Communications in Information and Systems
Volume 18, Number 3, 125–148, 2018

Dynamics of gene expression based on epigenetic modifications

Xiaopei Jiao and Jinzhi Lei

Gene expression is a basic process in life activities. Precise description of gene expression is essential for understanding many biological systems quantitatively. Here, we propose an analytical method to accurately calculate the mRNA distribution in equilibrium and non-equilibrium state based on a three-stage model. First, we consider a three-stage model of gene expression and obtain the mRNA distribution in equilibrium state under the absence of epigenetic modifications. Next, applying the characteristic line method of two-element PDE, we obtain accurate distribution function of non-equilibrium state which describes the evolution dynamics of the gene expression process. Moreover, based on the three-stage model, we construct a mathematical framework to illustrate the ergodic principle by which the time average is equivalent to the space average or ensemble average in a time-continuous dynamical system. We further consider the influence of DNA methylation in the transcription process. By considering methylation allocation during cell division and the influence for transition rate, we obtain analytic expression and make Gillespie random simulation for mRNA number in a cell population. The results reveal five types of diversified and novel mRNA distributions, which are highly consistent with single-cell sequence data. These results provide useful insights for our understanding of the gene expression process.

1. Introduction

Gene expression is a significant biochemical process in both prokaryotes and eukaryotes cells. With the development of advanced single-cell sequencing technology, cell-to-cell variance of gene expression is obvious among a cell

Key words and phrases: gene expression, mRNA distribution, three-stage model, ergodic principle, epigenetic modification.

population [1, 14]. People have been gradually aware of that the importance in gene expression heterogeneity among cells [16, 29]. Nevertheless, it remains difficult to quantitate the random process of gene expression.

Random fluctuations in gene expression exist universally in all kinds of organisms[19–21, 23, 24, 33]. Cells response to fluctuating environments by adjusting gene expression[5, 10]. Recently, such cell-to-cell variance can be observed by single cell RNA-sequencing techniques[4]. Cell-to-cell dynamical randomness and variability presumably result from burst-like stochastic transcription[11, 12]. There are several important factors to influence gene expression such as intrinsic factor[32], environmental factor, and epigenetic variability[27, 31]. Intrinsic factor results from discrete biochemical reactions in cells. Occurrence of a biochemical reaction is a random process because chemical molecules collide with each other randomly in some probability. Environmental factor means that gene expression can be influenced by tissue and environment around the cells through the surrounding signals from the niche [30]. The third factor is the epigenetic variability, which refers to epigenetic control of gene expression[2]. The effects of environmental factors can be studied experimentally, however, experiments can hardly be applied to the study of intrinsic fluctuation and epigenetic modifications.

Epigenetic regulation is a type of significant regulation that can induce variance in gene expression[2, 8, 25]. There are two main types of epigenetic modification, DNA methylation[9, 18] and histone modification[3], which are important for the regulation of gene expression. DNA methylation is a process by which methyl groups are added to the palindromic CpG (CG/GC) dinucleotides [13, 26]. DNA methylation level can inherit from mother cells to daughter cells during cell division. Gene expression in a cell can be described by a three-stage model[21], namely the central dogma[17], in which promoter of a gene can be switched between ON and OFF state. DNA methylation in the promoter region of a gene can modulate the transition rate between ON and OFF of the promoter through the remodeling of chromatin structure. Hence, DNA methylation can regulate the variance of gene expression at single cell level. However quantitative description of the heterogeneity due to DNA methylation remains unclear.

In this paper, we focus on the effects of intrinsic fluctuation and epigenetic modifications in a three-stage model of gene expression [21]. Major contribution of variance in protein level mainly comes from the transcription process[17], thus we can omit the translation process, and focus our study to transcription. This study was intended to provide a comprehensive and analytic model to show the distribution of mRNA in a cell population. In the rest of this paper, first we obtained analytic distribution of mRNA number

following the method proposed by Shahrezaei et al.[28]. Next, we extended the result of equilibrium state distribution and obtain accurate distribution in the non-equilibrium state. We further proved the ergodic theorem in the three-stage model. Finally, we considered the effect of DNA methylation in gene expression, and obtain five types of mRNA distributions, which are consistent to clinical data.

2. Results

2.1. Equilibrium state distribution of mRNA

The three-stage model considered here is illustrated at Fig. 1A. In this model, the promoter can transit between ON and OFF state with rate k_0 and k_1 respectively. Transcription starts with a rate v_0 when the promoter is at the ON state, the produced mRNA degrade with a rate d_0. This simple model of gene expression has been extensively studied in many works [6, 10, 15, 22, 28, 31, 34]. Here, we referred the method of characteristic line proposed in [28] to obtain the equilibrium state distribution. In [28], an approximation of the three-stage model of gene expression was obtained based on assumptions in the transcription and translation time scales. Here, we was intended to obtain the exact analytical distribution of mRNA numbers under general conditions.

Let P_n^0 and P_n^1 the probability of occurring n mRNAs when the promoter state is OFF and ON, respectively, the chemical master equation is given by

$$\begin{aligned} \frac{\partial P_n^0}{\partial t} &= k_1 P_n^1 - k_0 P_n^0 + d_0[(n+1)P_{n+1}^0 - nP_n^0], \\ \frac{\partial P_n^1}{\partial t} &= -k_1 P_n^1 + k_0 P_n^0 + v_0(P_{n-1}^1 - P_n^1) + d_0[(n+1)P_{n+1}^1 - nP_n^1]. \end{aligned} \tag{1}$$

We denote

$$\tau = d_0 t, \quad \kappa_0 = k_0/d_0, \quad \kappa_1 = k_1/d_0, \quad a = v_0/d_0, \tag{2}$$

and define the generating function

$$f^0(z,\tau) = \sum_{n=0}^{\infty} z^n P_n^0, \quad f^1(z,\tau) = \sum_{n=0}^{\infty} z^n P_n^1. \tag{3}$$

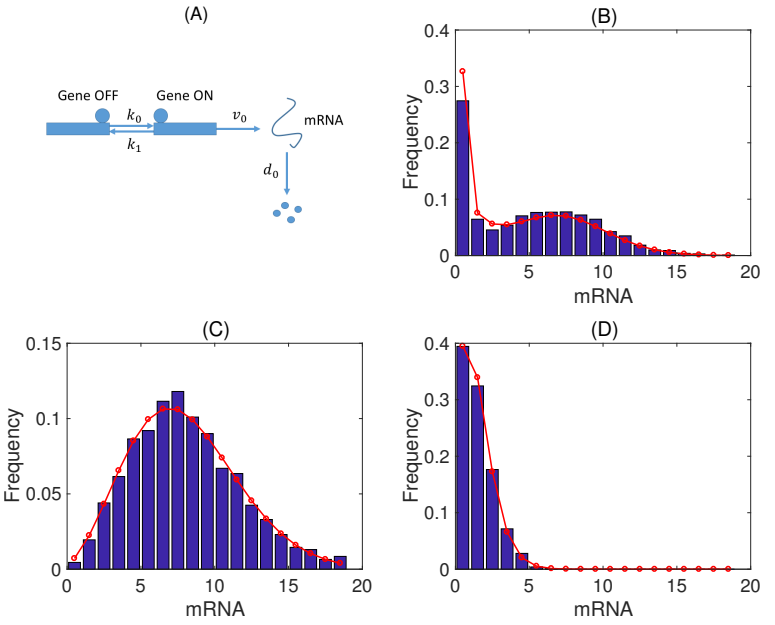

Figure 1. **Three-stage model and patterns of mRNA distribution.** (A) Illustration of the three-stage model. (B) Type I distribution, bimode distribution with two probability peaks. Here $k_0 = 0.2, k_1 = 0.2, v_0 = 8, d_0 = 1$. (C) Type II distribution, mono-mode distribution with a single peak at positive mRNA number. Here $k_0 = 4, k_1 = 4, v_0 = 15, d_0 = 1$. (D) Type III distribution, mono-mode distribution with a single peak at zero mRNA number. Here $k_0 = 3, k_1 = 3, v_0 = 2, d_0 = 1$. In the figures, blue bars represent simulation results by Gillespie algorithm, red lines are obtained by theoretical calculation.

Here $f^0(z,\tau)$ and $f^1(z,\tau)$ correspond to the generating functions of OFF and ON state of promoter, respectively. Let

$$F(z,\tau) = f^0(z,\tau) + f^1(z,\tau) = \sum_{n=0}^{\infty} z^n P_n, \quad P_n = P_n^0 + P_n^1.$$

the total generating function, then the distribution P_n can be obtained from the series expansion of the generating function $F(z,\tau)$.

From the master equation Eq. (1), we obtain the coupled partial differential equations

$$\begin{aligned}\frac{\partial f^0}{\partial \tau} &= \kappa_1 f^1 - \kappa_0 f^0 + \left(\frac{\partial f^0}{\partial z} - z\frac{\partial f^0}{\partial z}\right),\\ \frac{\partial f^1}{\partial \tau} &= -\kappa_1 f^1 + \kappa_0 f^0 + a(zf^1 - f^1) + \left(\frac{\partial f^1}{\partial z} - z\frac{\partial f^1}{\partial z}\right).\end{aligned} \tag{4}$$

At the stationary state, we set the derivatives with τ as 0, and denote $z - 1 = v$, and obtain the equation

$$\begin{aligned}v\frac{\partial f^0}{\partial v} &= \kappa_1 f^1 - \kappa_0 f^0,\\ v\frac{\partial f^1}{\partial v} &= -\kappa_1 f^1 + \kappa_0 f^0 + avf^1.\end{aligned} \tag{5}$$

These two equations can be solved by the method of series expansion, which give the total generating function (Appendix A)

$$F(v) = \sum_{n=0}^{\infty} \frac{(\kappa_0)_n a^n}{(\kappa_0 + \kappa_1)_n n!} v^n. \tag{6}$$

Here $(q)_n$ denotes the Pochhammer symbol

$$(q)_n = \begin{cases} 1, & n = 0,\\ q(q+1)\cdots(q+n-1), & n \neq 0.\end{cases} \tag{7}$$

The distribution P_k of mRNA number is given by the Taylor expansion of $F(z)$ at $z = 0$, i.e.,

$$P_k = \frac{1}{k!}\frac{\partial^k F}{\partial z^k}\Big|_{z=0}.$$

A careful calculation gives

$$\begin{aligned}P_k = {} & \frac{\Gamma(\kappa_0 + k)\Gamma(\kappa_0 + \kappa_1)}{\Gamma(\kappa_0)\Gamma(\kappa_0 + \kappa_1 + k)\Gamma(k+1)} a^k\\ & \times \sum_{n=0}^{\infty} \frac{\Gamma(\kappa_0 + k + n)\Gamma(\kappa_0 + \kappa_1 + n)}{\Gamma(\kappa_0 + n)\Gamma(\kappa_0 + \kappa_1 + k + n)n!}(-a)^n.\end{aligned} \tag{8}$$

Here $\Gamma(z)$ is the Gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$. The Eq. (8) gives the probability of occuring k mRNA at the stationary state. Fig. 1 shows the

distributions obtained from numerical simulation with Gillespie algorithm [7] and the above analytic expression, which show good agreement.

From Eq. (8), the probability P_k is dependent on the three parameters κ_0, κ_1, and a. We varied the parameters to investigate the possible distribution patterns. There are three type distributions depending on the parameter values (Fig. 1B-D): type I with bimode distribution with two probability peaks, type II of mono-mode distribution with peak probability at positive mRNA number, and type III of mono-mode distribution with peak probability at zero mRNA number. Hence, we re-obtained the results in [28]. We further investigated the parameters corresponding to the three type distributions, and find that the key parameters are $u = \kappa_0/a$ and $v = \kappa_1/a$ (Fig. 2). We have the type I distribution when both u and v are small. When u is small and v is adequately large, the distribution is type III, and we have the type II distribution when u is large.

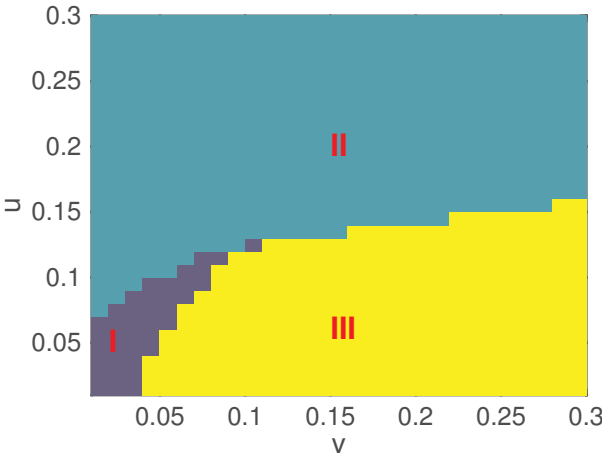

Figure 2. **Various distributions of mRNA number.** The parameters $u = \kappa_0/a$ and $v = \kappa_1/a$ are relative transition rates between ON and OFF state of the promoter. Here, we set $a = 8$ and $d_0 = 1$ fixed.

2.2. Non-equilibrium state distribution of mRNA

Now, to obtain the evolution of distribution, we need to solve the Eq. (4). To this end, we rewrite Eq. (4) as

$$\frac{\partial \vec{f}}{\partial \tau} + v\frac{\partial \vec{f}}{\partial v} = \begin{bmatrix} -\kappa_0 & \kappa_1 \\ \kappa_0 & av - \kappa_1 \end{bmatrix} \vec{f}. \tag{9}$$

where $\vec{f} = (f^0, f^1)^{\mathrm{T}}$. Applying the method of characteristic line and introduce a variable r, we obtain four differential equations

$$\begin{aligned} &\frac{d\tau}{dr} = 1, \\ &\frac{dv}{dr} = v, \\ &\frac{d\vec{f}}{dr} = \begin{bmatrix} -\kappa_0 & \kappa_1 \\ \kappa_0 & av - \kappa_1 \end{bmatrix} \vec{f}. \end{aligned} \tag{10}$$

Assuming that there are m mRNAs initially, so that

$$F(v, 0) = (1 + v)^m,$$

we solve the above equations to obtain the total generating function (Appendix B)

$$F(v, \tau) = (1 + ve^{-\tau})^m \frac{\sum\limits_{n=0}^{\infty} \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v^n}{\sum\limits_{n=0}^{\infty} \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v^n \mathrm{e}^{-n\tau}}. \tag{11}$$

It is easy to verify that the asymptotic solution when $\tau \to \infty$ gives the exact solution Eq. 6 for the stationary distribution.

In order to obtain the probability $P_k(\tau)$, we make some approximations for generating function. First, we ignore influence of denominator in Eq. (11), since $\mathrm{e}^{-n\tau}$ is infinitesimal of higher order than $\mathrm{e}^{-\tau}$. Moreover, we make an approximation $(1 + v\mathrm{e}^{-\tau})^m \approx 1 + mv\mathrm{e}^{-\tau}$. Thus, an asymptotic form of $F(v, \tau)$ is given by

$$F(v, \tau) \approx (1 + mv\mathrm{e}^{-\tau}) \sum_{n=0}^{\infty} \frac{(\kappa_0)_n a^n}{(\kappa_0 + \kappa_1)_n n!} v^n. \tag{12}$$

Next, we calculate the probability distribution for k mRNAs at τ time with initial value m as $P_k(\tau|m) = \frac{1}{k!} F^{(k)}(v, \tau)|_{v=-1}$, which gives

$$P_k(\tau|m) = (1 - m\mathrm{e}^{-\tau}) P_k^s + m\mathrm{e}^{-\tau} P_{k-1}^s, \tag{13}$$

where

$$P_k^s = \frac{1}{k!} \sum_{n=0}^{\infty} \frac{(\kappa_0)_{n+k} a^{n+k}}{(\kappa)_{n+k} n!} (-1)^n. \tag{14}$$

Here P_k^s is probability in the stationary distribution with k mRNAs.

Eq. (13) gives the evolution of the distribution function. When $\tau \to \infty$, we have $P_k(\tau|m) \to P_k^s$, and the distribution converges to the stationary distribution exponentially. Moreover, Eq. (13) implies that $P_k(\tau|m)$ is mainly related to the two neighbor items P_k^s and P_{k-1}^s at the stationary state. Fig. 3 shows the time evolution of the three type distributions.

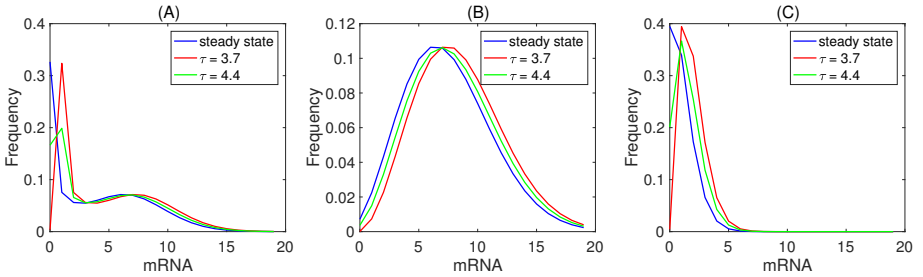

Figure 3. **Asymptotic behaviors of the three type mRNA distributions.** Here, we set the initial number of mRNA as 40 and examine the evolution of the probability distribution function. (A) Parameters: $k_0 = 0.2, k_1 = 0.2, v_0 = 8, d_0 = 1$. (B) Parameters: $k_0 = 4, k_1 = 4, v_0 = 15, d_0 = 1$. (C) Parameters: $k_0 = 3, k_1 = 3, v_0 = 2, d_0 = 1$. In these figures, red, green, and blue lines correspond to $\tau = 3.7, 4.4$, and the stationary state, respectively. The lines are calculated from Eq. (13).

2.3. Equivalence between time and ensemble distribution in gene expression dynamics

In statistical physics, there is a fundamental principle that time distribution and ensemble distribution are equivalent in major physical systems. Here, we show that in the three-stage model of gene expression, the ensemble distribution asymptotically approaches to the time distribution.

Fig. 4 shows the basic concepts of time and ensemble distributions. Time distribution is to trace a system and observe its time evolution. Hence, to obtain the mRNA distribution, we need to choose a long time window and collect mRNA value in this window. For ensemble distribution, we imagine that there exists an ensemble of independent systems starting with different initial mRNA number. At each time point, different system in the ensemble evolves to different mRNAs number. Hence, while we choose a certain time point and observe all system simultaneously, we are able to obtain the mRNA

distribution from this ensemble. This type of mRNA distribution is called ensemble distribution.

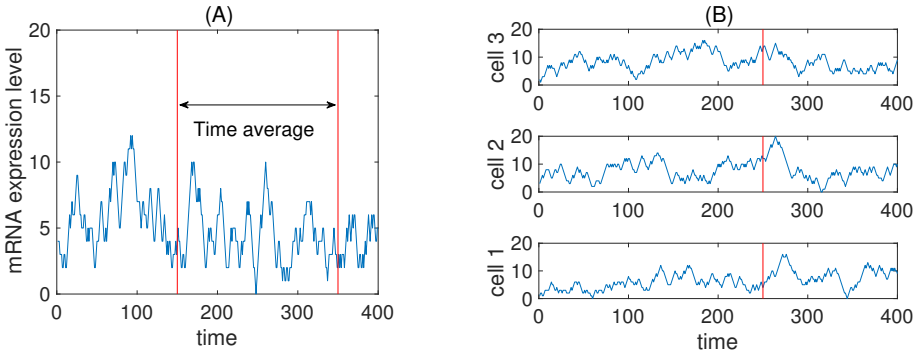

Figure 4. **Illustration of time and space distribution.** (A) Illustration of time distribution, which means that when we consider mRNA evolution in one cell, after some initial time, we pick a large time interval and obtain the frequency of mRNA in the time interval. (B) Illustration of ensemble distribution. Assume that we have many cells (the ensemble), each cell starts from different (randomly) initial values and evolves to a time shorter than the relaxation time. At this time point, we count the mRNA frequency among these cells and obtain the ensemble distribution.

The above exact solutions correspond to the ensemble statistics. Here, we asked whether these distribution functions can be used to describe the time distribution based on a single cell evolution. Fig. 5 shows an example of how time distribution converges to the ensemble distribution. We choose four time windows from a cell evolution to calculate the mRNA distributions. Fig. 5 shows that when the time window increasing, the mRNA distribution from one cell evolution converges to the theoretical stationary distribution. This example suggests that the above formulation of the ensemble distribution can be used to describe the time distribution of this system.

In ensemble distribution, the probability of occurring k mRNAs in total ensembles at time τ is not a deterministic value, but fluctuates around an average value depending on k. In another words, the distribution of mRNA itself obeys a probability distribution. Based on the three-stage model, we find that when the time increasing, average of the ensemble distribution of mRNA tends to a steady distribution which is the accurate distribution after infinitely long time. Proof of this result is given by the following theorem.

THEOREM. *Let $X(t, x_0)$ denotes the number of mRNA at time t of a cell with initial mRNA number x_0 at $t = 0$, and define a counting function*

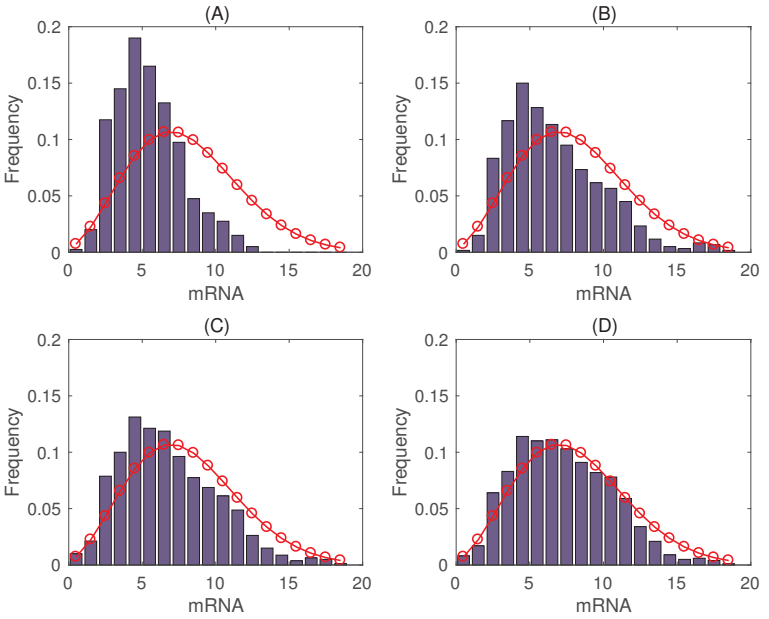

Figure 5. **Time distribution of a system under different time windows.** Four time windows: (A) $[0, 28.8]$, (B) $[0, 39.0]$, (C) $[0, 51.4]$, and (D) $[0, 60.5]$. Other parameters are taken from Fig. 1: $k_0 = k_1 = 4, d_0 = 15, v_0 = 1$. Gray bar is obtained from numerical results with the Gillespie method, and red line is theoretical stationary mRNA distribution Eq. (8).

$\chi(X, m)$ *as*

$$\chi(X,m) = \begin{cases} 1, & X = m \\ 0, & \text{other wise.} \end{cases} \tag{15}$$

Consider a single cell $X(t, x_0)$*, and an ensemble of* N *cells* $\Omega = \{X_i(t, x_i)\}_{i=1}^N$*. Then for any positive integer* m*, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \chi(X(i\Delta t, x_0), m) = \lim_{N\to\infty, \tau\to\infty} \frac{1}{N} \sum_{i=1}^{N} \chi(X_i(\tau, x_i), m), \tag{16}$$

where $\Delta t > 0$*. In the equation* (16)*, the left hand side represents the time average distribution, and the right hand side represents the ensemble average distribution.*

Proof. First, we calculate the ensemble average distribution and prove that the limit of the right hand side in Eq. (16) exists.

For the ensemble Ω of N cells with initial mRNA number $x_i (i=1, \cdots, N)$, we define the ensemble frequency ρ_e at time $\tau > 0$ as

$$\rho_e = \frac{1}{N} \sum_{i=1}^{N} \chi(X_i(\tau, x_i), m). \tag{17}$$

Thus, the probability distribution function

$$P(X_i(\tau, x_i) = m) = P_m(\tau | x_i). \tag{18}$$

Since the cells $X_i(\tau, x_i)$ in Ω are independent to each other, the mean value and variance of ρ_e can be calculated below:

$$E(\rho_e) = \frac{1}{N} \sum_{i=1}^{N} P_m(\tau | x_i). \tag{19}$$

$$D(\rho_e) = \frac{1}{N^2} \sum_{i=1}^{N} P_m(\tau | x_i)(1 - P_m(\tau | x_i)). \tag{20}$$

We note that $P_m(\tau | x_i)$ is given by Eq. (13). Now, we can analyze the limits of Eq. (19) and Eq. (20). When τ tends to infinity, we have

$$\lim_{\tau \to \infty} P_m(\tau | x_i) = P_m^s.$$

Substituting Eq. (13) into Eq. (19), we obtain

$$\begin{aligned} \lim_{\tau \to \infty} E(\rho_e) &= \frac{1}{N} \lim_{\tau \to \infty} \sum_{i=1}^{N} P_m(\tau | x_i) \\ &= \frac{1}{N} \sum_{i=1}^{N} \lim_{\tau \to \infty} (1 - x_i \mathrm{e}^{-\tau}) P_m^s + x_i \mathrm{e}^{-\tau} P_{m-1}^s = P_m^s. \end{aligned} \tag{21}$$

From Eq. 21, there exists a constant C, such that

$$|E(\rho_e) - P_m^s| \leq C \mathrm{e}^{-\tau}. \tag{22}$$

Hence, we obtain the exponential convergency when τ goes to infinity. Similarly, Eq. (20) yields

$$\lim_{\tau \to \infty} D(\rho_e) = \frac{1}{N} P_m^s (1 - P_m^s). \tag{23}$$

Hence, $N \to \infty$, the limitation in Eq. (23) tends to

$$\lim_{N\to\infty,\tau\to\infty} D(\rho_e) = 0. \tag{24}$$

Thus, we conclude that $\lim_{N\to\infty,\tau\to\infty} \rho_e = P_m^s$, i.e., the limit of the right hand side in Eq. (16) exists.

Next we calculate the time average distribution and prove that the limit of the left hand side in Eq. (16) exists.

It is easy to have the probability distribution

$$P(X(i\Delta t, x_0) = m) = P_m(i\Delta t|x_0). \tag{25}$$

When the time step Δt is large enough, we can ignore the correlation between $X(i\Delta t, x_0)$ and $X((i+1)\Delta t, x_0)$, and assume that $X(i\Delta t, x_0)$ and $X((i+1)\Delta t, x_0)$ are independent. Hence, denote the time frequency

$$\rho_t = \frac{1}{n}\sum_{i=1}^{n} \chi(X(i\Delta t, x_0), m),$$

the mean and variance of ρ_t are given by

$$\begin{aligned} E(\rho_t) &= \frac{1}{n}\sum_{i=1}^{n} P_m(i\Delta t|x_0), \\ D(\rho_t) &= \frac{1}{n^2}\sum_{i=1}^{n} P_m(i\Delta t|x_0)(1 - P_m(i\Delta t|x_0)). \end{aligned} \tag{26}$$

Again, applying Eq. (13), we obtain

$$\begin{aligned} P_m(i\Delta t|x_0) &= (1 - x_0 \mathrm{e}^{-i\Delta t})P_m^s + x_0 \mathrm{e}^{-i\Delta t} P_{m-1}^s \\ &= P_m^s + x_0(P_{m-1}^s - P_m^s)\mathrm{e}^{-i\Delta t}. \end{aligned} \tag{27}$$

Hence, substituting Eq. (27) into Eq. (26), we obtain

$$\begin{aligned} E(\rho_t) &= P_m^s + x_0(P_{m-1}^s - P_m^s)\frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{-i\Delta t} \\ &= P_m^s + x_0(P_{m-1}^s - P_m^s)\frac{\mathrm{e}^{-\Delta t}(1 - \mathrm{e}^{-n\Delta t})}{1 - \mathrm{e}^{-\Delta t}}\frac{1}{n}. \end{aligned} \tag{28}$$

Thus, there exists a constant C, such that

$$|E(\rho_t) - P_m^s| \le \frac{C}{n}. \tag{29}$$

Here, the convergence rate is in the order $O(n^{-1})$.

For the variance in Eq. (26), we apply the triangle inequality

$$|D(\rho_t)| \leq \frac{1}{n}\left(\left|\frac{\sum_{i=1}^{n} P_m(i\Delta t|x_0)}{n}\right| + \left|\frac{\sum_{i=1}^{n} P_m^2(i\Delta t|x_0)}{n}\right|\right) \leq \frac{C}{n}. \tag{30}$$

Here, we use convergence of $P_m(i\Delta t|x_0)$ when $i \to \infty$, and obtain the boundedness within the bracket in Eq. (30). From Eq. (29) and (30), we obtain

$$\lim_{n\to\infty} E(\rho_t) = P_m^s, \tag{31}$$

$$\lim_{n\to\infty} D(\rho_t) = 0. \tag{32}$$

Hence, have $\lim_{n\to\infty} \rho_t = P_m^s$.

Finally, we summarize all results and obtain

$$\lim_{n\to\infty} \rho_t = P_m^s = \lim_{N\to\infty,\tau\to\infty} \rho_e, \tag{33}$$

which gives

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi(X(i\Delta t, x_0), m) = \lim_{N\to\infty,\tau\to\infty} \frac{1}{N}\sum_{i=1}^{N} \chi(X_i(\tau, x_i), m). \tag{34}$$

This equation holds for any given positive integer m, and the Theorem is proved. □

We have proved that the equivalence between time distribution and ensemble distribution. Hence, the above formulation for the ensemble distribution can be applied to describe the time distribution of gene expression in a single cell. Fig. 6 shows the comparison between numerical simulation of an ensemble of finite cells and the theoretical results, which are consistent well to each other. Results for different time points and cell numbers in the ensemble are shown. Latter dynamics and more ensemble cells yield better consistence.

2.4. Distribution of mRNA with modifications in DNA methylation

Now, we consider how DNA methylation affects the stochasticity in gene expression. Here we consider the DNA methylation at the promoter region, and introduce a β-value β $(0 \leq \beta \leq 1)$ for the fraction of modified CpG

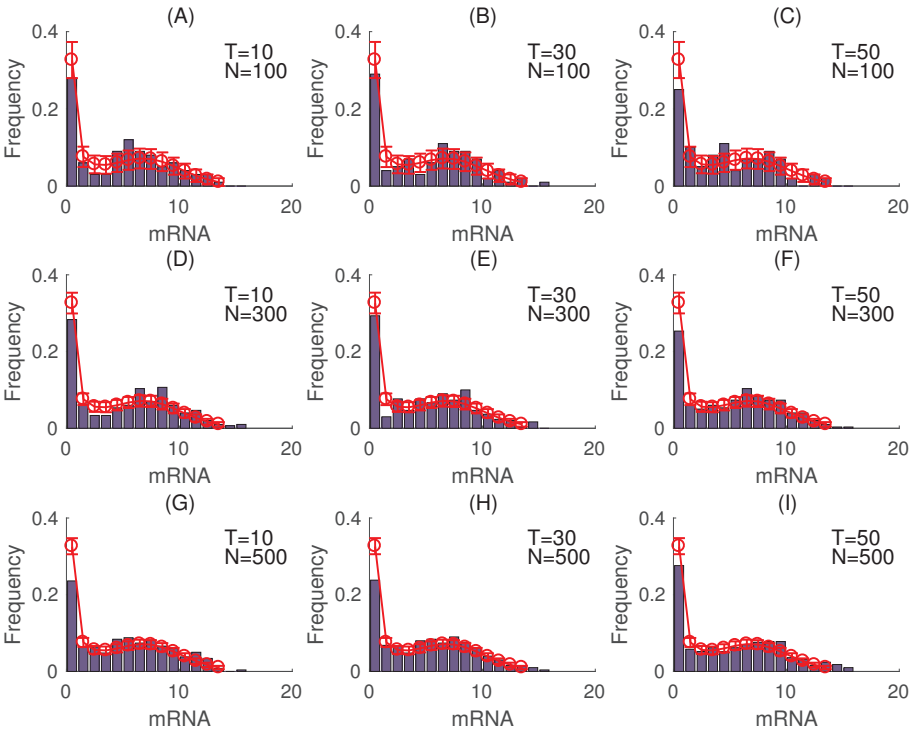

Figure 6. **mRNA distribution under different time and ensemble sizes.** In the figures, red lines represent theoretical mean value and error bars are scope of standard deviation. Gray bars are distribution obtained from numerical simulation. Here we set $k_0 = 0.2, k_1 = 0.2, v_0 = 8, d_0 = 1$, which corresponds to the type I distribution. From left to right, times are $T = 10, 30, 50$, from up to down, ensemble sizes are $N = 100, 300, 500$ respectively.

sites in the promoter region. DNA methylation affects the gene expression by alternating the switch rate between ON and OFF state, larger β-value tends to a more compact chromatin structure, hence decreases the rate k_0 and increases the rate k_1. We assume that

$$\begin{aligned} k_0 &= h_0 \mathrm{e}^{-\mu_0 \beta}, \\ k_1 &= h_1 \mathrm{e}^{-\mu_1 (1-\beta)}. \end{aligned} \tag{35}$$

The random changes of DNA methylation in cell division is an important source for perturbation of the gene expression. To model the random

transition of the DNA methylation state, let β_k the β-value of the k'th generation, we assume that the β-value at the next generation, β_{k+1}, obeys a Beta distribution $Beta(c_k, d_k)$ with parameters c_k and d_k depending on β_k, i.e.,

$$Prob(\beta_{k+1} = x) = \frac{\Gamma(c_k + d_k)}{\Gamma(c_k)\Gamma(d_k)} x^{c_k - 1}(1 - x)^{d_k - 1}. \tag{36}$$

With the distribution parameters c_k and d_k, the mean and variance of β_{k+1} are

$$\mathrm{E}(\beta_{k+1}) = \frac{c_k}{c_k + d_k}, \quad \mathrm{var}(\beta_{k+1}) = \frac{c_k d_k}{(c_k + d_k)^2(c_k + d_k + 1)}. \tag{37}$$

To obtain the relationship between c_k, d_k and β_k, we assume that there is a positive feedback of DNA methylation, so that the mean of the daughter cell β_{k+1} depends on the β-value of the mother cell β_k through

$$\langle \beta_{k+1} \rangle = \phi(\beta_k) = u + \frac{\beta_k^n}{\beta_k^n + v}. \tag{38}$$

Here, n, u, v are parameters, and we only consider one daughter cell after cell division. Hence, let $\mathrm{E}(\beta_{k+1}) = \langle \beta_{k+1} \rangle$, and assume

$$\mathrm{var}(\beta_{k+1}) = \frac{1}{m+1} \langle \beta_{k+1} \rangle (1 - \langle \beta_{k+1} \rangle),$$

where m represents the number of CpG sites in the promoter region, we have

$$c_k = m\langle \beta_{k+1} \rangle, \quad d_k = m(1 - \langle \beta_{k+1} \rangle). \tag{39}$$

Eq. (38) and (39) provide a way of determining c_k and d_k from β_k.

The above discussion provides a numerical scheme of simulating gene expression cross cell cycles with modifications in DNA methylation. We first initialize the β-value at the first cycle β_k ($k = 1$), which gives the rate k_0 and k_1 for first cycle gene expression. At the time of cell division, we calculate the coefficients c_k and d_k according to Eq. (39), and then find a β-value (β_{k+1})according to the Beta distribution density function Eq. (36). This newly obtained β-value enables us to perform the simulation for the next cycle, and so on.

Fig. 7 shows the distributions of mRNA numbers obtained by numerical simulations. Here, the distributions calculated from an ensemble of 2000

cells, each starts with randomly setting β-values. Results show five type distributions, with corresponding parameters given by Table 1. The five type distributions are consistent with the transcript variability in single mammalian cells [1].

Type	v_0	h_0	h_1	μ_0	μ_1	n
Type I	8	0.2	0.2	1	1	2.7
Type II	15	4	4	1	0.1	5
Type III	2	3	3	1	0.1	5
Type IV	17	11.9	40	0	10	3.1
Type V	17	11.9	40	0	10	2.7

Table 1. Parameters used for Fig. 7. In all types, we take $d_0 = 1, u = 0.05, v = 0.15, m = 60$.

To obtain better understanding of the simulation results, we derive the analytic formulations for the DNA distributions. The transition in the DNA methylation is regulated by the function ϕ so that

$$\langle \beta_{k+1} \rangle = \phi(\beta_k).$$

First, we assume that there is no random fluctuations in the β-value, so that the β-value during cell cycling is determined by the iteration $\beta_{k+1} = \phi(\beta_k)$. Hence, at the equilibrium state, β-values of all cells in the ensemble are determined by the stable fix points of $\beta^* = \phi(\beta^*)$.

From the Hill type function $\phi(\beta)$, we have two possible cases.

(1) **There are two fix point β_1^* and β_2^*.** In this case, the equation $\phi(x) = x$ gives three roots, $x_1 = \beta_1^*, x_2 = \beta_2^*$, and $x_1 < x_c < x_2$ (Fig. 8). Moreover, when we select initial β-values randomly from the interval $[0, 1]$. For any $x \in [0, 1]$, it is easy to have

$$\begin{aligned} &\lim_{n\to\infty} \phi^n(x) = \beta_1^*, \qquad \text{for } x \in (0, x_c) \\ &\lim_{n\to\infty} \phi^n(x) = \beta_2^*, \qquad \text{for } x \in (x_c, 1). \end{aligned} \tag{40}$$

Thus, at the stationary state, we have two subpopulation of cells, each with different β-value. Let the probability of having k mRNAs in the cells with given β-value as $P_k(\beta)$, then $P_k(\beta)$ is given by Eq. (8) with transition rate $k_0(\beta)$ and $k_1(\beta)$ by Eq. (35). Thus, the total probability

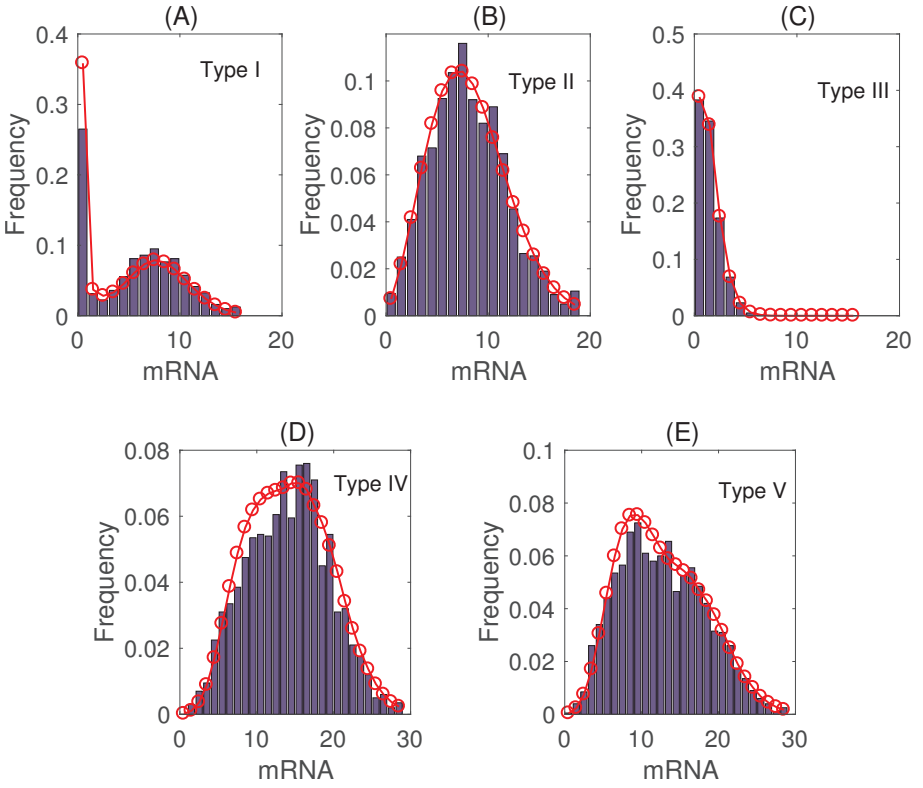

Figure 7. **The distribution of mRNA with random changes in DNA methylation.** Parameters for Type I - V distribution are shown at Table 1. Gray bars are obtained from numerical simulation, red lines are given by analytic formulations.

of having k mRNAs in all cells in the ensemble is

$$
\begin{aligned}
(41)\quad P_k^{tot} &= x_c P_k(\beta_1^*) + (1-x_c)P_k(\beta_2^*) \\
&= x_c \frac{\Gamma(\kappa_0(\beta_1^*)+k)\Gamma(\kappa_0(\beta_1^*)+\kappa_1(\beta_1^*))}{\Gamma(\kappa_0(\beta_1^*))\Gamma(\kappa_0(\beta_1^*)+\kappa_1(\beta_1^*)+k)\Gamma(k+1)} a^k \\
&\quad \times \sum_{n=0}^{\infty} \frac{\Gamma(\kappa_0(\beta_1^*)+k+n)\Gamma(\kappa_0(\beta_1^*)+\kappa_1(\beta_1^*)+k)}{\Gamma(\kappa_0(\beta_1 j)+k)\Gamma(\kappa_0(\beta_1^*)+\kappa_1(\beta_1^*)+k+n)n!}(-a)^n \\
&\quad + (1-x_c)\frac{\Gamma(\kappa_0(\beta_2^*)+k)\Gamma(\kappa_0(\beta_2^*)+\kappa_1(\beta_2^*))}{\Gamma(\kappa_0(\beta_2^*))\Gamma(\kappa_0(\beta_2^*)+\kappa_1(\beta_2^*)+k)\Gamma(k+1)} a^k \\
&\quad \times \sum_{n=0}^{\infty} \frac{\Gamma(\kappa_0(\beta_2^*)+k+n)\Gamma(\kappa_0(\beta_2^*)+\kappa_1(\beta_2^*)+k)}{\Gamma(\kappa_0(\beta_2^*)+k)\Gamma(\kappa_0(\beta_2^*)+\kappa_1(\beta_2^*)+k+n)n!}(-a)^n.
\end{aligned}
$$

This gives the analytic distribution functions for Types I, IV, V, which are shown by red lines in Fig. 7, and are in good agreement with numerical simulation.

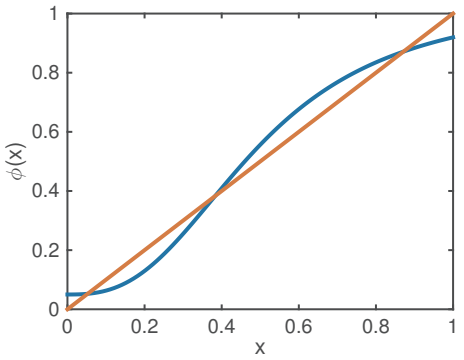

Figure 8. **The cross generation transition function** $\phi(x)$. Red line shows the reference $y = x$, blue line shows $y = \phi(x)$.

(2) **There is one fix point** β^*. In this case, similar to the above discussion, the mRNA distribution is given by the probability $P_k(\beta^*)$, i.e.,

$$P_k^{tot} = \frac{\Gamma(\kappa_0(\beta_*)+k)\Gamma(\kappa_0(\beta_*)+\kappa_1(\beta_*))}{\Gamma(\kappa_0(\beta_*))\Gamma(\kappa_0(\beta_*)+\kappa_1(\beta_*)+k)\Gamma(k+1)}a^k \times \sum_{n=0}^{\infty}\frac{\Gamma(\kappa_0(\beta_*)+k+n)\Gamma(\kappa_0(\beta_*)+\kappa_1(\beta_*)+k)}{\Gamma(\kappa_0(\beta_*)+k)\Gamma(\kappa_0(\beta_*)+\kappa_1(\beta_*)+k+n)n!}(-a)^n. \tag{42}$$

This gives the distribution function in Type II and III in Fig. 7.

Our results show that DNA methylation influences the dynamics of gene expression through the alternations in the β-value. Consequently, the transcriptome of mammalian cells shows extra mRNA distribution types due to the transition of DNA methylations. Moreover, the distributions are mainly determined by the transcription dynamics and the transition of β-value over cell division.

3. Discussion

In this study, we have studied the distribution of mRNA in gene expression through a three-stage model. Based on model analysis, we obtained the exact formulations for mRNA distribution at the stationary state and its time

evolution. Moreover, we proved the equivalence between the time distribution based on a single cell dynamics and the ensemble distribution of many cells at the stationary state. Finally, we considered the situation of gene expression with random transitions of DNA methylation in cell cycling. We show that under the influence of DNA methylation, the mRNA distributions of an ensemble of cells show 5 different type distributions, in agreement with the transcript variability in single mammalian cells experiments. A method of predicting the mRNA distribution through the transition dynamics of DNA methylation during cell cycle is proposed in our study.

The current study is basic and important for our understanding of the gene expression dynamics. Nowadays, single cell RNA-sequencing has been widely applied in studies of developmental biology and cancer research. These single cell RNA-sequencing data enable us to track the transcription level of each single cell, hence we can experimentally obtain the ensemble distributions of each gene. Thus, it is valuable to apply the theoretical results in this study to understand the single-cell RNA-sequencing data. The current study provides a scheme of combining theoretical study with experimental data to achieve a better understanding of each gene expression, especially the regulation of DNA methylation in shaping the mRNA distributions.

Appendix

A. Equilibrium state distribution of mRNA

To solve Eq. (5), we eliminate f^1 to obtain

$$v\frac{\partial^2 f^0}{\partial v^2} + [-av + (1 + \kappa_0 + \kappa_1)]\frac{\partial f^0}{\partial v} + -a\kappa_0 f^0 = 0. \tag{43}$$

Assuming $f^0 = \sum_n C_n v^n$, we substitute f^0 into the above equation, and compare the coefficients of v, we obtain

$$\begin{aligned} C_1 &= C_0 \frac{\kappa_0 a}{1 + \kappa_0 + \kappa_1}, \\ C_{n+1} &= C_n \frac{(\kappa_0 + n)a}{(n+1)(n+1+\kappa_0+\kappa_1)}. \end{aligned} \tag{44}$$

Thus, from the iteration about C_n, we have the expression of f^0.

$$f^0 = C_0 \sum_n \frac{(\kappa_0)_n a^n}{(1 + \kappa_0 + \kappa_1)_n n!} v^n. \tag{45}$$

Substituting Eq. (45) to Eq. (5), we have

$$f^1 = \frac{1}{\kappa} C_0 \sum_n \frac{(\kappa_0)_n a^n}{(1+\kappa_0+\kappa_1)_n n!}(n+\kappa_0)v^n. \tag{46}$$

Hence, the total generating function $F(z)$ is given by

$$F(v) = f^0 + f^1 = \frac{C_0(\kappa_0+\kappa_1)}{\kappa_1} \sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v^n. \tag{47}$$

Finally, from the normalization condition for $v = 0$, we have

$$F(v) = \sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v^n \tag{48}$$

B. Non-equilibrium state distribution of mRNA

To obtain the solution of Eq. (10), we solve the first two equations directly, and obtain

$$\begin{aligned} &\tau = r, \\ &v = v_0 \mathrm{e}^{\tau} = v_0 \mathrm{e}^{r}. \end{aligned} \tag{49}$$

Substituting Eq. (49) into Eq. (10), we obtain

$$v\frac{d\vec{f}}{dr} = \begin{bmatrix} -\kappa_0 & \kappa_1 \\ \kappa_0 & av - \kappa_1 \end{bmatrix} \vec{f}. \tag{50}$$

This is equivalent to Eq. (5), hence we have the exact solution Eq. (47), i.e.,

$$F(z) = \frac{C_0\kappa}{\kappa_1} \sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v^n, \tag{51}$$

where $v = z - 1$.

Now, we assume that there are m mRNAs in this cell initially, then

$$F(v_0) = z_0^m = (1+v_0)^m. \tag{52}$$

Hence

$$\frac{C_0\kappa}{\kappa_1} \sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!} v_0^n = (1+v_0)^m, \tag{53}$$

which gives

$$F(v,\tau)=\left[\sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!}v_0^n\right]^{-1}(1+v_0)^m\left[\sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!}v^n\right] \tag{54}$$

Finally, substituting $v_0 = v\mathrm{e}^{-r} = v\mathrm{e}^{-\tau}$ into Eq. (54), we have the generating function

$$F(v,\tau)=(1+v\mathrm{e}^{-\tau})^m\frac{\sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!}v^n}{\sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!}v^n\mathrm{e}^{-n\tau}}. \tag{55}$$

To calculate the distribution k mRNAs at τ time with initial value m, $P_k(\tau|m)$, based on the approximation Eq. (12)

$$F(v,\tau)\approx(1+mv\mathrm{e}^{-\tau})\sum_n \frac{(\kappa_0)_n a^n}{(\kappa_0+\kappa_1)_n n!}v^n, \tag{56}$$

we apply the derivative relation

$$\left(\sum_n \frac{(\kappa_0)_n a^n}{(\kappa)_n n!}v^n\right)^{(i)}=\sum_n \frac{(\kappa_0)_{n+i}a^{n+i}}{(\kappa)_{n+i}n!}v^n. \tag{57}$$

The k-order derivation of $F(v,\tau)$ is calculated as

$$\begin{aligned}F^{(k)}(v,\tau)&=\sum_{i=0}^{k}C_k^i(1+mv\mathrm{e}^{-\tau})^{(i)}\left(\sum_n \frac{(\kappa_0)_n a^n}{(\kappa)_n n!}v^n\right)^{(k-i)}\\&=(1+mv\mathrm{e}^{-\tau})\sum_n \frac{(\kappa_0)_{n+k}a^{n+k}}{(\kappa)_{n+k}n!}v^n\\&\quad+km\mathrm{e}^{-\tau}\sum_n \frac{(\kappa_0)_{n+k-1}a^{n+k-1}}{(\kappa)_{n+k-1}n!}v^n.\end{aligned} \tag{58}$$

Next, we have

$$P_k(\tau|m)=\frac{1}{k!}F^{(k)}(v,\tau)|_{v=-1}=(1-m\mathrm{e}^{-\tau})P_k^s+m\mathrm{e}^{-\tau}P_{k-1}^s, \tag{59}$$

where

$$P_k^s=\frac{1}{k!}\sum_n \frac{(\kappa_0)_{n+k}a^{n+k}}{(\kappa)_{n+k}n!}(-1)^n \tag{60}$$

is the stationary distribution for k mRNAs.

References

[1] N. Battich, T. Stoeger, and L. Pelkmans, *Control of transcript variability in single mammalian cells*, Cell **163** (2015), no. 7, 1596–1610.

[2] Y. Ben-Shahar, *Epigenetic switch turns on genetic behavioral variations*, Proc. Natl. Acad. Sci. USA **114** (2017), 201717376.

[3] H. Cedar and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms*, Nat. Rev. Genet. **10** (2009), no. 5, 295.

[4] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, *Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells*, Science **343** (2014), no. 6167, 193–196.

[5] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, *Stochastic gene expression in a single cell*, Science **297** (2002), no. 5584, 1183–1186.

[6] N. Friedman, L. Cai, and X. S. Xie, *Linking stochastic dynamics to population distribution: an analytical framework of gene expression*, Phys. Rev. Lett. **97** (2006), no. 16, 168302.

[7] D. T. Gillespie, *Exact Stochastic Simulation of Coupled Chemical Reactions*, J. Chem. Phys. **126** (2007), no. 12, 2350–312.

[8] R. Jaenisch and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals*, Nat. Genet. **33** (2003), 245.

[9] P. A. Jones, *Functions of DNA methylation: islands, start sites, gene bodies and beyond*, Nat. Rev. Genet. **13** (2012), no. 7, 484.

[10] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins, *Stochasticity in gene expression: from theories to phenotypes*, Nat. Rev. Genet. **6** (2005), no. 6, 451.

[11] N. Kumar, T. Platini, and R. V. Kulkarni, *Exact distributions for stochastic gene expression models with bursting and feedback*, Phys. Rev. Lett. **113** (2014), no. 26, 268105.

[12] N. Kumar, A. Singh, and R. V. Kulkarni, *Transcriptional bursting in gene expression: analytical results for general stochastic models*, PLoS Comput. Biol. **11** (2015), no. 10, e1004292.

[13] J. A. Law and S. E. Jacobsen, *Establishing, maintaining and modifying DNA methylation patterns in plant and animals*, Nat. Rev. Genet. **11** (2010), 204–220.

[14] J. T. Leek and J. D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis*, PLoS Genet **3** (2007), no. 9, e161.

[15] J. Lei, *Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters*, J. Theor. Biol **256** (2009), 485–492.

[16] J. Lei, S. A. Levin, and Q. Nie, *Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation*, Proc Nat Acad Sci USA **111** (2014), no. 10, E880–E887.

[17] J. J. Li and M. D. Biggin, *Statistics requantitates the central dogma*, Science **347** (2015), no. 6226, 1066–1067.

[18] L. D. Moore, T. Le, and G. Fan, *DNA methylation and its basic function*, Neuropsychopharmacology **38** (2013), no. 1, 23.

[19] B. Munsky, G. Neuert, and A. V. Oudenaarden, *Using gene expression noise to understand gene regulation*, Science **336** (2012), no. 6078, 183–187.

[20] J. R. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise*, Nature **441** (2006), no. 7095, 840.

[21] S. J. Park, S. Song, G. Yang, P. M. Kim, S. Yoon, J. Kim, and J. Sung, *The Chemical Fluctuation Theorem governing gene expression*, Nat. Commun. **9** (2018), no. 1, 297.

[22] J. Paulsson, *Summing up the noise in gene networks*, Nature **427** (2004), 415–418.

[23] A. Raj and A. V. Oudenaarden, *Nature, nurture, or chance: stochastic gene expression and its consequences*, Cell **135** (2008), no. 2, 216–226.

[24] J. M. Raser and E. K. O'shea, *Control of stochasticity in eukaryotic gene expression*, Science **304** (2004), no. 5678, 1811–1814.

[25] W. Reik, *Stability and flexibility of epigenetic gene regulation in mammalian development*, Nature **447** (2007), no. 7143, 425.

[26] S. Seisenberger, J. R. Peat, T. A. Hore, F. Santos, W. Dean, and W. Reik, *Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers*, Phil. Trans. R Soc. B **368** (2013), 20110330.

[27] V. Shahrezaei, J. F. Ollivier, and P. S. Swain, *Colored extrinsic fluctuations and stochastic gene expression*, Mol. Syst. Biol. **4** (2008), no. 1, 196.

[28] V. Shahrezaei and P. S. Swain, *Analytical distributions for stochastic gene expression*, Proc. Natl. Acad. Sci. USA **106** (2009), no. 1, 346–346.

[29] E. C. Small, L. Xi, J. Wang, J. Widom, and J. D. Licht, *Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity*, Proc. Natl. Acad. Sci. USA. **111** (2014), no. 24, E2462–E2471.

[30] E. N. Smith and L. Kruglyak, *Gene–environment interaction in yeast gene expression*, PLoS Biol **6** (2008), no. 4, e83.

[31] P. S. Swain, M. B. Elowitz, and E. D. Siggia, *Intrinsic and extrinsic contributions to stochasticity in gene expression*, Proc. Natl. Acad. Sci. USA **99** (2002), no. 20, 12795–12800.

[32] M. Thattai and A. V. Oudenaarden, *Intrinsic noise in gene regulatory networks*, Proc. Natl. Acad. Sci. USA **98** (2001), no. 15, 8614–8619.

[33] L. S. Tsimring, *Noise in biology*, Rep Prog Phys **77** (2014), no. 2, 026601.

[34] J. Zhang, L. Chen, and T. Zhou, *Analytical distribution and tunability of noise in a model of promoter progress*, Biophys J. **102** (2012), 1247–1257.

Zhou Pei-Yuan Center for Applied Mathematics
MOE Key Laboratory of Bioinformatics
Tsinghua University, Beijing 100084, P. R. China
E-mail address: jxp17@mails.tsinghua.edu.cn
E-mail address: jzlei@tsinghua.edu.cn