Probability Distributions for Stochastic Gene Expression with Multiple Binding Sites

Huahai Qiu\textsuperscript{1,2,*}

\textsuperscript{1}School of Mathematics and Computer, Research Center of Nonlinear Science, Wuhan Textile University, Wuhan 430073, P. R. China
\textsuperscript{2}Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P. R. China

(Received March 9, 2014; Revised July 25, 2014)

While regulation of intrinsic stochasticity in gene expression is essential for many cellular functions, there is considerable interest in understanding how different molecular mechanisms of gene expression impact variations in protein levels across a population of cells. Stochastic properties of the corresponding gene systems are usually examined by directly finding the probability distributions of mRNA or protein governed by chemical master equations. Here, we analyze four stochastic models of gene expression, each describing a representative topology of binding sites. The total mRNA or protein probability distributions are derived by the weight assignment method that we propose here. These distributions indicate that the total mRNA or protein number obeys a beta distribution or a linear superposition of two beta distributions in spite of differences in the local mRNA or protein numbers distributed in individual binding sites among the models. Our results reveal essential mechanisms of gene expression in terms of the probability distribution rather than finite-order moments, as was done in previous studies.

DOI: 10.6122/CJP.20140811A PACS numbers: 02.60.Cb, 87.15.ad, 87.15.ap

I. INTRODUCTION

Gene expression (a process for the information flow from DNA to protein through mRNA) is controlled by a set of biochemical reactions, thus gene regulatory networks can be regarded as biochemical reaction networks. Due to the low copy numbers of molecular species in the biochemical reactions, this inevitably leads to stochastic expressions of mRNA and protein within active cells. As a result, the abundance of a given protein changing in the same environment can exhibit differences between cellular populations with the same genetic information [1–3]. This stochasticity in gene expression has attracted extensive attention of researchers [4], but progress in the relevant research is slow. While recent advances in experimental methods allow direct observations of real-time fluctuations in gene expression levels in individual live cells, they also allow direct observations of static differences in gene expression levels across a population of cells [5–11]. It is not hard to imagine that these two kinds of measurements are closely related, analogous to the ergodic

\*Electronic address: qiuhuahai2006@163.com

http://PSROC.phys.ntu.edu.tw/cjp 

© 2014 THE PHYSICAL SOCIETY OF THE REPUBLIC OF CHINA
Gene expression in the transcription levels is mainly controlled by the cis-regulatory module and transcription factor protein binding sites [13, 14]. Transcription factors and other transcription components (e.g., RNA polymerase, TATA binding protein) interact with each other, thus affecting the transcription rate from DNA to mRNA, and further the translation rate from mRNA to protein. Like the interaction among any molecules, the binding of transcription factors to DNA loci is also a random event or process. More precisely, the states of the cis-regulatory module (either active or inactive) are randomly transitioning. In contrast to stochastic fluctuations caused by other factors (e.g., stochastic fluctuations in the nutrition supply, cell division, and regulatory input of transcription mechanisms), which is called extrinsic noise, the random source caused by the interactions among molecules is called intrinsic noise. Previous works showed that for simple gene expression models, the probability distribution of mRNA or protein can be analytically given. In fact, it has been shown that for a two-state or three-state gene expression model, the mRNA or protein number obeys a gamma distribution under some conditions [1–3, 15–20] (note: a gamma distribution is equivalent to a beta distribution in mathematics). Here, we point out that once the probability distributions are analytically given, then some common quantities, such as the average of species molecules (determined by the first-order moment), the noise intensity (defined as the ratio of standard variance over mean), and Fano factor (defined as the ratio of variance over mean), are completely determined. Therefore, deriving analytical distributions is significant for understanding the gene expression process as well as for revealing the corresponding stochastic mechanisms.

However, cis-regulatory modules in prokaryotic and eukaryotic cells may be complex. In general, the diversity of cis-regulatory architecture is reflected in the number of DNA binding sites, such as the ones of a single DNA binding site and multiple DNA binding sites. The latter case often reflects a cooperative effect in the process of gene regulation. This is because transcription factor proteins do not act directly on the DNA loci without interacting with other proteins, but rather act on it in a cooperative manner. In the case of the occurrence of multiple binding sites, the promoter activity states may stochastically transition to one another resulting in the stochastic synthesis of mRNA and protein. Incorporating the above complexity, many complex gene models have been established in recent years [6, 21], but they assume that the transcriptional rates are controlled mostly by the regulating function of gene expression (e.g., the Hill function [22, 23] or a special function fitting the experimental data [6, 21, 24, 25]), i.e., it is assumed that the change in transcription factor level is reflected immediately in the transcription rate. This assumption is reasonable in describing the static deterministic behavior of gene systems in some cases and is invalid in other cases. In particular, the assumption possibly leads to the transcription noise being significantly underestimated. In a word, the existing gene models cannot accurately describe the occurrence of stochastic events in the gene expression process, where the intrinsic noise would play an important role.

Aiming at shortcomings of the gene models in the literature, in this paper we establish stochastic models of gene expression, which consider the complexity of cis-regulatory modules in the promoter region. Specifically, we use the chemical master equation of stochastic
gene models for four representative topologies of the cis-regulatory module due to the regulation of transcription factors, where the active and inactive states of the promoter may stochastically transition. To solve these stochastic models, we introduce a new method called the weight assignment method, which converts the problems of finding steady-state distributions of mRNA or protein into Riccati equations that are solvable. As a result, we derive the analytical expressions for the static probability distributions in the models, and show that the total mRNA or protein number obeys a beta distribution or a linear superposition of two beta distributions. We point out that our results are significant from the viewpoint of gene expression dynamics and our method has broad applications in the stochastic analysis of biochemical reaction networks.

II. THE NEW METHOD DESCRIPTION: A SIMPLE EXAMPLE

In order to help understand our method of how to solve stochastic models describing gene expression with multiple binding sites for transcription factors, let us consider a simple gene model, where the promoter is assumed to have one active state and one inactive state. The corresponding biochemical reactions take the form

\[ \begin{align*}
I & \xrightarrow{\gamma} A, \\
A & \xrightarrow{\lambda} I, \\
A & \xrightarrow{\mu} A + X, \\
X & \xrightarrow{\delta} \phi,
\end{align*} \]

where \( I \) represents the gene inactive state, \( A \) represents the gene active state, and \( X \) stands for mRNA or protein; \( \gamma \) and \( \lambda \) represent the switching rates from \( I \) to \( A \) and from \( A \) to \( I \), respectively; \( \mu \) and \( \delta \) are the synthesis and degradation rates of \( X \), respectively. When the parameter \( \mu \) is large enough compared with the other parameters, we can treat the discrete variable \( X \) as a continuous variable, leading to the following ordinary differential equation (ODE):

\[ \frac{dx}{dt} = -\delta x + \mu f(t), \]

where \( x \) represents the concentration of \( X \) and \( f(t) \in \{0, 1\} \) represents the random telegraph signals of gene active or inactive states. In order to seek for the probability distribution of the steady-state \( X \), we can write the conditional probability density of \( X \) in the active and inactive states as

\[ \begin{align*}
P_A(t) &= Pr(f(t) = 1 \text{ and } y \in (x, x + dx)) : \text{ at time } t, \\
P_I(t) &= Pr(f(t) = 0 \text{ and } y \in (x, x + dx)) : \text{ at time } t.
\end{align*} \]

Note that the total probability density of \( X \) is the sum of two local probability densities, i.e., \( P = P_A + P_I \), and the flow of two conditional probability densities can be represented
by its production and degradation rates:

\[ J_A(x, t) = (\mu - \delta x) P_A(x, t), \]
\[ J_I(x, t) = -\delta x P_I(x, t). \]

Using the above equation and the conservation of probability condition, we have [12, 26]

\[
\begin{align*}
\frac{\partial P_A}{\partial t} + \frac{\partial J_A}{\partial x} &= \gamma P_I - \lambda P_A, \\
\frac{\partial P_I}{\partial t} + \frac{\partial J_I}{\partial x} &= -\gamma P_I + \lambda P_A.
\end{align*}
\]

(1)

This is a two-state stochastic model of gene expression in the case of continuous change. Now, we solve Eq. (1). Summing the two equations on both sides of Eq. (1) at steady state yields the following equation

\[ \frac{\partial}{\partial x} (\mu P_A - \delta x P) = 0. \]

This suggests that \( \mu P_A - \delta x P \) equals a constant. Because both \( P_A \) and \( P \) tend to zero as \( x \) increases, we can obtain \( \mu P_A - \delta x P = 0 \), or \( P_A = \frac{\delta x}{\mu} P \), which implies that \( P_I = (1 - \frac{\delta x}{\mu}) P \) due to \( P = P_A + P_I \). Furthermore, by Eq. (1) we have

\[ \frac{\partial}{\partial x} \left( -\delta x \left( 1 - \frac{\delta x}{\mu} \right) P \right) = \left( -\gamma \left( 1 - \frac{\delta x}{\mu} \right) + \lambda \frac{\delta x}{\mu} \right) P. \]

Solving this equation, we obtain

\[ P(x) = \left( \frac{\mu}{\delta} \right)^{1-\frac{\gamma}{\delta}+\frac{\lambda}{\delta}} \frac{\Gamma((\gamma + \lambda)/\delta)}{\Gamma(\gamma/\delta)\Gamma(\lambda/\delta)} x^{\gamma-1} \left( \frac{\mu}{\delta} - x \right)^{\lambda-1}, \]

where \( \Gamma(\cdot) \) represents the gamma function. When the deactivation rate \( \lambda \) is much greater than the activation rate \( \gamma \) [27–29], and slightly greater than the degradation rate \( \delta \) of \( X \), \( P(x) \) approaches to the gamma distribution:

\[ P(x) = \frac{\lambda/\mu}{\Gamma(\gamma/\delta)} ((\lambda/\mu)x)^{\gamma-1} e^{-\frac{\lambda}{\mu}x}. \]

(2)

This indicates that the \( X \) distribution is determined by only two lumped parameters \( \gamma/\delta \) and \( \lambda/\mu \).

We point out that the expressions \( P_A = \frac{\delta x}{\mu} P \) and \( P_I = (1 - \frac{\delta x}{\mu}) P \) are actually a kind of assignment of the total probability \( P \) over the binding sites. Motivated by this assignment, we can find analytical distributions in more general stochastic gene models. The corresponding method is called the weight assignment method.
III. RESULTS

The above gene model considers only the case that the promoter has two binding sites (active and inactive) in the cis-regulatory region. As pointed out in the introduction, this region may contain multiple binding sites (e.g., up to 18 in a eukaryotic cell) that altogether can form a complex pattern (called the transition pattern of the promoter). Here, we consider four representative transition patterns of the promoter structure, which are specified as follows:

Case I: the cis-regulatory module has three regulatory states, one of which (denoted by $D_0$) is the inactive state where a gene cannot be transcribed into mRNA or translated into a protein, and the other two (denoted by $D_1$ and $D_2$) are the active states where a gene can be transcribed into mRNA or translated into a protein. Transcription factors can bind to the DNA loci in order, and there is the transition between the active states;

Case II: the cis-regulatory module also has three regulatory states, one of which (denoted by $D_0$) is the inactive state and the other two (denoted by $D_1$ and $D_2$) are the active states, but there is no transition among the active states. For this case, transcription factor plays the role of promoting transcription or translation;

Case III is the same as Case II but there is transition among the active states;

Case IV: the cis-regulatory module has three regulatory states, one of which (denoted by $D_0$) is the inactive state and the other two (denoted by $D_1$ and $D_2$) are the active states. A gene can be transcribed into mRNA or translated into a protein under the assistance of polymerase, but transcription factors can collaboratively bind to the DNA loci, and induce transcription of DNA into mRNA or translation of mRNA into a protein.

The above four promoter architecture can find their prototype in the gene regulation of eukaryotic cells or prokaryotic cells [30–33]. However, in our mathematical modeling, we only consider a single step process, i.e., a gene is transcribed into mRNA, protein is produced when a gene is active. We point out that this simplification does not lose generality.

Case I

According to Fig. 1(a), we can accordingly write the master equation for the local probability densities [12, 26]

$$
\begin{align*}
\frac{\partial P_0}{\partial t} + \frac{\partial (-\delta x P_0)}{\partial x} &= \lambda P_1 - \gamma P_0, \\
\frac{\partial P_1}{\partial t} + \frac{\partial ((\mu - \delta x)P_1)}{\partial x} &= \gamma P_0 - (\gamma + \lambda)P_1 + \lambda P_2, \\
\frac{\partial P_2}{\partial t} + \frac{\partial ((\mu - \delta x)P_2)}{\partial x} &= \gamma P_1 - \lambda P_2,
\end{align*}
$$

(3)

where $x$ represents the concentration of $X$, $P_0$ represents the probability of $x$ when the gene is in the $D_0$ state, $P_1$ and $P_2$ represent the probability density of $x$ when the gene is in the $D_1$ and $D_2$ state, respectively.

Adding three equations on both sides of Eq. (3) and considering the steady-state distribution, we obtain

$$
\frac{\partial [\mu P_1 - (\mu - \delta x)P]}{\partial x} = 0,
$$
where \( P = P_0 + P_1 + P_2 \) represents the total probability of \( x \). Similar to the analysis of the case of the analyzed-above two-states gene model, we know \( P_0 = (1 - \frac{\delta x}{\mu})P \). In spite of this, from Eq. (3) at steady state we still cannot obtain the analytical solution. In fact, we have

\[
\frac{d}{dx} \left\{ a(x) \frac{d}{dx} \left[ a(x) \frac{d}{dx} (a(x)^2 P) \right] \right\} - \mu \frac{d}{dx} \left\{ a(x) \frac{d}{dx} \left[ a(x) \frac{d}{dx} (a(x)P) \right] \right\} + 2(\gamma + \lambda) \frac{d}{dx} \left\{ a(x) \frac{d}{dx} (a(x)^2 P) \right\} - \mu (\gamma + 2\lambda) \frac{d}{dx} \left\{ a(x) \frac{d}{dx} (a(x)P) \right\} + (\gamma^2 + \lambda \gamma + \gamma^2) \frac{d}{dx} [a^2(x)P] - \mu \lambda^2 \frac{d}{dx} [a(x)P] = 0,
\]

where \( a(x) = \mu - \delta x \). This is a third-order ODE from which it is difficult to obtain an analytical solution.

However, notice that we can assume \( P_1 = A(x)P, \ P_2 = B(x)P \), then

\[
A(x) + B(x) = \frac{\delta x}{\mu}. \tag{4}
\]

Therefore, by the first equation of Eq. (3), we have

\[
\log P = -\int \frac{\mu \lambda A + (\delta - \gamma)(\mu - \delta x) - \delta^2 x}{\delta x (\mu - \delta x)} dx. \tag{5}
\]

To obtain \( P \), the key is how to determine the function \( A \). By the third equation of Eq. (3), we obtain

\[
[-\delta B + (\mu - \delta x)B']P + (\mu - \delta x)BP' = (\gamma A - \lambda B)P.
\]

Combining Eq. (4) and Eq. (5) yields the following ODE:

\[
B' = -\frac{\mu \lambda}{\delta x (\mu - \delta x)} B^2 + \left[ \frac{\delta - \gamma}{\delta x} - \frac{\gamma}{\mu - \delta x} \right] B + \frac{\gamma \delta x}{\mu (\mu - \delta x)},
\]

or

\[
A' = \lambda \left( \frac{1}{\delta x} + \frac{1}{\mu - \delta x} \right) A^2 - \left[ \frac{\gamma - \delta}{\delta x} + \frac{2\lambda + \gamma}{\mu - \delta x} \right] A + \frac{\lambda}{\mu - \delta x} + \frac{\gamma - \lambda}{\mu}, \tag{6}
\]

Eq. (6) belongs to the familiar Riccati equation, which generally has no analytical solution. However, if \( \lambda \gg \gamma \) (the deactivation rate is much greater than the activation rate [27–29]), and \( \mu \) and \( \lambda \) are of the same order in quantity, then from Eq. (6), we can obtain the approximate solution

\[
A \approx \frac{\delta x}{\mu} - \frac{\delta}{\lambda}.
\]
Substituting it into Eq. (5) yields the approximate distribution

\[ P \sim x^2 \left( \mu - \delta x \right)^{1/2} - 2, \]

where \( P \) and \( x^2 \left( \mu - \delta x \right)^{1/2} - 2 \) differ by a constant (determined by the normalization condition for the probability density). As a result, the total probability density of \( x \) obeys a beta distribution. Therefore, by the weight assignment method, we find that the agreement between the stochastic numerical simulation and the analytical model is good (see Fig. 1(b)).

Note: if \( P_1 = A(x)P, \ P_2 = B(x)P \) is not set, then we cannot find the analytical expression for the total probability density.

**Case II**

According to Fig. 2(a), we can accordingly write the master equation for the local probability densities \([12, 26]\):

\[
\begin{align*}
\frac{\partial P_0}{\partial t} + \frac{\partial (-\delta x P_0)}{\partial x} &= \lambda (P_1 + P_2) - 2\gamma P_0, \\
\frac{\partial P_1}{\partial t} + \frac{\partial ([\mu - \delta x] P_1)}{\partial x} &= \gamma P_0 - \lambda P_1, \\
\frac{\partial P_2}{\partial t} + \frac{\partial ([\mu - \delta x] P_2)}{\partial x} &= \gamma P_0 - \lambda P_2,
\end{align*}
\]

where the meaning of the variables is the same that as in Case I. Adding the three equations on both sides of Eq. (8) and considering the steady-state distribution, we obtain

\[ \frac{\partial [-\mu P_0 + (\mu - \delta x)P]}{\partial x} = 0, \]

where \( P = P_0 + P_1 + P_2 \) represents the total probability density. Similar to the analysis of Case I, we can deduce \( P_0 = (1 - \frac{\delta x}{\mu})P \), then

\[ P_1 + P_2 = \frac{\delta x}{\mu} P. \]
FIG. 2: (a) Schematic diagram for Case II: the gene promoter has three regulatory states, one of which is inactive (the gene cannot be transcribed into mRNA or translated into protein) and the other two are active (the gene can be transcribed into mRNA or translated into protein), but no transition occurs between two active states. \( X \) represents mRNA or protein. (b) Probability density of \( X \). The solid line represents the theoretical prediction, whereas the open circles represent the stochastic simulation. Parameters: \( \gamma = 0.005, \ \lambda = 0.1, \ \mu = 0.7, \ \delta = 0.001. \)

FIG. 3: Schematic diagram for Case III: the gene promoter has three regulatory states, one of which is inactive (the gene cannot be transcribed into mRNA or translated into protein) and the other two are active (the gene can be transcribed into mRNA or translated into protein), but transitions can occur between two active states. \( X \) represents mRNA or protein.

Similar to the analysis of Case I again, we can deduce

\[
(\log P)' = \frac{2\gamma - \delta}{\delta x} - \frac{\lambda - \delta}{\mu - \delta x}.
\]

Thus, we obtain

\[
P \sim x^{\frac{2\gamma - \delta}{\delta}} (\mu - \delta x)^{\lambda - \delta} \delta,
\]

It indicates that the total probability in Case II is also a beta distribution (see Fig. 2(b)). This result is in accord with the previous result obtained in a two-state model.

**Case III**

According to Fig. 3, we can accordingly write the master equation for the local prob-
ability densities \[12, 26\]

\[
\begin{align*}
\frac{\partial P_0}{\partial t} + \frac{\partial ((-\delta x P_0))}{\partial x} &= \lambda(P_1 + P_2) - 2\gamma P_0, \\
\frac{\partial P_1}{\partial t} + \frac{\partial ((\mu - \delta x)P_1)}{\partial x} &= \gamma(P_0 + P_2) - 2\lambda P_1, \\
\frac{\partial P_2}{\partial t} + \frac{\partial ((\mu - \delta x)P_2)}{\partial x} &= \gamma P_0 + \lambda P_1 - (\lambda + \gamma)P_2,
\end{align*}
\]

(11)

where the meaning of the variables is the same as in Case I. Completely similar to the
analysis of Case I, we derive

\[ P_0 = \left(1 - \frac{\delta x}{\mu}\right) P, \]

where \( P = P_0 + P_1 + P_2 \) represents the total probability density. Therefore, we have

\[ P_1 + P_2 = \frac{\delta x}{\mu} P. \]  

(12)

Similar to the analysis of Case I again, we can deduce

\[ (\log P)' = \frac{2\gamma - \delta}{\delta x} - \frac{\lambda - \delta}{\mu - \delta x}. \]

Thus, we obtain

\[ P \sim x^{\frac{2\gamma - \delta}{s}} \left(\frac{\mu - \delta x}{\mu}\right)^{\frac{\lambda - \delta}{s}}. \]  

(13)

This indicates that the total mRNA or protein number in Case III also obeys a beta
distribution, but the local probability densities are different from those in Case II (the
analytical result is easy to derive).

**Case IV**

According to Fig. 4(a), we can accordingly write the master equation for the local
probability densities \[12, 26\]

\[
\begin{align*}
\frac{\partial P_0}{\partial t} + \frac{\partial ((-\delta x P_0))}{\partial x} &= \lambda(P_1 + P_2) - 2\gamma P_0, \\
\frac{\partial P_1}{\partial t} + \frac{\partial ((\mu - \delta x)P_1)}{\partial x} &= \gamma(P_0 + P_2) - 2\lambda P_1, \\
\frac{\partial P_2}{\partial t} + \frac{\partial ((\mu - \delta x)P_2)}{\partial x} &= \gamma P_0 + \lambda P_1 - (\lambda + \gamma)P_2,
\end{align*}
\]

(14)

where the meaning of variables is the same as in Case I. Completely similar to the
analysis in the above two kinds of cases, we can derive

\[ P_0 = \left(1 - \frac{\delta x}{\mu}\right) P, \]

where \( P = P_0 + P_1 + P_2 + P_3 \) represents the total probability density. Similar to the above
analysis, we assume \( P_i = A_i(x)P_i \), where \( i = 1, 2, 3 \). Then

\[ A_1(x) + A_2(x) + A_3(x) = \frac{\delta x}{\mu}. \]  

(15)
By this assumption and using the first equation of Eq. (14), we have

$$\log P = \int \frac{-\mu \lambda (A_1 + A_2) + (2\gamma - \delta)(\mu - \delta x) - \delta^2 x}{\delta(\mu x - \delta x^2)} \, dx. \quad \text{(16)}$$

To determine $P$, the key is how to determine $A_1 + A_2$. For convenience, we assume $B = A_1 + A_2$. Adding the second and third equations on both sides of Eq. (14), we obtain

$$[(\mu - \delta x)(P_1 + P_2)]' = 2\gamma(P_0 + P_3) - 2\lambda(P_1 + P_2).$$

By Eq. (15) combined with Eq. (16), then

$$B' = -\frac{\mu \lambda}{\delta x(\mu - \delta x)} B^2 - \left[ \frac{2\gamma - \delta}{\delta x} + \frac{2(\lambda + \gamma)}{\mu - \delta x} \right] B + \frac{2\gamma}{\mu - \delta x}. \quad \text{(17)}$$

This is also a Riccati equation, generally it does not have an analytical solution. However, if $\delta \gg \gamma$ and $x \ll \frac{\mu}{\delta}$, then Eq. (17) becomes a Bernoulli equation. With this assumption, we can derive

$$B = \frac{x(\mu - \delta x)^{2\lambda / \delta}}{(\mu / 2\delta)(\mu - \delta x)^{2\lambda / \delta} + c}.$$ 

Where $c$ is a positive constant and is determined by $A_2(x) + A_3(x)$ satisfying an ODE. Substituting $B$ into Eq. (16), we finally obtain

$$P \sim x^{-1}(\mu - \delta x)^{2\lambda / \delta} \left[ \frac{\mu}{2} + c\delta(\mu - \delta x)^{2\lambda / \delta} \right]. \quad \text{(18)}$$

Equation (18) indicates that the total number of mRNAs or proteins obeys the probability that is the overlapping of two beta distributions. In addition, we plot the probability distribution of $X$, which shows that the theoretical prediction is in accord with the numerical result (see Fig. 4(b)).

IV. CONCLUSION AND DISCUSSION

Gene expression is an inherently noisy process, involving a series of biochemical events, such as transcription, translation, and protein modification. This intrinsic stochasticity not only is important for cellular functions but also may be facilitated by the biological organism so that it actively carries out some cellular functions. Previous studies on gene expression concentrated on the influences of feedback on noise, including intrinsic, extrinsic, and total noise, trying to derive the quantitative relation between the noise intensity and feedback strength from the viewpoint of fluctuation-dissipation [1, 34–36], or on deriving distributions of mRNA or protein in two-state or three-stage gene models [16–20]. Except that experiments investigate distributions of mRNA or protein in the case of multiple binding sites, relevant theoretical studies are few. Here, we have studied four representative
FIG. 4: (a) Schematic diagram for Case IV: the gene promoter has three regulatory states, one of which is inactive (the gene cannot be transcribed into mRNA or translated into protein) and the other two are active. For the two active states, the gene can be transcribed into mRNA or translated into protein under the induction of transcription factor, respectively, and two transcription factors can collaboratively bind to the composite site $D_1 + D_2$, $X$ represents mRNA or protein. (b) Probability density of $X$. The solid line represents the theoretical prediction, whereas the open circles represent the stochastic simulation. Parameters: $\gamma = 0.00001$, $\lambda = 0.1$, $\mu = 0.7$, $\delta = 0.001$, $c = 2$.

gene models that are established based on the central dogma in biology. By analysis of these models, we find that the total mRNA or protein number in the case of multiple binding sites follows a beta distribution or a linear superposition of two beta distributions in most cases. In particular, we put forward a weight assignment method to solve the related chemical master equations, and derive analytical distributions in the four gene models. We point out that this method has a definite geometric meaning, in fact, such an assignment idea can find its prototype in a number of scientific fields such as mathematics, physics, chemistry, biology, sociology, and economy. The analytical expression for $P(t)$ in Case I can be derived by the limit $\lambda \gg \gamma$, but for Case II, III, and IV, the analytical expression for that which was derived does not need the limit $\lambda \gg \gamma$, moreover, the theoretical prediction is in accord with the numerical result (data not shown). First, for Case II and III, the local probability densities in Case III are different from those in Case II, but the total mRNA or protein number has the same analytical expression and both obey a beta distribution under any condition. Second, if $\delta \gg \gamma$ and $x \ll \frac{\delta}{\gamma}$, then the analytical expression for Case IV can be obtained. In addition, we used values of some parameters (e.g., $\lambda \gg \gamma$) according to the published parameter values for fundamental processes in gene expression [27–29].

For a biochemical network, there are a variety of ways to establish the corresponding stochastic model, such as the Chapman-Kolmogorov equation (CKE), chemical master equation with jump (JME), Fokker-Planck equation (FPE), Langevin equation (LGE), etc. Ullah et al. in Ref. [26] concluded the relationship among the equations, and showed that the CKE is more general (e.g., the other stochastic differential equations can be derived from the CKE). The models used in the paper were established based on the CKE, so the results obtained here have generality and reliability.

The central dogma in biology gives only a basic framework of gene expression regulation in cells, so our models that are established based on this dogma have limitations. With deep studies of gene regulatory systems, more complex molecular mechanisms of regulatory
ways have been revealed, which extend the notation of the traditional central dogma. This extension can be better used to interpret the finer and more diverse regulation of gene expression so as to satisfy different needs in the cellular development processes and the responses of cells to external stimuli. Although transcription and translation of most genes follow the basic regulatory framework described by the central dogma, distinct genes have their own regulatory ways that impact gene expression including the transcription of DNA into mRNA and translation of mRNA into protein. This complexity of regulation was only partially considered in this article. More details of regulation would lead to the conclusion that the mRNA or protein number obeys different distributions. The relevant studies are expected.

Acknowledgments

This work was supported by grants 91230204, 2014CB964703, 2010CB945400, 2011071120045, 11105235, 11301403, 11201353, 61202172.

References


