Modeling the stochastic dynamics of gene expression in single cells: A Birth and Death Markov Chain Analysis

Samik Ghosh, Preetam Ghosh, Kalyan Basu and Sajal K. Das
Biological Networking Research Group,
Department of Computer Science & Engineering
The University of Texas at Arlington, Arlington, TX 76010, USA.
{sghosh, ghosh, basu, das}@cse.uta.edu

Abstract— Fluctuations in protein number (noise) caused by the stochasticity in gene expression plays a central role in the dynamic behavior of cellular pathways. Deterministic models capture average cell population behavior and are limited in their relevance in modeling stochastic deviations of gene expression in single cells. In this paper, we develop a birth and death Markov chain model to capture the discrete molecular events of transcription and translation in prokaryotic cells. We derive mathematical models for the expression ‘burst frequency’ distribution as well as the number of protein molecules per burst. We validate our stochastic models with recent single cell experiments on the lacZ gene in Escherichia Coli and characterize expression noise sensitivity to biological parameters like gene activation ratio. Further, we build a discrete-event stochastic simulation system to study the transient dynamics of lacZ gene expression, quantifying the role of promoters in controlling the ‘burstiness’ of protein synthesis.

Keywords: gene expression, stochastic models, Markov chains, discrete event simulation

I. INTRODUCTION

One of the key goals in studying complex cellular pathways is the understanding of the dynamic interaction between the cell’s gene regulatory and metabolic networks. In particular, the dynamics of protein-coding genes play a vital role in controlling the expression patterns of other genes encoding regulatory proteins and metabolic enzymes. Continuous and deterministic differential equation-based models representing the discrete molecular events of transcription and translation have been traditionally used to study gene expression in cell populations [1]. However, the inherent stochasticity in molecular interactions limits the applicability of these models in studying single cell expression deviations and observed cell-to-cell variability [1], [3], [4].

Stochastic fluctuations in the expression pattern of genes have been mathematically studied by Arkin et. al [2], [5] and experimentally observed in [7], [8]. In particular, the ‘burstiness’ in protein production i.e. proteins are produced in random bursts of short duration rather than in a continuous manner, have been studied by various researchers [1], [2], [10-12]. The random fluctuations in the number of proteins, termed ‘noise’, stems from the interplay of a large number of factors: discrete, random nature of molecular interactions like promoter binding and transcription open-complex formation, low copy number of key transcriptional and translational machineries like RNA polymerase, transcription factors, ribosomal units etc., and the random nature of signals triggering gene expression. We provide an overview of stochastic models, primarily based on Monte Carlo simulation of the biochemical reactions involved in gene expression [10-12], in the next section.

In this paper, we model the molecular events of transcription and translation for prokaryotic cells as discrete space, continuous time Markov processes, computing the probability distribution of the two key parameters characterizing the expression of proteins from a gene [8]: (a) frequency of mRNA bursts per cell cycle, and (b) the number of proteins molecules per burst, the details of which are outlined in Section III.

We validate our model with recent experimental data obtained for ‘burst frequency’ and ‘burst size’ measurements for the lacZ gene in single cell experiments on the bacteria Escherichia Coli in Section IV. The validated model provides a useful mathematical tool for systematic analysis of the process. We use this validated model to study the sensitivity of the different factors (promoter
activation ratio, transcription and translation frequencies) on the frequency and burst
distributions in Section V. To understand the transient dynamics, we integrate our mathematical equations into a stochastic simulation of coupled gene transcription and translation events outlined in Section VI. Our simulation results provide biological insights into the role of promoter activation ratio in generating ‘protein bursts’ for the lacZ gene, specifically the role of increased promoter activity in controlling the ‘burstiness’ of protein production in E.Coli.

II. STOCHASTIC MODELS OF GENE EXPRESSION DYNAMICS

As mentioned in the previous section, gene expression is an inherently stochastic process, governed by random association and dissociation of transcription factors, RNAP polymerases, ribosomes and degradosomes (RNase E), etc. caused by the low copy number of molecular entities. In this section, we briefly overview Monte Carlo simulation and other stochastic models for prokaryotic gene expression, focusing on the molecular actors involved in transcription and translation events as depicted in Fig.1.

Specifically, the events of gene activation followed by transcription initiation, elongation and termination leading to the production of an mRNA molecule are considered. With respect to the translational machinery, the molecular events involved in translation initiation by ribosome binding/unbinding, mRNA decay by degradosome binding, translation elongation and termination are accounted for.

In a reaction model, the above events are represented as molecular reactions, identified by the reactants and products, together with the kinetic rate constants associated with each reaction [10], [11], [12], [14] (Refer to supplementary materials for a comprehensive list of reactions associated with gene expression). Once the reaction set is defined together with rate constants estimated from experimental and available data, stochastic Monte Carlo simulations (Gillespie algorithm [6]) are used to study their behavioral dynamics. McAdams and Arkin [2] combined a continuous model of transcription initiation with a stochastic model for subsequent processing and protein synthesis capturing the race between translation initiation and mRNA degradation. Kierzek et. al [9] systematically studied the effects of transcription and translation efficiencies for bacterial gene expression through Monte Carlo simulations. In [15], the authors have developed stochastic pi-calculus based techniques for studying gene expression dynamics. Paulsson [13] provides a comprehensive review of different models of gene expression and protein noise analysis, giving a common Fluctuation-Dissipation Theorem (FDT) based framework to encompass various models. Noise characterizations (intrinsic and extrinsic) in single cell gene expressions have also been in studied in [1,10].

In this paper, we consider the molecular process of prokaryotic gene expression as a complex stochastic system and abstract the process through probability distributions for the rate of mRNA and protein synthesis. While the estimation of the
various kinetic parameters involved in reaction models pose a challenge under different conditions, our stochastic model captures the available information to quantitatively characterize the gene expression process in terms of the average rates of mRNA synthesis and protein generation.

III. **Birth-Death Markov Chain Model of Gene Expression**

In this section, we build a stochastic model for gene expression, considering the molecular actors affecting expression dynamics. Stochastic behavior arises in cellular reactions from the individual randomness of molecules in motion and the activation energy required forming the complex when two molecules come close to each other. Our modeling approach is motivated by the fact that the system dynamics of such processes can be captured by algebra of random numbers. To make the process mathematical tractable, we use the following set of assumptions:

- **a.** The effect of promoter arrangement, as in an operon, or regulon or tandem promoters has not been considered.
- **b.** The transcription machinery is not limited by the number of RNA polymerase (RNAP), transcription factors etc. available in the system.
- **c.** The process of ribosome assembly is not considered and translation is assumed to be initiated by an active ribosome. The translation process is not rate limited by the number of ribosome or amino acids present in the system.

**A. Modeling Transcriptional Dynamics**

In modeling the dynamics of prokaryotic transcription, we consider the key processes involved (as outlined in Fig.1): gene activation and deactivation, transcription initiation and RNAP recruitment, mRNA chain elongation, and finally transcription termination. We observe the ‘concurrency’ involved in the transcriptional process [45], i.e. each gene is simultaneously transcribed by several RNAP molecules, which can be visualized as a combed structure, with the gene-encoding region of the DNA forming the backbone, and the transcripts of increasing length (left-to-right) forming the teeth (refer supplementary materials) [14]. Thus, in order to calculate the average number of mRNA transcripts, i.e. the burst frequency distribution, we model the system in terms of the number of RNAP molecules attached to the gene.

<table>
<thead>
<tr>
<th>Biological Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$ : Promoter binding rate</td>
<td></td>
</tr>
<tr>
<td>$\lambda$ : Promoter unbinding rate</td>
<td></td>
</tr>
<tr>
<td>$mR_{on}$ : RNAP binding rate</td>
<td></td>
</tr>
<tr>
<td>$mR_{off}$ : RNAP unbinding rate</td>
<td></td>
</tr>
<tr>
<td>$mR_{init}$ : Transcription initiation rate</td>
<td></td>
</tr>
<tr>
<td>$mR_{elong}$ : Transcription elongation Rate</td>
<td></td>
</tr>
<tr>
<td>$mR_{elong}$ : Transcription elongation Rate</td>
<td></td>
</tr>
<tr>
<td>$P_{on}$ : Ribosome binding rate</td>
<td></td>
</tr>
<tr>
<td>$P_{off}$ : ribosome unbinding rate</td>
<td></td>
</tr>
<tr>
<td>$P_{init}$ : Translation initiation rate</td>
<td></td>
</tr>
<tr>
<td>$P_{elong}$ : Translation elongation Rate</td>
<td></td>
</tr>
<tr>
<td>$P_{elong}$ : Translation elongation Rate</td>
<td></td>
</tr>
<tr>
<td>$mRNA_{decay}$ : degradosome binding rate</td>
<td></td>
</tr>
<tr>
<td>$m_{decay}$ : mRNA decay rate</td>
<td></td>
</tr>
<tr>
<td>$p_{decay}$ : protein decay constant</td>
<td></td>
</tr>
</tbody>
</table>
Let $X_{mR}(t)$ denote the number of RNAP molecules attached to the gene-coding region of the DNA for a particular gene at time $t$. $X_{mR}(t)$ can take values in the discrete state space $S_{mR} = \{0,1,\ldots,N_{\text{RNAP}}^{\text{max}}\}$ where,

$$N_{\text{RNAP}}^{\text{max}} = \left\lfloor \frac{L_{\text{gene}}}{\text{RNAP footprint}} \right\rfloor$$  \hspace{1cm} (1)

and, $L_{\text{gene}}$ = length of the gene (in base-pair length)

$\text{RNAP footprint}$ = promoter clearance, or distance between two RNAP molecules.

At any time $t$, the system starts at state 0, i.e., no RNAP is attached to the gene. Once the promoter is activated and an RNAP molecule is recruited, the system moves to state 1. Now, assuming at some time $t$, the system is in state $k$, i.e. $k$ RNAP molecules are attached to the gene, the following events can occur, captured in Fig.2:

(i) The gene may be activated again and another RNAP molecule is recruited, taking the system to state $k+1$

(ii) An RNAP molecule reaches the terminator sequence releasing an mRNA transcript for translation, taking the system to state $k-1$

(iii) A repressor may exclude further RNAP molecules from binding to the promoter region, deactivating the gene and taking the system to an inactive state ($D_{\text{gene}}$).

We now consider a continuous parameter, discrete space Markov chain \cite{18,19},

$\chi_{mR} = \{X_{mR}(t), \hspace{0.5cm} t \in T\}$ \hspace{0.5cm} and \hspace{0.5cm} $T = \{t: 0 \leq t < \infty\}$

where $X_{mR}(t)$ is the state of the system at time $t$, with a finite state space $S_{mR}$.

For initial solution, let us neglect the dead state $D_{\text{gene}}$, and use standard birth and death modeling technique to derive the probability of states

$\chi_{mR} = \{X_{mR}(t), \hspace{0.5cm} t \in T\}$ \hspace{0.5cm} and \hspace{0.5cm} $T = \{t: 0 \leq t < \infty\}$

where $X_{mR}(t)$ is a birth-death Markov chain where the intensity matrix \cite{24} $Q$ is defined as,

$$Q = \begin{pmatrix}
-\lambda_{mR} & \lambda_{mR} & 0 & 0 & \cdots \\
\mu_{mR} & -\lambda_{mR} - \mu_{mR} & \lambda_{mR} & 0 & \cdots \\
0 & \mu_{mR} & -\lambda_{mR} - \mu_{mR} & \lambda_{mR} & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots 
\end{pmatrix}$$  \hspace{1cm} (7)

$\lambda_{mR}$ is defined as the birth-rate while $\mu_{mR}$ is defined as the death rate of the process, which is depicted in Fig.3 along with the possible states. The birth and death rates encompass the dynamics of the different biological actors affecting transcription and are defined as follows:

(i) Birth rate ($\lambda_{mR}$): defines the rate at which new RNAP molecules arrive in the system, which is given by, $\lambda_{mR} = \frac{1}{T_{\text{on}}}$ where, $T_{mR}$ is the sum of the average time taken for gene activation, RNAP-promoter binding, open-complex formation and the time taken for the previous RNAP molecule to clear the promoter for the next RNAP molecule (promoter clearance time), as obtained from the parameters in Table 1, i.e.,

$$T_{mR} = T_{\text{activation}} + T_{\text{binding}} + T_{\text{init}} + T_{\text{clearance}}$$  \hspace{1cm} (8)

$$T_{mR} = \frac{1}{\lambda^*} + \frac{1}{k_{\text{on}}} + \frac{1}{k_{\text{init}}} + \frac{\text{RNAP footprint}}{k_{\text{elong}}}$$  \hspace{1cm} (9)
(ii) Death Rate ($\mu_i^{mR}$) : defines the rate at which RNAP molecules leave the system, i.e. the rate at which mRNA transcripts are released and is governed by the time taken by the RNAP molecule to elongate the chain and reach the terminator sequence. Now, at the $i^{th}$ state, any one of the $i$ RNAP molecules can release a transcript with probability $p$ and $i$ RNAP molecules are working in parallel to produce mRNA molecules. Thus,

$$\mu_i^{mR} = i \times \frac{1}{T_{\text{el long}}} \times p$$

where $T_{\text{el long}} = \frac{L_{\text{gene}}}{k_{\text{el long}}}$

(11)

With the birth and death rates computed as above, the birth-death process can be characterized by the set of differential-difference equations for the state probabilities (obtained from the Chapman-Kolmogorov forward equations [20]),

$$\frac{dp_{0}^{mR}(t)}{dt} = -\lambda_0^{mR}(p_{0}^{mR}(t) + \sum_{j=1}^{\infty} p_{j}^{mR}(t)) + \mu_1^{mR}p_{1}^{mR}(t)$$

(12)

and

$$\frac{dp_{j}^{mR}(t)}{dt} = -\lambda_j^{mR}p_{j}^{mR}(t) + \mu_{j+1}^{mR}p_{j+1}^{mR}(t) + \mu_0^{mR}p_0^{mR}(t)$$

(13)

where $p_{j}^{mR}(t) =$ probability of being in state $j$ at time $t$. Applying the stochastic balance procedure [20], the stationary state probabilities can be obtained from eqn. 12 and 13 for the chain in Fig.3 as,

$$P_{n}^{mR} = \left(\frac{\lambda_n^{mR}}{\mu_n^{mR}}\right)^{n-1} \frac{1}{n!\left(\mu_n^{mR}\right)^n}$$

where

$$\lambda_0^{mR} = \lambda_{\text{mr}}, \mu_n^{mR} = \mu_{\text{mr}}, n = 1, 2, \ldots N_{\text{RNAP}}$$

(14)

and

$$P_0^{mR} = \frac{1}{\sum_{n=1}^{N_{\text{RNAP}}} \left(\frac{\lambda_n^{mR}}{\lambda_{\text{mr}}^{mR}}\right)^{n-1} \frac{1}{n!\left(\lambda_{\text{mr}}^{mR}\right)^n}}$$

(15)

In the above formulation, the cessation of the transcription process by a repressor molecule has not been considered. As mentioned earlier, the repressor binding can prevent access of the RNAP molecule to the promoter region, thereby taking the system to an inactive state, $D_{\text{gene}}$. As this can happen from any state in the system, the modified Markov chain is represented by Fig. 4.

The Markov chain in Fig. 4 represents a special birth-death process with killing [15, 17]. Based on Karlin and McGregor’s integral representation for the transition probabilities ($Q$ matrix), it has been shown in [16] that the representation can be extended in case where transition to the death state occurs from any other state, as in Fig. 4. Also, [17] shows that as long as killing is possible from a finite number of states and the criteria of certain absorption and positive decay parameter are maintained, existence of quasi-stationary distributions for a birth-death process with killing are maintained. This particular chain can now be analyzed in a quasi-stationary manner as shown in [17] and reduced to a simple birth-death process with the following mapping for the birth-death process representing the chain,

$$\mu_0^{mR} = 0, \quad \lambda_0^{mR} = \lambda_{\text{mr}}^{mR} + \lambda_{\text{el}}^{mR}, \quad \mu_i^{mR} = \frac{(\lambda_{i-1}^{mR})\mu_{i-1}^{mR}}{\lambda_{i-1}^{mR}}$$

(16)

Using the above transformations for the modified birth and death rates in eqn. 13 and 14, the individual state probabilities, $P_{n}^{mR}$ can be computed. Now, $P_{n}^{mR}$ gives the probability of $n$ RNAP molecules in the system producing transcripts at the rate $\mu_n^{mR} = n \times \frac{1}{T_{\text{el long}}} \times p$.

Thus, the average rate of transcript synthesis, $R_{\text{mr}}$ is given by,

$$\overline{R}_{\text{mr}} = \sum_{n=0}^{N_{\text{RNAP}}} P_n^{mR} \mu_n^{mR}$$

(16)

and the variance is given by,

$$\sigma_{\text{mr}} = \sum_{n=0}^{N_{\text{RNAP}}} (\mu_n^{mR} - \overline{R}_{\text{mr}})^2 P_n^{mR}$$

(17)
The first and second moments [19] i.e. mean and variance respectively, specified by eqn. 16 and 17 characterize the probability distribution of mRNA synthesis rate or frequency of transcript generation. In the next sub-section, we develop the model for the distribution of the number of proteins synthesized from an mRNA molecule.

B. Modeling Translation Dynamics

Once an mRNA molecule is transcribed, the translation machinery is recruited to initiate the synthesis of proteins. Although translation can be initiated even before the transcription machinery has released the mRNA in prokaryotic cell, we consider the recruitment of the translation machinery after the complete transcription of an mRNA molecule in this model.

We observe a commonality in the role of the ribosomal unit in translation as played by the RNAP molecule in transcription. Translation is initiated by a ribosome macromolecule binding to the ribosome binding site (RBS) on the mRNA. The ribosome then reads out the genetic code from mRNA in three-letter codons corresponding to the amino acids, assembling them into a growing chain of amino acids which subsequently fold into a functional protein.

Concurrency is also observed [14] in the translation process, with multiple ribosomes (forming polysomes) reading out simultaneously from a transcript. Another characteristic in the translation machinery is the competition between a ribosome and a degradosome (RNaseE) molecule to bind to the RBS and initiate translation or decay of mRNA respectively, as shown in Fig. 5(a).

Based on the above observations and the fact that the number of ribosomes attached to an mRNA governs the count of protein molecules synthesized before decay of the mRNA, we consider \( X_p(t) \), the number of ribosomes attached to the mRNA at time \( t \). The maximum number of ribosomes (ribosome load) simultaneously processing an mRNA depends on the length of the gene and the inter-ribosome spacing (\( L_{\text{ribosome}} \)) and is given by, \( N_{\text{ribosome}} = \left\lfloor \frac{L_{\text{gene}}}{L_{\text{ribosome}}} \right\rfloor \) i.e. \( X_p(t) \) takes values in the discrete state space, \( S_{\text{protein}} = \{0,1,\ldots,N_{\text{ribosome}}\} \). Analyzing the events which can occur at a state \( X_p(t) = k \) as in the case of transcription (as shown in Fig.5(b)), we have,

(i) After the time taken for the ribosome to move \( L_{\text{ribosome}} \) base pairs, another ribosome can get attached to the mRNA depending on the length of the gene and the inter-ribosome spacing and takes the system to state \( k+1 \).

(ii) A previously attached ribosome can complete the process of protein generation and gets released from the system bringing it to state \( k-1 \).

(iii) An RNaseE molecule can get attached to the RBS and start the decay process taking the system to a ‘death’ or ‘killing’ state.

We now define a Markov chain, \( \chi_p = \{X_p(t), \ t \in T\} \) and \( T = \{t: 0 \leq t < \infty\} \)
where \( X_p(t) \) is the state of the system at time \( t \). As in the case of transcription, the Markov chain can be analyzed as a birth-death process with a killing or decay state, for translation. The main differences are the computation of birth and death rate of the translation process and the computation of the maximum possible states of the system.

(i) Computation of birth rate \( \lambda'_i \): The birth-rate captures the rate at which the system moves forward in state, i.e., another ribosome unit binds to the RBS successfully. This event occurs after the time taken for the ribosome unit bound to the RBS in the previous \((k-1)\)th state to move \( L_{\text{ribosome}} \) units along the mRNA chain to clear the next ribosome for association at the RBS, together with the time taken for the ribosome to bind to the RBS and initiate translation. Thus,

\[
\lambda'_i = \frac{1}{T_{\text{binding}} + T_{\text{clearance}}} = \frac{1}{k^{p}_{\text{on}} + L_{\text{ribosome}}/k^{p}_{\text{long}}} \quad \text{(18)}
\]

It may be noted here that, \( \lambda'_0 = \frac{1}{T_{\text{binding}}} = k^{p}_{\text{on}} \), since ribosome clearance time is not required for the first ribosome binding to the RBS.

(ii) Computation of death rate \( \mu^p \): The death event signifies a ribosomal unit successfully completing the synthesis of a protein molecule and getting released from the system. This can be computed on the same lines at the birth-rate with the observation that the ribosome here has to traverse 3 more codons to generate the last amino acid in the chain and stop the synthesis of the protein. At the \( i \)th state, any one of the \( i \) proteins can successfully complete protein chain synthesis with probability \( p_{\text{ribosome}} \). The assembly of the protein in the ribosomal complex is progressing in parallel, so the rate of protein production will be multiplication of the number of parallel production stations (ribosome units) to the rate of an individual station to complete production of a protein. Thus, death rate is given by,

\[
\mu^p = i \times \frac{1}{T_{\text{long}}^{p}} \times p_{\text{ribosome}} \quad \text{where} \quad T_{\text{long}}^{p} = L_{\text{gene}}/k^{p}_{\text{long}} + 3 \quad \text{(19)}
\]

(ii) Decay Rate (\( k_{\text{mRNA}} \)): The decay rate is defined by the binding rate of the RNase E molecule to the RBS which triggers the mRNA decay process (as outlined in Table 1).

Based on the above rates, the stationary state probability of the states, \( P^p_n \), as,

\[
P^p_n = \left( \frac{\lambda^p_n (\lambda^p)^{n-1}}{n! (\mu^p)^n} \right) P^p_0
\]

where \( \lambda^p_n = \lambda^p_n, \mu^p_n = \mu^p, n = 1,2,..N_{\text{ribosome}} \quad \text{(20)} \)

and

\[
P^p_0 = \frac{1}{\sum_{n=1}^{N_{\text{ribosome}}} \frac{(\lambda^p)^{n-1}}{n! (\mu^p)^n}} \quad \text{(21)}
\]

Once the state probabilities are computed, we observe that at the \( n \)th state, the rate of protein synthesis is given by \( \mu^p_n \) and the probability of this synthesis rate at that state is \( P^p_n \). The average
rate of protein synthesis, \( \bar{R}_p \), can thus be computed as
\[
\bar{R}_p = \sum_{n=0}^{N_{\text{mRNA}}} P_n^p \mu_n^p \tag{22}
\]
and the variance is given by,
\[
\sigma_p^2 = \sum_{n=0}^{N_{\text{mRNA}}} (\mu_n^p - \bar{R}_p)^2 P_n^p \tag{23}
\]

Based on the above equations, we characterize the probability distribution of the number of proteins available from a mRNA or the burst size distribution for protein synthesis.

C. Combined Model of Protein Synthesis

As mentioned earlier, the process of gene expression is marked by coupling between the processes of transcription and translation. Thus, the probability distribution for protein synthesis is a combined process arising out of the two stochastic processes of transcription and translation elucidated in the previous sub-sections. Now, the probability distribution of protein arrival, specifically the time between two protein molecules being synthesized will depend on the number of mRNA molecules present in the system. In order to compute the stationary probability of \( k \) mRNA molecules being present in the system, we consider the transcript arrival and decay process as a birth-death Markov chain with infinite state space, where the arrival rate is given by \( \tilde{R}_{\text{mR}} \) computed in eqn. 16, and death rate is the mRNA decay rate or \( \mu_{\text{decay}} \).

Solving the Markov chain, the probability of \( k \) mRNA transcript in the system is given by
\[
P_k^\text{transcript} = \frac{\left( \tilde{R}_{\text{mR}} \right)^n}{n! \left( \mu_{\text{decay}} \right)^n} P_0^\text{transcript}
\]
\[
P_0^\text{transcript} = \frac{1}{\sum_{n=1}^{\infty} \left( \frac{\tilde{R}_{\text{mR}}}{\mu_{\text{decay}}} \right)^n} \tag{24}
\]

Now, each of these \( k \) transcripts has a corresponding protein synthesis distribution with average rate \( \bar{R}_p \) given by eqn. 22. Thus, we can define the distribution for the time-interval between two protein molecules (inter-arrival time) released in the combined system by an Erlang distribution [19] with shape parameter \( k \) and rate parameter \( R_p \), i.e. \( F_{\text{erlang}}(x, k, \bar{R}_p) \). Specifically, the CDF of protein inter-arrival time is given by,
\[
F_{\text{inter}}(\tau_p) = \sum_{k=1}^{\infty} P_k^\text{transcript} \cdot F_{\text{erlang}}(\tau_p, k, \bar{R}_p) \tag{25}
\]

D. Modeling Noise Dynamics

The quantification of the fluctuation in the number of proteins produced is a key component in understanding the stochasticity of gene expression. Generically, this fluctuation can be defined as the ratio of the variance over mean squared which allows separation of noise sources [13]. If \( P(t) \) is the protein concentration at time \( t \), then the protein noise \( \eta(t) \) is given as [10],
\[
\eta(t) = \frac{\left\langle P(t)^2 \right\rangle - \left\langle P(t) \right\rangle^2}{\left\langle P(t) \right\rangle^2}
\]
\[
\left\langle P(t) \right\rangle = \text{ensemble average} \tag{26}
\]

In [13], the author has elucidated that the classification of noise components in terms of ‘extrinsic’ or ‘intrinsic’ depends on the definition of the system versus the environment. Thus, the noise contribution of the transcription machinery is intrinsic to transcriptional noise while extrinsic to translation noise. Based on our probability distribution for burst frequency and size, we can characterize the total protein noise (\( \eta_{\text{protein}} \)) as sum of transcriptional noise (\( \eta_{\text{transcription}} \)) and translational noise (\( \eta_{\text{translation}} \)), given as,
\[
\eta_{\text{protein}} = \eta_{\text{transcription}} + \eta_{\text{translation}}
\]
where,
\[
\eta_{\text{transcription}} = \frac{\sigma_{\text{mR}}}{(R_{\text{mR}})^2} \quad \text{and} \quad \eta_{\text{translation}} = \frac{\sigma_p}{(R_p)^2}
\]

In the next section, we estimate the burst frequency and size distributions based on our model and validate it with experimental data obtained from single cell experiments in \textit{E.Coli}. 


IV. MODEL VALIDATION

In this section, we validate our stochastic models for gene expression with experimental data obtained from recent single cell experiments. Specifically, we estimate the average rate of mRNA synthesis and protein generation from our model and validate the distributions with actual experiments data reported in [7], [8] wherein measurements at the single cell level were conducted to obtain average transcript synthesis rate and protein number distributions for the lacZ gene expression in *E.Coli* cells. We briefly overview the lac operon system in *E.Coli*, before presenting the model parameters and validation results.

**A. The lac operon experimental system**

The lac operon, which encodes a set of genes for the lactose permease, has been extensively used in experimental and analytical systems. The native lacZ gene is the first in the operon and is translated into monomers composing of the catalytic protein, β-galactosidase and galactose. The native operon is positively regulated in the presence of lactose and negatively regulated by glucose. The lac operon also encodes LacY (lactose permease) and LacA (galactosidase acetyltransferase) apart from LacI which encodes a regulatory protein. The lac operon, together with the gene products is depicted in Fig.6. We focus on the first gene of the operon, lacZ, to validate our model against experimental data obtained on single cell experiments for protein synthesis for the lacZ system [7, 8].

![Fig.6 Birth-death with killing state markov chain for transcription](image)

**Table 2. Model parameter values**

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Value</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter Clearance</td>
<td>30-60bp</td>
<td>Based on reported values in [15]</td>
</tr>
<tr>
<td>base-pair length (lacZ)</td>
<td>3075</td>
<td>Ecocyc database</td>
</tr>
<tr>
<td>Activation rate</td>
<td>&lt; 1.6x 10^-4/s</td>
<td>Based on reported values in [12]</td>
</tr>
<tr>
<td>Inactivation rate</td>
<td>0.0083/s</td>
<td>Based on reported values in [12]</td>
</tr>
<tr>
<td>Transcription initiation rate</td>
<td>0.1/s</td>
<td>Based on reported values in [15]</td>
</tr>
<tr>
<td>Open-complex Formation rate</td>
<td>1/s</td>
<td>Based on values reported in [15]</td>
</tr>
<tr>
<td>Transcript elongation rate</td>
<td>30-80 bp/s</td>
<td>Based on reported values in [10]</td>
</tr>
<tr>
<td>Ribosome clearance</td>
<td>80-120 bp</td>
<td>Based on reported values in [10]</td>
</tr>
<tr>
<td>Translation elongation rate</td>
<td>15-45bp/s</td>
<td>Based on reported values in [10,15]</td>
</tr>
<tr>
<td>Ribosome binding rate</td>
<td>10^9/M/s</td>
<td>Based on reported values in [10]</td>
</tr>
<tr>
<td>Degradosome binding rate</td>
<td>0.1-0.006/s</td>
<td>Based on reported values in [10]</td>
</tr>
<tr>
<td>mRNA decay half life</td>
<td>1.5min</td>
<td>Based on reported values in [10,8]</td>
</tr>
<tr>
<td>Protein half life</td>
<td>3hrs</td>
<td>Based on reported values in [10]</td>
</tr>
<tr>
<td>Cell cycle Time</td>
<td>60mins</td>
<td>Based on reported values in [10,8,9]</td>
</tr>
</tbody>
</table>

**B. Model parameter estimation and validation**

The stochastic model for gene expression elucidated in the previous section computes the distributions for mRNA synthesis (burst frequency) and protein synthesis (burst size) characterized in terms of their mean and variance. The mathematical expression for the average rates (eqn. 16 and 22) involve different biological parameters which are obtained from available literature data to estimate the model parameters (Table 2 lists the different model parameters). Based on single cell experiments of gene expression in *E.Coli*, the average number of mRNA bursts was estimated as 1.2 bursts/cell cycle in [7, 8]. The model estimate for $R_{\text{mRNA}}$, as given in eqn. 16, using the parameter values in Table 2 was calculated as 1.03 ± 0.85 burst/cell cycle time. Based on the fluorescent $\beta$ -
galactosidase reporter molecules, details of which are provide in [8], the authors obtain an average burst size of 4.2 ± 1.8 protein molecules. Now, the authors also showed that a burst of proteins occurred from a single mRNA (0.037 ± 0.013) where the average life-time of the mRNA is 1.5 ± 0.2 mins. Thus, assuming that 4.2 protein molecules are produced from a single mRNA during its life time, we obtain an average protein generation experimental rate of 0.046/s for $\tilde{R}_p$ using eqn. 22. Based on $\tilde{R}_p$ and the associated probability, $P_n^{mR}$, Fig.7 shows the probability distribution for burst size of proteins computed from the model and compared with experimental data [8]. Thus, the validated protein burst size distribution and mean rate estimated from the model provides a mathematical tool for systematically studying the effect of different parameters, which we elucidate next.

V. MODEL ANALYSIS

We use our validated model to analyze the dynamics of gene expression, specifically steady state sensitivity analysis of the model with respect to different biological parameters. We focus on the sensitivity of mRNA synthesis rate on gene activation ratio and transcription initiation rates based on our model.

(i) Activation Ratio: The activation ratio, defined as the ratio of the gene activation rate to the deactivation rate ($\frac{\lambda}{\tilde{\lambda}}$), quantifies the strength of the transcription factor. Fig. 8\(^1\) shows the effect of increasing activation ratio on the transcript production rate and the corresponding transcriptional noise. As seen from the plot, increasing activation ratio increases the efficiency the transcriptional machinery thereby increasing the average rate of transcript production $\tilde{R}_p$ and decreasing transcriptional noise, $\eta_{\text{transcription}}$.

(ii) Transcription Initiation Rate: The rate of binding of the RNAP molecule to the promoter region exerts control on the rate of transcription by increasing the efficiency of the transcription process. Thus, as seen from Fig. 9, increase in the rate of transcription initiation increases the average rate of transcript synthesis and decreases $\eta_{\text{transcription}}$. It may be noted here that the activation ratio has a greater effect on the efficiency of transcription as seen from the slope of the curves in Fig.8 and Fig.9 and noted in earlier work [9].

In studying the effect of different molecular actors on the translational machinery, we focus on the role of the ribosomal units and its ‘competition’ with the degradosome (RNase E) molecule.

\(^1\) All plots have error bars for the computed average values as well as arrows demarcating the region where experimental data values are validated for the \textit{lacZ} gene.
(iii) **Ribosome clearance**: The clearance region of the ribosome on the mRNA controls the number of ribosomal units (ribosome load) that can concurrently translate an active mRNA molecule. By increasing the spacing between two ribosome molecules, the load decreases, thereby decreasing the protein synthesis rate and increasing translational noise $\eta_{\text{transl}}$, depicted in Fig. 10.

(iv) **Ribosome and Degradosome binding rates**: As mentioned earlier, the degradosome competes with the ribosome for binding to the RBS on the mRNA molecule. Thus, increasing the degradosome binding rate will increase the mRNA degradation event rate thereby decreasing the protein synthesis rate. However, the dynamics of the overall competition are controlled by the ribosome binding rate and the degradosome binding rate. In Fig. 11, we show the interplay of these two effects as a surface plot. As seen from the graph, for the $\text{lacZ}$ gene parameters, the ribosome exerts greater control on the protein synthesis rate compared to the degradosome.
VI. SIMULATION STUDY

While a static sensitivity analysis captures the steady state behavior of a biological process, gene expression in this case, it is pertinent to consider a dynamical system where transcription and translation are coupled and the protein noise ($\eta_{protein}$) is manifested through the interaction of these coupled mechanisms. Using the parameterized probability distributions of transcription and translation, captured in eqn. 16-17 and eqn. 22-23 respectively, we build a discrete event simulator (details in Additional Materials and [20]) which allows a time-course study of mRNA and protein profiles for transcription and translation events. We conduct simulation studies to characterize ‘protein bursts’ and identify the molecular actors controlling ‘burstiness’.

(a) Simulating translation bursts: Simulations was conducted with the experimentally validated transcriptional burst frequency and protein burst size distributions reported in the previous subsections. Fig. 12 shows the event dynamics along with the protein and total protein noise profiles. As seen from the plots, the simulation shows the burstiness of protein production over a 3 cell cycle time periods as experimentally reported in [8,9]. The low number of mRNA molecules during the simulated time shows the rarity of transcription events while the noise profile reflects the fluctuations in protein count with the bursts.

(b) Simulating disappearance of bursts: The event dynamics in Fig. 12 show that translation bursts are caused by the rare transcription events (i.e. large value of $\tau_{mR}$). Thus, we conducted simulations with increasing burst frequency of transcription (i.e. decreasing $\tau_{mR}$) by increasing the activation ratio of the promoter to study the dynamics of protein synthesis. As observed from Fig 13, increasing promoter strength increases the number of mRNA molecules, thereby reducing the burstiness of protein synthesis along with the noise profile. The simulation studies highlight the role of promoter strength in controlling transcription events and governing the nature of bursts in protein synthesis for the lacZ gene.
VII. DISCUSSION

Bacterial gene expression involves a complex process of interaction between multiple molecular actors acting individually or in a concurrent fashion. These actors contribute to the temporal fluctuations in the number of proteins produced in a cell. In this paper, we have modeled the stochastic process of gene expression, incorporating the effect of various actors in parameterized probability distributions for mRNA and protein synthesis. The parameterized distributions help in systematically analyzing the sensitivity of the noise in protein production to the different molecular actors.

The model can be enhanced to incorporate the impact of different molecular mechanisms not considered in the current scope, particularly the effect of promoter arrangements, number of regulators, switch-like signal propagation and the chromosomal positioning of genes [11]. Our current work is focused on expanding the stochastic model for gene expression and incorporating it in a genome-scale discrete event simulation of regulatory and metabolic network dynamics in a bacterial cell.

VIII. ADDITIONAL MATERIALS

Supplementary materials containing the derivation of quasi-steady state markov model with killing state, are available online at the following link: http://crewman.uta.edu/~sghosh/suppl.pdf.

Further sensitivity analysis of the model together with simulation results on various *in silico* experiments are also provided.

REFERENCES


Fig. 13 Simulation results showing the effect of increased burst frequency on protein synthesis


