Protein noise and distribution in a two-stage gene-expression model extended by an mRNA inactivation loop*

Candan Çelik[1], Pavol Bokes[1,2], and Abhyudai Singh[3]

[1] Department of Applied Mathematics and Statistics, Comenius University, Bratislava 84248, Slovakia
candan.celik@fmph.uniba.sk
[2] Mathematical Institute, Slovak Academy of Sciences, Bratislava 81473, Slovakia
pavol.bokes@fmph.uniba.sk
[3] Department of Electrical and Computer Engineering, University of Delaware, Newark, Delaware 19716, USA
absingh@udel.edu

Abstract. Chemical reaction networks involving molecular species at low copy numbers lead to stochasticity in protein levels in gene expression at the single-cell level. Mathematical modelling of this stochastic phenomenon enables us to elucidate the underlying molecular mechanisms quantitatively. Here we present a two-stage stochastic gene expression model that extends the standard model by an mRNA inactivation loop. The extended model exhibits smaller protein noise than the original two-stage model. Interestingly, the fractional reduction of noise is a non-monotonous function of protein stability, and can be substantial especially if the inactivated mRNA is stable. We complement the noise study by an extensive mathematical analysis of the joint steady-state distribution of active and inactive mRNA and protein species. We determine its generating function and derive a recursive formula for the protein distribution. The results of the analytical formula are cross-validated by kinetic Monte-Carlo simulation.

Keywords: Stochastic gene expression · Master equation · Analytical distribution · Generating function · Stochastic simulation

1 Introduction

As many other biochemical mechanisms, gene expression in which protein synthesis occurs is inherently stochastic due to random fluctuations in the copy number of gene products, e.g. proteins [7]. From the viewpoint of biochemical reactions, in simplest formulations, gene expression consists of two main steps: transcription and translation. While RNA polymerase enzymes produce mRNA

* CÇ is supported by the Comenius University grant for doctoral students Nos. UK/106/2020 and UK/100/2021.

molecules in the former, protein synthesis takes place by ribosomes in the latter, each reaction corresponding to the production and decay of relevant species. Additionally, the two-stage model can be extended by the regulation of transcription factors, which affect gene expression by modulating the binding rate of RNA polymerase [3].

Over the last decades, the two-stage model of gene expression has been extensively studied to understand how the stochastic phenomenon in cellular processes takes place [14, 17, 18]. Specifically, quantifying the number of species in terms of probability distributions has become an interesting and challenging endeavour due to the subtleties involved in finding a solution to the underlying problem. On the other hand, the fluctuations in mRNA and protein levels are considered as a major source of noise, leading to cell-to-cell variability in gene regulatory networks [12, 15, 16]. The noise emerges from different sources, namely *intrinsic* and *extrinsic* noise [23, 25]; yet, structural elements such as stem-loops can also contribute to noise by binding to an untranslated region of mRNA [6]. The untranslated regions of mRNAs often contain these stem-loops that can reversibly change configurations making individual mRNAs translationally active/inactive.

Numerous modelling approaches have been proposed that are based on deterministic and stochastic frameworks, and recently also hybrid ones as a combination of the preceding two [5, 10, 21]. Only a few of those provide an explicit solution to the two-stage gene-expression model [4, 18]; most of the studies are based on Monte Carlo simulations, which are usually computationally expensive.

As a generalisation of the two-state model, some studies in the literature consider a set of multiple gene states and investigate the dynamics of stochastic transitions among these states [11, 26]. Nevertheless, to the best of our knowledge, none of these studies takes an mRNA inactivation into account. Here we extend the two-stage model by an MRNA inactivation loop, by which we mean that after transcription species can switch between active and inactive states. In other words, there exists a pair of reversible chemical reactions occurring at constant rates by turning active mRNA species into inactive ones, and vice versa. Subsequently, the active mRNA is translated, while the inactive mRNA stays dormant. The schematic of reactions describing the model is given in (1). Here we thereafter refer to the aforementioned model as *the extended model*.

This paper is organised as follows. In Section 2, the stationary means of active mRNA, inactive mRNA, and protein are obtained from a deterministic formulation the model; the master equation of the stochastic model is formulated, and transformed into a partial differential equation for the generating function. In Section 3, the partial differential equation is transformed into one for the factorial cumulant generation function and a power series solution is found; recursive expressions for the coefficients — the factorial cumulants of the three molecular species — are thereby provided. In Section 4, the protein Fano factor is expressed in terms of the first two factorial cumulants, and the noise-reduction effect of the mRNA inactivation loop is analysed. The generating function of the stationary distribution of active mRNA, inactive mRNA and protein amounts is represented in the special-function form in Section 5. The marginal protein and

active and inactive mRNA distributions are derived in Section 6. The paper is concluded in Section 7.

2 Model formulation

The extended model involves three species, mRNA, inactive mRNA (imRNA for short), and protein, and consists of the reactions

$$\begin{aligned} &\emptyset \underset{\gamma_1}{\overset{\lambda_1}{\rightleftharpoons}} \text{mRNA}, \quad \text{mRNA} \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} \text{imRNA}, \quad \text{imRNA} \xrightarrow{\tilde{\gamma}_1} \emptyset, \\ &\text{mRNA} \xrightarrow{\lambda_2} \text{mRNA} + \text{protein}, \quad \text{protein} \xrightarrow{\gamma_2} \emptyset. \end{aligned} \tag{1}$$

The reactions in (1) correspond to mRNA transcription and decay, mRNA activation and inactivation, inactive mRNA decay, protein translation, and protein decay, respectively.

Due to the linearity of kinetics in (1), the mean levels of the mRNA (m), inactive mRNA ($\tilde{m}$) and protein (n) exactly satisfy the system of deterministic rate equations

$$\begin{aligned} \frac{\mathrm{d}\langle m\rangle}{\mathrm{d}t} &= \lambda_1 - (\gamma_1+\alpha)\langle m\rangle + \beta\langle\tilde{m}\rangle, \\ \frac{\mathrm{d}\langle\tilde{m}\rangle}{\mathrm{d}t} &= \alpha\langle m\rangle - (\tilde{\gamma}_1+\beta)\langle\tilde{m}\rangle, \\ \frac{\mathrm{d}\langle n\rangle}{\mathrm{d}t} &= \lambda_2\langle m\rangle - \gamma_2\langle n\rangle. \end{aligned} \tag{2}$$

Setting time derivatives in (2) to zero, and solving the resulting algebraic system, the stationary means are obtained as

$$\langle m\rangle = \frac{\lambda_1}{\gamma_1^{\mathrm{eff}}}, \quad \langle\tilde{m}\rangle = \frac{\alpha}{\tilde{\gamma}_1+\beta}\langle m\rangle, \quad \langle n\rangle = \frac{\lambda_2}{\gamma_2}\langle m\rangle, \tag{3}$$

for the mRNA, inactive mRNA, and protein respectively, where

$$\gamma_1^{\mathrm{eff}} = \gamma_1 + \frac{\alpha\tilde{\gamma}_1}{\tilde{\gamma}_1+\beta} \tag{4}$$

denotes the effective rate of mRNA decay. Owing to the linearity of reaction rates, one can find a closed system of differential equations not only for means, but also for higher-order moments [19, 22]; however these equations are typically less revealing than the mean dynamics. Here we take a different approach and quantify the protein noise as a by-product of a generating-function analysis in Section 4.

The probability $p_{m,\tilde{m},n}(t)$ of having m mRNA, $\tilde{m}$ inactive mRNA, and n protein molecules at time t satisfies the chemical master equation

$$\begin{aligned}\frac{\mathrm{d}p_{m,\tilde{m},n}}{\mathrm{d}t} &= \lambda_1(p_{m-1,\tilde{m},n} - p_{m,\tilde{m},n}) + \alpha((m+1)p_{m+1,\tilde{m}-1,n} - mp_{m,\tilde{m},n}) \\ &+ \tilde{\gamma_1}((\tilde{m}+1)p_{m,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}) + \lambda_2 m(p_{m,\tilde{m},n-1} - p_{m,\tilde{m},n}) \\ &+ \gamma_2((n+1)p_{m,\tilde{m},n+1} - np_{m,\tilde{m},n}) + \gamma_1((m+1)p_{m+1,\tilde{m},n} - mp_{m,\tilde{m},n}) \\ &+ \beta((\tilde{m}+1)p_{m-1,\tilde{m+1},n} - \tilde{m}p_{m,\tilde{m},n}).\end{aligned} \tag{5}$$

Equating the left-hand side of (5) to zero yields the steady-state master equation

$$\begin{aligned}0 =& \lambda_1(p_{m-1,\tilde{m},n} - p_{m,\tilde{m},n}) + \alpha((m+1)p_{m+1,\tilde{m}-1,n} - mp_{m,\tilde{m},n}) \\ &+ \tilde{\gamma_1}((\tilde{m}+1)p_{m,\tilde{m}+1,n} - \tilde{m}p_{m,\tilde{m},n}) + \lambda_2 m(p_{m,\tilde{m},n-1} - p_{m,\tilde{m},n}) \\ &+ \gamma_2((n+1)p_{m,\tilde{m},n+1} - np_{m,\tilde{m},n}) + \gamma_1((m+1)p_{m+1,\tilde{m},n} - mp_{m,\tilde{m},n}) \\ &+ \beta((\tilde{m}+1)p_{m-1,\tilde{m+1},n} - \tilde{m}p_{m,\tilde{m},n}),\end{aligned} \tag{6}$$

We additionally require that the normalising condition

$$\sum_{m,\tilde{m},n} p_{m,\tilde{m},n} = 1 \tag{7}$$

hold.

We aim to find the moments of the probability distribution $p_{m,\tilde{m},n}$ by using the generating function approach [8]. In order to solve (6)–(7), we employ the probability generating function

$$G(x,y,z) = \sum_{m,\tilde{m},n} x^m y^{\tilde{m}} z^n p_{m,\tilde{m},n} \tag{8}$$

for the probability distribution $p_{m,\tilde{m},n}$. Multiplying (6) by the factor $x^m y^{\tilde{m}} z^n$ and summing over $m, \tilde{m}$ and n yields

$$\begin{aligned}\lambda_1(1-x)G =& (\lambda_2 x(z-1) + \gamma_1(1-x) + \alpha(y-x))\frac{\partial G}{\partial x} \\ &+ (\tilde{\gamma_1}(1-y) + \beta(x-y))\frac{\partial G}{\partial y} + \gamma_2(1-z)\frac{\partial G}{\partial z}.\end{aligned} \tag{9}$$

Equation (9) is subject to

$$G(1,1,1) = 1, \tag{10}$$

which is implied by the normalisation condition (7).

3 Factorial cumulant generating function

In order to find a particular solution to (9)–(10), we change the variables according to

$$x = 1+u, \quad y = 1+v, \quad z = 1+w, \quad G = \exp(\varphi), \tag{11}$$

and obtain that the factorial cumulant generating function [9] $\varphi = \varphi(u, v, w)$ is a solution of the inhomogeneous linear partial differential equation (PDE),

$$\lambda_1 u = (-\lambda_2(1+u)w + \gamma_1 u + \alpha(u-v))\frac{\partial \varphi}{\partial u} + (\tilde{\gamma_1} v + \beta(v-u))\frac{\partial \varphi}{\partial v} + \gamma_2 w \frac{\partial \varphi}{\partial w} \quad (12)$$

subject to

$$\varphi(0,0,0) = 0. \quad (13)$$

In order to solve (12)–(13) we shall employ the ansatz

$$\varphi(u,v,w) = \varphi_{00}(w) + u\varphi_{10}(w) + v\varphi_{01}(w). \quad (14)$$

We immediately obtain the partial derivatives

$$\frac{\partial \varphi}{\partial u} = \varphi_{10}(w), \quad \frac{\partial \varphi}{\partial v} = \varphi_{01}(w), \quad \frac{\partial \varphi}{\partial w} = \varphi'_{00}(w) + u\varphi'_{10}(w) + v\varphi'_{01}(w). \quad (15)$$

Inserting (15) into (12) and rearranging the terms yields an inhomogeneous system of ODEs

$$\begin{aligned} &\gamma_2 w\varphi'_{00} - \lambda_2 w\varphi_{10} = 0, \\ &\gamma_2 w\varphi'_{10} + (\gamma_1 + \alpha - \lambda_2 w)\varphi_{10} - \beta\varphi_{01} = \lambda_1, \\ &\gamma_2 w\varphi'_{01} + (\tilde{\gamma_1} + \beta)\varphi_{01} - \alpha\varphi_{10} = 0. \end{aligned} \quad (16)$$

Let us assume that the functions φ_{00}, φ_{10}, and φ_{01} are of the power series form, i.e.,

$$\varphi_{00}(w) = \sum_{k=0}^{\infty} a_k w^k, \quad \varphi_{10}(w) = \sum_{k=0}^{\infty} b_k w^k, \quad \varphi_{01}(w) = \sum_{k=0}^{\infty} c_k w^k. \quad (17)$$

The coefficients a_k, b_k, and c_k give the factorial cumulants of the joint molecular distribution [9]. Note that $a_0 = 0$ follows immediately from the normalisation condition (13). Evaluating the derivatives in (17) and substituting into (16), we obtain the following recurrence equations:

$$a_k = \frac{\lambda_2}{k\gamma_2} b_{k-1}, \quad k \geq 1, \quad (18)$$

$$(\gamma_1 + \alpha)b_0 - \beta c_0 - \lambda_1 + \sum_{k=1}^{\infty} (\gamma_2 k b_k + (\gamma_1 + \alpha)b_k - \lambda_2 b_{k-1} \beta c_k) w^k = 0, \quad (19)$$

$$(\tilde{\gamma_1} + \beta)c_0 - \alpha b_0 + \sum_{k=1}^{\infty} (\gamma_2 k c_k + (\tilde{\gamma_1} + \beta)c_k - \alpha b_k) w^k = 0. \quad (20)$$

Since we consider (17) as a solution to (12) then all the coefficients in (19)–(20) must be zero. Thus, we get

$$(\gamma_1 + \alpha + \gamma_2 k)b_k - \lambda_2 b_{k-1} - \beta c_k = 0, \quad (21)$$

$$(\tilde{\gamma_1} + \beta + \gamma_2 k)c_k - \alpha b_k = 0, \quad (22)$$

for b_k and c_k. Solving the algebraic system (21)–(22) in b_k, $k \geq 1$, yields

$$(\gamma_2^2 k^2 + \gamma_2(\tilde{\gamma_1} + \gamma_1 + \beta + \alpha)k + \tilde{\gamma_1}\gamma_1 + \gamma_1\beta + \tilde{\gamma_1}\alpha)b_k = \lambda_2(\tilde{\gamma_1} + \beta + k\gamma_2)b_{k-1},$$

i.e.

$$b_k = \frac{\lambda_2(\tilde{\gamma_1} + \beta + k\gamma_2)}{\gamma_2^2 k^2 + \gamma_2(\tilde{\gamma_1} + \gamma_1 + \beta + \alpha)k + \tilde{\gamma_1}\gamma_1 + \gamma_1\beta + \tilde{\gamma_1}\alpha} b_{k-1}, \tag{23}$$

where the zeroth term of the sequence b_k is obtained, by equating the terms out of the sums in (19) and (20) to zero, as

$$b_0 = \frac{\lambda_1(\tilde{\gamma_1} + \beta)}{(\gamma_1 + \alpha)(\tilde{\gamma_1} + \beta) - \beta\alpha} = \frac{\lambda_1}{\gamma_1^{\text{eff}}}. \tag{24}$$

Equation (24) thus rederives the stationary mRNA mean (3) by means of factorial cumulant analysis; similarly, c_0 and a_1 can be identified as the stationary imRNA and protein means. Thus, the sequence b_k can be calculated iteratively from (23) starting from the initial condition (24). Having calculated b_k, the sequence a_k and c_k can be evaluated via (18) and (22). In Section 5, we will utilise these formulas to obtain a special-function representation of the generating function. Before doing that, we show that the first two terms of these sequences determine protein variability.

4 Protein variability

As outlined in the previous section, the first-order cumulants b_0, c_0, and a_1 ($a_0 = 0$ by normalisation condition), coincide with the stationary mRNA, imRNA, and protein mean values. In this section, we use the second-order cumulants to describe the stationary noise in our model. The noise in mRNA and imRNA is Poissonian (see Section 6 for details) and therefore uninteresting: we focus on the protein noise.

This section is divided into two parts: the first expresses the Fano factor in terms of the first and second order cumulants (and is independent of the specifics of the current model); the second part uses the formula to analyse the noise reduction effect of the inactivation loop.

Expressing the Fano factor in terms of the cumulants. The generating function is expanded by the Taylor formula as

$$G(1,1,z) = G(1,1,1) + \frac{\partial G}{\partial z}(1,1,1)(z-1) + \frac{1}{2}\frac{\partial^2 G}{\partial z^2}(1,1,1)(z-1)^2 + \mathcal{O}(z-1)^3. \tag{25}$$

Differentiating (8) with respect to z and setting $(x,y,z) = (1,1,1)$ links the derivatives of the generating function to the factorial moments:

$$\frac{\partial G}{\partial z}(1,1,1) = \langle m \rangle, \quad \frac{\partial^2 G}{\partial z^2} = \langle m(m-1) \rangle. \tag{26}$$

Inserting (10) and (26) into (25), we have

$$G(1,1,z) = 1 + \langle m\rangle(z-1) + \frac{\langle m(m-1)\rangle}{2}(z-1)^2 + \mathcal{O}(z-1)^3. \tag{27}$$

On the other hand, (11), (14), and (17) imply

$$G(1,1,z) = \exp\left(a_1(z-1) + a_2(z-1)^2 + \mathcal{O}(z-1)^3\right) \tag{28}$$

$$= \left(1 + a_1(z-1) + \frac{a_1^2}{2}(z-1)^2\right)\left(1 + a_2(z-1)^2\right) + \mathcal{O}(z-1)^3 \tag{29}$$

$$= 1 + a_1(z-1) + \left(a_2 + \frac{a_1^2}{2}\right)(z-1)^2 + \mathcal{O}(z-1)^3. \tag{30}$$

Comparing (27) and (28) gives

$$\langle m\rangle = a_1, \quad \langle m(m-1)\rangle = 2a_2 + a_1^2.$$

The Fano factor,

$$\mathrm{F} = \frac{\langle m^2\rangle}{\langle m\rangle} - \langle m\rangle = \frac{\langle m(m-1)\rangle}{\langle m\rangle} + 1 - \langle m\rangle = \frac{2a_2}{a_1} + 1, \tag{31}$$

is thus expressed in terms of the first two factorial cumulants a_1 and a_2.

Noise reduction by mRNA inactivation loop. Substituting (18) and (23) into (31) and simplifying gives

$$\mathrm{F} = 1 + \frac{b_1}{b_0} = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1 + \frac{\alpha(\gamma_2+\tilde{\gamma}_1)}{\gamma_2+\tilde{\gamma}_1+\beta}}. \tag{32}$$

Formula (32) gives the steady-state protein Fano factor as function of the model parameters (degradation rate constants $\gamma_1, \tilde{\gamma}_1, \gamma_2$ of active/inactive mRNA and protein; inactivation/activation rate constants α, β; translation rate constant λ_2).

In order to compare the protein noise in the current model to that exhibited by the classical two-stage model (without the inactivation–activation loop) we define the baseline Fano factor as

$$\mathrm{F}_0 = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1^{\mathrm{eff}}} = 1 + \frac{\lambda_2}{\gamma_2 + \gamma_1 + \frac{\alpha\tilde{\gamma}_1}{\tilde{\gamma}_1+\beta}}, \tag{33}$$

which can be obtained from (32) by first setting $\alpha = 0$ (no inactivation) and then replacing the mRNA decay rate γ_1 by its effective value (4). Adjusting the mRNA decay rate maintains the same species means in the baseline model like in the full model extended by the inactivation loop.

The protein variability formulae (32) and (33) can equivalently be expressed in terms of the squared coefficient of variation [13, 20] $\mathrm{CV}^2 = \mathrm{F}/\langle n\rangle$ and $\mathrm{CV}_0^2 =$

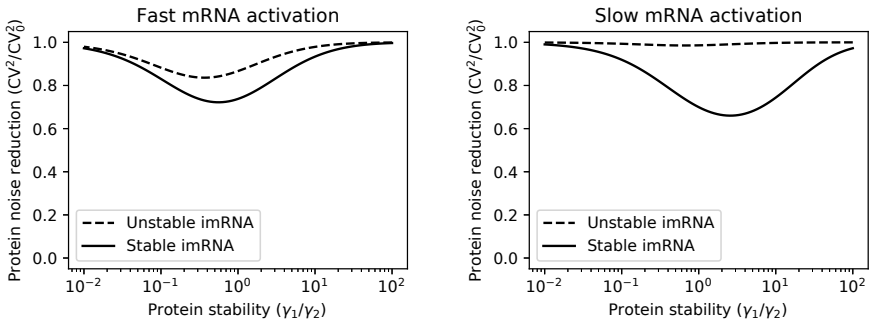

Fig. 1: Fractional protein noise reduction by the mRNA inactivation loop as function of protein stability. The ordinate gives the protein noise (the squared coefficient of variation) in the two-stage model extended by the mRNA inactivation loop relative to the protein noise in a baseline two-stage model without the mRNA inactivation loop (adjusting the mRNA decay rate to obtain the same species means). The protein mean is set to $\langle n\rangle = 500$; the mRNA mean is $\langle m\rangle = 10$; the imRNA decay rate is either the same as that of active mRNA ($\tilde{\gamma}_1 = \gamma_1$; dashed line) or set to zero ($\tilde{\gamma}_1 = 0$; solid line). The inactivation and activation rates are $\alpha = 3$, $\beta = 3$ (left panel) or $\alpha = 1$, $\beta = 0.1$ (right panel); we thereby set $\gamma_1 = 1$ without loss of generality.

$\mathrm{F}_0/\langle n\rangle$. Combining (3) and (32)–(33), we find

$$\mathrm{CV}^2 = \frac{1}{\langle n\rangle} + \frac{1}{\langle m\rangle}\frac{\gamma_2}{\gamma_2+\gamma_1+\frac{\alpha(\gamma_2+\tilde{\gamma}_1)}{\gamma_2+\tilde{\gamma}_1+\beta}}, \tag{34}$$

$$\mathrm{CV}_0^2 = \frac{1}{\langle n\rangle} + \frac{1}{\langle m\rangle}\frac{\gamma_2}{\gamma_2+\gamma_1+\frac{\alpha\tilde{\gamma}_1}{\tilde{\gamma}_1+\beta}} \tag{35}$$

for the protein coefficient of variation and its baseline value (no activation loop).

Comparing (34) to (35), we see that $\mathrm{CV}^2 < \mathrm{CV}_0^2$, allowing us to conclude that the inclusion of the mRNA inactivation loop decreases protein noise. However, the two coefficients will be very close in many parameter regimes; the necessary conditions for observing a significant difference are given by

$$\tilde{\gamma}_1 \lesssim \min\{\beta, \gamma_2\}, \quad \max\{\gamma_1, \gamma_2\} \lesssim \alpha, \tag{36}$$

where by "$\lesssim$" we mean smaller than or of similar magnitude. Thus, in order to obtain significant reduction of noise, we require that an individual active mRNA molecule be more likely to be inactivated than degraded, and that an individual inactive mRNA molecule be more likely to be activated than degraded. Additionally, we require that inactive mRNA be more stable than protein (which is possible if inactivation protects the mRNA from decay).

One particular consequence of the necessary conditions (36) is that the fraction protein noise reduction, $\mathrm{CV}^2/\mathrm{CV}_0^2$, is a non-monotonous function of protein

stability: it tends to one for highly unstable or highly stable proteins, and is less than one for proteins of optimal stability (cf. Figure 1). The optimal value of protein stability critically depends on the rate constant β of mRNA activation. In case of fast mRNA activation, the optimum noise reduction is achieved by unstable proteins (less stable than mRNA; Figure 1, left panel. In case of slow mRNA activation, the optimum can be achieved by stable proteins (Figure 1, right panel). However, slow activation ($\beta \ll 1$) imposes, via (36), a stringent condition on the stability of inactivated mRNA. Indeed, the right panel of Figure 1 demonstrates that there is hardly any reduction of noise if the inactive mRNA is unstable.

In the next section, we go beyond the mean and noise statistics (the first and second order factorial cumulants), using the higher order cumulants to find a special-function representation of the generating function of the joint distribution of mRNA, imRNA, and protein copy numbers.

5 Special-function representation

Factorising the second-order polynomial in k in the denominator of (23) gives

$$b_k = \lambda_2 \frac{\tilde{\gamma_1} + \beta + k\gamma_2}{\gamma_2^2 (k + r_1)(k + r_2)} b_{k-1} \quad \text{for } k \geq 1, \tag{37}$$

where

$$r_{1,2} = \frac{\gamma_1 + \alpha + \tilde{\gamma_1} + \beta \pm \sqrt{(\tilde{\gamma_1} + \beta - \gamma_1 - \alpha)^2 + 4\beta\alpha}}{2\gamma_2}.$$

Note that the sequence b_k in (37) can be rewritten as

$$b_k = b_0 \frac{(1+\tau)_k}{(1+r_1)_k (1+r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1, \tag{38}$$

where we set $\tau = (\tilde{\gamma_1} + \beta)/\gamma_2$ for the sake of simplicity and the polynomial

$$(x)_k = x(x+1)(x+2)\dots(x+k-1), \quad (x)_0 = 1$$

represents the rising factorial, also called the Pochammer symbol.

We next find the remaining sequences a_k and c_k. Inserting (38) into (18) gives

$$a_k = \frac{b_0 r_1 r_2}{\tau} \frac{(\tau)_k}{k(r_1)_k (r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1. \tag{39}$$

Similarly, substituting (38) into (22) yields

$$c_k = \frac{\alpha b_0}{\tilde{\gamma_1} + \beta} \frac{(\tau)_k}{(1+r_1)_k (1+r_2)_k} \left(\frac{\lambda_2}{\gamma_2}\right)^k, \quad k \geq 1, \tag{40}$$

where $c_0 = \frac{\alpha b_0}{\tilde{\gamma_1} + \beta}$, which can be obtained by combining (20) and (24).

Having found the sequences in (17), we next return to the original variables in (11) to obtain the generating function of the stationary distribution of active mRNA, inactive mRNA, and protein amounts, which is given by

$$G(x,y,z) = \exp\left(\sum_{k\geq 1} a_k (z-1)^k + (x-1)\sum_{k\geq 0} b_k (z-1)^k + (y-1)\sum_{k\geq 0} c_k (z-1)^k\right). \tag{41}$$

Equation (41) can be rewritten as

$$\begin{aligned} G(x,y,z) = \exp\Bigg(&\frac{b_0\lambda_2}{\gamma_2}\int_1^z {}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right)ds \\ &+b_0(x-1){}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right) \\ &+\frac{\alpha b_0}{\tilde{\gamma_1}+\beta}(y-1){}_2F_2\left(\begin{matrix}1, \tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right)\Bigg) \end{aligned} \tag{42}$$

in terms of the generalised hypergeometric functions defined by [2]

$${}_pF_q\left(\begin{matrix}a_1, \ldots, a_p\\ b_1, \ldots, b_q\end{matrix}; \tilde{z}\right) = \sum_{n=0}^{\infty} \frac{(a_1)_n \ldots (a_p)_n}{(b_1)_n \ldots (b_q)_n}\frac{\tilde{z}^n}{n!}. \tag{43}$$

Equation (41) provides the sought-after special function representation of the joint generating function. In the following section, we focus on specific one-dimensional sections of the joint generating function that give the generating functions of the three marginal distributions.

6 Marginal distributions

In this section, we use the analytic formula (42) for the generating function to determine the marginal active and inactive mRNA, and protein distributions. To do so, we first set $y = z = 1$ in (42) and obtain

$$G(x) = G(x,1,1) = \exp(b_0(x-1))$$

for the marginal active mRNA distribution. Similarly, setting $x = z = 1$ in (42) yields the marginal inactive mRNA distribution

$$G(y) = G(1,y,1) = \exp\left(\frac{\alpha b_0}{\tilde{\gamma_1}+\beta}(y-1)\right).$$

Finally, we set $x = y = 1$ in (42) and get the marginal protein generating function $G(z)$ as

$$G(z) = G(1,1,z) = \exp(\psi(z)),$$

where ψ is given by

$$\psi(z) = \frac{b_0\lambda_2}{\gamma_2}\int_1^z {}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right)ds. \tag{44}$$

In order to obtain the marginal protein distribution, we exploit its generating function

$$p_{\cdot,\cdot,n} = \left.\frac{\mathrm{D}^n(G(z))}{n!}\right|_{z=0}, \tag{45}$$

where D stands for the differential operator d/dz and $p^{st}_{\cdot,\cdot,z}$ gives the probability of having z protein molecules and any number of active and inactive amount of mRNA. The first derivative of the composite function $G(z)$ in (45) is obtained by chain rule as

$$\frac{\mathrm{d}G(z)}{\mathrm{d}z} = G(z)\frac{\mathrm{d}\psi(z)}{\mathrm{d}z}. \tag{46}$$

For the n-th derivative, we evaluate the $(n-1)$th derivative of (46) according to the Leibniz rule, thus we have

$$\mathrm{D}^n(G(z)) = \sum_{i=0}^{n-1}\binom{n-1}{i}\mathrm{D}^i(G(z))\mathrm{D}^{n-i}(\psi(z)). \tag{47}$$

Next, we determine the rth–r is an arbitrary positive integer–derivative of the function $\psi(z)$, which is given by

$$\mathrm{D}^r(\psi(z)) = b_0\left(\frac{\lambda_2}{\gamma_2}\right)^r\frac{(r-1)!(1+\tau)_{r-1}}{(1+r_1)_{r-1}(1+r_2)_{r-1}}{}_2F_2\left(\begin{matrix}r, \tau+r\\ r_1+r, r_2+r\end{matrix}; \frac{\lambda_2}{\gamma_2}(z-1)\right), \tag{48}$$

in which we used the formula

$$\frac{\mathrm{d}^s}{\mathrm{d}\tilde{z}^s}{}_pF_q\left(\begin{matrix}a_1, \ldots, a_p\\ b_1, \ldots, b_q\end{matrix}; \tilde{z}\right) = \frac{\prod_{i=1}^p(a_i)_s}{\prod_{j=1}^q(b_j)_s}{}_pF_q\left(\begin{matrix}a_1+s, \ldots, a_p+s\\ b_1+s, \ldots, b_q+s\end{matrix}; \tilde{z}\right)$$

for the s-th derivative of the generalised hypergeometric function ${}_pF_q$. Inserting the derivatives in (48) into (47), taking $z = 0$, and rearranging the resulting equation according to (45) gives the formula for the marginal protein probabilities

$$\begin{aligned} p_{\cdot,\cdot,n} =& \frac{b_0\lambda_2}{n\gamma_2}\sum_{i=0}^{n-1}\left(\frac{\lambda_2}{\gamma_2}\right)^{n-i-1}\frac{(1+\tau)_{n-i-1}}{(1+r_1)_{n-i-1}(1+r_2)_{n-i-1}}\\ &\times {}_2F_2\left(\begin{matrix}n-i, \tau+n-i\\ n-i+r_1, n-i+r_2\end{matrix}; -\frac{\lambda_2}{\gamma_2}\right)p_{\cdot,\cdot,i}, \end{aligned} \tag{49}$$

where the first term of the series is given by

$$p_{\cdot,\cdot,0} = G(0) = \exp\left(-\frac{b_0\lambda_2}{\gamma_2}\int_0^1 {}_2F_2\left(\begin{matrix}1, 1+\tau\\ 1+r_1, 1+r_2\end{matrix}; \frac{\lambda_2}{\gamma_2}(s-1)\right)ds\right). \tag{50}$$

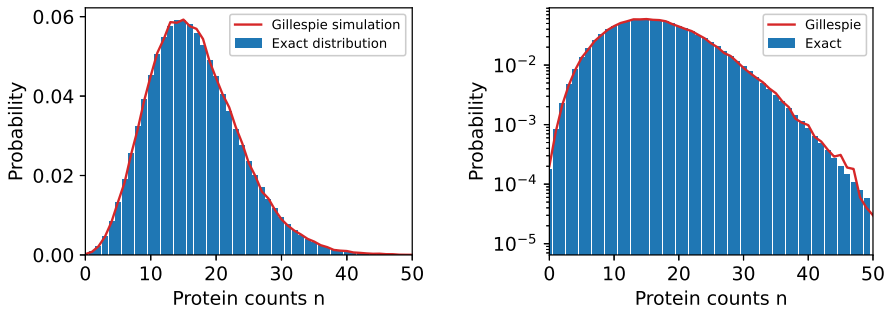

Fig. 2: *Left:* Comparison of the probability mass function (49) of the marginal protein distribution and the probability calculated by Gillespie's stochastic simulation algorithm (the solid line). *Right:* A logarithmic scale plot of the probability, out of 10^5 repeats, obtained by the two approaches. *Parameter values:* The kinetic parameters are: $\lambda_1 = 5$, $\alpha = \gamma_1 = \beta = \tilde{\gamma}_1 = \gamma_2 = 1$, $\lambda_2 = 5$.

In order to calculate and compare the marginal protein probabilities (49) with those obtained by stochastic simulations based on Gillespie's algorithm, we implement the recursive formula (49) in a high-level programming language, Python, together with using its numerical computing library NumPy and plotting library Matplotlib. The probabilities in (49) are calculated iteratively starting from its first term given by (50) up to $n = 50$. In Figure 2, the right panel compares the theoretical probability distribution (49) (blue bars) with the one obtained using stochastic simulations (solid line) at the timepoint $t = 100$, while the left panel shows the same comparison but on a logarithmic scale. The number of Gillespie iterations was set to 10^5 in the Python package GillesPy2 [1]. The initial number of active and inactive mRNA and protein was set to 5. A Python routine `mpmath.hyp2f2` used to calculate the generalised hypergeometric function ${}_2F_2$ in (49)–(50).

7 Conclusion

In this paper, we analysed a formulation of the two-stage model for gene expression that extends the classical version [4, 24] by an mRNA inactivation loop. The principal results of our analysis are the characterisation of the mean and noise behaviour, as well as the underlying probability distribution. The principal tool is the factorial cumulant generating function and the factorial cumulant expansion.

The incorporation of the mRNA inactivation loop into the classical two-stage model for gene expression reduces the protein noise. However, in order for the reduction be substantial, several restrictions on the parameter rates have to be in place. In particular, the protein cannot be too stable or unstable, but its stability

has to be optimally chosen. The resulting optimal value of protein stability is typically unrealistically low (lower than mRNA stability, in particular). In order to obtain an optimal stability that is greater than mRNA stability, one has to assume that inactivation protects the mRNA from degradation and activation is slow. Thus, our noise analysis points towards a potential role of the mRNA inactivation loop in gene expression noise control; at the same time, it delineates the limits of its application.

In addition to the noise analysis, we provide a comprehensive classification of the underlying probability distributions. Unsurprisingly, the distributions of the active/inactive mRNA are Poissonian. On the other hand, the protein distribution is highly non-trivial, and is characterised in terms of the generalised hypergeometric series. The characterisation is used to derive a recursive expression for the protein probability mass function. The recursive formula is found to be consistent with kinetic Monte-Carlo simulation (by means of the Gillespie direct method).

In summary, the paper provides a systematic mathematical analysis of an mRNA–protein model for gene expression extended by an inactive mRNA species, and hints at possible functional roles of mRNA inactivation loop in the control of low copy number gene-expression noise.

Bibliography

[1] Abel, J.H., Drawert, B., Hellander, A., Petzold, L.R.: Gillespy: A python package for stochastic model building and simulation. IEEE Life Sci. Lett. **2**, 35–38 (2016).

[2] Abramowitz, M., Stegun, I.A., Romer, R.H.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. American Journal of Physics **56**(10), 958–958 (1988).

[3] Bartman, C.R., Hamagami, N., Keller, C.A., Giardine, B., Hardison, R.C., Blobel, G.A., Raj, A.: Transcriptional burst initiation and polymerase pause release are key control points of transcriptional regulation. Molecular cell **73**(3), 519–532 (2019).

[4] Bokes, P., King, J.R., Wood, A.T.A., Loose, M.: Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. J. Math. Biol. **64**(5), 829–854 (2012).

[5] Bokes, P., King, J.R., Wood, A.T.A., Loose, M.: Transcriptional bursting diversifies the behaviour of a toggle switch: hybrid simulation of stochastic gene expression. Bulletin of mathematical biology **75**(2), 351–371 (2013).

[6] Dacheux, E., Malys, N., Meng, X., Ramachandran, V., Mendes, P., McCarthy, J.E.G.: Translation initiation events on structured eukaryotic mRNAs generate gene expression noise. Nucleic Acids Research **45**(11), 6981–6992 (2017).

[7] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S.: Stochastic gene expression in a single cell. Science **297**(5584), 1183–1186 (2002).

[8] Gardiner, C.: Stochastic Methods: A Handbook for the Natural and Social Sciences. Springer Series in Synergetics, Springer-Verlag, Berlin Heidelberg, 4th ed. (2009).

[9] Johnson, N.L., Kemp, A.W., Kotz, S.: Univariate Discrete Distributions. John Wiley & Sons, 3rd ed. (2005).

[10] Kurasov, P., Mugnolo, D., Wolf, V.: Analytic solutions for stochastic hybrid models of gene regulatory networks. J. Math. Biol. **82**(1), 1–29 (2021).

[11] Li, J., Ge, H., Zhang, Y.: Fluctuating-rate model with multiple gene states. J. Math. Biol. **81**(4), 1099–1141 (2020).

[12] Munsky, B., Neuert, G., van Oudenaarden, A.: Using gene expression noise to understand gene regulation. Science **336**(6078), 183–187 (2012).

[13] Paulsson, J.: Summing up the noise in gene networks. Nature **427**(6973), 415–418 (2004).

[14] Pendar, H., Platini, T., Kulkarni, R.V.: Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes. Physical Review E **87**(4), 042720 (2013).

[15] Raser, J.M., O'Shea, E.K.: Noise in Gene Expression: Origins, Consequences, and Control. Science **309**(5743), 2010–2013 (2005).

[16] Sanchez, A., Choubey, S., Kondev, J.: Regulation of noise in gene expression. Annual Review of Biophysics **42**, 469–491 (2013).

[17] Schnoerr, D., Sanguinetti, G., Grima, R.: Approximation and inference methods for stochastic biochemical kinetics–a tutorial review. Journal of Physics A: Mathematical and Theoretical **50**(9), 093001 (2017).
[18] Shahrezaei, V., Swain, P.S.: Analytical distributions for stochastic gene expression. P. Natl. Acad. Sci. USA (2008).
[19] Singh, A., Hespanha, J.P.: Approximate Moment Dynamics for Chemically Reacting Systems. IEEE Transactions on Automatic Control **56**(2), 414–418 (2011).
[20] Singh, A., Bokes, P.: Consequences of mRNA Transport on Stochastic Variability in Protein Levels. Biophysical Journal **103**(5), 1087–1096 (2012).
[21] Singh, A., Hespanha, J.P.: Stochastic hybrid systems for studying biochemical processes. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **368**(1930), 4995–5011 (2010).
[22] Soltani, M., Vargas-Garcia, C.A., Singh, A.: Conditional moment closure schemes for studying stochastic dynamics of genetic circuits. IEEE transactions on biomedical circuits and systems **9**(4), 518–526 (2015).
[23] Swain, P.S., Elowitz, M.B., Siggia, E.D.: Intrinsic and extrinsic contributions to stochasticity in gene expression. P. Natl. Acad. Sci. USA **99**(20), 12795–12800 (2002).
[24] Thattai, M., Oudenaarden, A.v.: Intrinsic noise in gene regulatory networks. P. Natl. Acad. Sci. USA **98**(15), 8614–8619 (2001).
[25] Thomas, P.: Intrinsic and extrinsic noise of gene expression in lineage trees. Scientific Reports **9**(1), 474 (2019).
[26] Zhou, T., Liu, T.: Quantitative analysis of gene expression systems. Quantitative Biology **3**(4), 168–181 (2015).