Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations

Tomáš Gedeon[1]
Department of Mathematical Sciences,
Montana State University, Bozeman MT, USA

Pavol Bokes
Department of Applied Mathematics and Statistics
Comenius University, Mlynská dolina, Bratislava 842 48, Slovakia

[1]Corresponding author. Address: Wilson Hall 2-233, Department of Mathematical Sciences, Montana State University, Bozeman, MT 59715, USA , Tel.: (406)994-5359, Fax: (406)994-1789

Abstract

Recent experimental results indicate that in single *E. coli* cells the fluctuations in mRNA level are uncorrelated with those of protein. On the other hand, a basic two-stage model for prokaryotic gene expression suggests that there ought to be a degree of correlation between the two. Therefore, it is important to investigate realistic modifications of the basic model that have the potential to reduce the theoretical level of the correlation. In this work we focus on translational delay, reporting that its introduction into the two-stage model reduces the cross-correlation between instantaneous mRNA and protein levels. Our results indicate that the experimentally observed sample correlation coefficient between mRNA and protein levels may increase if the protein measurements are shifted back in time by the value of the delay.
Key words: Single cell; mRNA-protein correlation; noise; delay

Introduction

Uncertainty and noise are present in all cellular signals [1]. The abundances of mRNA molecules and proteins, which can be measured at a single-cell level thanks to novel experimental techniques, have been demonstrated to fluctuate widely both in time and between different, even isogenic, cells [2]. These fluctuations have been measured with single-molecule precision, first for a limited number of genes [3, 4] and later on a whole-genome scale [5].

The progress in experimental techniques has been accompanied by the development of theoretical models for stochastic gene expression. Mathematical modeling provides a framework to quantify the essential features of the biophysical mechanisms involved in gene expression, leading to quantitative predictions which can then be compared to experimental measurements. Such comparisons play a crucial role in identifying the origins of noise and its implications in genetic regulatory circuitry [2].

Taniguchi *et al.* [5] quantified with single-molecule sensitivity the expression of a large portion of the proteome and transcriptome in individual *E. coli* cells. Among other results they reported that for individual genes the fluctuations in mRNA levels are uncorrelated with those of protein. This result is both important and surprising. It is important as it implies that little can be learnt about the deviation of the protein level from the mean value by making an mRNA measurement. Methods like fluorescence *in situ* hybridisation (FISH), cDNA chips [6] and mRNA sequencing [7] are widely available for measuring mRNA in single cells. Protein abundance measurements are harder to obtain, requiring either a fluorescent protein strain library, or a specifically designed antibody for each protein. Therefore it would be helpful if mRNA measurements can be used as a proxy for protein abundance. The lack of correlation between mRNA and protein fluctuations implies that such an approach has its limitations.

The observed lack of correlation is surprising because mRNA serves as a template for protein production: the more mRNA molecules present in the cell, the faster the production of proteins. The apparent paradox can in part be explained by short mRNA half-lives in prokaryotes [5]: short half-life implies fast turnover, making the mRNA copy number become rapidly independent of the amount of protein they gave rise to. Investigating whether the lack of correlation can be attributed solely to the difference between mRNA and protein half-lives, the experimental sample correlation coefficient was compared to theoretical predictions based on simple mathematical descriptions of stochastic gene expression [5, SI]. First, the authors of [5] used what is sometimes referred to as the two-stage model [8, 9], in which

transcription and translation, represented by elementary chemical reactions, are complemented by elementary degradation mechanisms. For typical parameter values for prokaryotic gene expression, the theoretical correlation coefficient between the mRNA and protein levels was not small enough to fully explain the apparent lack of correlation in experimental results. In search of mechanisms that may bring the theoretical value of the correlation coefficient down, various sources of extrinsic noise were introduced in the two-stage model and their impact on the theoretical value of the coefficient was examined [5, SI]. Extrinsic noise that affects translation rate was identified as one that has the potential to drive this value down.

The primary aim of this paper is to quantify the effect of a qualitatively different mechanism, delay in protein synthesis, on the correlation between the instantaneous protein and mRNA levels. The process of translation introduces a delay in protein production, during which the statistics of mRNA copy number decouple from the amount of protein whose translation they initiated. We show that physiological translational delays have a profound impact on correlations between instantaneous mRNA and protein levels.

Delayed transcription and translation have been studied by several authors. In systems with negative feedback, they have been implicated in driving biochemical oscillations responsible for the upkeep of cellular rhythms [10]. Both deterministic [11] and stochastic [12, 13] modelling approaches have been applied to investigate the phenomenon. The two-stage model for stochastic constitutive gene expression, which we introduce below, has recently been extended by delays in [14, 15].

Results

We consider the two-stage model [9] for gene expression extended by a delay in translation,

$$\begin{aligned} \emptyset &\xrightarrow{k_1} \mathrm{mRNA}, \\ \mathrm{mRNA} &\xrightarrow{\gamma_1} \emptyset, \\ \mathrm{mRNA}(t) &\xrightarrow{k_2} \mathrm{mRNA}(t) + \mathrm{protein}(t+d), \\ \mathrm{protein} &\xrightarrow{\gamma_2} \emptyset, \end{aligned} \tag{1}$$

in which k_1 is the rate of transcription, k_2 is the translation rate constant, and γ_1 and γ_2 are the mRNA and protein degradation rate constants. Translation is a delayed reaction which, having been initiated at time t, takes d units of time to be completed.

Using the model for delayed protein synthesis, we give an explicit formula for the correlation coefficient between the level of mRNA and that of protein as

$$\rho = e^{-\gamma_1 d}\rho_{\text{const}}, \tag{2}$$

where ρ_{const} is the correlation between mRNA and protein abundances in the absence of translational delay [5, SI]. This formula shows that while the mRNA degradation rate γ_1 is certainly a key factor, it is the synergy of the mRNA turnover rate and the translational delay d that causes the lack of correlation between mRNA and protein.

Modeling gene expression from a typical constitutive reporter, Taniguchi *et. al.* [5] assumed the mRNA lifetime $1/\gamma_1$ of 5 min, protein lifetime $1/\gamma_2$ of 180 minutes, and protein production within the range $k_2 = 0.6$–60 per minute, obtaining $\rho_{\text{const}} = 0.13$–$0.16$.

By (2), the correlation coefficient ρ_{const} is reduced if a delay in protein synthesis is incorporated into the model. The elongation of the growing amino acid chain and post-translational modification both contribute toward the delay. The average gene length in *E. coli* is 1100 bases [16], encoding a chain of 366 amino acids (aa); the yellow fluorescent protein (YFP) contains 238 aa, so a chimeric reporter protein, such as that used in [5], will contain approximately 600 aa. With an elongation rate of 12 aa/s [17, 18], we can expect a delay of 50 s incurred in the process of elongation. Conservatively, we take 30 s as an estimate of the contribution of elongation toward the delay (see the Discussion for more detail).

An amino acid chain becomes a mature protein during the process of post-translational modification. The delay due to this process can be substantial and has previously been implicated in driving slow circadian oscillations [19]. In the specific example of the chimeric reporter, the time required for YFP maturation is 7 min [20], implying an overall translational delay of $d = 7.5$ min, which, by the formula (2), implies a reduced correlation coefficient

$$\rho = e^{-1.5}\rho_{\text{const}} = 0.22\rho_{\text{const}} = 0.029\text{–}0.035,$$

which is within the bounds 0.01 ± 0.03 determined experimentally [5]. Note that our estimate of the translational delay $d = 7.5$ min is conservative, and a larger d will make observed correlations smaller. For instance, $d = 10$ min will produce ρ in the range $\rho = 0.018$–0.021. We conclude that the inclusion of translational delay explains the lack of correlation between mRNA and protein abundance in single cells.

Simulations

One of the most appealing features of the model (1) for delayed protein synthesis is its simplicity, which allows for exact analysis and leads to explicit formulae such as (2). The main reason why the model can be solved explicitly is that the protein trajectory of the delayed model is obtained merely by shifting that of the model without delay forward in time by the value d, while the mRNA trajectory is unchanged by the delay (see Figure 1). A thorough justification of the time-shift approach and a discussion of its limitations is provided later on in the paper. The classical Gillespie algorithm can thus be used for generating the trajectories of the model (1).

In both examples presented in Figure 1 we used the mRNA lifetime of $1/\gamma_1 = 5$ min, while we estimated the protein lifetime by the value $1/\gamma_2 = 30$ min of the generation time of *E. coli* in the nutritionally rich LB medium [21]. We selected a translational delay $d = 12$ min, which is larger than our conservative estimate for the YFP system. These values have been chosen to illustrate and emphasize our main points, not to represent a most common, or median, situation. Nevertheless, these values are in biologically plausible ranges.

In order to illustrate the robustness of our conclusions with respect to changes in transcription and translation rate constants (and hence with respect to gene-to-gene variation in absolute mRNA and protein levels), we first consider a low-copy scenario with $k_1 = 0.04\text{min}^{-1}$ and $k_2 = 0.8\text{min}^{-1}$ (Figure 1a) and then a situation with a stronger transcription and translation, $k_1 = 0.24\text{min}^{-1}$ and $k_2 = \frac{10}{3}\text{min}^{-1}$ (Figure 1b). The former scenario represents a non-essential protein, or alternatively a transcription factor, while the latter can be thought of as an ubiquitous product of a housekeeping gene [5].

In both examples, the mRNA trajectory and the non-delayed protein trajectory (shown in grey) tend to peak at the same times, which is indicative of a degree of correlation. On the other hand, the delayed protein trajectory (shown in black) often achieves maxima at times when there is no mRNA molecule in the system, suggestive of little cross-correlation. This is confirmed by our formula (2), which implies that the correlation coefficient is reduced by the factor $\mathrm{e}^{-\gamma_1 d} = \mathrm{e}^{-2.4} \approx 0.09$ once the delay is introduced.

Our model predicts that if we shift the experimentally observed protein trajectory (the black one in Figure 1) backwards in time (down to the grey trajectory), the correlation with the mRNA level increases. Comparing the increased correlation level to the original would enable the experimenter to separate the individual effects of translational delay and fast mRNA turnover

on the overall lack of correlation. At this point we should add, that while it is possible to measure protein abundance in live cells as a function of time [5], a majority of existing methods for measuring mRNA abundance in individual cells require fixation of cell culture. To our knowledge, the experimental measurement of both the abundance of mRNA and proteins in single cells as a function of time has not been done. However, new experimental techniques based on RNA aptamers that bind fluorophores resembling the fluorophore in GFP [22] promise to be able to continuously monitor mRNA abundance in single live cells.

An alternative approach to visualising fluctuations in gene expression, complementary to plotting real-time dynamics, is to determine at a single time point the mRNA and protein levels across a large population of isogenic cells. The fraction of cells which contain m molecules of mRNA and n molecules of protein is approximately equal to the probability $p_{m,n}$ of observing these numbers at the given time in any of these identical cells. The presented model for delayed protein synthesis is exactly solvable, and we use the solution derived later on in the paper to determine the individual probabilities.

Figure 2 shows the resulting distributions for the two examples whose trajectories we examined above. For each example, we first show the distribution resulting in the absence of delay (which we obtain if the grey protein trajectories in Figure 1 are considered instead of the black ones), and then the distribution that follows if the delay is introduced. This gives us a total of four panels, each detailing the marginal mRNA and marginal protein distributions (the dark-shaded charts) and the protein distribution conditioned on a given number of mRNA molecules (the light-shaded charts). Both examples imply that without delay, the protein level is distinctly dependent on the number of mRNA observed: measuring mRNA abundance informs the observer, even if not conclusively, about the deviation of protein level from the population average. On the other hand, if the delay is included, the conditional protein distributions resemble the marginal protein distribution and depend very little on the number of mRNA molecules. The delayed model predicts that by making an mRNA measurement on a single cell, we cannot make inferences on the deviation of the protein level from the average.

Analysis of the model

In this section we provide the analytical background for the results in the previous sections. We will show that

1. The protein trajectory of the model with delay is obtained by shifting forward in time by the value d the protein trajectory of the model without delay.

2. The correlation coefficient ρ between mRNA and protein abundances in the delayed model is given by

$$\rho = \mathrm{e}^{-\gamma_1 d}\sqrt{\frac{k_2\gamma_2}{(k_2+\gamma_1+\gamma_2)(\gamma_1+\gamma_2)}}. \tag{3}$$

3. The steady-state distribution of mRNA and protein levels can be characterized by an exact analytic formula for its generating function.

Shifting protein dynamics

We show that given that if (a) the translational delay d is identical for all protein molecules and constant in time, (b) protein levels do not affect the frequency of the initiation events (i.e. there is no feedback in the system) and (c) incomplete protein molecules are not subject to decay, then the protein trajectory of the delayed model is obtained by shifting the protein trajectory of the non-delayed model forward in time by the value d of the translational delay. The mRNA trajectory $M(t)$ in the absence of autoregulatory feedback is the same regardless of the value of translational delay.

Due to translational delay, any protein molecule is produced d units of time later than it was initiated. Since there is no feedback, the translation initiation times in the delayed model are the same — in the sense of equal distribution — as the protein synthesis times in the model without delay. Thus, if we determine (stochastically generate) these time points, and then add to them the additional time lag, we obtain the synthesis time points for the delayed model.

During the period between translation initiation and the completion of synthesis, the protein acquires its primary, secondary and tertiary structure. It is only after these structures are defined that the protein is a mature molecule fully able to exert its biological function by interaction via specific reaction channels with other elements. We treat protein degradation as one of such reaction channels, thereby making an assumption that proteins are subject to decay only once their synthesis has been completed. Then the lifetime, i.e. the period between the synthesis completion and degradation, of any protein molecule, will be the same in the delayed model and the model without delay, implying that the absolute values of degradation times

in the delayed model are obtained by adding the delay constant d to the degradation times generated for the model without delay.

Thus, if t_i, where $i = 1, 2, \ldots$, represent the time points at which a protein molecule is synthesized or degraded — i.e. the time points when the protein number changes — in the model without the translational delay, then the values $t_i + d$ give the time points at which the change occurs in the delayed model. This implies that the entire protein trajectory of the delayed model is obtained by shifting that of the non-delayed model forward in time by the value of the translational delay, i.e.

$$N_d(t) = N(t - d), \tag{4}$$

where $N_d(t)$ and $N(t)$ are the protein trajectories of the delayed and non-delayed models, respectively.

The result (4) says that the protein dynamics is affected by translational delay only to the extent that everything happens d units of time later. The steady-state distribution of the protein level does not depend on the size of the delay (just as that of the mRNA level does not). However, the correlation between the steady-state mRNA and protein fluctuations is by (4) equal to the correlation between $M(t)$ and $N(t-d)$, which is dependent on the value of the time lag. The correlation coefficient between $M(t)$ and $N(t - d)$ is determined below, where we also characterize the joint distribution of these two stochastic quantities.

The main result of this section, equation (4), is hardly surprising; yet it is important to appreciate that the result holds only in the case of constitutive gene expression and for constant, deterministic delay. If an autoregulatory loop was included, the times of translation initiation could depend on the protein level, and hence on the delay parameter that affects the time when this level changes. Hence the simple time shift (4) would be inadequate.

If we considered the case of non-constant, stochastic, translational delay, then each initiation instance would be followed by a distinct time lag. In such case the shape of the trajectory of the delayed model will differ from that of the model without delay. Equation (4) would not be justified, and the distribution of the protein level would depend on the delay in a non-trivial fashion. In view of these complications we regard the simplicity of our modeling assumptions as a major advantage that enables us to obtain a simple yet revealing characterization (4).

The correlation coefficient

The stationary means of mRNA and protein counts in the model without delay are given, cf. [9], by

$$\langle M(t)\rangle = \frac{k_1}{\gamma_1}, \quad \langle N(t)\rangle = \frac{k_1 k_2}{\gamma_1\gamma_2},$$

and the stationary variances and the covariance, cf. [9], are

$$\begin{aligned}\mathrm{Var}(M(t)) &= \frac{k_1}{\gamma_1}, \quad \mathrm{Cov}(M(t),N(t)) = \frac{k_1 k_2}{\gamma_1(\gamma_1+\gamma_2)},\\ \mathrm{Var}(N(t)) &= \frac{k_1 k_2}{\gamma_1\gamma_2}\left(1+\frac{k_2}{\gamma_1+\gamma_2}\right).\end{aligned}$$

Our aim is to find the covariance, and subsequently the correlation coefficient, between $M(t)$ and $N(s)$, where $t \geq s$. By the previous section these are equal to the covariance and the correlation coefficient between the mRNA and protein levels in the two-stage model with translational delay $d = t - s$. Let

$$P(m,t;n,s) = \mathrm{Prob}(M(t)=m, N(s)=n), \quad t \geq s,$$

be the probability of having m mRNA molecules at time t and n protein molecules at an earlier time s. For a fixed value of s, the probability $P(m,t;n,s)$, as function of t, satisfies the master equation for mRNA dynamics,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P(m,t;n,s) = {}& k_1(P(m-1,t;n,s) - P(m,t;n,s))\\ &+ \gamma_1((m+1)P(m+1,t;n,s) - mP(m,t;n,s)),\end{aligned} \tag{5}$$

on the right-hand side of which the first term gives the probability mass transfer due to transcription, occurring with rate k_1, while the second term gives the transfer of probability due to mRNA degradation, which takes place with rate γ_1 per molecule.

The covariance between $M(t)$ and $N(s)$, where $t \geq s$, satisfies

$$\begin{aligned}\mathrm{Cov}(M(t),N(s)) &= \langle M(t)N(s)\rangle - \langle M(t)\rangle\langle N(s)\rangle\\ &= \sum_{m,n} mnP(m,t;n,s) - \frac{k_1^2 k_2}{\gamma_1^2\gamma_2}.\end{aligned} \tag{6}$$

Differentiating (6) with respect to t and subsequently using the master equation (5) yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Cov}(M(t),N(s)) &= \sum_{m,n} mn\frac{\mathrm{d}}{\mathrm{d}t}P(m,t;m,s)\\ &= \sum_{m,n}(k_1-\gamma_1 m)nP(m,t;n,s)\\ &= k_1\langle N(s)\rangle - \gamma_1\langle M(t)N(s)\rangle = -\gamma_1\mathrm{Cov}(M(t),N(s)).\end{aligned}$$

Therefore

$$\mathrm{Cov}(M(t),N(s)) = \mathrm{Cov}(M(s),N(s))\mathrm{e}^{-\gamma_1(t-s)} = \frac{k_1k_2\mathrm{e}^{-\gamma_1(t-s)}}{\gamma_1(\gamma_1+\gamma_2)},$$

and the correlation coefficient between $M(t)$ and $N(s)$ is given by

$$\begin{aligned}\rho(M(t),N(s)) &= \frac{\mathrm{Cov}(M(t),N(s))}{\sqrt{\mathrm{Var}(M(t))\mathrm{Var}(N(s))}}\\ &= \mathrm{e}^{-\gamma_1(t-s)}\sqrt{\frac{k_2\gamma_2}{(k_2+\gamma_1+\gamma_2)(\gamma_1+\gamma_2)}}.\end{aligned}$$

Setting $d = t - s$ in the above equation yields the desired correlation coefficient (3) between the mRNA and protein fluctuations in the two-stage gene expression model extended by delayed product synthesis.

The joint distribution

The steady state joint distribution of mRNA and protein levels, i.e. the joint distribution of $M(t)$ and $N(s)$, where $s = t - d$, is comprised of the probabilities $P(m,t;n,s)$, which were introduced in the previous section. If we are faced with a complex distribution, such as this one will turn out to be, it is often more tractable to specify instead of the individual probabilities an auxiliary function, known as the generating function

$$G(x,t;y,s) = \sum_{m,n} x^m y^n P(m,t;n,s). \tag{7}$$

In principle, any probability $P(m,t;n,s)$ can be recovered from a generating function by differentiating it m times with respect to x and n times with respect to y and then taking $x = y = 0$. In practice, however, this approach would be fraught with numerical inaccuracy, and a more subtle method, such as described in Appendix B, is required.

For the model with delay, the generating function is given by the formula (see Appendix A for derivation)

$$G(x,t;y,s) = \exp\left(\alpha\beta\int_1^y M(1,1+\lambda,\beta(s-1))\mathrm{d}s \right. \\ \left. + \alpha\delta(x-1)M(1,1+\lambda,\beta(y-1)) + \alpha(1-\delta)(x-1)\right), \tag{8}$$

where

$$\lambda = \frac{\gamma_1}{\gamma_2}, \quad \alpha = \frac{k_1}{\gamma_1}, \quad \beta = \frac{k_2}{\gamma_2}, \quad \delta = \mathrm{e}^{-\gamma_1(t-s)} = \mathrm{e}^{-\gamma_1 d}, \tag{9}$$

and $M(a,b,z)$ is Kummer's function [23]. Kummer's function has repeatedly appeared in exact characterisations of probability distributions arising from stochastic models for gene expression [24–28], and its implementation is provided by most platforms for numerical computing. Expressions such as (8) can be used for evaluation of the underlying probability distribution, here $P(m,t;n,s)$, by means of the discrete Fourier transform (see Appendix B for details).

Of the auxiliary parameters in (9), the ratio λ compares the protein half-life to that of mRNA, α gives the mean mRNA copy number and β is the ratio between the mean protein and mean mRNA numbers. The factor δ provides the correction due to translational delay to the cross-correlation between mRNA and protein (cf. Equation 2). Without translational delay we have $\delta = 1$, and the generating function (8) reduces to that of the original two-stage model [28]. In the opposite limit of very large delays, when $\delta \approx 0$, the generating function (8) factorises into a product of two terms, one of which is a function of x only (and corresponds to the Poisson distribution of mRNA levels), and the other depends only on y (and corresponds to the marginal protein distribution). This confirms our expectation that for large delays the fluctuations in mRNA and protein counts are statistically independent.

Discussion

In the paper we examined a two-stage model for constitutive gene expression extended by a deterministic delay in translation. The inclusion of the delay of a realistic length was shown to reduce the theoretical value of the mRNA–protein cross-correlation to experimentally observed levels.

The problem of finding the correlation coefficient in the delayed model was shown to be equivalent to the problem of determining (a part of) the

autocorrelation function for the model without delay. The autocorrelation function can be readily determined for any chemical system composed only of first-order reactions [29]; therefore the presented approach is applicable to a wide range of models for constitutive gene expression.

Finding exact characterizations of copy number distributions for first-order stochastic chemical systems is not as straightforward as determining means and variances of the reacting species, or covariances and correlations between them. Nevertheless, an exact characterization of the distribution is known for the two-stage model without delay and in this paper it was generalized to the delayed case. Other systems for which exact characterizations have been provided in the absence of delay could be treated similarly.

Our focus in this paper was exclusively on the translational delay. We claim that transcriptional delay has no effect on mRNA-protein correlations. Since for the purposes of mRNA-protein correlations the only role of mRNA is its ability to produce protein, the mRNA joins the ranks of *active* mRNA's at the time when a ribosome is able to bind it. In prokaryotes, nascent mRNAs can be translated, and so the mRNA is activated (for the purposes of protein production) after the first few bases at the $5'$ end of the mRNA molecule that code for the ribosome binding site (RBS) are transcribed [30]. In eukaryotes the mRNA is post-transcriptionally modified — capped, spliced and transported to the cytoplasm — before it can be translated. Only after this time the mRNA can be added to the pool of active mRNA's. Since we are interested in the correlation between the abundance of active mRNA and proteins produced from this mRNA after a translational delay d, the transcriptional delay plays no role. We can conceptually model the overall translational delay as a sum

$$d = d_{\text{mod}} + d_{\text{el}},$$

where d_{mod} is the delay incurred by post-translational modification and d_{el} is the total elongation time. These constituents will vary from protein to protein; d_{el} will not only depend on the length of the mRNA but also on frequency of ribosomal pausing while d_{mod} can range widely.

Note that the delay d is most likely different from the delay Δ which results from experimental measurements. Delay Δ is the time by which one would have to shift the protein abundance time series backward (or the mRNA abundance time series forward) to recover the maximal correlation between active mRNA abundance and the abundance of protein. The delay Δ can be modeled as

$$\Delta = d + d_{prot} - d_{mRNA},$$

where d_{prot} is the additional time between protein maturation and the production of the protein-detecting signal, and d_{mRNA} is the time between mRNA activation (in the above sense) and the production of the mRNA-detecting signal. If the protein is detected by a gfp signal, then d_{prot} is the additional time it takes to fold the gfp protein after the primary protein had been folded. Note that d_{prot} can be positive or a negative number. The same comment applies to d_{mRNA}. If the mRNA is detected before mRNA is activated (for example by the MS2 system) then d_{mRNA} is negative.

While the precise number depends on the protein in question, delay in gene expression is a well-documented phenomenon with a number of implications for the regulation of cellular function. We conclude that a potential for reducing the mRNA–protein fluctuations belongs to that list.

Appendix A

We provide details regarding the derivation of the formula (8) for the generating function of the joint mRNA and protein number distribution in the delayed model. Multiplying the master equation (5) by $x^m y^n$ and summing over m and n, we find that the generating function satisfies a linear partial differential equation of the first order,

$$\frac{\partial G}{\partial t} + \gamma_1(x-1)\frac{\partial G}{\partial x} = k_1(x-1)G,$$

which can be solved by the method of characteristics, yielding solution

$$G(x,t;y,s) = G(1+\mathrm{e}^{-\gamma_1(t-s)}(x-1),s;y,s)\mathrm{e}^{\alpha(1-\mathrm{e}^{-\gamma_1(t-s)})(x-1)}. \qquad \text{(A1)}$$

This expression gives the joint generating function of $M(t)$ and $N(s)$, where $t \geq s$, in terms of the joint generating function of $M(s)$ and $N(s)$ (both taken at the same time-point). The latter can be found in [28]. Substituting the result therein into (A1), we arrive at (8).

Appendix B

We present a method for determining from the generating function (8) the stationary probability distribution $P(m,t;n,s=t-d)$, which we refer to as $p_{m,n}$ throughout this appendix. The probabilities $p_{m,n}$ can be found by expanding the generating function (8) into a power series in x and y; in what follows we show how this can be done numerically. While this method has

already been applied in [28] for the model without translational delay, in the exposition below we provide more detail on the nature of the numerical error (aliasing) incurred by the method than was given in [28].

We take two, typically large, positive integers M and N and consider the following values of the generating function,

$$g_{k,l} = G(\mathrm{e}^{\frac{2\pi \mathrm{i} k}{M}}, t; \mathrm{e}^{\frac{2\pi \mathrm{i} l}{N}}, s), \quad k = 0, \ldots, M-1, \quad l = 0, \ldots, N-1,$$

where i is the imaginary unit. The terms $g_{k,l}$ can be computed from (8) using an implementation of Kummer's function and a suitable integration routine. By (7), we have

$$g_{k,l} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} p_{m,n} \mathrm{e}^{2\pi \mathrm{i}(\frac{mk}{M} + \frac{nl}{N})} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \hat{p}_{m,n} \mathrm{e}^{2\pi \mathrm{i}(\frac{mk}{M} + \frac{nl}{N})}, \tag{B1}$$

where

$$\hat{p}_{m,n} = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} p_{m+kM,n+lN}, \quad m = 0, .., M-1, \quad n = 0, .., N-1,$$

is an "aliased" distribution. By (B1), the matrix $\hat{p}_{m,n}$ is the 2-dimensional discrete Fourier transform (DFT) of the matrix $g_{k,l}$. The DFT can be computed efficiently using the fast Fourier transform algorithm [31]. If M and N are sufficiently large, then we can approximate $p_{m,n}$ by the aliased values $\hat{p}_{m,n}$ for $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$ to obtain the desired stationary probability distribution of mRNA and protein counts in the model with translational delay.

The work of TG was partially supported by National Science grant DMS-0818785, National Science CMMI grant 0849433 and National Institutes of Health R01 grant 1R01AG040020-01. PB gratefully acknowledges the support of the Slovak Research and Development Agency (contract no. APVV-0134-10).

References

1. Elowitz, M., A. Levine, E. Siggia, and P. Swain, 2002. Stochastic gene expression in a single cell. *Science* 297:1183–6.

2. Larson, D., R. Singer, and D. Zenklusen, 2009. A single molecule view of gene expression. *Trends Cell Biol.* 19:630–7.

3. Cai, L., N. Friedman, and X. Xie, 2006. Stochastic protein expression in individual cells at the single molecule level. *Nature* 440:358–62.

4. Yu, J., J. Xiao, X. Ren, K. Lao, and X. Xie, 2006. Probing gene expression in live cells, one protein molecule at a time. *Science* 311:1600–3.

5. Taniguchi, Y., P. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. Xie, 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329:533–8.

6. Tietjen, I., J. Rihel, Y. Cao, G. Koentges, L. Zakhary, and K. Dulac, 2003. Single-cell transcriptional analysis of neuronal progenitors. *Neuron* 38:161–5.

7. Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. Tuch, A. Siddiqui, K. Lao, and S. M. A., 2009. Single-cell transcriptional analysis of neuronal progenitors. *Nat. Methods* 6:377.

8. Shahrezaei, V., and P. Swain, 2008. Analytical distributions for stochastic gene expression. *P. Natl. Acad. Sci. USA* 105:17256.

9. Thattai, M., and A. van Oudenaarden, 2001. Intrinsic noise in gene regulatory networks. *P. Natl. Acad. Sci. USA* 98:151588598.

10. Novák, B., and J. Tyson, 2008. Design principles of biochemical oscillators. *Nat. Rev. Mol. Cell Biol.* 9:981–91.

11. Monk, N., 2003. Oscillatory expression of Hes1, p53, and NF-κB driven by transcriptional time delays. *Curr. Biol.* 13:1409–13.

12. Barrio, M., K. Burrage, A. Leier, and T. Tian, 2006. Oscillatory regulation of Hes1: discrete stochastic delay modelling and simulation. *PLoS Comput. Biol.* 2:e117.

13. Galla, T., 2009. Intrinsic fluctuations in stochastic delay systems: Theoretical description and application to a simple model of gene regulation. *Phys. Rev. E* 80:021909.

14. Lafuerza, L., and R. Toral, 2011. Role of delay in the stochastic creation process. *Phys. Rev. E* 84:021128.

15. Jia, T., and R. Kulkarni, 2011. Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Phys. Rev. Lett.* 106:58102.

16. Sundararaj, S., A. Guo, B. Habibi-Nazhad, P. Rouani, M.and Stothard, M. Ellison, and D. Wishart, 2004. The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res.* 32 (Database issue).

17. Young, R., and H. Bremer, 1976. Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. *Biochem. J.* 160:185–94.

18. Roussel, M., and R. Zhu, 2006. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys. Biol.* 3:274–84.

19. Gallego, M., and D. Virshup, 2007. Post-translational modifications regulate the ticking of the circadian clock. *Nat. Rev. Mol. Cell Biol.* 8:139–48.

20. Xiao, J., J. Elf, G. Li, J. Yu, and X. Xie, 2008. Imaging Gene Expression in Living Cells at the Single-Molecule Level. *In* Single Molecule Techniques A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

21. Bernstein, J., A. Khodursky, P. Lin, S. Lin-Chao, and S. Cohen, 2002. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *P. Natl. Acad. Sci. USA* 99:9697–702.

22. Paige, J., K. Wu, and S. Jaffrey, 2011. RNA mimics of green fluorescent protein. *Science* 333:642–6.

23. Abramowitz, M., and I. Stegun, 1972. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards, Washington, D.C.

24. Peccoud, J., and B. Ycart, 1995. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* 48:222–34.

25. Hornos, J., D. Schultz, G. Innocentini, J. Wang, A. Walczak, J. Onuchic, and P. Wolynes, 2005. Self-regulating gene: an exact solution. *Phys. Rev. E* 72:051907.

26. Innocentini, G., and J. Hornos, 2007. Modeling stochastic gene expression under repression. *J. Math. Biol.* 55:413–31.

27. Raj, A., C. Peskin, D. Tranchina, D. Vargas, and S. Tyagi, 2006. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4:e309.

28. Bokes, P., J. King, A. Wood, and M. Loose, 2011. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *J. Math. Biol.* DOI: 10.1007/s00285–011–0433–5.

29. Lestas, I., J. Paulsson, N. Ross, and G. Vinnicombe, 2008. Noise in gene regulatory networks. *IEEE T. Circuits-I* 53:189–200.

30. Kierzek, A., J. Zaim, and P. Zielenkiewicz, 2001. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J. Biol. Chem.* 276:8165–72.

31. Cooley, J., P. Lewis, and P. Welch, 1970. The fast Fourier transform algorithm: Programming considerations in the calculation of sine, cosine and Laplace transforms. *J. Sound Vib.* 12:315–37.

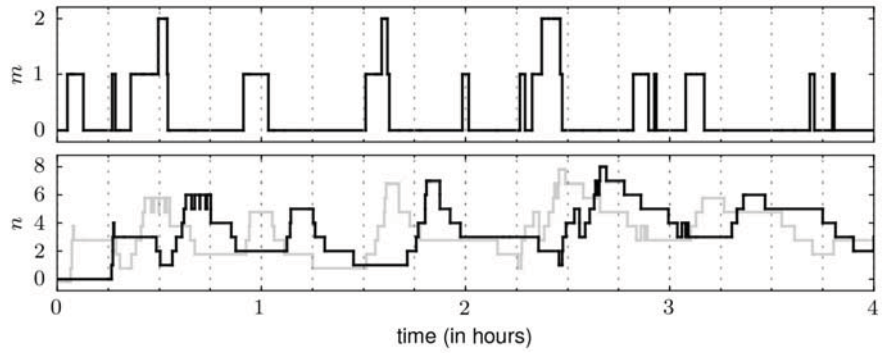
Figure Legends

Figure 1.

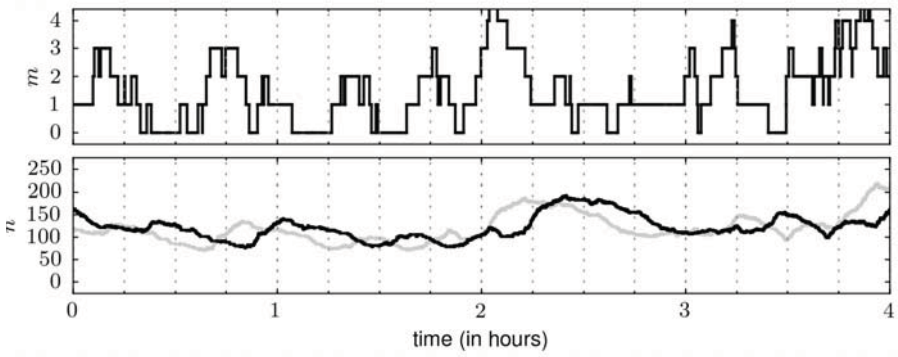
Gillespie simulations of the two-stage model with delayed translation for (a) a low-copy or (b) an ubiquitous protein. The parameter values are detailed in the main text. For each case we show sample paths of mRNA abundance (top), protein abundance assuming instantaneous production (bottom, in grey), and protein abundance resulting if translational delay is included (bottom, in black).

Figure 2.

The mRNA and protein distributions for a low copy and an ubiquitous protein with or without translational delay. The parameter values for (a) and (b) are those that were previously used for the trajectory of Figure 1a, while panels (c) and (d) share parameter values with Figure 1b. In each panel, the marginal mRNA distribution, $p_{m,\cdot} = \sum_{n=0}^{\infty} p_{m,n}$, and the marginal protein distribution, $p_{\cdot,n} = \sum_{m=0}^{\infty} p_{m,n}$, flank the conditional protein distributions $p_{n|m} = p_{m,n}/p_{m,\cdot}$.

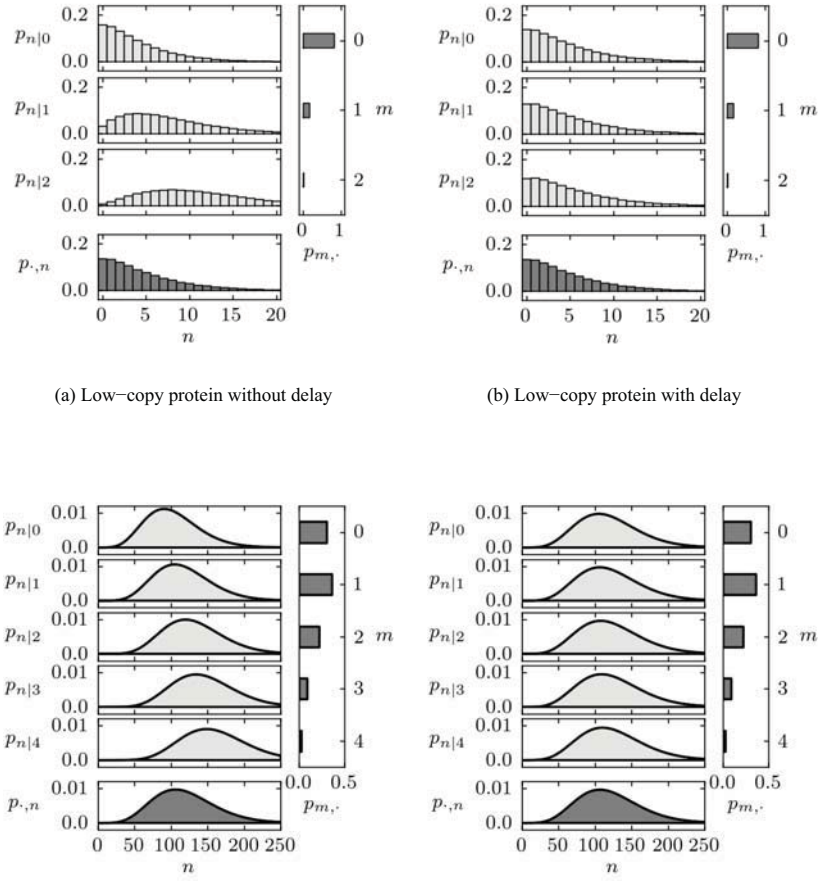


(a) Trajectories of a low−copy protein



(b) Trajectories of an ubiquitous protein

Figure 1



(a) Low−copy protein without delay

(b) Low−copy protein with delay

(c) Ubiquitous protein without delay

(d) Ubiquitous protein with delay

Figure 2