Inferring distributions from observed mRNA and protein copy counts in genetic circuits

Komlan Atitey, Pavel Loskot and Paul Rees

College of Engineering, Swansea, United Kingdom

E-mail: komlan.atitey@swan.ac.uk and p.loskot@swan.ac.uk

Abstract

Defining distributions of molecule counts produced in the cell can elucidate stochastic dynamics of the underlying biological circuits. For genetic circuits, only a few distributions of messenger RNA and protein counts were reported in literature, so the task is to decide which of these candidate distributions best fit the observed data. In this paper, we present a statistical method to infer distributions of mRNA and protein counts from observed data. The main advantage of this method is that it does not require any prior assumptions or knowledge about underlying chemical reactions. In particular, a given distribution is fitted to the observed copy counts using a histogram with optimized bin sizes in order to reduce the fitting error. The goodness of fit is evaluated by Kolmogorov-Smirnov and chi-square statistical tests to accept or reject the hypothesis that observed molecule counts were generated from given distribution. The distribution fitting also yields the values of distribution parameters, or they can be estimated using the Bayes theorem. These parameters appear to be themselves random processes. The presented statistical framework for analyzing the observed mRNA and protein copy counts is illustrated for a simulated model of lac genetic circuit in Escherichia coli. For reaction rates assumed in the model, the results in literature predict that mRNA and protein counts at steady-state are gamma distributed. Our analysis shows that both mRNA and protein in the lac circuit model can be considered gamma distributed in at least 70% of times from the initial state until steady-state. The shape and scale parameters of observed gamma distributions are also gamma distributed, giving rise to double stochastic processes. More importantly, as shown previously, the distribution parameters are functions of transcription and translation rates, so presented statistical framework can be used to estimate or optimize reaction rates in biochemical systems.

Keywords: Bayesian inference, gamma distribution, gene expression, goodness of fit, Markov chain Monte Carlo sampling

1. Introduction

Proteins are the most versatile building blocks of biological circuits. Protein engineering has many industrial and biomedical applications [1]. The biological cells rely on complex networks of protein-to-protein interactions to carry out various living functions [2, 3] including responding to information signals from the extracellular environment [4, 5]. A classic example of such biochemical signal processing is chemotaxis of Escherichia coli (E. coli) which is mediated by a well-characterized signal transduction network [6]. Despite growing knowledge about molecular components of cellular circuits, their dynamics are much less understood [7]. These circuits can be often considered as having modular structures [8-10]. The modularity allows reusing the same sub-parts to design similar biological circuits with different functionality such as amplifiers, switches and oscillators. More importantly, to guide the design of these circuits in synthetic biology applications, it is crucial to have accurate statistical description of the underlying stochastic protein production and processing [13-16]. Such knowledge is presently still limited due to protein versatility in function, dynamics and interactions [1]. There is even less knowledge about production statistics of the corresponding messenger

RNA (mRNA) which is a key component to translate the protein. Moreover, mRNA has been recently suggested as a novel target for controlling the protein production [11, 12].

Theoretical modeling of stochasticity in gene circuits has been subject to intense investigations to elucidate the effects of stochastic noise on the gene expression [17-19]. The stochasticity of protein and mRNA synthesis in the cell can be described by chemical master equation (CME) [20]. Solving the CME yields a time evolution of mRNA and protein distributions. However, in most cases, the CME is mathematically intractable, so it is solved numerically using simulations [21-23]. Mathematical models and simulation algorithms can also inform experimental techniques [20].

The existing studies of gene expression circuits usually assume random bursts of proteins with exponentially distributed number of molecules. The assumption of protein lifetime being longer than mRNA lifetime has been made in models considered in [24-27] in order to simplify the model analysis. Such assumption was shown to yield gamma distributed protein synthesis at steady-state. In addition, the observed steady state protein distributions are not symmetric, so they are poorly characterized by their mean and variance [28]. Simple 2-stage and 3-stage models of gene expression were solved analytically in [28] and [29] to provide time evolutions of reaction rates dependent protein distributions. It is shown that protein distribution can vary significantly depending on specific reaction rates and initial molecule counts. However, none of these works considered the statistics of parameters of protein distributions such as scale and shape in case of gamma distribution.

Predictions of protein distributions presented in [28] and [29] are limited by knowledge of reaction rates. In laboratory experiments, such knowledge is at best limited or not available at all. In addition, gene regulatory networks of studied biological systems may not be fully known or not easily approximated by 2-stage or 3-stage model of gene expression. In such scenarios, it is necessary to obtain empirical distributions from observed molecule counts, since simple sample mean and sample variance are not representative of asymmetric distributions [28].

In silico experiments can readily produce large amount of data of molecule counts. Consequently, our aim is to investigate a statistical methodology to identify the distribution of molecule count time series which best describes the observed data. The main advantage of obtaining the distribution empirically is that it does not require any assumptions, and can be used even when the underlying reaction rates are not known, and when the regulatory reaction networks are complex. Equivalently to selecting the best distribution from a set of candidate distributions, we evaluate the hypothesis that the selected distribution is a good fit to the observed data, since none of the candidate distributions may be a good fit.

Our numerical experiments were carried out for a lac circuit model of E. coli fully specified in [40]. This model contains a single positive feedback loop with 14 chemical species interacting in a network of 23 chemical reactions. Stochastic simulations were performed in the Lattice Microbe software [31] using the Gillespie algorithm [30-32]. The stochastic traces of mRNA and protein counts are statistically independent, and reflect the cell-to-cell variability in otherwise identical cell populations [33]. The simulations were run over the span of cell half lifetime which is about 1 hour for E. coli in order to guarantee the existence of steady-state for both mRNA and protein production. The data from simulations are then processed to infer time dependent molecule distributions with their parameter values. For reaction rates considered, the works [24-29] predicts that mRNA and protein counts in the model of lac genetic circuit in E. coli should be gamma distributed. Our objective is to verify this prediction for mRNA and protein during the transition from initial state and also at steady state.

We use goodness of fit to measure how well the selected distribution fits the histogram of observed molecule counts. In particular, the histogram at each observation time instant is optimized in order to reduce the fitting error before running the Kolmogorov-Smirnov and chi-square statistical tests [34, 35]. These tests yield the significance levels for testing the hypothesis that observed data are from a given distribution. We found that the assumed gamma distribution has time varying shape and scale parameters which themselves appear to be random processes. The joint distribution of shape and scale parameters can be statistically inferred using a Bayesian framework [36, 37]. The bivariate posterior distribution of scale and shape parameters conditions on observed mRNA and protein counts can be visualized using Markov chain Monte Carlo (MCMC) methods such as the bivariate Metropolis-Hasting (BMH) sampler [38, 39]. Finally, we also measured auto-correlation of shape and scale random processes to infer their statistical dependency across time and to estimate the correlations between mRNA and protein productions.

2. Methods

2.1 Statistical description of mRNA and protein abundances in gene expression

The reaction rates in the lac circuit considered are kept constants and set to default values specified in [40]. It is predicted in [28] and [29] that, for these values of transcription and translation rates, and mRNA and protein degradation rates, the protein synthesis in steady-state should yield gamma distribution. We produced 10,000 independent time trajectories of mRNA and protein counts

over 1 hour of cell half lifetime to obtain statistically meaningful amount of data for further analysis [31, 49].

It has been established that gene expression can be modeled as a three-stage process consisting of transcription, translation, and switching of the promoter between active and inactive states [29, 41]. Recent single-cell studies confirmed stochasticity of gene expression [42-44], and the bursting nature of mRNA and protein synthesis [45] where intermittent production bursts are separated by periods of inactivity [46]. A conceptual model of gene circuit is given in Figure 1. The full model of the simulated lac circuit is presented in [40]. Even though the model in Figure 1 and to some extent also more complex model from [40] are biologically simplistic [47], both models are useful to generate time dependent data of mRNA and protein counts for developing statistical methods of data analysis to elucidate dynamics of stochastic systems including gene expression in prokaryotic cells [48].

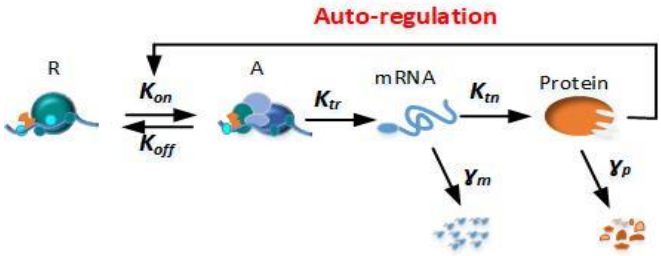

Figure 1. The gene expression model. (R) The repressed promoter when the repressor binds to the operator. (A) An active promoter when the RNA polymerase binds to the promoter. Each step represents several biochemical reactions which are associated with transitions between two promoter states (repressed and active). The mRNA and the protein production during transcription and translation, respectively, are followed by their degradation. K_{off}, K_{on},K_{tr}, K_{tn}, γ_m, and γ_p are the rate constants associated with the these steps as indicated. The auto-regulation step controls the protein production. The reaction steps involve binding and dissociation events which are occurring randomly.

Ignoring spatial heterogeneity in the cell, time evolution of mRNA and protein counts can be obtained by solving the corresponding CME. The full state of CME contains copy counts or concentrations of all chemical species [30, 31, 49], however, we are only interested in abundances of mRNA and protein. The complexity of CME for the lac circuit model considered necessitates stochastic simulations. At each observation time, the probability distributions of mRNA and protein counts are estimated using optimized histograms. We note that while exact molecule counts are readily available from in silico experiments, it is never the case in in vivo and in vitro experiments.

Denote as $y_1, \ldots, y_k$ the time series of mRNA or protein copy counts observed over time interval $k \in [0, T]$ where T is the maximum simulation time which is assumed to be half lifetime of E. coli, i.e., *T=3600* s [50]. For n independent simulation replicas, our observation data are, $\{y_{i,k\in[0,T]}\}_{i=1}^{n}$. The task is to estimate the probability density function[1] (PDF) f_k of random variables y_k including their parameter values at all times k.

2.2 Time dependent PDF estimation of mRNA and protein counts from observed data

We use histogram as an unbiased and consistent estimator of PDF [51]. The histogram counts the fraction of samples that fall into the predefined bins. The number and width of the bins affect the accuracy of PDF estimation. Thus, when the bins are too narrow, the histogram may have large variations between neighboring bins with a number of empty bins in between. On the other hand, the histogram has poor resolution, if the bins are too wide. In both cases, the accuracy of fitting PDF to the histogram is reduced [52]. It is therefore desirable, especially for smaller values of n (i.e., amount of data), to optimize the bins to achieve better accuracy [53].

Denote as $R_k = y_{k,max} - y_{k,min}$ the value range of data considered at specific time k. For N equal size bins, the bin width is, $h_k = R_k/N = const$. The optimum bin sizes minimize the cost function [53, 54]:

$$h_k^* = \arg\min_{h_k} C_{n,k}(h_k) \tag{1}$$

Provided that $q_{k,i}$ are relative frequencies assigned to the i-th bin at time k, the optimum bin sizes in (1) are obtained by the following procedure.

1. Divide the observation range R_k into N disjoint equal size bins of width h_k, and count the bin values $q_{k,i}$ assuming all n data samples at time k.
2. Compute the sample mean and variance of $q_{k,i}$ as:

$$\bar{q}_k = \frac{1}{N}\sum_{i=1}^{N} q_{k,i} \tag{2}$$

$$v_k = \frac{1}{N-1}\sum_{i=1}^{N}(q_{k,i} - \bar{q}_k)^2 \tag{3}$$

3. Finally, iteratively minimize the cost function by adjusting the number and width of the bins as:

$$C_{n,k}(h_k) = \frac{2\bar{q}_k - v_k}{(nh_k)^2} \tag{4}$$

Once the optimized histogram has been obtained, it can be fitted with the selected PDF expression [55]. It is very useful to automate the whole process of obtaining and fitting

[1] Strictly speaking, probability mass function (PMF) should be considered as PDF is a continuous approximation of discrete molecule counts.

the histogram, since this procedure needs to be repeated for all time instances of interest.

As discussed above, we can assume that the production bursts of mRNA and protein yield exponentially distributed number of molecules, and assuming the protein lifetime being much larger than the mRNA lifetime, the steady-state protein counts y are gamma distributed [24]:

$$p(y) = \frac{y^{a-1}e^{-y/b}}{\Gamma(a)b^a} \quad (5)$$

where the shape parameter $a = K_{tr}/\gamma_p$ is equal to the number of mRNA molecules produced per cell cycle, and the scale parameter $b = K_{tn}/\gamma_m$ represents the number of protein molecules produced per translation burst from one mRNA molecule. Consequently, we can statistically test whether gamma distribution is a good description of observed protein counts also in transition from the initial states, and whether gamma distribution can be also assumed to describe the mRNA production.

2.3 Statistical tests for fitting gamma distribution to observed mRNA and protein data

We consider the Kolmogorov-Smirnov test and the chi-square test to measure the goodness of fit of gamma distributions to mRNA and protein abundances at different time instances.

Kolmogorov-Smirnov statistical test

The Kolmogorov-Smirnov (K-S) test is a statistical measure to compare two cumulative distribution functions (CDFs). In particular, let $\hat{F}_k(y)$ be the CDF of gamma distribution to be statistically compared with the empirical CDF $F_k(y)$ obtained from the observed data [56, 57]. The empirical CDF $F_k(y)$ for the observed random molecule counts $y_{k,1}, \dots, y_{k,n}$ is computed as:

$$F_k(y) = \frac{I_k(y)}{n} \quad (6)$$

where n is the sample size, and $I_k(y)$ counts the number of samples smaller than y. The K-S test statistic D_k is defined as the maximum absolute difference between the empirical CDF $F_k(y)$ and the hypothetical CDF $\widehat{F}_k(y)$, i.e.:

$$D_k = \sup_y \left\{ \left| F_k(y) - \widehat{F}_k(y) \right| \right\} \quad (7)$$

The null hypothesis that the observed data can be described by the hypothetical CDF is rejected, provided that the statistic D_k is larger than a critical value obtained from the K-S table of significant values [58, 59]; otherwise, the null hypothesis is accepted. In addition, for each K-S test, we also determine the level of significance that the null hypothesis is true.

Chi-square statistical test

In order to detect any bias of the K-S test, we also use the chi-square test to decide whether gamma distribution is a good description of observed data. The empirical distributions of observed data at selected time instances are again obtained by first optimizing the bin sizes of the histogram, and then calculating the empirical CDF (6). The time dependent two-sided chi-square statistic is [60-62]:

$$x_k^2 = \sum_i \frac{\left(O_{k,i} - E_{k,i}\right)^2}{E_{k,i}} \quad (8)$$

where $O_{k,i}$ and $E_{k,i}$ are the observed and the expected relative frequencies, respectively, for the optimized bins computed at time k. For two-sided test, the chi-square statistic x_k^2 is compared with the tabulated upper-tail and lower-tail critical values [62, 63]. Provided that the chi-square statistic (8) is between these critical values, the corresponding significance level represents the level of acceptance of the null hypothesis that observed data can be described by the hypothetical CDF.

2.4 Time-varying shape and scale parameters of gamma distributed mRNA and protein counts

By fitting gamma distribution to observed mRNA and protein counts at different times, we obtain time series for shape and scale parameters of gamma distribution. These parameters appear as other random processes with their own PDF. Since these distributions have positive support and are positively skewed, we again assume that they are both gamma distributed.

Denote as $f_k(y_k|\alpha_k,\beta_k)$ the gamma distribution of mRNA or protein counts in the cell at time k where α_k and β_k are the shape and scale parameters, respectively. The parameters α_k and β_k are assumed to be gamma distributed, i.e., $\alpha_k \sim \text{Gamma}(a_k, b_k)$ and $\beta_k \sim \text{Gamma}(u_k, v_k)$. The task is to use the observed mRNA or protein counts $y_k = [y_{k,1}, \dots, y_{k,n}]$ at time k across n simulation traces to infer the joint PDF of shape and scale parameters α_k and β_k while assuming their marginal gamma distributions:

$$\pi(\alpha_k) = \frac{b_k^{a_k}}{\Gamma(a_k)} \alpha_k^{a_k-1} \exp(-\alpha_k b_k) \propto \alpha_k^{a_k-1} \exp(-\alpha_k b_k) \quad (9)$$

$$\pi(\beta_k) = \frac{v_k^{u_k}}{\Gamma(u_k)} \beta_k^{u_k-1} \exp(-\beta_k v_k) \propto \beta_k^{u_k-1} \exp(-\beta_k v_k) \quad (10)$$

Since the data across multiple simulation traces are independent, the likelihood of shape and scale parameters is given by the product:

$$p(y_k \mid \alpha_k, \beta_k) = \prod_{i=1}^{n} f_k\left(y_{k,i} \mid \alpha_k, \beta_k\right) = \prod_{i=1}^{n} \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} y_k^{\alpha_k - 1} \exp\left(-\beta_k y_{k,i}\right) \quad (11)$$

The likelihood function can be rewritten as:

$$p(y_k \mid \alpha_k, \beta_k) = \frac{\beta_k^{n\alpha_k}}{\Gamma(\alpha_k)^n} \prod_{i=1}^{n} y_{k,i}^{\alpha_k - 1} \exp\left(-\beta_k \sum_{i=1}^{n} y_{k,i}\right) \quad (12)$$

Using the Bayes theorem, the joint posterior PDF of shape and scale parameters can be calculated as:

$$p(\alpha_k, \beta_k \mid y_k) = p(y_k \mid \alpha_k, \beta_k)\pi(\alpha_k, \beta_k) / p(y_k) \quad (13)$$

By ignoring the proportionality factor $p(y_k)$, and assuming the independence of scale and shape parameters, so their joint prior distribution is the product of marginal distributions (9) and (10), we obtain the joint posterior distribution of scale and shape parameters conditioned on the observed mRNA or protein counts:

$$p(\alpha_k, \beta_k \mid y_k) \propto \frac{\beta_k^{n\alpha_k}}{\Gamma(\alpha_k)^n} \prod_{i=1}^{n} y_{k,i}^{\alpha_k - 1} \exp\left(-\beta_k \sum_{i=1}^{n} y_{k,i}\right) \times \alpha_k^{a_k - 1} \exp\left(-\alpha_k b_k\right) \beta_k^{u_k - 1} \exp\left(-\beta_k v_k\right) \quad (14)$$

Numerically evaluating the distribution (14) has complexity $O(n^2)$. Since n is typically very large, we can visualize the distribution (14) using the BMH sampler [38]. We can then investigate the convergence and mixing properties of the BMH sampler to generate samples α_k and β_k having the density $p(\alpha_k|\beta_k, y_k)$, and $p(\beta_k|\alpha_k, y_k)$, respectively.

2.5 Shape and scale parameters of gamma distribution as random processes

At steady state, it was shown in [24] and [25] that shape and scale parameters of gamma distribution are equal to the mean number of protein bursts per cell cycle, and the mean number of protein molecules produced per burst, respectively. However, our numerical experiments reveal that both shape and scale parameters appear to be random during the transition to steady-state, giving rise to double-stochastic processes of both mRNA and protein counts. For default constant values of mRNA and protein degradation rates, in the transition phase before steady-state, we can assume time-varying shape and scale parameters to be linearly dependent on equivalent time varying transcription and translation rates as:

$$a_k \approx \frac{1}{\gamma_p} K_{tr,k} \text{ and } b_k \approx \frac{1}{\gamma_m} K_{tn,k} \quad (15)$$

where the time dependence is removed at steady state as one would expect. Hence, for measured values of shape a_k and scale b_k during the transition phase, we can measure the equivalent time varying translation rates $K_{tr,k}$ and $K_{tn,k}$, respectively.

In addition to conditional joint bivariate distribution of scale and shape parameters (13), we investigate correlations of these parameters in time. As shown in [64], the auto-correlation plots can be used generally to infer the level of randomness in stochastic processes. The autocorrelations are plotted as functions of time lag which is defined as time difference between consecutive mRNA or protein production events. Since both shape and scale processes are found to be correlated in time, so do the productions of mRNA and protein in the genetic circuit. The autocorrelation plots are also indicative of when mRNA and protein production reaches steady-state.

3. Results

3.1 Fitting gamma distribution to measured mRNA and protein counts

We first obtain optimized histograms of mRNA and protein counts as the estimates of their distributions [65] as described in Methods. At each time instant, the observed value ranges and so do the optimum bin sizes are different. The number of samples to create the histogram is equal to the number of simulation replicas which is set to *n=10,000*. The mRNA and protein counts are recorded once per second from time *T=0* s until the time *T=3600* s, so that *2×3601* histograms are produced in total.

Examples of histograms for mRNA and protein counts generated from simulations of the lac circuit in E. coli at 6 selected time instances are compared in Figure 2 and Figure 3, respectively. We observe that mRNA distributions are usually heavily skewed with long tails whereas the corresponding protein distributions at the same time instances are skewed much less. In addition to fitting the PDF of gamma distribution to these histograms, we also produced cumulative histograms defined by Eq. (6) in order to investigate fitting of the CDF of gamma distribution to the empirically obtained CDF which is shown in last row of Figure 2 and Figure 3, respectively.

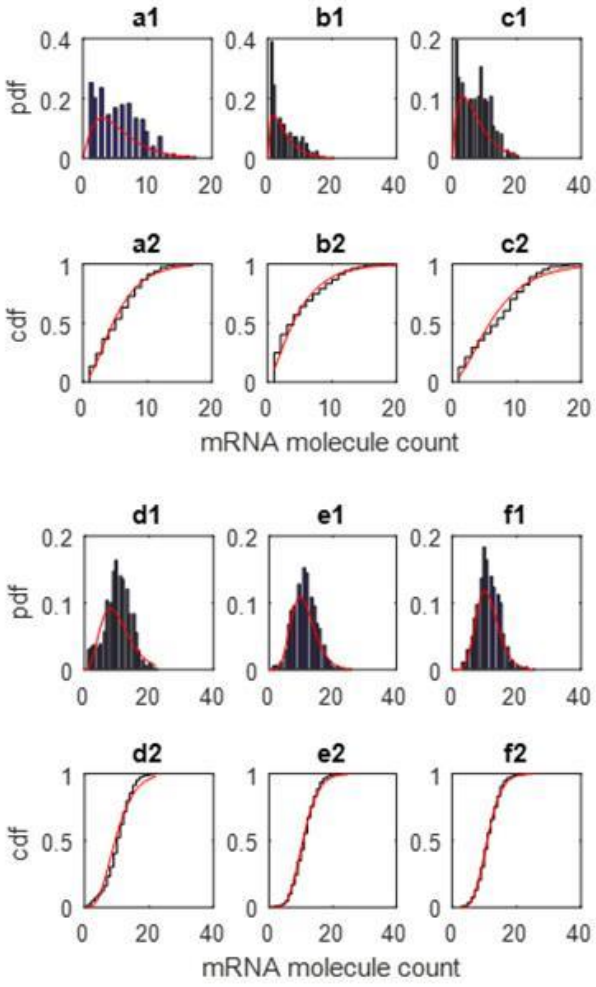

Figure 2. The distributions of mRNA counts synthesized in the lac circuit of E. coli. (a), (b), (c), (d), (e), and (f) corresponds to 100, 300, 600, 1000, 2000, and 3000 s, respectively. (1) The PDF fitting (red color) of gamma distribution to the histogram (blue color) of 10,000 independent samples of mRNA molecule counts. (2) The CDF fitting (black color) to the empirically computed CDF (red color).

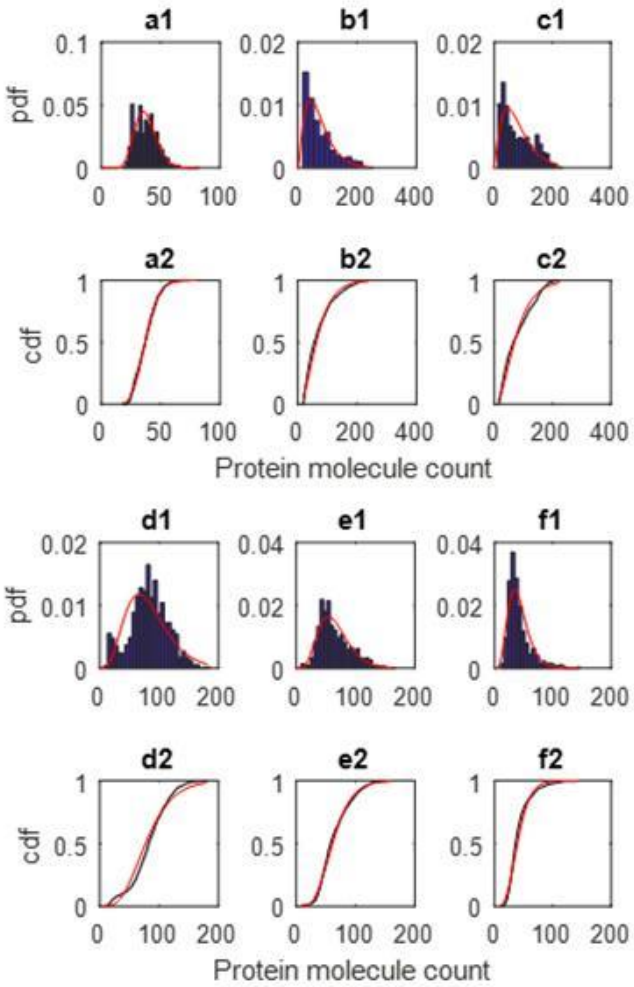

Figure 3. The distributions of protein molecule counts synthesized in the lac circuit of E. coli. (a), (b), (c), (d), (e), and (f) corresponds to 100, 300, 600, 1000, 2000, and 3000 s, respectively. (1) The PDF fitting (red color) of gamma distribution to the histogram (blue color) of 10,000 independent samples of protein molecule counts. (2) The CDF fitting (black color) to the empirically computed CDF (red color).

3.2 Goodness of fit tests

We now report the results of K-S and C-S statistical tests to assess the goodness of fitting gamma distribution to the measured histograms [66]. Specifically, at each time instant, we evaluated the significance level of the null hypothesis that the observed data are drawn from gamma distribution. Figure 4 shows the measured significance levels assuming the significance threshold *0.05* [56].

As shown in Figure 4 A1, only 980 K-S tests in time intervals [1310, 1440] (s) and [1475, 2290] (s) have the significance levels below the threshold value of 0.05. Thus, only about 26% of the K-S tests did not allow us to accept the null hypothesis that mRNA counts are gamma distributed. Assuming the significance levels in Figure 4 A2 for protein counts, only 1040 K-S tests in time interval [955, 1990] (s) (i.e., about 29% of all tests) did not allow us to accept the null hypothesis.

Recall that C-S test compares the empirically observed relative frequency and the expected relative frequency. The measured significance levels are compared with the tabulated critical values to decide whether the null hypothesis can be accepted. As for K-S test, the default threshold value for the significance level was set to 0.05 [56]. Figure 4 B1 shows that only 765 tests or 21% of all tests of mRNA counts in time interval [1310, 2075] (s) did not allow us to accept the null hypothesis. Finally, the significance levels in Figure 4 B2 for protein counts revealed that 1005 or 28% of the tests in time interval [1090, 2106] (s) did not allow us to accept the null hypothesis.

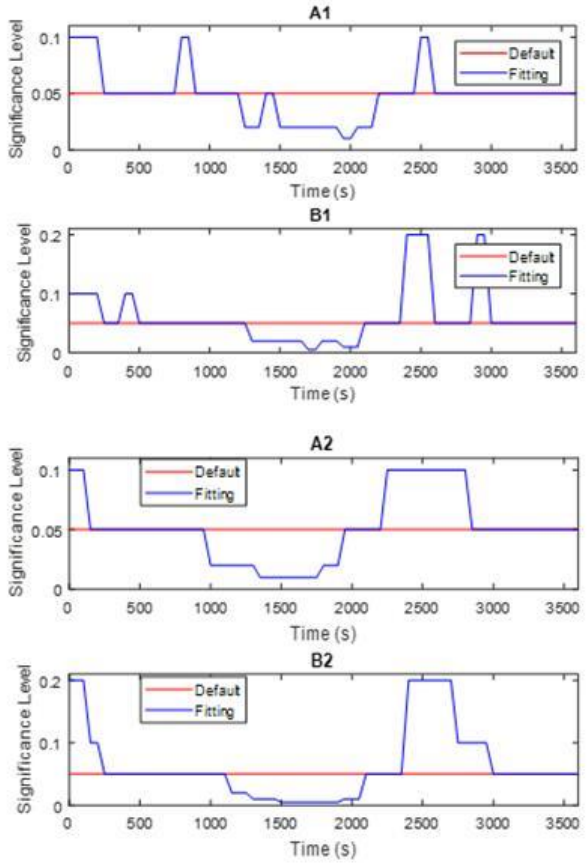

Figure 4. The goodness of fit tests. (A) Chi-square test and (B) Kolmogorov-Smirnov test. (1) mRNA molecule counts and (2) protein molecule counts. (Red color) Default significance level of *0.05* and (blue color) the measured significance levels for the null hypothesis.

In summary, in all cases considered, at least 70% of all statistical tests performed confirm that mRNA and protein counts are gamma distributed. Moreover, we observe from Figure 4 that towards the end of the cell cycle, the significance levels are equal to the significance threshold; thus, at steady-state, we can both reject or accept the null hypothesis. Provided that these cases are also included in acceptance of the null hypothesis, the probabilities of mRNA or protein being gamma distributed increase to 85, 83, 84 and 76%, respectively, for 4 cases in Figure 4.

3.3 Bayesian analysis of gamma distribution parameters

In order to visualize the conditional bivariate distribution of gamma distribution scale and shape parameters, we set n equal to 25, and use the BMH sampler to draw 10,000 samples from the joint PDF of shape and scale according to Eq. (14) assuming $a_k = u_k = 0.25$ and $b_k = v_k = 0.025$. The initial values for the BMH sampler are $\alpha_k = 10^{-2}$ and $\beta_k = 10^{-4}$. The generated histograms are shown in Figure 5 and Figure 6 assuming gamma distributed mRNA and protein counts, respectively. We observe that the corresponding distributions are positively skewed at smaller time instances, i.e., for times 250, 650, 1050, and 1450 s in Figure 5. Moreover, towards the end of the cell half lifetime, the histograms show lower tails, i.e., at time 2650 and 3050 s in Figure 5 which is due to mRNA degradation at the end of the cell half lifetime. More importantly, despite the assumption used in (14) that the priors of shape and scale parameters are independent, the concentration of samples along the diagonal in Figure 5 indicates that the posteriors of the shape and scale parameters are strongly correlated. Moreover, the bivariate distributions of scale and shape can be bimodal as shown in Figure 5 for time 1450 s, and also in Figure 6 for all times considered. Unlike in Figure 5, the correlations between shape and scale in Figure 6 appear to vary from highly correlated to much less correlated at different time instances.

The convergence of the BMH sampler can be checked, for instance, by observing the sample means. The BMH sampler appears to be very sensitive to the initial values. The sample means are computed using a sliding window of 1000 and 4500 samples of shape and scale, respectively, and they are shown in Figure 7. We found that the BMH sampler needs to produce at least 1000 samples for mRNA and 3000 samples for protein in order to generate a stationary distribution. The correlations between shape and scale parameters can be also deduced from Figure 7.

The default values of key reaction rates of transcription, translation, mRNA degradation and protein degradation are summarized in Table 1 [40]. These values yields theoretical shape and scale parameters of protein production [24, 25]:

$$\alpha = K_{tr}/\gamma_p = 600 \text{ and } \beta = K_{tn}/\gamma_m = 4.$$

Table 1. Default values of the key reaction rates.

Reaction	**Rate** (s^{-1})
Transcription	k_{tr} = 1.26e-01
Translation	k_{tn} = 4.44e-02
mRNA degradation	γ_m = 1.11e-02
Protein degradation	γ_p = 2.1e-04

The theoretical values of shape and scale parameters can be estimated from Figure 7 as the long-term mean values. In particular, the estimated shape value at steady state is 554 and 667 assuming gamma distributed mRNA and protein, respectively, whereas the estimated scale values are 3.9 for mRNA nad 4.9 for protein.

The empirical auto-correlations of scale and shape parameters are compared in Figure 8. For protein synthesis, both scale and shape parameters have correlated values over the span of as many as 185 observation samples whereas this values is reduced to about 100 samples in case of mRNA synthesis. Hence, the observed counts of mRNA and protein, respectively, are highly correlated over 10's of samples.

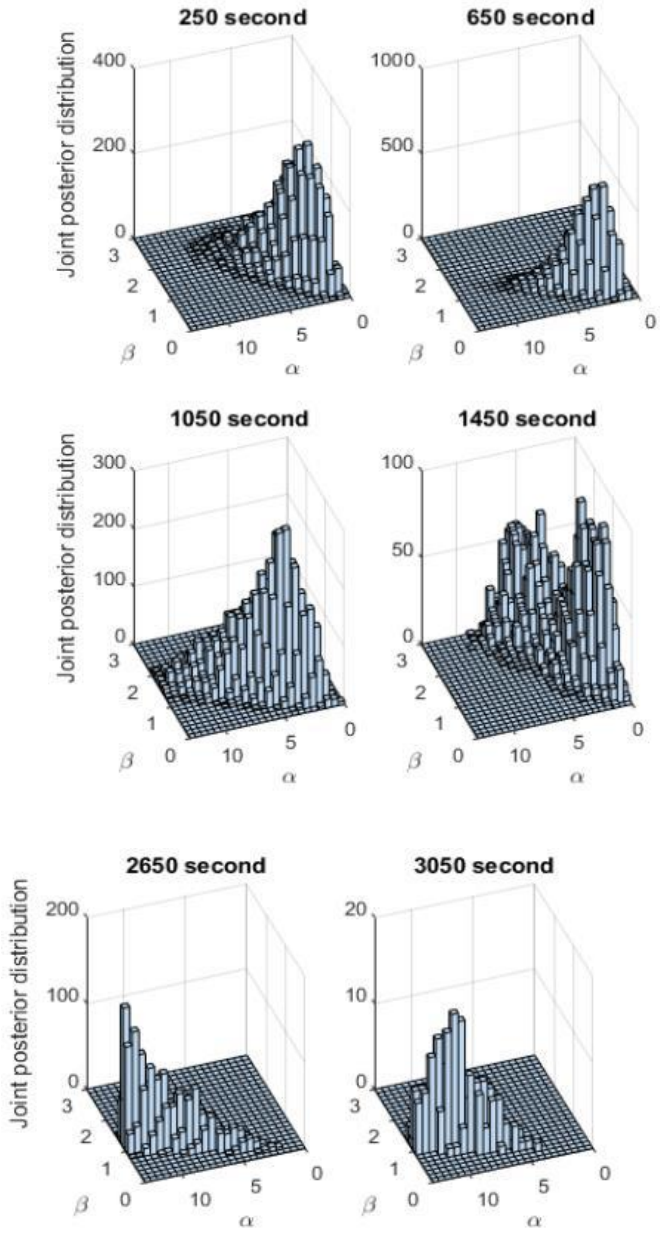

Figure 5. The histograms of 10,000 random scale and shape samples generated using the BMH sampler of the bivariate posterior distribution for the case of mRNA synthesized by the lac circuit in E. coli. The time instances considered are the same as those in Figures 2 and 3.

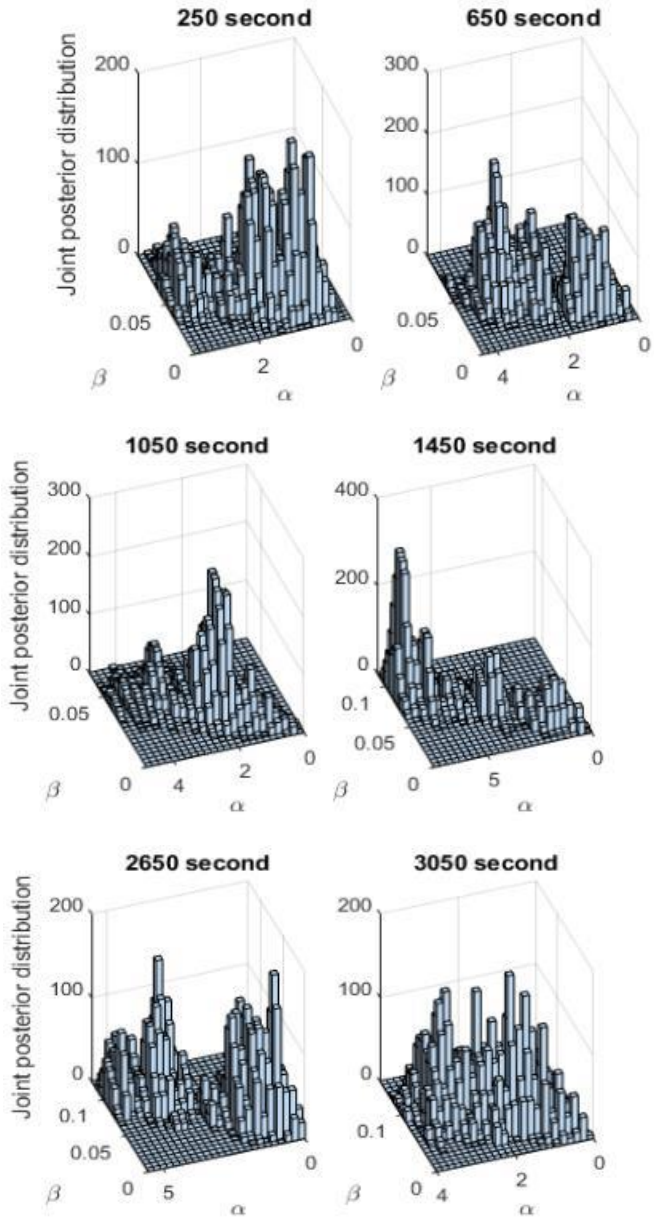

Figure 6. The histograms of 10,000 random scale and shape samples generated using the BMH sampler of the bivariate posterior distribution for the case of protein synthesized by the lac circuit in E. coli. The time instances considered are the same as those in Figures 2 and 3.

4. Discussion

Our statistical analysis of simulated time dependent mRNA and protein counts produced by the lac circuit revealed that gamma distribution is a good fit for both mRNA and protein counts in over 70% of times from initial state to steady state. The scale and shape parameters of gamma distributed mRNA and protein counts can be also considered to be gamma distributed. The Bayes theorem was used to find the posterior bivariate distributions of scale and shape parameters conditioned on the observed counts. In order to visualize joint PDF of scale and shape parameters, the BMH sampler was implemented to obtain the corresponding bivariate histograms.

The product of shape and scale parameters is equal to the mean of gamma distribution, and very similar relationship can be obtained for the mode of gamma distribution. In general, gamma distribution is unimodal while the bivariate distributions of scale and shape parameters can be bimodal. Since the bivariate distributions of scale and shape are not circularly symmetric, these parameters controlling the properties of gamma distribution are strongly correlated.

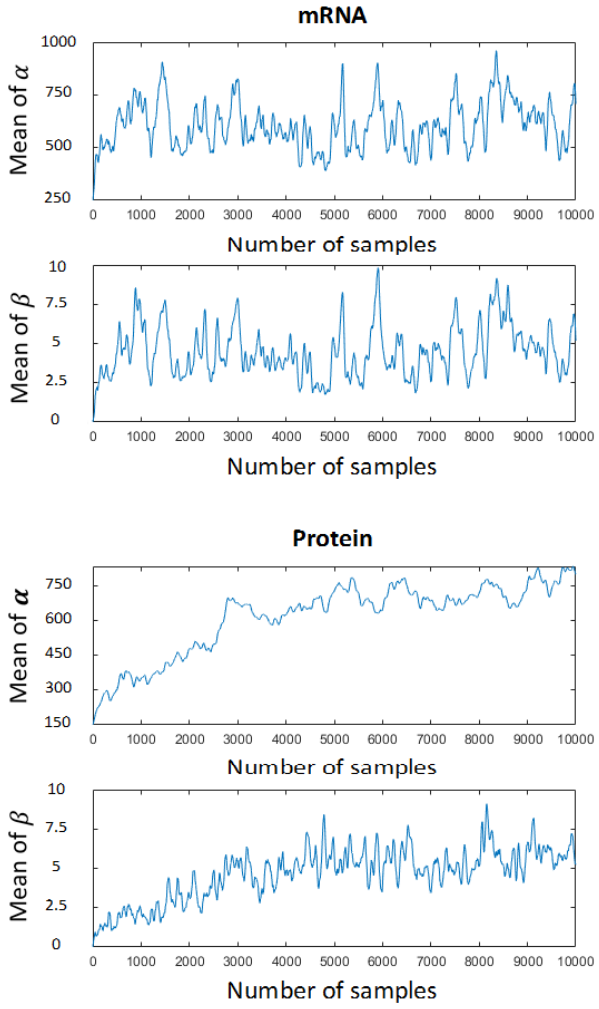

Figure 7. The convergence of the moving average sample means of the scale and shape parameters for the mRNA and protein production using a sliding window of 1000 and 4500 samples, respectively. However, the BMH sampler requires to through away the first 1500 and 3500 samples, respectively, in order to start generating a stationary distribution.

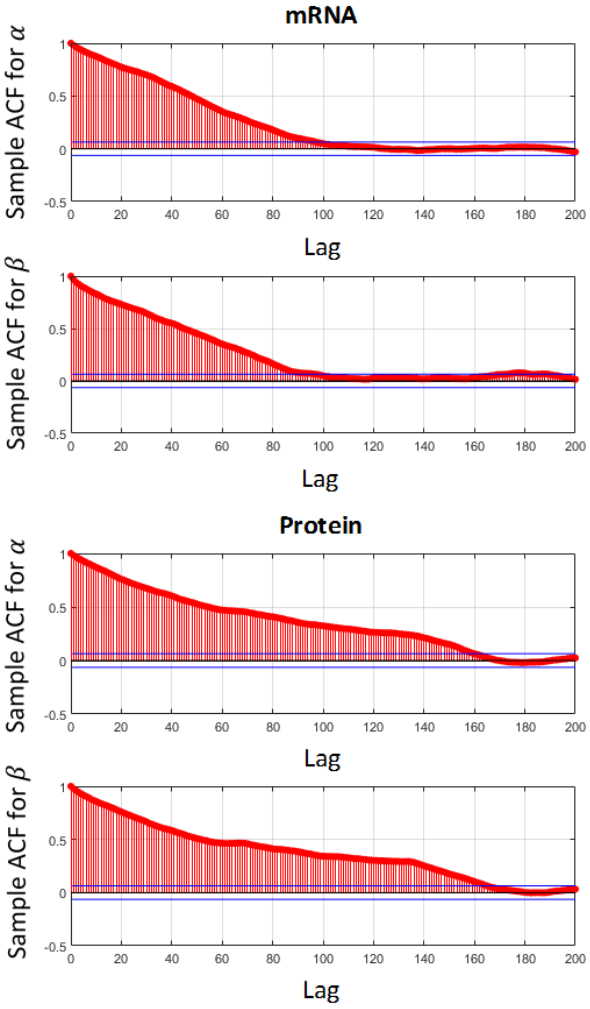

Figure 8. The estimated autocorrelation of shape and scale parameters representing mRNA and protein production, respectively.

In general, gamma distribution is commonly used to model randomness in living systems such as pausing times and other stochastic phenomena in biological circuits. Gamma distribution can be also used as conjugate prior, so that both the prior and the posterior distributions of gamma distribution parameters are gamma distributed.

Biological significance of shape and scale parameters were considered in [24] and [28]. These parameters for gamma distributed mRNA depend linearly on the ratio of transcription and degradation rates [26][32]. This explains the observed auto-correlation values which were obtained using our histogram analysis. We also observed that mRNA transcription tends to be more bursty than the subsequent protein translation with the latter appearing to be more evenly spread over time. Since protein synthesis is the most energy-consuming process in proliferating living cells, understanding what controls protein abundances is one of the key questions in molecular biology and biotechnology [67]. A number of previous research works have suggested that mRNA distribution can be the target for controlling distributions of protein synthesis [68-70]. A Bayesian approach could be used to infer mRNA distribution from empirically obtained protein distributions using observed molecule counts.

It may be useful to also consider other distributions whether they may provide better fit than the gamma distribution over larger periods of time. The main reason we considered gamma distribution is that it has been assumed in many papers previously. One class of suitable distributions to consider are truncated distributions. For instance, we could assume truncated Gaussian distribution with additional one or two parameters defining the truncation interval.

Furthermore, due to a linear relationship between the shape parameter and transcription rate, and the scale parameter and translation rate, the estimated scale and shape parameters can be used to infer the corresponding gene expression rates. Prior reaching the steady-state, scale and shape parameters appear to be random processes. One can define time–varying transcription and translation rates assuming constant mRNA and protein degradation rates as suggested in Eq. (15). The empirical autocorrelation values of scale and shape processes in the transition phase prior to steady-state points to highly correlated values of mRNA and protein production across 10's of observed samples.

Unlike mRNA synthesis, protein production in the lac circuit appears to be inversely affected by values of scale and shape parameters. The protein distributions were found to be less skewed and more Gaussian-like (Figure 6) whereas mRNA distributions remained heavily skewed (Figure 5). This could be explained by the fact that sufficient protein synthesis requires abundance of the corresponding mRNA molecules to ensure enough translation events. Thus, mRNA and protein copy counts are highly correlated. However, if protein degrades more slowly, so that proteins from different translation bursts can co-exist in the cell, their distribution is less skewed, and mRNA and protein counts are less correlated.

Our analysis of observed mRNA and protein counts does not require knowledge of reaction rates nor the structure of genetic circuit considered. The analysis is numerically efficient and is well suited to process large amount of data form in silico experiments. In vitro and in vivo experiments, on the other hand, are likely to produce much less data while these data can be also noisy. The histogram estimators of molecule count distributions may suffer from large estimation errors when fitting a distribution to a few noisy data. In such cases, it is possible to compute likelihood or a posteriori probabilities of observed molecule counts for several candidate distributions, and decide which one is the best fit to the observed data.

In the transition from initial to steady state, the observed molecule counts represent a non-stationary random process. Such processes are often doubly-stochastic meaning that their distribution parameters are themselves random. We have observed this phenomenon assuming gamma distributed mRNA and protein counts in the lac circuit model of E. coli. Parameters of these gamma distributions appear to be themselves gamma distributed. However, a better strategy to model the molecule count distributions may be to approximate the distribution parameters by time-dependent deterministic functions. Thus, assuming a given distribution with time varying parameters to model time evolution of molecule counts during the transition phase before reaching steady state may provide computationally efficient models for describing stochastic dynamics of genetic circuits. More importantly, assuming models obtained empirically from measured data can yield mathematical expressions more amenable to further mathematical and statistical analysis than trying to analytically solve CME.

The measured autocorrelations of gamma distribution parameters in the lac circuit show that mRNA and protein counts are both highly correlated in time. Understanding the correlations in mRNA and protein synthesis is, in general, useful in designing synthetic biological circuits with more predictable properties, inferring model parameters, suppressing observation noise, optimizing production, also in experiment design [71]. For instance, maximizing the recombinant protein synthesis is of great interest for industrial production of pharmaceuticals and biofuels [72]. Since overproduction of the recombinant protein imposes a significant stress on the host organism [73, 74], knowing a time evolution of protein distribution can be exploited to balance the production while consuming the host's resources. This may even lead to choice of a different host organism or to changes of growth conditions.

5. Conclusion

We have presented a statistical methodology to obtain distribution of mRNA and protein in transition from initial state as well as in steady state. The method does not require any assumptions. The method was illustrated to investigate whether mRNA and protein counts in the model of lac circuit in E. coli can be considered as gamma distributed. Using optimized histogram and two statistical tests, we found that both mRNA and protein counts can be considered to be gamma distributed in at least 70% of times from the initial state until steady state. In addition, shape and scale parameters of gamma distribution are themselves gamma distributed. The Bayes theorem and the BMH sampler were used to further study the gamma distribution parameters.

We observed that shape and scale parameters are statistically correlated, and their joint PDF is often bimodal. These parameters have been previously linked to ratios of key reaction rates in the genetic circuit. Here, we have considered these definitions in transition phase prior reaching steady state by assuming equivalent time-varying transcription and translation rates.

Acknowledgements

The work of Komlan Atitey was supported by the Zienkiewicz Scholarship award at Swansea University.

References

[1] Grunberg R, and Serrano L 2010 Strategies for protein synthetic biology *Nucleic Acids Research* **38**, 8 2663-2675

[2] Lyons S, Xu W, Medford J, and Prasad A 2014 Loads bias genetic and signaling switches in synthetic and natural systems *Plos computational biology* **10** 3 1-16

[3] Ochieng P J Kusuma W A, and and Haryanto T 2017 Detection of protein complex from protein-protein interaction network using Markov clustering *Journal of Physics in International symposium on bioinformatics, chemometrics and metabolomics*

[4] Bray D 1995 Protein molecules as computational elements in living cells *Nature* **376** 307

[5] Szathmary E, Jordan F, and Pal C 2001 Can gene explain biological complexity? *Science* **292** 1315

[6] Bray D 2002 Bacterial chemotaxis and the question of gain *PNAS* **99** 1 7-9

[7] Guet C C, Elowitz B, Hsing W, and Leibler S 2002 Combinatorial synthesis of genetic networks *Science* **296** 1466

[8] Costa L, Rodrigues F A, and Cristino A S 2008 Complex networks: the key to systems biology *Genetics and molecular biology* **31** 3 591-601

[9] Hintze A, and Adami C 2008 Evolution of complex modular biological *Plos computational biology* **4** 2 0001-0012008

[10] Mol M, Bejugam P R, and Singh S 2014 Synthetic biology at the interface of functional genomics *Briefings in functional genomics* **14** 3 180-188

[11] Walters R, and Parker R 2014 Is there quality control of localized mRNAs? *Cell Biology* **204** 6 863-868

[12] Elgart V, Jia T, Fenley A T, and Kulkarni R 2011 Connecting protein and mRNA burst distributions for stochastic models of gene expression *IOP Phys. Bio* **8** 046001

[13] Kwon Y, and Jewett M C 2014 High-throughput preparation methods of crude extract for robust cell-free protein synthesis *Nature scientific reports* **5**

[14] OECD 2014 Emerging policy issues in synthetic biology *OECD Publishing*

[15] Currin A, Swainston N, Day P J, and Kell D B 2014 Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently *Royal society of chemistry* **44** 1172-1239

[16] Mondal S, Kallianpur M V, Udgaonkar J B, and Krishnamoorthy G 2015 Molecular crowding causes narrowing of population heterogeneity and restricts internal dynamics in a protein *IOP: Methods and applications in fluorescence,* **5** 2016 2015

[17] Paulson J 2005 Models of stochastic gene expression *Physics of Life Reviews* **2** 157-175

[18] Shahrezaei V, and Swain P S 2008 The stochastic nature of biochemical networks *Current opinion in biotechnology* **19** 369-374

[19] Cheong R, Pliwal S, and Levchenko A 2010 Models at the single cell level *Systems biology and medicine* **2** 34-48

[20] Khanin R, and Higham D J 2008 Chemical master equation and langevin regimes for a gene transcription model *Theoretical computer science* **408** 31-40

[21] Smadbeck P, and Kaznessis A 2015 Chemical master equation closure for computer-aided synthetic biology *Methods molecular biology* **1244** 179-191

[22] Kazeev V, Khammash M, Nip M, and Schwab C 2014 Direct solution of the chemical master equation using quantized tensor trains *Plos computational biology* **10** 3

[23] Dattani J, and Barahona M 2018 Stochastic models of gene transcription with upstream drives: exact solution and sample path characterization *The royal society*

[24] Friedman N, Cai L, and Xie X S 2006 Linking stochastic dynamics to population distribution: an analytical framework of gene expression *Physical review letters* **97** 168302

[25] Co A D, Lagomarsino M C, Caselle M, and Osella M 2017 Stochastic timing in gene expression for simple regulatory strategies *Nucleic Acids Research* **45** 3 1069-1078

[26] Bernstein J A, Khodursky A B, Lin P, Lin-Chao S, and Cohen S N 2002 Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays *PNAS* **99** 15 9697-9702

[27] Bokes P, King J R, and Wood A T 2013 Transcriptional bursting diversifies the behaviour of a toggle switch: Hybrid simulation of stochastic gene expression *Bull Math Bio* **2013** 75 351-371

[28] Shahrezaei V, and Swain P S 2008 Analytical distributions for stochastic expression *PNAS* **105** 45 17256-17261

[29] Iyer-Biswas S, Hayot F, and Jayaprakash C 2009 Stochasticity of gene products from transcriptional pulsing *Physical review* **79** 031911

[30] McQuarrie D A, and Gillespie D T 1967 Stochastic approach to chemical kinetics *Appl. Prob.* **4** 413-478

[31] Roberts E, Stone J E and Luthey-Schulten Z 2013 Lattice microbes: high-performance stochastic simulation method for the reaction-diffusion master equation *Comput Chem* **34** 3 245-255

[32] Atitey K, Loskot P, and Rees P 2018 Determining the transcription rates yielding steady-state production of mRNA in the lac genetic switch of Echerichia coli *Computational biology*

[33] Raj A, and Oudenaarden A 2008 Stochastic gene expression and its consequences *Cell* **135** 2 216-226

[34] Zeng X, Wang D, and Wu J 2015 Evaluating the three methods of goodness of fit test for frequency analysis *Journal of risk analysis and crisis response* **5** 3 178-187

[35] Jantschi L, and Bolboaca S D 2018 Computation of probability associated with Anderson-Darling statistic *Mathematics* **6** 88 1-17

[36] Tsionas E G 2003 Bayesian inference for multivariate gamma distributions *Statistics and computing* **14**223-233

[37] Atitey K and Cang Y 2015 A novel prediction algorithm in Gaussian-Mixture probability hypothesis density filter for target tracking *in International conference on Image and Graphics* China 373-393

[38] Steyvers M 2011 Computational statistics with matlab, http://psiexp.ss.uci.edu/research/teachingP205C/205C.pdf

[39] Ravenzwaaij D, Cassey P, and Brown S D 2016 A simple introduction to Markov Chain Monte Carlo *Journal of Mathematical Psychology* **25** 1143-154

[40] Roberts E, Magis A, Ortiz J O, Baumeister W, and Luthey-Schulten Z 2011 Noise contributions in an inducible genetic switch: A whole-cell simulation study *Plos computational Biology* **7** 3 e1002010

[41] Jia C, Qian H, Chen M, and Zhang M Q 2018 Relaxation rates of gene expression kinetics reveal the feedback signs of autoregulatory gene networks *The journal of Chemical physics,* **148** 9

[42] Paulsson J, and Ehrenberg M 2000 Random signal fluctuations can reduce random fluctuations in regulated components *Physical review letters* **84** 23 5447-5449
[43] Taniguchi Y, Choi P J, Li G, Chen H, Babu M, Hearn J, Emili A, and Xie X S 2010 Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells *Science* **329** 5991 533-538
[44] Heerden J H, Kempe H, Doerr A, Maarleveld T, Nordholt N, and Bruggeman F J Statistics and simulation of growth of single bacterial cells: illustrations with B. subtilis and E. coli *Nature, scientific reports* **7**, 16094
[45] Cai L, Friedman N, and Xie X S 2006 Stochastic protein expression in individual cells at the single molecule level *Nature* **440**
[46] Ohno M, Karagiannis P, and Taniguchi Y 2014 Protein expression analyses at the single cell level *Molecules* **19** 2014 13932-13947
[47] Kaoern M, Elston T C, Blake W J, and Collins J J 2005 Stochasticity in gene expression: from theories to phenotypes *Nature* **6** 461-464
[48] Orphanides G, and Reinberg D 2002 A unified theory of gene expression *Cell* **108** 439-451
[49] Andrews S S, Dinh T, and Arkin A P 2009 Stochastic models of biological processes *Encyclopedia of Complexity and System Science* **9** 8730-8749
[50] Nath K, and Koch A L 1970 Protein degradation in Echerichia coli *Journal of Biological chemistry* **245** 11 2889-2900
[51] Chatla S B, Chen C, and Shmueli G 2017 Selected topics in statistical computing *Semantic computing* **1** 45-62
[52] Scott D W 1979 On optimal and data-based histograms *Biometrika* **66** 3 605-10
[53] Shimazaki H, and Shinomoto S 2007 A method for selecting the bin size of a time histogram *Neural computation,* **19** 1503-1527
[54] Shimazaki H, and Shinomoto S 2007 A recipe for optimizing a time-histogram *Advances in Neural Information Processing Systems* **19** 2007 1289-1296
[55] Ramberg J S, Dudewicz E J, Tadikamalla P R, and Mykytka E F 1979 A probability distribution and its uses in fitting data *Technometrics* **21** 2 201-214
[56] Jantschi L and Bolboaca S D 2009 Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Cramer-von-Misses and Jarque-Bera statistics *Horticulture* **66** 2 691-697
[57] Evans D L, Drew J H, and Leemis L M 2008 The Distribution of the Kolmogorov–Smirnov, Cramer–von Mises, and Anderson–Darling test statistics for exponential populations with estimated parameters," *Communications in statistics, Simulation and computation* **37** 1396-1421
[58] Razali N M 2011 Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests *Journal of statistical modeling and analysis* **2** 1 21-33
[59] Karadag O, and Aktas S 2016 Goodness of fit tests for generalized gamma distribution *in AIP Conference proceedings* **1738** 1
[60] Franke T M, Ho T, and Christie C A The Chi-square test: often used and more often misinterpreted *American journal of evaluation* **33** 3 448-458
[61] Bolboaca S D, Jantschi L, Sestras A F, Sestras R E, and Pamfil D C 2011 Pearson-Fisher Chi-Square Statistic Revisited *Information* **2** 528-545
[62] Kulinskaya E 2008 On Two-Sided P-Values for Non-Symmetric Distributions *Mathematics and statistics theory arXiv: 0810.2124*
[63] Massey A, and Miller S J 2006 Tests of hypotheses using statistics *Mathematics Department, Brown University, Providence* **RI** 2912
[64] Box G E, and Jenkins G M 1976 Time series analysis Mathematics Holden-Day
[65] Guha S, Koudasy N, and Shimz K 2001 Data-streams and histograms *in In Proc. of the 2001 Annual ACM Symp. on Theory of Computing* 471-475
[66] Abd-Elfattah A M, Hala A F, Omima A M 2010 Goodness of fit tests generalized Frechet distribution *Australian lournal of basic an applies Sciences* **4** 2 286-301
[67] Lahtvee P, Sanchez B J, Kasvandik A S, Elsemman I E, and Nielsen F G 2017 Absolute Quantification of Protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast *Cell systems* **4** 495-507
[68] Branduardi P 2016 Synthetic biology for cellular remodelling to elicit industrially relevant microbial phenotypes *Synthetic biology for cellular remodelling* **2**
[69] Decker C J, and Parker R 2012 P-bodies and stress granules: possible roles in the control of translation and mRNA degradation *Cold Spring Harb Perspect Biol* **4** 9
[70] Abil Z, Gumy L F, Zhao H, and Hoogenraad C 2017 Inducible control of mRNA transport using reprogrammable RNA-binding proteins *Synthetic biology* **6** 950-956
[71] Chizzolini F, Forlin M, Martín N Y, Berloffa G, Cecchi D, and Mansy S S 2017 Cell-Free translation is more variable than transcription *ACS Synthetic biology* **6** 638-647
[72] Dragosits M, Nicklas D, and Tagkopoulos I 2012 A synthetic biology approach to self-regulatory recombinant protein production in Escherichia coli *Journal of biological engineering* **6** 2
[73] Ramström O 2016 The Nobel Prize in Chemistry 2016; Award Ceremony Speech *Nobel Prizes and Laureates 2016*
[74] Wurm F M 2004 Production of recombinant protein therapeutics in cultivated mammalian cells *Nature biotechnology* **22** 11 1393-1398