distribution.

We note from Eq (4) that, for Poisson arrivals, i.e. $K_g = 1$, and geometrically distributed burst, i.e. $\sigma_{m_b}^2 = \langle m_b \rangle(\langle m_b \rangle - 1)$, the equations for the noise and mean have only two unknown burst parameters, k_b and $\langle m_b \rangle$. In this case, experimental measurements of the first two moments of the steady-state distribution are sufficient to estimate the burst parameters, as has been done in multiple studies. However, when the arrival process is non-Poisson or if the burst distribution deviates from a geometric distribution, measurements of the first two steady-state moments are not sufficient for estimating the burst parameters. This observation motivates the need for analytical expressions for the higher moments which we turn to next.

We now derive analytical expressions for the third moment, specifically the skewness parameter. For mRNAs, the exact expression for skewness γ_{m_s} is given by

$$\begin{aligned}
\frac{\gamma_{m_s}\sigma_{m_s}^3}{m_s} &= 1 + \langle m_s \rangle\langle m_b \rangle\mathcal{K}_1(\mu_m) + 2\langle m_b \rangle^2\mathcal{K}_2(\mu_m, \langle m_b \rangle) \\
&+ (\sigma_{m_b}^2 + \langle m_b \rangle^2 - \langle m_b \rangle)\mathcal{K}_3(\mu_m, \langle m_b \rangle) \\
&+ \frac{\langle m_b(m_b - 1)(m_b - 2)\rangle}{3\langle m_b \rangle},
\end{aligned} \quad (7)$$

where we have defined

$$\begin{aligned}
\mathcal{K}_1(\mu_m) &= K_g(2\mu_m) - K_g(\mu_m), \\
\mathcal{K}_2(\mu_m, \langle m_b \rangle) &= \frac{K_g(\mu_m) - 1}{4}\left(\frac{3}{\langle m_b \rangle} + K_g(2\mu_m) - 1\right), \\
\mathcal{K}_3(\mu_m, \langle m_b \rangle) &= \frac{3}{2\langle m_b \rangle} + \frac{K_g(\mu_m) + K_g(2\mu_m)}{2} - 1.
\end{aligned} \quad (8)$$

For proteins, we obtain in the burst-limit ($\mu_m \gg \mu_p$),

$$\begin{aligned}
\frac{\gamma_{p_s}\sigma_{p_s}^3}{p_s} &= 1 + (A_1^p)^2\left[\frac{\langle p_s \rangle}{b}\mathcal{K}_1(\mu_p) + 2\mathcal{K}_2(\mu_p, A_1^p)\right] \\
&+ A_2^p\mathcal{K}_3(\mu_p, A_1^p) + \frac{A_3^p}{3A_1^p},
\end{aligned} \quad (9)$$

where the functions $\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_2$ are given by Eq (8), A_k^p is defined by $A_k^p = d^k A^p(z)/dz^k|_{z=1}$ and, using Eq (1) [54], we obtain the parameters A_k^p as:

$$
\begin{aligned}
A_1^p &= \langle m_b \rangle \langle p_b \rangle, \\
A_2^p &= \langle m_b \rangle (\sigma_{p_b}^2 - \langle p_b \rangle) + (\sigma_{m_b}^2 + \langle m_b \rangle^2) \langle p_b \rangle^2, \\
A_3^p &= \langle p_b \rangle^3 \langle m_b(m_b - 1)(m_b - 2) \rangle + 3 \langle m_b(m_b - 1) \rangle \langle p_b \rangle \\
&\quad \langle p_b(p_b - 1) \rangle + \langle m_b \rangle \langle p_b(p_b - 1)(p_b - 2) \rangle.
\end{aligned}
\tag{10}
$$

Similarly, expressions for higher order moments of protein and mRNA steady-state distributions can be obtained iteratively. The corresponding expressions for the kurtosis are provided in the S1 Text.

The analytical results derived above for proteins are exact in the burst limit, which assumes that proteins are produced instantaneously from all the mRNAs in a burst. Going beyond the burst limit (i.e. not limited to $\mu_m \gg \mu_p$), exact results for the higher moments of the protein steady-state distribution will, in general, depend on the details of the kinetic scheme for gene expression. However, we can derive approximate analytical expressions for general schemes by requiring that: a) the results reduce to the exact results in the burst limit and b) they match the exact results for the two-stage model of gene expression. For the two-stage model, exact results for the first four moments have been derived by Bokes et. al [55]. Comparing these exact results with our results derived in the burst limit, we observe that results of [55] can be reproduced by a suitable scaling of the burst-size parameters A_k^p. For example, the exact expression for the noise is obtained by the following scaling [43]:

$$
\left(\frac{\sigma_{p_s}^2}{\langle p_s \rangle^2} - \frac{1}{\langle p_s \rangle} \right) \rightarrow \left(\frac{\sigma_{p_s}^2}{\langle p_s \rangle^2} - \frac{1}{\langle p_s \rangle} \right) \frac{1}{1 + \frac{\mu_p}{\mu_m}}.
\tag{11}
$$

Similarly, for the expression for skewness, the parameters A_2^p and A_3^p are scaled as:

$$
A_2^p \rightarrow A_2^p \frac{1}{1 + \frac{\mu_p}{\mu_m}} \quad \text{and} \quad A_3^p \rightarrow A_3^p \frac{1}{\left(1 + \frac{\mu_p}{\mu_m}\right)\left(1 + 2\frac{\mu_p}{\mu_m}\right)}.
\tag{12}
$$

As shown in Fig 2 (for the random telegraph model) the analytical expressions using this approach are in good agreement with results from simulations [56].

It is noteworthy that the results derived are valid for a general class of kinetic schemes of gene expression. For a specific kinetic scheme, we can determine the corresponding waiting-time distribution for the arrival process and the burst distributions for mRNA and proteins. Substituting these results in the equations derived leads to the corresponding expressions for the moments of the steady-state distribution. The results obtained can thus provide insight into how specific kinetic schemes of gene expression (e.g. combining promoter-based regulation and post-transcriptional regulation) can be used to impact the noise and higher moments of steady-state distributions.

Estimation of burst parameters

The results derived for the steady-state moments indicate that, if the burst arrival process is not a Poisson process, then it is no longer accurate to estimate burst parameters based on measurements of mean and variance only, as has been done in previous studies [13]. In the following, we present approaches for estimating burst parameters in the general case.

We begin by considering the general kinetic scheme shown in Fig 3. This form for the kinetic scheme is supported by recent experiments in mammalian cells which suggest the presence of
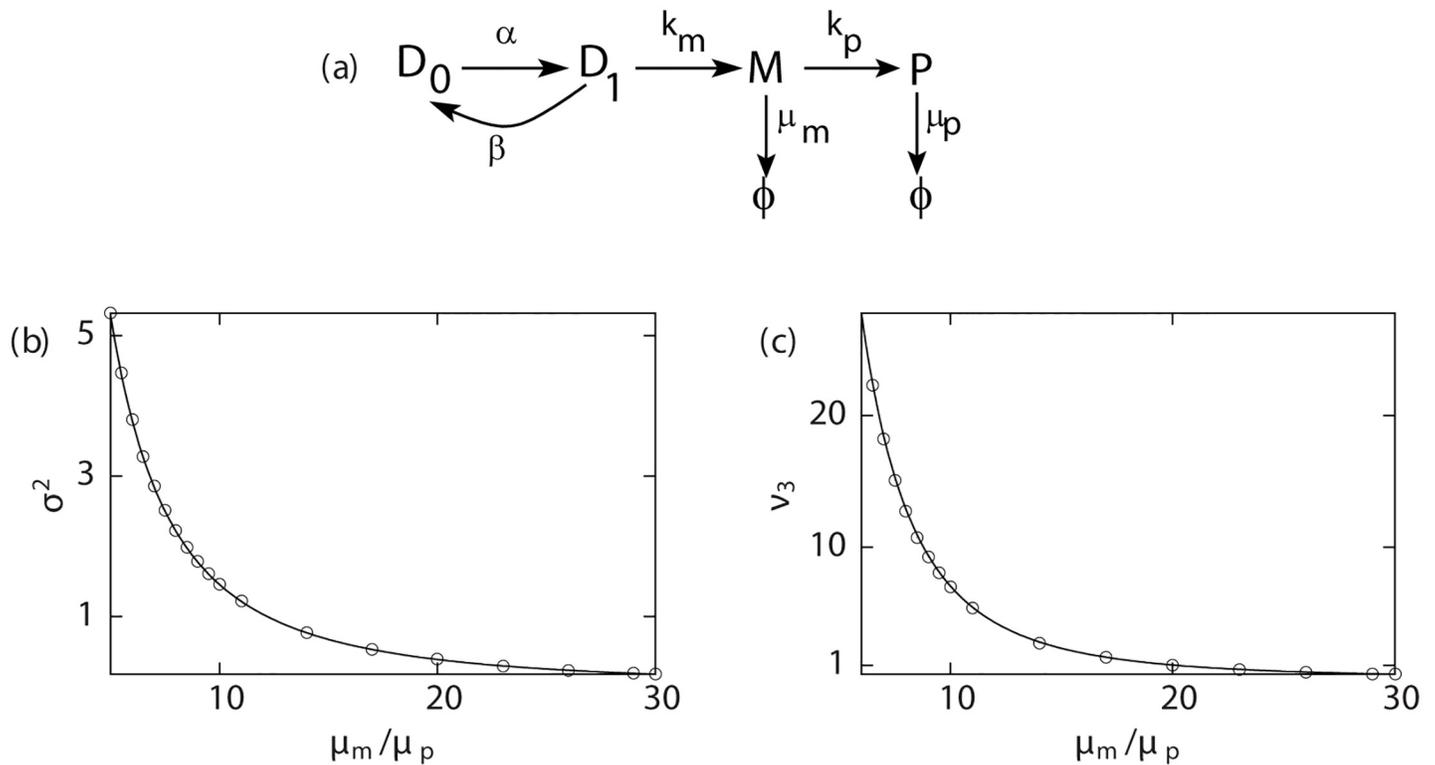
Fig 2. Steady state moments for proteins. (a) Kinetic scheme for the two-state random telegraph model. For this model, steady state variance (scaled by 10^{-5}) and third central moment v_3 (scaled by 10^{-6}) of proteins as a function of μ_m/μ_p are plotted in (b) and (c) respectively: lines represent analytic estimates and points correspond to the simulation results. Parameters are: $\alpha = 0.5$, $\beta = 0.25$, $k_m = 2$, $\langle m_b \rangle = 5$, $k_p = 0.5$.

doi:10.1371/journal.pcbi.1004292.g002

multiple rate-limiting steps between transition of the promoter from OFF to ON state [45, 57]. However, as observed in these experiments, a promoter in the ON state switches to the OFF state by a single rate-limiting step. We model the promoter switching from OFF to ON state by a general waiting-time distribution, $g(t)$. The switching rate from ON to OFF state is given by β.
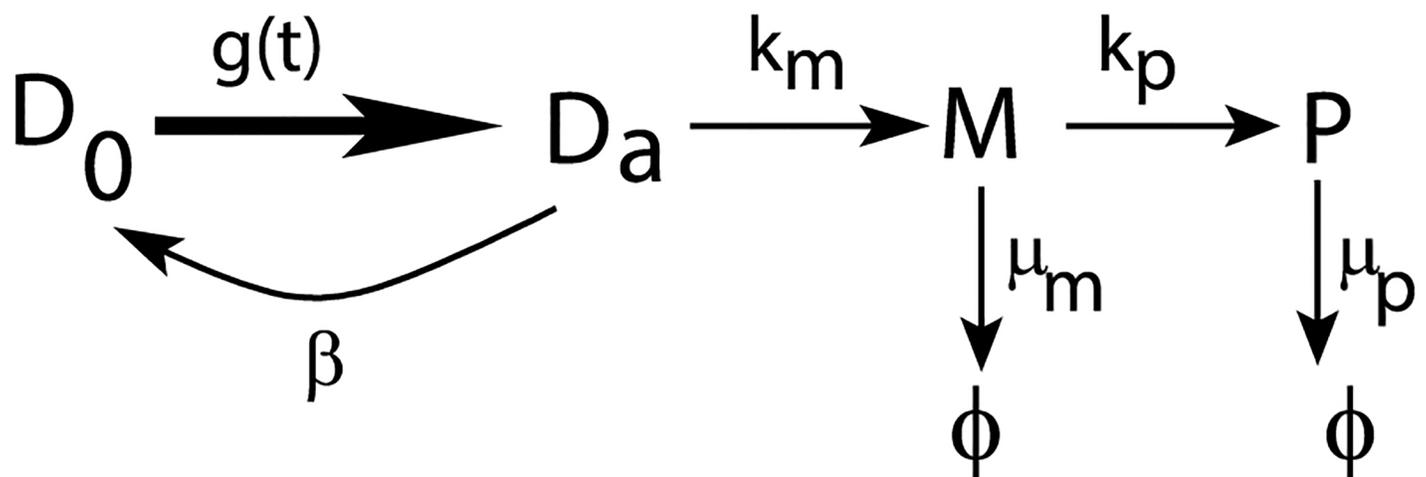
Fig 3. Schematic representation of the general kinetic scheme with promoter switching. Thick line from inactive state D_0 to active state D_a represents a general kinetic scheme with $g(t)$ as the waiting-time distribution for the promoter to switch to the ON state.

doi:10.1371/journal.pcbi.1004292.g003

Burst parameters from the sequence-size function. To extract burst parameters for the general scheme considered above, we first note that bursts are generated due to the interplay of two time scales, one that corresponds to production of mRNAs (when the gene is active) while the other one corresponds to the waiting-time between production events (when the gene is in inactive state). For bursty gene expression, we expect a clear separation of time-scales between the characteristic time periods for these two cases. Following [41], it is convenient to define a sequence-size function,

$$\phi(\tau) = \frac{1}{1 - \int_0^\tau f(t)dt},$$ (13)

where $f(t)$ is the waiting-time distribution for the arrival of a *single* mRNA starting with the promoter in the ON state. For a fixed τ, the sequence-size function can be used to categorize time intervals larger than τ as separating bursts. Correspondingly, the term $1 - \int_0^\tau f(t)dt$ represents the fraction of all mRNA arrivals that correspond to the arrivals produced in a single burst; thus ϕ provides the corresponding mean burst size. For bursty gene expression with a separation of time-scales, for a specific choice of $\tau = \tau_x$, the sequence-size function can be related to the actual mean burst size. If $f(t)$ can be measured, then determination of τ_x can result in accurate estimates of the burst parameters such as mean burst size and frequency. In the following, we discuss how to determine τ_x for the general class of arrival processes considered in Fig 3.

The key insight is based on the observation that, due to the separation of time scales within bursts and between consecutive bursts, determination of τ_x can be done by using a simple two-state model as shown in Fig 2a. Even though the actual waiting time distribution between bursts ($g(t)$) may differ from the exponential distribution for the two-state model, the short-time behavior of the sequence-size function will be indistinguishable between the two cases (given separation of time-scales). If τ_x can be connected to the short-time behavior, then analytical expressions for the sequence-size function $\phi(\tau)$ for the two-state model can be used to estimate τ_x and thereby the mean burst size.

For the two-state model, we find that burst size can be determined using a specific τ_x, which corresponds to an inflexion point where the curvature of $\phi(\tau)$ changes its sign. Specifically, for the two-state model, we obtain $f(t)$ by taking inverse Laplace transform of $f(s)$ given by Eq (23). In the burst-limit, i.e., $\alpha/\beta \to 0$, we find that the sequence-size function, using Eq (13), is given by

$$\phi(\tau) = \frac{(k_m + \beta)e^{\tau(k_m+\beta)}}{k_m + \beta e^{\tau(k_m+\beta)}},$$ (14)

and the value of τ at which $\phi(\tau)$ exhibits inflexion is

$$\tau_x = \frac{1}{k_m + \beta} \ln \frac{k_m}{\beta}, \qquad k_m > \beta.$$ (15)

The sequence size function $\phi(\tau)$ at this point ($\tau = \tau_x$) is given by:

$$\phi(\tau_x) = \frac{1}{2}\left(1 + \frac{k_m}{\beta}\right) = \frac{1}{2}(1 + \langle m_b \rangle),$$ (16)

Thus, the procedure for determination of the mean burst size $(1 + \langle m_b \rangle)$, given $f(t)$, is as follows:

1. Obtain the sequence-size function $\phi(\tau)$ from $f(t)$. For bursty synthesis, $\phi(\tau)$ will have an inflexion point.

2. The mean burst size $(1 + \langle m_b \rangle)$ is simply twice the value of the the sequence-size function $\phi(\tau)$ at the inflexion point, τ_x.

This approach has been validated using stochastic simulations for multiple promoter models with correspondingly complex waiting-time distributions between bursts.

Estimation of $f(t)$ from steady-state moments.   The procedure outlined in the previous section assumes that the waiting-time distribution $f(t)$ can be determined. However, this can be challenging experimentally, thus it is desirable to develop approaches for estimating $f(t)$ based on measurements of steady-state distributions.

To proceed in this direction, let us first obtain a relation connecting the two waiting-time distributions $f(t)$ (for single mRNA arrival) and $g(t)$ (for burst arrival). In Fig 3, we note that when the promoter is in the active state, D_a, it can make multiple trips to D_0 before producing mRNA. Whenever gene is in D_a state, it can either create mRNA or can switch back to D_0 state. Gene in D_a state can produce mRNA either in a single step, i.e., without switching back to D_0 state, or by making multiple trips to D_0 before producing mRNA. Denoting the number of trips made before producing mRNA by q, we obtain that the Laplace transform of the waiting-time distribution $f(t)$ is given by

$$f_L(s) = \frac{k_m}{\beta + k_m} \sum_{q=0}^{\infty} \left(\frac{\beta}{\beta + k_m} \right)^q [g_L(s)]^q \left(\frac{k_m + \beta}{k_m + \beta + s} \right)^{q+1}, \tag{17}$$

which leads to:

$$f_L(s) = \frac{k_m}{k_m + s + [1 - g_L(s)]\beta}. \tag{18}$$

In order to determine $f_L(s)$, we will assume a specific functional form for $g_L(s)$. We consider that $g_L(s)$ is given by the following rational function,

$$g_L(s) \equiv g_n^m(s) = \frac{1 + a_1 s + a_2 s^2 \cdots a_m s^m}{1 + b_1 s + b_2 s^2 \cdots b_n s^n}, \quad n > m. \tag{19}$$

This form for the Laplace transform of the waiting-time distribution is consistent with known waiting-time distributions for phase-type processes [54] and thus is valid quite generally.

Once we have an explicit form for $f_L(s)$, the next step is to determine the parameters, k_m, β, $a_1 \ldots a_m$, and $b_1 \ldots b_n$. Thus, in general, we need $m+n+2$ measurements to estimate these parameters if we use $g(s) = g_n^m(s)$. The simplest case, $g_L(s) = g_1^0(s)$, implies the presence of one kinetic step from inactive state to active state, with rate $1/b_1$, and so it corresponds to the standard two-state random telegraph model. For this simple kinetic scheme, we can find the parameters, k_m, β, and b_1, and hence $f_L(s)$ and the sequence size function by using three measurements associated with either mRNAs or proteins.

The form, $g_L(s) = g_1^0(s)$, is exact for the two-state random telegraph model. Using the expressions obtained for the first four steady-state moments, we can derive an analytic condition that determines whether the underlying mechanism can be represented by $g_1^0(s)$ (see Supplementary S2 Text). However, if the arrival process is complex and involves multiple rate-limiting steps, then $g_1^0(s)$ will not be an accurate representation of the underlying kinetic process. In such cases, we need to use $g_L(s)$ of higher order. The next step in this iterative process is to take $g_L(s) = g_2^0(s)$. This form of $g_L(s)$ is valid if there are only two rate-limiting steps in the promoter transition from OFF to ON state. For kinetic schemes that involve more than two steps, it will serve as an approximate reduced representation. Interestingly, it turns out that even if $g_L(s) = g_2^0(s)$ is not a correct representation of the underlying kinetic process, this

reduced representation works very well as far as estimating burst size is concerned. In Fig 4, we have illustrated the effectiveness of this approach for a complex kinetic scheme for the promoter transition from OFF to ON state. The figure also illustrates the effectiveness of the approach outlined in the previous subsection for determining the mean burst size using the sequence-size function $\phi(\tau)$.

While the reduced representation, $g_L(s) = g_2^0(s)$, works reasonably well for estimating burst size, with additional data, it is possible to extend the process further. The iterative procedure we propose is as follows:

1. Start with the simplest form $g_1^0(s)$ and use three moments associated with either mRNA or proteins (or both) to find $f_L(s)$ as discussed above. Then this $f_L(s)$ can be used to get analytic predictions for higher moments [48].

2. If these analytic predictions are consistent with the corresponding experimental observations then $g_1^0(s)$ provides a reasonable representation of the underlying kinetic scheme, else a representation using more complex kinetic schemes is required.

3. To address more complex kinetic schemes, we iteratively change $g_L(s)$ from $g_1^0(s)$ to $g_2^0(s)$...and so on, and iterate the steps outlined to determine the underlying $f_L(s)$. However, we note that for uncovering more complex kinetic scheme we need additional measurements to estimate $f_L(s)$. If moment measurements are possible at different mRNA/protein degradation rates, then these additional measurements can be used to estimate $f_L(s)$ and hence the corresponding mean transcriptional burst size.

Effect of extrinsic noise on burst estimation. The burst estimation approach discussed in the preceding section assumes that the dominant contribution comes from intrinsic sources of fluctuations. However, extrinsic noise [1, 12], e.g. arising from different concentration of cellular components such as RNA polymerase, can also contribute significantly to the observed variations. It is thus of interest to examine how the proposed burst parameter estimation procedure works if we also consider sources of extrinsic noise.

To explore the effects of such fluctuations, we consider the model shown in Fig 5. In this kinetic scheme, the activation of gene from OFF to ON state involves two sequential steps, with rates α_1 and α_2. To include extrinsic fluctuations in the model, we consider that the rate of transcription k_m is a Log-normally distributed random variable with mean $\langle k_m \rangle$ and standard deviation σ_{k_m}. For a given value of σ_{k_m}, we determine the mean burst size following the procedure outlined above: i.e. by taking $g(s) = g_2^0(s)$ and then using the simulation values for the first four steady-state moments of mRNAs to estimate the unknown parameters (b_1, b_2, k_m, β), and hence the burst size. By varying σ_{k_m} we study how the estimated burst size $\langle m_b \rangle_\sigma$ deviates from the one without extrinsic noise, $\langle m_b \rangle_0$. As can be seen in Fig 5, for smaller values of σ_{k_m}, the estimated burst size $\langle m_b \rangle_\sigma$ is reasonably close to $\langle m_b \rangle_0$, however, as expected, $\langle m_b \rangle_\sigma$ shows monotonic deviations from $\langle m_b \rangle_0$ for larger values of σ_{k_m}.

Signatures for non-Poisson arrivals

The analytical expressions derived for the steady-state moments for mRNAs and proteins can also be used to make inferences about the burst arrival process based on steady-state measurements. Since multiple studies assume that the burst arrival process is characterized by an exponential waiting-time distribution, it would be useful to determine if this assumption is invalid using measurements of steady-state distributions. As shown below, we can obtain conditions for the same using the results derived for higher moments.
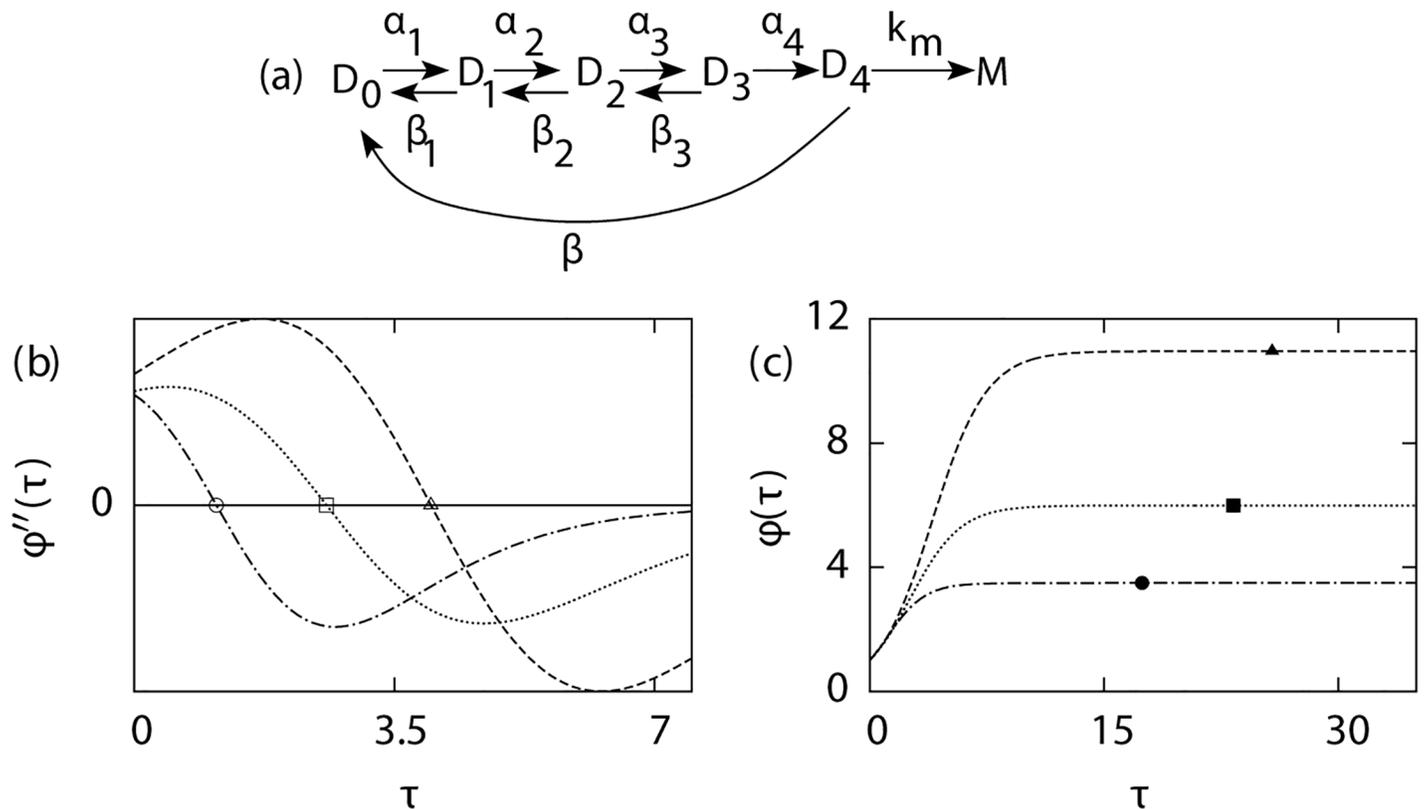
Fig 4. Estimation of mean burst size from sequence size function $\phi(\tau)$. For the transcriptional scheme shown in (a), the variations of $\phi''(\tau)$ and $\phi(\tau)$ as a function of time τ (scaled by 10^3) are shown in (b) and (c) respectively. The three lines correspond to three different values of β, 50 (dashed line), 100 (dotted line) and 200 (dashed-dotted line), while keeping $k_m = 500$: Exact burst size for these three cases are 11, 6 and 3.5, respectively. Estimated mean burst size has been indicated by filled symbols and the inflexion points in the sequence size function are shown by empty symbols. Other parameters: $\alpha_1 = 1, \alpha_2 = 0.5, \alpha_3 = 0.25, \alpha_4 = 0.75, \beta_1 = 0.1, \beta_2 = 0.2, \beta_3 = 0.5$.

In the following, we will first focus on the cases that the mRNA burst distribution is conditional geometric and the protein burst distribution is geometric, which is consistent with multiple experimental observations. As discussed, choosing the conditional geometric distribution for mRNAs allows us to consider both single mRNA arrivals and geometric mRNA bursts in one framework. Since experiments can provide measurements of both mRNA and protein steady-state distributions, it is useful to have conditions for the arrival process using either mRNA data or protein data or both mRNA and protein data. Based on these three possibilities, we present three different conditions in the following.

Using moments of mRNA steady-state distributions. Let us first consider the case where we have only measurements of the mRNA steady-state distribution. We note that for Poisson arrivals $K_g(\mu_m) = K_g(2\mu_m) = 1$, and using the expressions for mean and noise from Eqs (2) and (4) we get, $F_m = \langle m_b \rangle$, where $F_m = \sigma_{m_s}^2 / \langle m_s \rangle$ is the mRNA Fano factor. Further, using this in the equation for skewness, Eq (7), we derive the following condition that must be satisfied if the arrival of mRNA bursts is a Poisson process:

$$\mathcal{D}_m \equiv \frac{\gamma_{m_s} \sigma_{m_s}^3}{\langle m_s \rangle \left[3(F_m - 1)\{1 + \frac{2}{3}(F_m - 1)\} + 1 \right]} - 1 = 0. \tag{20}$$

Thus $\mathcal{D}_m \neq 0$ is a signature of non-Poisson arrival processes. Since the above prescription is
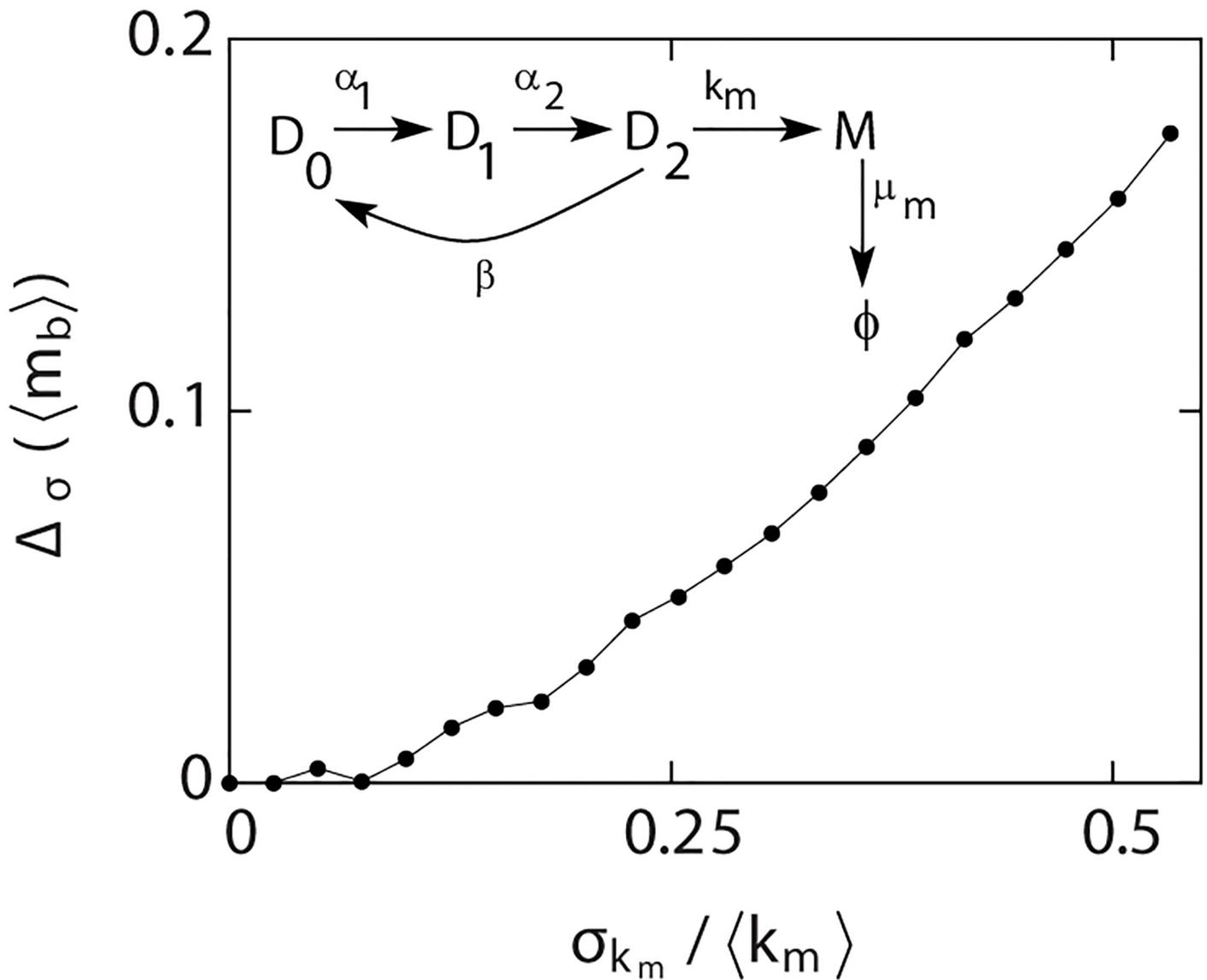
Fig 5. Effects of extrinsic noise on burst estimation. For the transcriptional scheme shown in the inset, the relative error $\Delta_\sigma(\langle m_b \rangle) = (\langle m_b \rangle_0 - \langle m_b \rangle_\sigma)/\langle m_b \rangle_0$ is plotted. Parameters as $\alpha_1 = 1$, $\alpha_2 = 0.5$, $\beta = 50$, $\langle k_m \rangle = 500$ and $\mu_m = 1$.

based on experimentally measurable quantities such as $\langle m_s \rangle$, $\sigma^2_{m_s}$, γ_{m_s} and μ_m, it can be used to determine if the assumption of a Poisson arrival process is invalid.

Using moments of protein steady-state distributions. We next consider the case where we have access to only the protein steady-state distribution. The steps followed are similar to those outlined for the mRNA case. For Poisson arrivals, $K_g(\mu_p) = 1$, and using Eqs (2) and (11) we get

$$b = (F_p - 1)\left(1 + \frac{\mu_p}{\mu_m}\right),$$

where $F_p = \sigma^2_{p_s}/\langle p_s \rangle$ is the protein Fano factor. Substituting this in the expression for protein

skewness, Eq (9) with the scaled A_2^p and A_3^p given by Eq (12), we arrive at the following condition for Poisson arrivals.

$$\mathcal{D}_p \equiv \frac{\gamma_{p_s}\sigma_{p_s}^3}{\langle p_s \rangle \left[3(F_p - 1)\left\{ 1 + \frac{2}{3}\left(\frac{\mu_m + \mu_p}{\mu_m + 2\mu_p} \right)(F_p - 1) \right\} + 1 \right]} - 1$$

$$= 0 \qquad (21)$$

Again, non-zero value of $\mathcal{D}_p$ is a signature of non-Poisson arrivals.

Using both mRNA and protein steady-state distributions. Finally, if we have both mRNA and protein steady-state distribution measurements available, then the condition for Poisson arrivals can be obtained by combining measurements of second moments of mRNA and protein distributions as follows: Using Eqs (2),(4) and (11), we get,

$$\mathcal{D}_{mp} \equiv \frac{F_m}{\langle m_s \rangle \mu_m} - \frac{(\mu_m + \mu_p)(F_p - 1)}{\mu_m \langle p_s \rangle \mu_p}$$

$$= \frac{1}{2k_b}\left[K_g(\mu_m) - K_g(\mu_p) \right], \qquad (22)$$

which vanishes for Poisson arrival of mRNA bursts. Thus non-zero values of $\mathcal{D}_{mp}$ indicate non-Poisson arrival of mRNA bursts. Interestingly, for this condition there is no need to assume that the mRNA burst distribution is geometric. That is, the condition holds true for arbitrary mRNA burst distributions. Also, the condition does not require measurement of third moments.

Signatures for a simple kinetic scheme. To illustrate the prescription derived for determining non-Poisson arrival processes, we consider a specific kinetic scheme, Fig 2a. For this kinetic scheme, the mRNA arrival time distribution in the Laplace domain is given by (Eq (S3–9) Supplementary S3 Text)

$$f_L(s) = \frac{k_m(\alpha + s)}{k_m(\alpha + s) + s(\alpha + \beta + s)}. \qquad (23)$$

Using this in Eq (5) we find the gestation factor, K_g, and hence the mean, Fano factor and skewness for both mRNAs and proteins. Finally, we derive exact analytic expressions for $\mathcal{D}_m$, $\mathcal{D}_p$ and $\mathcal{D}_{mp}$ from Eqs (20), (21) and (22) respectively. The expression for $\mathcal{D}_m$ reads

$$\mathcal{D}_m = \frac{2k_m\beta(1 - \langle m_b \rangle)\theta\left(1 + \frac{(1+\alpha)(\alpha+\beta)\langle m_b \rangle k_m}{\theta(\langle m_b \rangle - 1)} \right)}{(2 + \alpha + \beta)(\theta + \beta k_m)((2\langle m_b \rangle - 1)\theta + 2k_m\langle m_b \rangle\beta)}, \qquad (24)$$

where

$$\theta = (\alpha + \beta)(\alpha + \beta + 1), \qquad (25)$$

and we have set $\mu_m = 1$ for simplicity. As expected, we note that $\mathcal{D}_m$ vanishes for the Poisson arrival processes, i.e., either when β is zero, or when the switching rates α and β are very large compared to the rate of transcription, k_m. The general expression for $\mathcal{D}_p$ is complicated. However, to gain insight about the arrival process, we can write down a simpler expression for $\mathcal{D}_p$ in
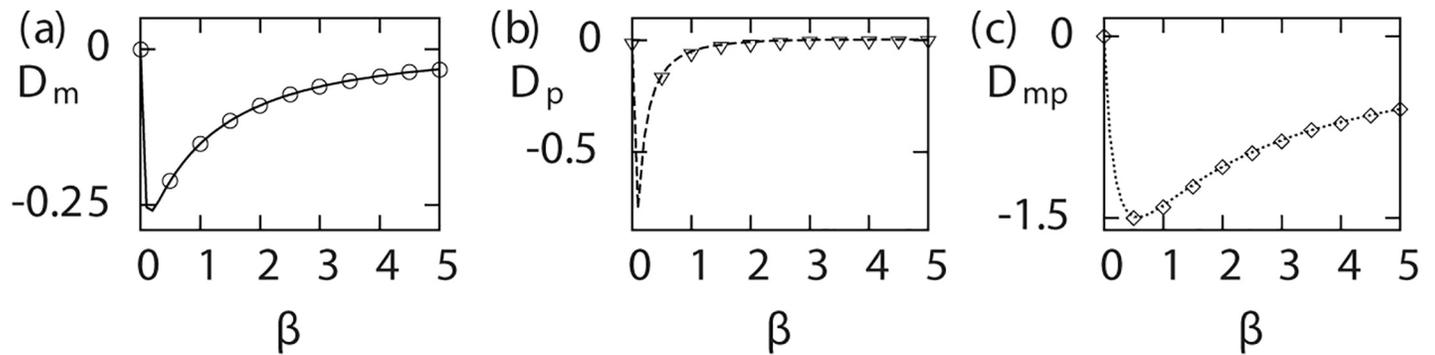
Fig 6. Signatures for non-Poisson arrival. The quantities $\mathcal{D}_m$, $\mathcal{D}_p$ and $\mathcal{D}_{mp}$ are plotted for the model shown in Fig 2a as a function of *off* rate β. Analytic estimates are shown by lines whereas points correspond to the simulation results with parameters: $\alpha = 0.25$, $k_m = 2$, $\langle m_b \rangle = 5$, $k_p = 0.5$, $\mu_m = 1$, $\mu_p = 0.01$.

the burst limit, $\mu_m = 1 \gg \mu_p$:

$$\mathcal{D}_p = -\frac{2\langle m_b \rangle^2 k_m^2 k_p^2 \alpha \beta}{(\alpha + \beta)^4 + 3\langle m_b \rangle k_p (\alpha + \beta)^2 \psi + 2\langle m_b \rangle^2 k_p^2 \psi^2}, \tag{26}$$

where

$$\psi = k_m \beta + (\alpha + \beta)^2. \tag{27}$$

Again, for Poisson arrival processes $\mathcal{D}_p$ vanishes. Finally, we obtain an analytic expression for $\mathcal{D}_{mp}$, which is given by

$$\mathcal{D}_{mp} = \frac{\beta(\mu_p - \mu_m)}{\alpha(\alpha + \beta + \mu_m)(\alpha + \beta + \mu_p)}, \tag{28}$$

and as expected, we note that $\mathcal{D}_{mp}$ vanishes for Poisson arrivals and is negative for $\mu_p < \mu_m$. In Fig 6, we have plotted the three analytic expressions together with simulation results as a function of β.

Signatures for non-geometric bursts

As discussed in the previous section, it is widely assumed that the mRNA burst distribution can be represented by a conditional geometric distribution (i.e. including both single mRNA arrivals and geometrically distributed burst arrivals). While this assumption is consistent with multiple experimental observations, for general kinetic schemes the possibility of non-geometric mRNA burst distributions has to be considered. Thus, it is of interest to examine if the results obtained can be used to determine if the mRNA burst distribution deviates from a conditional geometric distribution.

To address the possibility of non-geometric mRNA burst distributions, let us first consider that the random variable corresponding to the mRNA burst distribution (m_b) has a conditional geometric distribution. That is, the probability that a burst produces n mRNA molecules is given by

$$\mathcal{P}(m_b = n) = (1 - p)^{n-1} p, \tag{29}$$

where $0 < p \leq 1$, and $n = 1, 2, 3 \ldots \infty$. This distribution leads to

$$\sigma_{mb}^2 = \langle m_b \rangle (\langle m_b \rangle - 1). \tag{30}$$

Using Eqs ([2]) and ([30]) in Eq ([4]), and denoting $F_m = \sigma^2_{m_s}/\langle m_s \rangle$ as the Fano factor of mRNA copy number, Eq ([4]) can be rewritten as

$$F_m(\mu_m) = \frac{\langle m_b \rangle}{2} \left[1 + K_g(\mu_m) \right]. \tag{31}$$

Similarly, using the burst size distribution from Eq ([29]), the skewness in Eq ([7]) is given by

$$\begin{aligned} \frac{\gamma_{m_s} \sigma^3_{m_s}}{\langle m_s \rangle} &= 1 + \langle m_s \rangle \langle m_b \rangle \mathcal{K}_1(\mu_m) + 2\langle m_b \rangle^2 \mathcal{K}_2(\mu_m, \langle m_b \rangle) \\ &+ 2(\langle m_b \rangle - 1)[(1 + \mathcal{K}_3(\mu_m, \langle m_b \rangle))\langle m_b \rangle - 1]. \end{aligned} \tag{32}$$

We note that Eq ([32]) connects experimentally measurable moments of the steady-state distribution to the parameters $K_g(\mu_m)$, $K_g(2\mu_m)$ and $\langle m_b \rangle$. Furthermore, note that Eq ([31]) can be recast as $K_g(\mu_m) = (2F_m(\mu_m)/\langle m_b \rangle) - 1$. Now, considering a change in the degradation rate from μ_m to $2\mu_m$ (keeping the mean burst size, $\langle m_b \rangle$ invariant), we obtain

$$K_g(2\mu_m) = (2F_m(2\mu_m)/\langle m_b \rangle) - 1. \tag{33}$$

Using the above in Eq ([32]), we get an expression connecting experimentally measurable quantities associated with moments of the mRNA steady-state distribution. The resulting expression is:

$$\begin{aligned} \mathcal{G}_m &\equiv \frac{\gamma_{m_s} \sigma^3_{m_s} \langle m_s \rangle^{-1}}{2F_m(2\mu_m)(\langle m_s \rangle - 1) + F_m(1 - 2\langle m_s \rangle + 2F_m(2\mu_m))} \\ &= 1. \end{aligned} \tag{34}$$

We note that the above expression has been derived by making just one assumption, namely, the mRNA burst distribution is a conditional geometric distribution. The derived expression thus indicates that a combination of experimentally measurable quantities has to deviate from 1 if the mRNA burst distribution deviates from a conditional geometric distribution. Thus the analytical results derived provide a signature for deviation from conditional geometric mRNA bursts using measurements of the first three moments of the mRNA steady-state distribution.

The main requirement for using the above relation is that measurements of mRNA steady-state distribution can be carried out at two different rates of the mRNAs μ_m and $2\mu_m$. Given that mRNA degradation rates can be tuned experimentally, a straightforward strategy to ensure that the degradation rate is tuned to twice the original value ($2\mu_m$) is to compare the mean mRNA levels at μ_m and $2\mu_m$. Given these measurements, a value of $\mathcal{G}_m \neq 1$ implies that bursts are not distributed geometrically. The strength of this result lies in the fact that it holds for general arrival processes for mRNA bursts with arbitrary waiting-time distributions.

Let us consider a specific simple model to illustrate the condition derived above. First, let the arrival process for mRNA bursts be a Poisson process. For this, arrival time distributions of mRNA bursts in the time domain, t, and in the Laplace domain, s, are given by

$$f(t) = k_b e^{-k_b t}, \qquad f_L(s) = k_b/(k_b + s), \tag{35}$$

where k_b is the rate of arrival of mRNA bursts. For the mRNA burst distribution, let us assume that it is given by the negative binomial distribution, i.e.

$$\mathcal{P}(m_b = n) = \frac{(n + r - 1)!}{n!(r-1)!} p^n (1-p)^r, \tag{36}$$

where $0 < p \leq 1$, $r \geq 1$, and $n = 0, 1, 2, 3 \ldots \infty$. For $r = 1$, the above reduces to the geometric distribution and therefore we expect $G_m = 1$ in this limit. Using the expressions for the moments derived in the previous section, we obtain an explicit expression for G_m (Supplementary S3 Text):

$$\mathcal{G}_m = \frac{1}{3}\left(-\frac{p+1}{pr+1} + \frac{4}{p(r-1)+2} + 2\right). \tag{37}$$

Notice that for the geometric bursts ($r = 1$) we get $\mathcal{G}_m = 1$, as expected. However, for non-geometric bursts, deviations of $\mathcal{G}_m$ from 1 are observed (also see Fig S3–1 in Supplementary S3 Text). Two additional examples of microscopic models for non-geometric bursts (the two state random telegraph model and a model with three promoter states where mRNAs are produced from two states) are discussed in the.

The preceding analysis can be extended to protein steady-state distributions to derive a similar condition for deviations from geometric burst distributions in terms of steady state moments associated with proteins (see Supplementary S4 Text).

Discussion

In this paper we study stochastic gene expression models with a general renewal-type arrival process for mRNAs. By mapping such a generic model of gene expression to systems studied in queueing theory, we derive analytical expressions for the moments for mRNA and protein steady-state distributions. While the focus of this work is on using approaches drawn from queueing theory, it is noteworthy that the kinetic scheme defined in Fig 1 can also be analyzed using the general theory of branching processes with immigration [58]. In future work, it would be of interest to explore potential connections between complementary approaches to such models based on branching processes and queueing theory.

While previous studies [37, 43] have focused on protein noise, in the present work we derive analytic expressions for higher order moments of both mRNA and protein steady-state distributions. For arbitrary kinetic schemes, the results obtained determine how the moments of steady-state distributions depend on model parameters. They elucidate how different sources (promoter-based regulation, transcriptional bursting, post-transcriptional regulation) combine to determine the overall noise and higher moments. Furthermore, the results derived show how parameters of interest (such as mean protein production rate k_p) can be estimated for general models (i.e. without making any assumptions about specific features of the models).

The expressions derived for the moments can also be used to infer if the arrival process for mRNAs is non-Poisson or if the mRNA burst distribution deviates from the geometric distribution. Correspondingly, we obtain analytic conditions that provide signatures for non-Poisson arrivals of mRNA bursts and for non-geometric mRNA burst distributions. These conditions involve relations between combinations of of experimentally measurable quantities and can thus be tested by using measurements of either mRNA steady-state distributions or protein steady-state distributions or both. Apart from obtaining insights into the statistics of the arrival process, we can use the results derived for steady-state moments for accurately estimating burst parameters using an iterative approach. Notably, the results and the approaches developed in this work are valid for general models of gene expression i.e., given the general assumptions made, they do not depend on the specifics of the kinetic schemes.

It is important to note that the burst parameter estimation approaches presented in this paper rely on the accurate measurements of higher order moments, such as skewness or kurtosis. This, in turn requires that we have relatively large sample sizes. For example, simulations of two state random telegraph model (see Supplementary S5 Text) indicate that for the standard

error in skewness to be below 10%, the sample size should be ~ 1000. Current experimental limitations on measurements of mRNA distributions (e.g. using RNA FISH) do not allow for such large sample sizes and thus do not lead to accurate computation of skewness or kurtosis. While accurate measurements of higher moments are not readily available in the existing data, it is hoped that our results will provide motivation for carrying out the corresponding experiments in future. The combination of these experimental results with our theoretical approaches can be used in obtaining accurate representations of the arrival process and burst parameters for a wide range of cellular systems.

Supporting Information

S1 Text. Derivation of steady-state moments for mRNAs and proteins.
(PDF)

S2 Text. Condition for the two-state random telegraph model.
(PDF)

S3 Text. Illustrative examples for condition identifying non-geometric bursts.
(PDF)

S4 Text. Condition for non-geometric bursts using protein steady-state moments.
(PDF)

S5 Text. Estimation of error in skewness due to finite sample sizes.
(PDF)

Author Contributions

Conceived and designed the experiments: NK AS RVK. Performed the experiments: NK. Analyzed the data: NK AS RVK. Contributed reagents/materials/analysis tools: NK RVK. Wrote the paper: NK RVK. Conceived and designed the research: NK AS RVK. Developed the analytical approaches: NK RVK. Carried out stochastic simulations: NK. Discussed the results at all stages and also their implications: NK AS RVK. Wrote the main paper: NK RVK. Wrote the Supplementary Information: NK. Read, commented on and edited the manuscript: NK AS RVK.

References

1. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297(5584):1183–1186. PMID: 12183631

2. Kærn M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. Nature Reviews Genetics. 2005; 6(6):451–464. doi: 10.1038/nrg1615 PMID: 15883588

3. Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. Science. 2005; 309(5743):2010–2013. doi: 10.1126/science.1105891 PMID: 16179466

4. Sanchez A, Choubey S, Kondev J. Regulation of noise in gene expression. Annual review of biophysics. 2013; 42:469–491. doi: 10.1146/annurev-biophys-083012-130401 PMID: 23527780

5. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010; 467(7312):167–173. doi: 10.1038/nature09326 PMID: 20829787

6. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008; 135(2):216–226. doi: 10.1016/j.cell.2008.09.050 PMID: 18957198

7. Larson DR. What do expression dynamics tell us about the mechanism of transcription? Current opinion in genetics & development. 2011; 21(5):591–599.

8. Junker JP, van Oudenaarden A. Every Cell Is Special: Genome-wide Studies Add a New Dimension to Single-Cell Biology. Cell. 2014; 157(1):8–11. doi: 10.1016/j.cell.2014.02.010 PMID: 24679522

9. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012; 336(6078):183–187. doi: 10.1126/science.1216379 PMID: 22499939

10. Golding I. Decision making in living cells: lessons from a simple system. Annual review of biophysics. 2011; 40:63–80. doi: 10.1146/annurev-biophys-042910-155227 PMID: 21545284

11. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. Nature genetics. 2006; 38(6):636–643. doi: 10.1038/ng1807 PMID: 16715097

12. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature. 2006; 441(7095):840–846. doi: 10.1038/nature04785 PMID: 16699522

13. Weinberger L, Voichek Y, Tirosh I, Hornung G, Amit I, Barkai N. Expression noise and acetylation profiles distinguish HDAC functions. Molecular cell. 2012; 47(2):193–202. doi: 10.1016/j.molcel.2012.05.008 PMID: 22683268

14. Kumar N, Platini T, Kulkarni RV. Exact distributions for stochastic gene expression models with bursting and feedback. Physical Review Letters. 2014; 113(26):268105. doi: 10.1103/PhysRevLett.113.268105 PMID: 25615392

15. Tsimring LS. Noise in biology. Reports on Progress in Physics. 2014; 77(2):026601. doi: 10.1088/0034-4885/77/2/026601 PMID: 24444693

16. Hinczewski M, Thirumalai D. Cellular Signaling Networks Function as Generalized Wiener-Kolmogorov Filters to Suppress Noise. Phys Rev X. 2014 Oct; 4:041017. Available from: http://link.aps.org/doi/10.1103/PhysRevX.4.041017.

17. Balázsi G, van Oudenaarden A, Collins JJ. Cellular decision making and biological noise: from microbes to mammals. Cell. 2011; 144(6):910–925. doi: 10.1016/j.cell.2011.01.030 PMID: 21414483

18. Suter DM, Molina N, Naef F, Schibler U. Origins and consequences of transcriptional discontinuity. Current opinion in cell biology. 2011; 23(6):657–662. doi: 10.1016/j.ceb.2011.09.004 PMID: 21963300

19. Coulon A, Chow CC, Singer RH, Larson DR. Eukaryotic transcriptional dynamics: from single molecules to cell populations. Nature Reviews Genetics. 2013; 14(8):572–584. doi: 10.1038/nrg3484 PMID: 23835438

20. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell. 2005; 123(6):1025—1036. Available from: http://www.sciencedirect.com/science/article/pii/S0092867405010378. doi: 10.1016/j.cell.2005.09.031 PMID: 16360033

21. Chubb JR, Trcek T, Shenoy SM, Singer RH. Transcriptional pulsing of a developmental gene. Current biology. 2006; 16(10):1018–1025. doi: 10.1016/j.cub.2006.03.092 PMID: 16713960

22. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS biology. 2006; 4(10):e309. doi: 10.1371/journal.pbio.0040309 PMID: 17048983

23. So Lh, Ghosh A, Zong C, Sepúlveda LA, Segev R, Golding I. General properties of transcriptional time series in Escherichia coli. Nature genetics. 2011; 43(6):554–560. doi: 10.1038/ng.821 PMID: 21532574

24. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010; 329(5991):533–538. doi: 10.1126/science.1188308 PMID: 20671182

25. Zong C, So Lh, Sepúlveda LA, Skinner SO, Golding I. Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene. Molecular systems biology. 2010; 6(1). doi: 10.1038/msb.2010.96 PMID: 21119634

26. Sanchez A, Golding I. Genetic determinants and cellular constraints in noisy gene expression. Science. 2013; 342(6163):1188–1193. doi: 10.1126/science.1242975 PMID: 24311680

27. Dar RD, Razooky BS, Singh A, Trimeloni TV, McCollum JM, Cox CD, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proceedings of the National Academy of Sciences. 2012; 109(43):17454–17459. doi: 10.1073/pnas.1213530109

28. Singh A, Razooky BS, Dar RD, Weinberger LS. Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. Molecular systems biology. 2012; 8(1).

29. Gefen O, Gabay C, Mumcuoglu M, Engel G, Balaban NQ. Single-cell protein induction dynamics reveals a period of vulnerability to antibiotics in persister bacteria. Proceedings of the National Academy of Sciences. 2008; 105(16):6145–6149. doi: 10.1073/pnas.0711712105

30. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell. 2005; 122(2):169–182. doi: 10.1016/j.cell.2005.06.006 PMID: 16051143

31. Zeng L, Skinner SO, Zong C, Sippy J, Feiss M, Golding I. Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection. Cell. 2010; 141(4):682—691. Available from: http://

www.sciencedirect.com/science/article/pii/S0092867410003521. doi: 10.1016/j.cell.2010.03.034 PMID: 20478257

32. Wernet MF, Mazzoni EO, Çelik A, Duncan DM, Duncan I, Desplan C. Stochastic spineless expression creates the retinal mosaic for colour vision. Nature. 2006; 440(7081):174–180. doi: 10.1038/nature04615 PMID: 16525464

33. Ochiai H, Sugawara T, Sakuma T, Yamamoto T. Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. Scientific reports. 2014;4. doi: 10.1038/srep07125

34. Senecal A, Munsky B, Proux F, Ly N, Braye FE, Zimmer C, et al. Transcription factors modulate c-Fos transcriptional bursts. Cell reports. 2014; 8(1):75–83. doi: 10.1016/j.celrep.2014.05.053 PMID: 24981864

35. Cai L, Friedman N, Xie XS. Stochastic protein expression in individual cells at the single molecule level. Nature. 2006; 440(7082):358–62. doi: 10.1038/nature04599 PMID: 16541077

36. Yu J, Xiao J, Ren X, Lao K, Xie XS. Probing gene expression in live cells, one protein molecule at a time. Science. 2006; 311(5767):1600–1603. doi: 10.1126/science.1119623 PMID: 16543458

37. Pedraza JM, Paulsson J. Effects of molecular memory and bursting on fluctuations in gene expression. Science. 2008; 319(5861):339–343. doi: 10.1126/science.1144331 PMID: 18202292

38. Zhang J, Zhou T. Promoter-mediated Transcriptional Dynamics. Biophysical journal. 2014; 106(2):479–488. doi: 10.1016/j.bpj.2013.12.011 PMID: 24461023

39. Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. Theoretical population biology. 1995; 48(2):222–234. doi: 10.1006/tpbi.1995.1027

40. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences. 2008; 105(45):17256–17261. doi: 10.1073/pnas.0803850105

41. Dobrzyński M, Bruggeman FJ. Elongation dynamics shape bursty transcription and translation. Proceedings of the National Academy of Sciences. 2009; 106(8):2583–2588. doi: 10.1073/pnas.0803507106

42. Skupsky R, Burnett JC, Foley JE, Schaffer DV, Arkin AP. HIV promoter integration site primarily modulates transcriptional burst size rather than frequency. PLoS computational biology. 2010; 6(9):e1000952. doi: 10.1371/journal.pcbi.1000952 PMID: 20941390

43. Jia T, Kulkarni RV. Intrinsic Noise in Stochastic Models of Gene Expression with Molecular Memory and Bursting. Phys Rev Lett. 2011 Feb; 106:058102. Available from: http://link.aps.org/doi/10.1103/PhysRevLett.106.058102. doi: 10.1103/PhysRevLett.106.058102 PMID: 21405439

44. Xu X, Kumar N, Krishnan A, Kulkarni RV. Stochastic modeling of dwell-time distributions during transcriptional pausing and initiation. In: Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on. IEEE; 2013. p. 4068–4073.

45. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011; 332(6028):472–474. doi: 10.1126/science.1198817 PMID: 21415320

46. Harper CV, Finkenstädt B, Woodcock DJ, Friedrichsen S, Semprini S, Ashall L, et al. Dynamic analysis of stochastic transcription cycles. PLoS biology. 2011; 9(4):e1000607. doi: 10.1371/journal.pbio.1000607 PMID: 21532732

47. Elgart V, Jia T, Kulkarni RV. Applications of LittleÕs Law to stochastic models of gene expression. Physical Review E. 2010; 82(2):021901. doi: 10.1103/PhysRevE.82.021901

48. Liu L, Kashyap BRK, Templeton JGC. On the GIX/G/Infinity system. Jour Appl Prob. 1990; 27(3):671–683. doi: 10.2307/3214550

49. Bokes P, King JR, Wood AT, Loose M. Multiscale stochastic modelling of gene expression. Journal of mathematical biology. 2012; 65(3):493–520. doi: 10.1007/s00285-011-0468-7 PMID: 21979825

50. Ingram PJ, Stumpf MPH, Stark J. Nonidentifiability of the Source of Intrinsic Noise in Gene Expression from Single-Burst Data. PLoS Comp Biol. 2008; 4(10). doi: 10.1371/journal.pcbi.1000192

51. Cookson NA, Mather WH, Danino T, Mondragón-Palomino O, Williams RJ, Tsimring LS, et al. Queueing up for enzymatic processing: correlated signaling through coupled degradation. Molecular systems biology. 2011; 7(1). doi: 10.1038/msb.2011.94 PMID: 22186735

52. Mather WH, Cookson NA, Hasty J, Tsimring LS, Williams RJ. Correlation resonance generated by coupled enzymatic processing. Biophysical journal. 2010; 99(10):3172–3181. doi: 10.1016/j.bpj.2010.09.057 PMID: 21081064

53. Little JD. A proof for the queuing formula: L = λ W. Operations research. 1961; 9(3):383–387. doi: 10.1287/opre.9.3.383

54. Ross SM. Introduction to Probability Models, Ninth Edition. Orlando, FL, USA: Academic Press, Inc.; 2006.

55. Bokes P, King JR, Wood AT, Loose M. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. Journal of mathematical biology. 2012; 64(5):829–854. doi: 10.1007/s00285-011-0433-5 PMID: 21656009

56. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry. 1977; 81(25):2340–2361. doi: 10.1021/j100540a008

57. Daigle BJ, Soltani M, Petzold LR, Singh A. Inferring Single-Cell Gene Expression Mechanisms using Stochastic Simulation. Bioinformatics. 2015;p. btv007.

58. Karlin S. A first course in stochastic processes. Academic press; 2014.