Abstract— The stochastic nature of gene expression can lead to significant cell-to-cell variability in the time at which a certain protein level is attained. This is reflected in the timing of cellular events triggering at critical protein thresholds as well. A problem of interest is to understand how cells regulate gene expression to ensure precise timing of important events. To this end, we consider a gene expression model assuming constitutive expression in translation bursts. We also assume the proteins to be stable. The event timing is formulated as a first-passage time (FPT) problem and stochasticity in FPT for this model is quantified. We also investigate the effect of auto-regulation, a control mechanism often present in cells, on the stochasticity of FPT. In particular, we ask: given FPT threshold of proteins and mean FPT, what form of auto-regulation minimizes variance in FPT? Our results show that the objective is best achieved by having no auto-regulation. Moreover, a smaller mean burst size would result into lower stochasticity. We discuss our results in context of lysis time of E. coli cells infected by a λ phage virus. An optimal lysis time provides evolutionary advantage to λ phage, suggesting a possible regulation to minimize its stochasticity. Our results are consistent with previous studies showing there is no auto-regulation of the protein responsible for lysis. Moreover, congruent with experimental evidences, our analysis predicts that the expression of the lysis protein should show a small burst size.

I. INTRODUCTION

Gene expression is the process of transcription of genetic information to mRNAs, and translation of each mRNA to proteins. As the copy number of species involved in the process is small, the probabilistic nature of biochemical reactions reflects as random fluctuations in gene expression. These fluctuations are referred to as gene expression noise/stochasticity/variability [1]–[6]. Gene expression stochasticity affects the cell in two contrasting ways. On one hand it can lead genetically identical cells to different cell fates, helping the cells in responding to the ever-changing environment [7]–[12]. On the other hand, several diseased states are attributed to gene expression variability [13]–[15], and therefore cells aim to minimize it [16], [17]. Accordingly, different regulatory mechanisms to control random fluctuations exist in the cells [18]–[25]. Auto-regulation wherein the transcription rate is a function of protein count is an example of one such mechanism. Its effect on stochasticity in gene expression has been a subject of several studies [23]–[26].

After onset of gene expression, its stochasticity consequently manifests as randomness in the time at which a certain protein level is reached [27]. This implies that the timing of a cellular event which triggers at a critical protein level is stochastic in nature [27]–[29]. For instance, lysis time for an E. coli cell infected by a λ phage virus is stochastic [30]. Lysis of the cell takes place when holin, the protein responsible for lysis, reaches a critical concentration [30]–[32]. Moreover, such events can be modeled as First-passage time (FPT) problems which is defined as the time it takes for a stochastic process to reach a given threshold for the first time [32]–[34].

Further, it has been suggested that optimality in lysis time provides evolutionary advantage to λ phage virus [35]–[39]. This indicates there could be some regulation of gene expression to ensure lysis at the optimal time, with minimum stochastic fluctuations. In this work, we study variability in FPT at a single-cell level and investigate the effect of auto-regulation on stochasticity of FPT. Particularly, we seek answer to the question: given the mean FPT (corresponding to optimal lysis time, for instance), what auto-regulatory feedback will lead to minimum variability in the FPT? Typically, variability is quantified as coefficient of variation (CV) squared which is ratio of variance over mean squared. However, since we are fixing the mean FPT here, minimizing $CV^2$ of FPT is equivalent to minimizing variance in FPT.

Our analysis deals with a standard gene expression model consisting of transcription, translation, and mRNA degradation. We assume that proteins do not degrade because the lysis protein in λ phage (called holin) is stable as observed in [40]. Along the lines of [32], we find expressions for statistical moments of FPT for this model, and discuss their implications with respect to minimizing variance in FPT for given mean FPT. Next, we introduce auto-regulation in the above model and derive the moments for FPT. Then, we deduce the expression for optimal feedback function that minimizes the variance in FPT with its mean being fixed. We show that no auto-regulation is optimal strategy for this purpose. The results are also validated by carrying out simulations. Also, various notations used in this work are tabulated in Table I (see page 2).

II. FIRST-PASSAGE TIME FOR GENE EXPRESSION MODEL WITHOUT REGULATION

In this section, we describe an unregulated gene expression model (as shown in Fig. 1) and define FPT for it. Then, we find expressions for the moments of FPT and discuss the implications of these expressions in context of minimizing variance of FPT, for fixed mean FPT and threshold.
TABLE I
DESCRIPTION OF NOTATIONS USED IN THIS WORK

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_m$</td>
<td>Transcription rate for unregulated gene expression model</td>
</tr>
<tr>
<td>$k_p$</td>
<td>Translation rate</td>
</tr>
<tr>
<td>$\gamma_m$</td>
<td>mRNA degradation rate</td>
</tr>
<tr>
<td>$P$</td>
<td>Probability</td>
</tr>
<tr>
<td>$B_i$</td>
<td>Burst size after $i^{th}$ transcription event</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Parameter of geometric distribution corresponding to translational bursts</td>
</tr>
<tr>
<td>$b$</td>
<td>Mean of protein burst size</td>
</tr>
<tr>
<td>$x(t)$</td>
<td>Protein count at time $t$</td>
</tr>
<tr>
<td>$x_i$</td>
<td>Protein count after $i^{th}$ burst</td>
</tr>
<tr>
<td>$k_m(x_i)$</td>
<td>Transcription rate for auto-regulated gene expression model after $i^{th}$ transcription event</td>
</tr>
<tr>
<td>$X$</td>
<td>Threshold for protein count</td>
</tr>
<tr>
<td>$N$</td>
<td>Minimum number of transcription events for protein count to reach the threshold $X$</td>
</tr>
<tr>
<td>$T_i$</td>
<td>Waiting time for $i^{th}$ transcription event</td>
</tr>
<tr>
<td>$f_X(n)$</td>
<td>Probability mass function for minimum number of transcription events to reach the threshold $X$</td>
</tr>
<tr>
<td>$f_X(j)$</td>
<td>Probability mass function for protein count after $i$ transcription events</td>
</tr>
</tbody>
</table>

A. Model Formulation

In our model, transcription of mRNAs from the gene occurs at a rate $k_m$, translation of proteins from each mRNA occurs at a rate $k_p$, and each mRNA degrades at a rate $\gamma_m$. As mentioned earlier, we assume proteins to be stable. Consistent with experimental data, the mRNA arrival process is considered to be Poisson, thereby transcription events occurring at exponentially distributed times intervals. To further simplify the model, we assume each mRNA molecule degrades instantaneously after producing a burst of random number of protein molecules [41]–[44]. Congruent with experimental and theoretical evidences, we assume that protein burst follows a geometric distribution, and the mean burst size is given by $b = k_p/\gamma_m$ [45], [46]. Thus, the simplified model considers gene expression wherein each burst event (equivalent to transcription event) occurs at an exponentially distributed time with parameter $k_m$, and size of burst follows a geometric distribution with mean $b$.

Let us denote the size of $i^{th}$ burst by random variable $B_i$ and the parameter of its distribution by $\mu$. The probability mass function, therefore, can be written as:

$$P(B_i = k) = \mu (1 - \mu)^{k-1}, \mu \in [0, 1], k \in \{0, 1, 2, \ldots\}. \quad (1)$$

The mean burst size, $b$, can be expressed as:

$$\langle B_i \rangle = b = \frac{1 - \mu}{\mu}. \quad (2)$$

Further, let protein count after $n$ transcription events be denoted as $x_n$. It can be expressed as a sum of random variables $B_i$:

$$x_n = \sum_{i=1}^{n} B_i. \quad (3)$$

Being sum of independent and identically distributed geometric random variables, $x_n$ has a negative binomial distribution with parameters $n$ and $\mu$. The probability mass function of $x_n$, denoted as $f_x(x_n)$, can be expressed as:

$$f_x(x_n) = \mathbb{P}(\sum_{i=1}^{n} B_i = j) = \binom{n+j-1}{n-1} \mu^n (1-\mu)^j. \quad (4)$$

Also, the cumulative distribution function is given by [47]:

$$\mathbb{P}\left(\sum_{i=1}^{n} B_i \leq j\right) = 1 - I_{1-\mu}(j+1,n), \quad (5)$$

where $I_{1-\mu}(j+1,n)$ is regularized incomplete beta function:

$$I_{1-\mu}(j+1,n) = \sum_{l=j+1}^{n+j} \binom{n+j}{l} (1-\mu)^l \mu^{j+n-l}, \quad (6)$$

and satisfies the following property:

$$I_{1-\mu}(j+1,n) = 1 - I_{\mu}(n, j+1). \quad (7)$$

Thus, we have determined the distribution of the protein count. The FPT for the protein count to reach a certain threshold can now be defined.

B. Expression for First-Passage Time

For a random process corresponding to protein count, $x(t)$, with $x(0) = 0$, the FPT, for a threshold $X$ is defined as:

$$FPT := \inf\{t : x(t) \geq X\}, \quad X \in \{1, 2, 3, \ldots\}. \quad (8)$$

Because in our model, the protein count changes only when a transcription event occurs, we can calculate the minimum number of transcription events, $N$, it takes for the protein count to reach the threshold $X$ and define the FPT as sum of inter-burst arrival times.

Let us denote the waiting time for $i^{th}$ transcription event, i.e., time between arrival of $(i-1)^{th}$ and $i^{th}$ bursts by random variable $T_i$. Then, FPT can be expressed as:

$$FPT = \sum_{i=1}^{N} T_i, \quad (9)$$

3048
where $N$ is given by the following equation:

$$N = \inf(n : x_n \geq X), \quad n \in \{1, 2, \ldots\}, \quad X \geq 1.$$  \hspace{1cm} (10)

Note that in Eq. (9), $T_i$ are independent, and identically distributed exponential random variables with parameter $k_m$. We denote this by $T_i \sim \exp(k_m)$. Also, each of $T_i$ is independent of $N$. Using this definition of FPT, its first two moments can be derived.

C. Moments of First-Passage Time

Using standard results from probability theory, one may write the mean and variance of FPT as follows:

$$\langle FPT \rangle = \langle N \rangle \langle T_i \rangle, \quad \text{(11a)}$$

$$\text{Var}(FPT) = \langle N \rangle \text{Var}(T_i) + \langle N \rangle \langle T_i \rangle^2. \quad \text{(11b)}$$

This implies that expressions for first two moments of $T_i$ and $N$ are required to determine statistical moments of FPT in Eq. (11a)–(11b).

1) First Two Moments of $N$: The cumulative distribution function for $N$ defined in Eq. (10) can be written as:

$$P(N \leq n) = P(x_n \geq X), \quad \text{(12a)}$$

$$= 1 - P(x_n \leq X - 1). \quad \text{(12b)}$$

Since $x_n$ is a negative binomial distribution, we have:

$$P(N \leq n) = 1 - (1 - I_{1\mu}(X, n)), \quad \text{(13a)}$$

$$= I_{1\mu}(X, n). \quad \text{(13b)}$$

Using the property of incomplete beta function mentioned in Eq. (7), we get:

$$P(N \leq n) = 1 - I_{1\mu}(n, X). \quad \text{(14)}$$

Comparing with Eq. (4) and Eq. (5), the probability mass function corresponding to Eq. (14) can be written as:

$$f_N(n) = \binom{n + X - 2}{n - 1} (1 - \mu)^{n-1} \mu^X, \quad n \in \{1, 2, \ldots\}, \quad X \geq 1. \quad \text{(15)}$$

First two statistical moments of the distribution in Eq. (15) are given by:

$$\langle N \rangle = \frac{\mu X}{1 - \mu} + 1 = \frac{X}{b} + 1, \quad \text{(16a)}$$

$$\text{Var}(N) = \langle N^2 \rangle - \langle N \rangle^2 = \frac{\mu X}{(1 - \mu)^2} = \frac{X}{b} + 1. \quad \text{(16b)}$$

2) First Two Moments of $T_i$: Since $T_i \sim \exp(k_m)$, its statistical moments are given by:

$$\langle T_i \rangle = \frac{1}{k_m}, \quad \text{(17a)}$$

$$\text{Var}(T_i) = \langle T_i^2 \rangle - \langle T_i \rangle^2 = \frac{1}{k_m^2}. \quad \text{(17b)}$$

We now have expressions for first two moments of $T_i$ and $N$. The expressions for first two moments of FPT in terms of model parameters can, therefore, be written as:

$$\langle FPT \rangle = \left(\frac{X}{b} + 1\right) \frac{1}{k_m} \approx \frac{X}{b k_m}, \quad \text{(18)}$$

$$\text{Var}(FPT) = \frac{X (2b + 1) b^2}{b^2 k_m^2} \approx \frac{X}{b^2 k_m} (1 + 2b). \quad \text{(19)}$$

where the approximations are valid when $X \gg b$. It can be observed that a smaller mean burst size $b$ would result in smaller variance of FPT. The mean FPT can be kept fixed by a commensurate change in the transcription rate, $k_m$. Therefore, the variance can independently be reduced by a lower mean burst size $b = k_p / \gamma_m$. This means adopting a high transcription rate $k_m$, and a low translation rate $k_p$ (and/or having a higher degradation rate $\gamma_m$ for the mRNAs) results in a lower variance in FPT without affecting its mean. In the following section, expressions for statistical moments of FPT are derived when the transcription rate depends on the protein count.

III. INTRODUCING AUTO-REGULATION IN GENE EXPRESSION MODEL

To investigate the effect of auto-regulation on statistical moments of FPT, we assume that transcription rate is a function of protein count, i.e., it changes after each transcription event. Similar to previous section, we derive expressions for moments of inter–burst arrival times $T_i$, and minimum number of transcription events $N$ in order to derive the expression for FPT moments defined in Eq. (9). It is important to point out that protein burst size is determined by $k_p / \gamma_m$ and is, therefore, independent of the transcription rate. So the distribution of $N$ to reach a certain threshold $X$ remains same as gene expression model without any regulation discussed in previous section. However, distribution of each $T_i$ is different and depends upon the corresponding rate of transcription. The first two moments of $T_i$ and FPT can now be derived.

A. Inter–burst arrival time for auto-regulatory gene expression model

Let the transcription rate after arrival of $i^{th}$ burst be denoted by $k_m(x_i)$. It may be noted that if protein count after any burst event is known, arrival time for the next burst will be exponentially distributed. Therefore, the distribution of each $T_i$ can be modeled as a conditional exponential distribution. More specifically, we can write:

$$T_i \sim \exp(k_m(x_{i-1}) | x_{i-1}), \quad \text{(20)}$$

where $T_i$, and $x_{i-1}$ respectively denote the arrival time for $i^{th}$ burst and protein count after the $(i - 1)^{th}$ burst. The expressions for mean and variance of $T_i$ can be calculated as follows.

1) Mean: Before arrival of the first burst, there are no protein molecules, i.e., $x_i = 0$ for $i = 1$. Therefore, we can write the mean for arrival time for the first burst as:

$$\langle T_1 \rangle = \frac{1}{k_m(0)}. \quad \text{(21)}$$

For $i \in \{2, 3, 4\ldots\}$, the corresponding arrival times would be conditionally exponential, implying:

$$\langle T_i \rangle = \sum_{j=0}^{\infty} \frac{1}{k_m(j)} P(x_{i-1} = j) = \sum_{j=0}^{\infty} \frac{1}{k_m(j)} f_{x_{i-1}}(j). \quad \text{(22a)}$$
2) Second Order Moments: Adopting similar approach as above, we derive the expressions for second order moments of $T_i$. For $i = 1$, we have:

$$\langle T_i^2 \rangle = \frac{2}{k_m^2(0)}.$$  
(23a)

For $i \in \{2,3,4,...\}$:

$$\langle T_i^2 \rangle = \sum_{j=0}^{\infty} \frac{2}{k_m^2(j)} \mathbb{P}(x_{i-1} = j) = \sum_{j=0}^{\infty} \frac{2}{k_m^2(j)} f_{x_{i-1}}(j).$$  
(23b)

Therefore the expression for variance of $T_i$:

$$\text{Var}(T_i) = \frac{1}{k_m^2(0)} \langle T_i \rangle^2.$$  
(24a)

For $i \in \{2,3,4,...\}$, the expression for $\text{Var}(T_i)$ will be

$$\text{Var}(T_i) = \sum_{j=0}^{\infty} \frac{2}{k_m^2(j)} f_{x_{i-1}}(j) - \left[ \sum_{j=0}^{\infty} \frac{1}{k_m(j)} f_{x_{i-1}}(j) \right]^2.$$  
(24b)

B. FPT for auto-regulatory gene expression model

Having derived the expressions for moments of inter-bursts arrival times, we see how the introduction of auto-regulation influences the expressions for FPT moments. The expressions for statistical moments of FPT are presented in theorem–proof format. In developing the proofs, we make use of the fact that each $T_i$ will be independent of $N$. Also, $T_i$ are independent of each other. However, they are not identically distributed like the unregulated gene expression case discussed in previous section.

**Theorem 1 (Mean and Second Moment of FPT):** For the FPT defined in Eq. (9), the mean and second order moment of FPT are given by following expressions:

$$\langle \text{FPT} \rangle = \sum_{n=1}^{\infty} \sum_{i=1}^{n} \langle T_i \rangle f_N(n),$$  
(25)

$$\langle \text{FPT}^2 \rangle = \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} \text{Var}(T_i) + \frac{n}{\sum_{i=1}^{n} \langle T_i \rangle} \right) f_N(n),$$  
(26)

where $f_N(n)$ is defined in Eq. (15), $N$ is given by Eq. (16a), $\langle N^2 \rangle$ can be deduced from Eq. (16b), $\langle T_i \rangle$ is given by Eq. (21), (22a) and $\text{Var}(T_i)$ is given by Eq. (24a), (24b).

**Proof:** Both expressions can be deduced by conditioning arguments. For mean FPT, one may write

$$\langle \text{FPT} \rangle = \sum_{n=1}^{\infty} \langle \text{FPT} | N = n \rangle \mathbb{P}(N = n),$$  
(27a)

$$= \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} T_i \right) f_N(n) = \sum_{n=1}^{\infty} \sum_{i=1}^{n} \langle T_i \rangle f_N(n).$$  
(27b)

Similarly for the second order moment

$$\langle \text{FPT}^2 \rangle = \sum_{n=1}^{\infty} \langle \text{FPT}^2 | N = n \rangle f_N(n),$$  
(28a)

$$= \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} T_i \right) f_N(n),$$  
(28b)

$$= \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} T_i^2 + \sum_{i=1}^{n} \sum_{j<i} T_i T_j \right) f_N(n).$$  
(28c)

Since $T_i^2$ are independent of each other, and $T_j$ are independent of $T_i$ for each $j \neq i$, we can write:

$$\langle \text{FPT}^2 \rangle = \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} \langle T_i^2 \rangle + \sum_{i=1}^{n} \sum_{j<i} \langle T_i T_j \rangle \right) f_N(n).$$  
(28d)

Using $\text{Var}(T_i) = \langle T_i^2 \rangle - \langle T_i \rangle^2$, we have:

$$\langle \text{FPT}^2 \rangle = \sum_{n=1}^{\infty} \left( \sum_{i=1}^{n} \text{Var}(T_i) + \frac{n}{\sum_{i=1}^{n} \langle T_i \rangle} \right) f_N(n).$$  
(28f)

This completes the proof.

So far we have developed analytical expressions for mean and variance of FPT when there is an auto-regulatory feedback to transcription rate from protein count. In the next section, we discuss the optimal auto–regulation function to minimize FPT variance for a constant mean FPT and threshold.

IV. MINIMIZING VARIANCE IN FIRST-PASSAGE TIME FOR GIVEN MEAN FPT

We present the form of auto–regulatory feedback function $k_m(x_{i-1}), \ i \in \{1,2,3,...\}$ that gives minimum variance in FPT, given the mean FPT and event threshold are fixed. The result is stated as a theorem. We have omitted the proof of the theorem due to space constraints. It is deduced by assuming each transcription rate as a perturbation from a fixed transcription rate. Using the constraint that mean FPT is fixed, it can be proved that these perturbations have to be equal to zero in order to minimize the variance of FPT.

**Theorem 2 (Optimal feedback for minimum variance):** Let the FPT be defined as Eq. (9), and its mean and variance, respectively, given by Eq. (25) and Eq. (26). The optimal transcription rates as a function of protein count that minimizes the variance of FPT for a given mean of FPT and event threshold are given by the following expression:

$$k_m(x_{i-1}) = \frac{1}{\langle \text{FPT} \rangle} \left( \frac{X}{b} + 1 \right), \ \forall i \in \{1,2,3,...\},$$  
(29)

where $\langle \text{FPT} \rangle$ denotes the mean FPT, $X$ is the FPT threshold while $b$ represents the mean burst size.

This result was verified via Monte Carlo simulations using Gillespie’s SSA [48]. We did not specifically assume that production of protein is geometrically distributed thereby relaxing the burst approximation. To simulate, we considered three separate cases: no feedback, negative feedback and positive feedback. The positive feedback is implemented using Hill function as follows:

$$k_m(j) = k_{\max} \left( r + (1 - r) \frac{(jc)^H}{1 + (jc)^H} \right),$$  
(30)

where $k_{\max}$ is the maximum transcription rate, $r$ represents fraction of $k_{\max}$ corresponding to the basal transcription rate (minimum transcription rate), $H$ denotes the Hill coefficient while $c$ is coefficient proportional to the binding affinity.
of the proteins (when \( j = 1/c, \ k_m(j) = k_{\text{max}}(1 + r)/2 \)). The negative feedback is implemented in similar fashion using the following function:

\[
k_m(j) = k_{\text{max}} \left( r + (1 - r) \frac{1}{1 + (jc)^{H}} \right).
\]

We carried out the simulations for several sets of parameters assuming a fixed event threshold. Rest of the model parameters were chosen to keep the mean FPT approximately equal. In all of them, we found that no–feedback case has minimum variance in FPT. Simulation results for 10000 realizations are shown in Fig. 2. We note that the variance is minimum in no–feedback case, validating our theoretical claims for this set of parameter values.

V. DISCUSSION

In this work, we studied variability in event timing at a single–cell level. We considered a standard gene expression model without protein degradation. Next, we formulated the FPT problem for this model and derived the formulas for statistical moments of FPT. Further, we introduced auto-regulation in the gene expression wherein the transcription rate is a function of protein count. We derived the formulas for moments of FPT in this case as well, and demonstrated that for a given mean of FPT, the variance in FPT is minimized when there is no auto–regulation of gene expression. The result was verified with simulations as well.

The result can be connected to the \( \lambda \) phage lysis time. Due to existence of optimal lysis time [35], [36], one can hypothesize that the phage would like to kill the cell at that time with as much precision as possible. As a result, it should resort to a strategy that would minimize the lysis time variance. A significant contribution in the total variation in lysis time comes from the time it takes for the critical protein (holin) concentration to build up which can be modeled as FPT [30]–[32]. Hence there should not be any protein–dependent feedback regulation of transcription rate in the expression of holin. Previous studies reveal that expression from the late promoter in \( \lambda \) phage, which produces holin, indeed has no evidence of auto–regulation [49]–[51]. This observation underscores the hypothesis that the phage \( \lambda \) tries to minimize the lysis time stochasticity around the optimal.

Recall that in no auto–regulation case too, the variance of FPT can be independently decreased by lowering the mean burst size \( b \). Other studies also reveal that in case of \( \lambda \) phage, the burst size is indeed small [30], [35]. Also, antiholin, another protein expressed from the same promoter that expresses holin, binds to holin to decrease the effective burst size [32], [52].

In this paper, there is an underlying assumption of protein being stable. However, since this is not true in general, we plan to use a gene expression model with protein degradation and carry out a similar analysis in our future work. This can be further extended to more generalized gene expression models wherein the promoter can also switch between on and off states [12], [42], [53].

ACKNOWLEDGMENT

AS is supported by the National Science Foundation Grant DMS-1312926, University of Delaware Research Foundation (UDRF) and Oak Ridge Associated Universities (ORAU).

REFERENCES


