Mathematical models of gene expression

Philippe Robert

INRIA de Paris
2 rue Simone Iff, 75589 Paris, France
e-mail: Philippe.Robert@inria.fr

Abstract: In this paper we analyze the equilibrium properties of a large class of stochastic processes describing the fundamental biological process within bacterial cells, *the production process of proteins*. Stochastic models classically used in this context to describe the time evolution of the numbers of mRNAs and proteins are presented and discussed. An extension of these models, which includes elongation phases of mRNAs and proteins, is introduced. A convergence result to equilibrium for the process associated to the number of proteins and mRNAs is proved and a representation of this equilibrium as a functional of a Poisson process in an extended state space is obtained. Explicit expressions for the first two moments of the number of mRNAs and proteins at equilibrium are derived, generalizing some classical formulas. Approximations used in the biological literature for the equilibrium distribution of the number of proteins are discussed and investigated in the light of these results. Several convergence results for the distribution of the number of proteins at equilibrium are in particular obtained under different scaling assumptions.

Contents

1. Introduction

The *gene expression* is the process by which the genetic information is synthe-sized into a functional product, the proteins. These macro-molecules are the crucial agents of functional properties of cells. They play an important role in most of the basic biological processes within the cell, either directly, or as a component of complex macro-molecules such as polymerases or ribosomes. The information flow from DNA genes to proteins is a fundamental process, common to all living organisms.

We analyze this fundamental process in the context of prokaryotic cells, like bacterial cells or archaeal cells. The *cytoplasm* of these cells is not as struc-tured as eukaryotic cells, like mammalian cells for example, so that most of the macro-molecules of these cells can potentially collide with each other. This key biological process can be, roughly, described as resulting of multiple encoun-ters/collisions of several types of macro-molecules of the cell: *polymerases* with DNA, *ribosomes* with *mRNAs*, or *proteins* with DNA, ... An additional feature of this process is that it is consuming an important fraction of energy resources of the cell, to build chains of amino-acids or chains of nucleotides in particular.

The fact that the cytoplasm of a bacterial cell is a disorganized medium has important implications on the internal dynamics of these organisms. Numer-ous events are triggered by random events associated to thermal noise. When the external conditions are favorable, these cells can nevertheless multiply via division at a steady pace. In each cell, around 4000 different types of proteins are produced, with different concentrations: from 5 elements per cell for some rare proteins to 100,000 per cell for some ribosomal proteins. Despite of the noisy environment, the protein production process can handle such exponential growth with these constraints. The cost of the translation phase, the last step of the protein production process, is estimated to account for 50% of the energy consumption of a rapidly growing bacterial cell, see Russell and Cook [34] for example.

An important question in this context is of understanding and quantifying the role of the parameters of the cell in the production of proteins when large fluctuations may occur, as in the case of abundant proteins. Given the high level of energy consumed by protein production, these large stochastic fluctuations should be, at least, "controlled" to avoid a shortage of resources for example. Note nevertheless that, outside "pure" fluctuations, random events involving

macro-molecules with low copy-numbers may play an important role in the time evolution of cells, in the case of change of environment of the cell for example so that the cells may adapt. Some components of the "noise" in the gene expression may thus have a functional role. See Eldar and Elowitz [10], Lestas et al. [22] and Norman et al. [28].

From the point of view of the design, the cell can be seen as a system where the processing power is expressed in terms of numerous chemical reactions occurring in a noisy environment. An important difference with classical, non-biological, systems, like computer networks, the possibility of *explicitly* storing information used to regulate the production process is not as clear as for these systems. See Burrill and Silver [7] on this topic. This is a really challenging aspect of these biological systems.

1.1. The protein production process

We give a quick, simplified, sketch of the steps involved in the production of proteins. See Watson et al. [41] for a much more detailed description of these complex processes.

— *Gene Activation/Deactivation*.
A gene is a contiguous section of the DNA associated to a functional property of the cell. When a gene is *active*, a macro-molecule of the cell, an *RNA polymerase*, one of the macro-molecules moving in the cytoplasm, can possibly bind to the DNA at the promoter, i.e. the beginning of the section of DNA for the gene. The binding to the gene is subject to the various (random) fluctuations of the medium. The gene may also be *inactive* and in this case no binding is possible. The reason is that some macro-molecules, like proteins or mRNAs, of the cell can "block", via chemical bonds, the gene so that polymerases cannot access it. This is one of the regulation mechanisms of the cell. When random perturbations break these bindings, the gene becomes again active.

— *Transcription*.
When an RNA polymerase is bound to an active gene, it starts to make a copy of this gene. The product which is a sequence of nucleotides is a *messenger RNA*, or *mRNA*. The time during which nucleotides are added sequentially is the *elongation* phase of the production of the mRNA. When the full sequence of nucleotides of the mRNA has been successively assembled, the mRNA is released in the cytoplasm. It has a finite lifetime, being degraded by other macro-molecules.

— *Translation*.
The step is achieved through another large macro-molecule: the *ribosome*. A ribosome is also moving within the cytoplasm. When it encounters an mRNA, it can also bind to it via chemical reactions. In this case it builds a chain of amino-acids using the mRNA as a template to produce a *protein*. This is the *elongation* phase of the production of the protein. As for mRNAs, proteins have also a finite lifetime. See Figure 1.
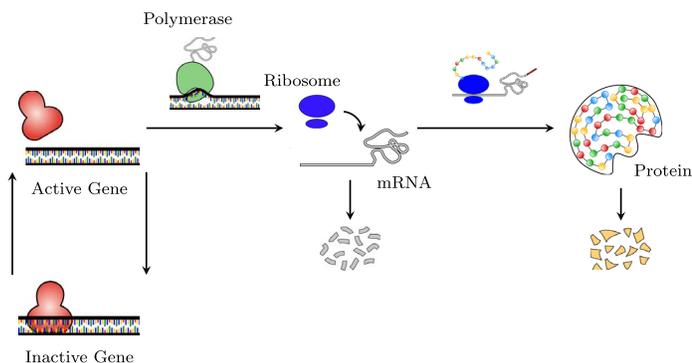
FIG 1. *A Simplified View of the Protein Production Process.*
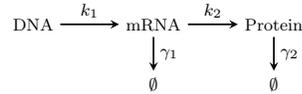
1.2. Mathematical models

The above description of the protein production process is clearly simplified. The processes associated to actions of polymerases and ribosomes are composed of several steps. The way ribosomes bind to mRNAs in particular, see Chapter 14 of Watson et al. [41] for example. Once a polymerase binds to the gene, the messenger RNA chain is built through a series of specific stages, in which the polymerase recruits one of the four nucleotides in accordance to the DNA template. Additionally, a dedicated proof reading mechanism takes place during this process. There is a similar description for the translation step. The binding events of polymerases to gene or of ribosomes to mRNAs are also due to a sequence of specific steps.

Because of the disorganized medium of the bacteria, the protein production process is a highly stochastic process. The randomness is partially due to the thermal excitation of the environment. It drives the diffusion of the main components, mRNAs and ribosomes within the cytoplasm and it also impacts the pairing of cellular components diffusing through the cytoplasm. It can cause the spontaneous rupture of such pairs, before either transcription or translation can start.

The problem is of understanding the mechanisms used by the cell to produce a large number of proteins with very different concentrations in such a random, "noisy", context. The main goals of a mathematical analysis in the biological literature are generally the following.

1. Estimate impact on variance of the number of proteins of assumptions on

   — Activation/Deactivation rates of the gene;

   — Transcription/Translation rates. Polymerases may bind more easily to some genes. A similar phenomenon holds for ribosomes and mRNAs. This is generally mathematically represented via a "rate" of binding: transcription rate for polymerases on gene and translation rate for ribosomes on mRNAs, k_1 and k_2 in the following;

— The distributions of lifetimes of mRNAs and proteins, the death rate of mRNAs, (resp., proteins), is denoted by γ_1, (resp., γ_2).

$$\text{DNA} \xrightarrow{k_1} \text{mRNA} \xrightarrow{k_2} \text{Protein}$$
$$\downarrow \gamma_1 \qquad\qquad \downarrow \gamma_2$$
$$\emptyset \qquad\qquad\qquad \emptyset$$

FIG 2. *Basic Chemical Reactions.*

2. Determine the parameters which achieve minimal variance of concentration of a type of protein with a fixed average concentration. When an analytical formula for the variance is available, one can determine, in principle, the parameters that minimize it, with the constraint of that its average is fixed. The general idea is of determining if the current parameters of the cell are adjusted, depending of the environment, to minimize this variance.
3. Estimate biological parameters for experiments. Analytical formulas for the average and the variance of the number of proteins can also be used to estimate some biological parameters. See Taniguchi [37].

The protein production process is a fundamental biological process, a process at the basis of "life", justifying the extensive efforts to understand and to quantify, via mathematical models, the role of its different components: polymerases, ribosomes, mRNAs and proteins. It also serves as a generic example of a typical biological process: a component of type A is produced with components of type B, but via the creation of another component of type C, which may also be used in the regulation of the process.

As a general remark, the mathematician should always keep in mind that she/he is studying an incredibly complex system, the molecular biology of gene expression. It has evolved over a time span of *four billion* years. A mathematical model may take into account only partial aspects of the sophisticated mechanisms into play. In this presentation we focus on the basic parameters of the transcription and translation phases.

Some history.
The basic principles of molecular biology, starting with Crick and Watson [42], Brenner et al. [5] and Jacob and Monod [15], have been discovered in the 1950–1960's. At that time, measuring concentrations of different types of macromolecules within the cell was not really possible with the experimental tools available. Efficient methods to estimate these concentrations, like the *fluorescence microscopy*, have been available much later at the beginning of the 1990's. This is probably the main motivation of mathematical models at the end the 1970's in the pioneering works Berg [2] and Rigney [32] to study the fluctuations of the concentration of proteins within the cell. Analytical expressions for the mean and the variance of the number of proteins in a cell have been derived for some simplified models. These results were later extended in the context of Markov chain theory. They are recalled in Section 2. The formulas obtained can

give some insight on the role of the parameters of the cell like the transcription rate of the gene or the average lifetime of some macro-molecules on these fluctuations. Interestingly, this is one of the few examples in applied mathematics when mathematical models have been used *before* measurements could be done. There is a huge literature on the stochastic analysis of this process using a large range of mathematical methods. For a detailed review of these works, see Paulsson [29]. We focus here on exact formulas that can be obtained for specific stochastic models.

State space of the production process.
In most mathematical analyses, the production process of proteins of a fixed gene is investigated. There is an implicit assumption in this case that the cell allocates a fixed fraction of its resources to the production of the mRNAs, proteins of this gene, independently of the production process of other proteins. With this hypotheses the production process of this type of protein can be analyzed in isolation of the other processes within the cell. This is a convenient given the complexity of the interactions. In this case, the production process is usually described as a three-dimensional Markov process $(I(t), M(t), P(t))$ where, for $t \geq 0$,

— $I(t) \in \{0, 1\}$ describes the state of the gene at time t, active or inactive;
— $M(t)$ is the number of mRNAs at this instant;
— $P(t)$ is the number of proteins.

For these models, the duration of elongation of mRNAs or of proteins is not taken into account. For example, the transition $M(t) \to M(t)+1$, the production of an mRNA, is, implicitly, associated to the binding of a polymerase to the gene. This amounts to neglect the elongation time of an mRNA. A glance at the numbers of Section A reveals that, in average, the time to build an mRNA of 3000 nucleotides is 35 seconds, which is not small compared to its lifetime, of the order of 2 minutes. For different types of mRNAs, the average length is 1000, see Figure 5 of Section A in the Appendix.

Another assumption, the price of the Markov property, is that the duration of the transitions have an exponential distribution. This hypothesis for the duration of time to have a binding polymerase-gene or ribosome-mRNA can be justified with the current parameters of the cell. The assumption is more questionable for lifetimes of mRNAs and, in particular, of proteins. The lifetimes of proteins are comparable to the duration of time between divisions of the cell. The interpretation of death/degradation is more difficult in such a case. See the discussion on the assumptions on the distribution of lifetimes.

The Markovian description of the protein production process gives the possibility of using classical results of Markov theory, for the convergence to equilibrium as well as for an analytical characterization of equilibrium points.

Numbers of copies or concentration?
The representation of the levels of proteins and mRNAs by *numbers*, i.e. integers, is convenient for the mathematical analysis, in particular to investigate the

stochastic fluctuations. For some aspects, it is nevertheless more natural to express the various quantities in terms of *concentrations* rather than numbers. For example, if $\mathcal{R}$, $\mathcal{M}$ and $\mathcal{P}$ denotes respectively the chemical species ribosomes, mRNAs and proteins, the production and destruction of a protein is expressed as, using notations from chemistry,

$$\begin{cases} \mathcal{M}+\mathcal{R} \longrightarrow \mathcal{P}+\mathcal{M}+\mathcal{R}, \\ \qquad \mathcal{P} \longrightarrow \emptyset. \end{cases}$$
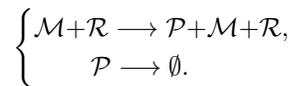
Let $[\mathcal{A}](t)$ denote the concentration of species $\mathcal{A}$, $\mathcal{A}\in\{\mathcal{R}, \mathcal{M}, \mathcal{P}\}$, at time $t{\geq}0$. This quantity is the ratio of the number of macro-molecules in the cell of type $\mathcal{A}$ to the volume of the cell at time t. The *law of mass action* of chemical physics gives the associated *Michaelis-Menten* kinetics equations for the concentration of the various macro-molecules, they are analytically expressed by a deterministic differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}[\mathcal{P}](t) = k_{\mathrm{on}}[\mathcal{M}](t)[\mathcal{R}](t) - \nu[\mathcal{P}](t).$$

See van Kampen [40] and Chapter 6 of Murray [26] for example. The parameter k_{on} is the rate at which a given ribosome binds to a given mRNA and ν is the exponential growth rate of the volume of the cell. The concentration of proteins decreases mostly because of volume growth. This is the *dilution effect*. In a Markovian context, see Section 2, the number of copies of a given type of protein is used rather than its concentration. In this case the parameter ν is interpreted as a death rate of proteins, which is less natural in some way. In this presentation, we will nevertheless use the discrete representation with numbers. It is more convenient to describe stochastic phenomena involving a finite, but not too large, number of macro-molecules, like for mRNAs. See Anderson and Kurtz [1] for an introduction on stochastic modeling of biological systems.

Organization of the paper.
Due to its historical importance, the analysis of the first moments of the equilibrium of Markovian models is presented in Section 2. Section 3 introduces a fundamental marked Poisson point process used to describe the whole protein production process. It is an extension based on the model of Fromion et al. [13]. A result of convergence to equilibrium is proved via a "coupling from the past" method. An important representation of the number of mRNAs and proteins in terms of the Poisson process is established. See Theorem 3.1.

It should be noted that we do not include negative/positive feedback mechanisms in the models of gene expression, i.e. when proteins/mRNAs, or some other macro-molecules, may lock/unlock the gene of some proteins under some circumstances. We are here mainly focused on the "historical" model of gene expression. The main technical reason is that the marked Poisson point process approach presented in this paper does not seem to work for this class models. They are generally investigated via scaling methods. See Mackey [25] for example.

In Section 4, (resp., Section 5), general formulas for the mean and variance of the number of mRNAs (resp., of proteins), at equilibrium are established, generalizing the classical formulas of Markovian models. An important part of the biological literature is devoted to the analysis of these first moments, this is the main motivation of these two sections. The results obtained are an extension, with somewhat simpler technical arguments, of results of Fromion et al. [13]. When the gene is always active, the equilibrium of the detailed description of the state of mRNAs is characterized by Proposition 4.4. A formula for the generating function of the equilibrium distribution of the number of proteins is established in Proposition 5.2. The joint distribution of the number of mRNAs and of proteins at equilibrium is also investigated in Section 5.

In Section 6 the various approximations used in the biological literature are revisited in the light of the results proved in the previous sections. They are formulated as new scaling results when one of the parameters goes to infinity: switching rate of the state of the gene, average lifetime of proteins, translation rate, ... generally under the constraint that the average number of proteins is fixed. Several convergence results for the equilibrium distribution of the number of proteins are obtained in this way. The methods rely on the results of Section 5, probabilistic arguments and some technical estimates. A *central limit theorem*, which seems to be new, is established in Proposition 6.1. A study on the impact of the elongation phase of proteins on the variability of the protein production process concludes this section. This topic is rarely addressed in mathematical models of the biological literature. It is shown that, under some statistical assumptions, increasing the variability of the elongation phase in a stochastic model may have the surprising effect of reducing the variance of the equilibrium distribution of the number of proteins. A consequence of this observation is that the choice of exponential distribution for the distribution of the duration of the elongation phase of proteins, as it would be natural in a Markovian context, may lead to underestimate the "real" variance.

The central result of this approach is Theorem 3.1, it gives an explicit representation of the number of proteins at equilibrium in terms of a functional of a marked Poisson point process. The main advantages are that

1. it is not necessary to solve invariant measure equations as in the Markovian approaches;
2. Formulas for Poisson processes, recalled in the Appendix, give additional insight on the equilibrium distribution of the number of proteins;
3. General distributions, instead of exponential distributions, for the duration of elongation times and lifetimes of mRNAs and proteins can be included in the model.

Mathematical notations and conventions.
Throughout this presentation, for the sake of clarity, we will frequently do the following abuse of notations. If X is some integrable random variable on $\mathbb{R}_+$

1. $X(\mathrm{d}x)$ will denote its distribution on $\mathbb{R}_+$;

2. (X_n) an i.i.d. sequence of random variables with this distribution;
3. F_X a random variable whose distribution has density

$$x \mapsto \frac{\mathbb{P}(X \geq x)}{\mathbb{E}(X)},$$

with respect to Lebesgue's measure on $\mathbb{R}_+$. We denote by $(F_{X,n})$ an i.i.d. sequence with this distribution.

Note that, with Fubini's Theorem, for $a \geq 0$,

$$\mathbb{E}\left((X-a)^+\right) = \int_a^{+\infty} \mathbb{P}(X \geq u) \, \mathrm{d}u = \mathbb{E}(X)\mathbb{P}(F_X \geq a), \tag{1.1}$$

where $b^+ = \max(b, 0)$, $b \in \mathbb{R}$. Additionally, a simple calculation of the Laplace transform shows that the relation $F_X \overset{\mathrm{dist}}{=} X$ holds if and only if the law of X is exponential.

Throughout the paper, we will use the following notation:

— For mRNAs, $L_1(\mathrm{d}w)$ is the distribution of their lifetimes and $E_1(\mathrm{d}v)$ the distribution of their elongation times.
— For proteins, $L_2(\mathrm{d}z)$ is the distribution of their lifetimes and $E_2(\mathrm{d}y)$ the distribution of their elongation times.

The index 1 refers to mRNAs and 2 to proteins. With our convention, items (1) and (2) above, (L_1) and (E_1), (resp. $(L_{1,n})$ and $(E_{1,n})$), denote random variables, (resp. i.i.d. sequences of random variables) associated to the distribution $L_1(\mathrm{d}w)$, (resp. $E_1(\mathrm{d}v)$). And similarly for proteins, for the random variables (L_2) and (E_2) and the i.i.d. sequences $(L_{2,n})$ and $(E_{2,n})$.

2. The classical Markovian three-step model

In this section, Markovian models of the protein production process are presented. Although these models have some limitations in terms of modeling this is an important class of models to study the stochastic fluctuations associated to gene expression. This is in fact the main stochastic model of gene expression used in the biological literature from the early works of Rigney [32], Berg [2] to more recent studies Thattai and van Oudenaarden [38], Shahrezaei and Swain [35]. See Paulsson [29] for a survey. See also Chapter 6 of Bressloff [6] and Chapter 4 of Mackey et al. [25]. These models are still popular in the biological literature.

A more general modeling of the stochasticity of gene expression is presented and discussed in Section 3. It is analyzed in the rest of the paper. In this section we will consider the time evolution of the number of mRNAs and of proteins associated to a gene G_0 in a given cell (bacterium). The statistical assumptions are the following:

Ph. Robert

State of Gene   mRNAs   Proteins
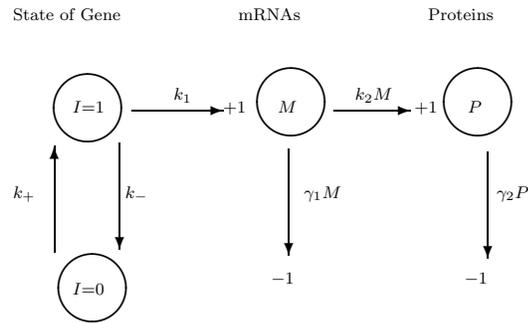
FIG 3. *Classical Three-Step Model for Protein Production.*

1. If the gene is inactive, in state 0 say, (resp., active, in state 1), it becomes active, (resp., inactive) after an exponentially distributed amount of time with parameter k_+ (resp., k_-).
2. If the gene is active, the mRNAs of G_0 are produced according to a Poisson process with parameter k_1 and their lifetimes are exponentially distributed with parameter γ_1.
3. During its lifetime an mRNA of G_0 produces proteins according to a Poisson process with parameter k_2 and the lifetime of a protein is exponentially distributed with parameter γ_2.

For this model, the parameter k_1 can be seen as the rate at which a polymerase binds on an active gene, but also the rate at which an mRNA is produced. The steps of binding and elongation are thus reduced to a single step. The same remark holds for the production of proteins. The model of Section 3 distinguishes these two steps.

For $t \geq 0$, we denote $I(t) \in \{0, 1\}$ the state of gene at time t and $M(t) \in \mathbb{N}$, (resp., $P(t) \in \mathbb{N}$), is the number of mRNAs, (resp., of proteins), at this instant. It is easily checked that the process $(Z(t)) \overset{\text{def}}{=} (I(t), M(t), P(t))$ is an irreducible Markov process in the state space $\mathcal{S}_0 \overset{\text{def}}{=} \{0, 1\} \times \mathbb{N}^2$. The non-zero elements of its Q-matrix $Q = (q(z, z'), z, z' \in \mathcal{S}_0)$ outside the diagonal are given by, for $z = (i, x, y) \in \mathcal{S}_0$,

$$\begin{cases} q(z, z+e_1) = k_+ \mathbb{1}_{\{i=0\}}, & q(z, z-e_1) = k_- \mathbb{1}_{\{i=1\}}, \\ q(z, z+e_2) = k_1 \mathbb{1}_{\{i=1\}}, & q(z, z-e_2) = \gamma_1 x, \\ q(z, z+e_3) = k_2 x, & q(z, z-e_3) = \gamma_2 y, \end{cases}$$

where e_i, $i = 1, 2, 3$, are the unit vectors of $\mathcal{S}_0$. See Figure 3. As a functional operator Q, it can be defined as

$$Q(f)(z) = \sum_{z' \in \mathcal{S}_0} q(z, z') f(z') = \sum_{z' \in \mathcal{S}_0 \setminus \{z\}} q(z, z')(f(z') - f(z)), \qquad z \in \mathcal{S}_0,$$

for some function f on $\mathcal{S}_0$. The starting point of analyses of the literature is always the classical system of Kolmogorov's equations associated to this Markov

process

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[f(Z(t))\right] = \mathbb{E}\left[Q(f)(Z(t))\right],$$

for some set of test functions f, like the indicator function of x, for $x \in \mathcal{S}_0$. They are generally referred to as *the master equation* in this literature. Its fixed point is the invariant distribution of the Markov process.

Proposition 2.1. *The Markov process $(I(t), M(t), P(t))$ has a unique invariant distribution $(\pi(z), z \in \mathcal{S}_0)$ with finite second moments*

$$\sum_{(i,x,y) \in \mathcal{S}_0} \left(x^2 + y^2\right) \pi(i, x, y) < +\infty.$$

Proof. The proof is skipped. This result is in fact established in the next sections in a much more general context. See the proofs of Propositions 4.2 and 5.1. $\square$

This result on the first two moments is used in the following to derive an explicit expression of them. A much more general result holds in fact since the equilibrium of both variables $(M(t))$ and $(P(t))$ has a finite exponential moment, i.e. there exists $\eta > 0$ such that

$$\sum_{(i,x,y) \in \mathcal{S}_0} \left(e^{\eta x} + e^{\eta y}\right) \pi(i, x, y) < +\infty.$$

See Gupta et al. [14] for example.

There does not seem that there exists a closed form expression for the invariant distribution $(\pi(i, x, y))$ of the equilibrium equations. See Relation (2.2) below for the generating function of P when the gene is always active. It turns out that, nevertheless, one can get explicit expressions for the moments of this distribution. Related formulas of the kind have been the main and essentially, the only, rigorous results of mathematical models of gene expression starting from the first studies in 1977. They are still used in the quantitative analyses of the biological literature. See the supplementary material of Taniguchi et al. [37] for example.

Proposition 2.2. *If (I, M, P) is a random variable whose law is the invariant distribution of the Markov process $(I(t), M(t), P(t))$, then*

$$\mathbb{E}(I) = \delta_+ \stackrel{\mathrm{def}}{=} \frac{k_+}{k_+ + k_-}, \quad E(M) = \delta_+ \frac{k_1}{\gamma_1} \ and \ \mathbb{E}(P) = \delta_+ \frac{k_1 k_2}{\gamma_1 \gamma_2},$$

and, for the variances, if $\Lambda = k_+ + k_-$,

$$\frac{\mathrm{Var}(M)}{\mathbb{E}(M)} = 1 + (1 - \delta_+) \frac{k_1}{\Lambda + \gamma_1},$$

$$\frac{\mathrm{Var}(P)}{\mathbb{E}(P)} = 1 + \frac{k_2}{\gamma_1 + \gamma_2} + (1 - \delta_+) \frac{k_1 k_2}{\gamma_1 + \gamma_2} \frac{\Lambda + \gamma_1 + \gamma_2}{(\Lambda + \gamma_1)(\Lambda + \gamma_2)}. \tag{2.1}$$

The ratio of the variance and of the mean given for M and P is called the *Fano factor*, it can be seen as a measure of the dispersion of the probability distribution. In the biological literature it is interpreted as a measure of deviation from the Poisson distribution. Recall that the Fano factor of a Poisson distribution is 1.

Proof. The proofs in the literature are generally based on the differential equations satisfied by the generating function of the distribution of the three-dimensional process. See Leoncini [21]. A simple approach, avoiding manipulations of generating functions, is presented, it is based on some natural flow equations at equilibrium.

The equilibrium equations for the invariant distribution π of the Markov process $(I(t), M(t), P(t))$ can be expressed as

$$\int_{\mathcal{S}_0} Q(f)(i, x, y) \, \mathrm{d}\pi(i, x, y) = \mathbb{E}[Q(f)(I, M, P)] = 0,$$

for any function on $\mathcal{S}_0$ such that $Q(f)$ is integrable with respect to π. If, for $z \in \mathcal{S}_0$, by taking f the indicator function of z, this gives the usual balance equations.

We will use the class of functions $f(i, x, y) = i^a x^b y^c$, with $i \in \{0, 1\}$, $a, b \in \{0,1,2\}$ to get an appropriate system of linear equations to get these moments. Proposition 2.1 shows that $Q(f)$ is integrable for any function f of this class.

The equations

$$
\begin{aligned}
k_+ \mathbb{E}(1-I) - k_- E(I) = 0, && f(i, x, y) = i, \\
k_1 \mathbb{E}(I) - \gamma_1 \mathbb{E}(M) = 0, && f(i, x, y) = x, \\
k_2 \mathbb{E}(M) - \gamma_2 \mathbb{E}(P) = 0, && f(i, x, y) = y
\end{aligned}
$$

give immediately the formulas for the first moments.

A second system of linear equations

$$
\begin{aligned}
k_1 \mathbb{E}((2M+1)I) + \gamma_1 \mathbb{E}(M(1-2M)) = 0, && f(i, x, y) = x^2, \\
k_2 \mathbb{E}(M(2P+1)) + \gamma_2 \mathbb{E}(P(1-2P)) = 0, && f(i, x, y) = y^2, \\
k_1 \mathbb{E}(IP) - \gamma_1 \mathbb{E}(MP) + k_2 \mathbb{E}(M^2) - \gamma_2 \mathbb{E}(MP) = 0, && f(i, x, y) = xy, \\
k_+ \mathbb{E}((1-I)M) - k_- \mathbb{E}(IM) + k_1 \mathbb{E}(I) - \gamma_1 \mathbb{E}(IM) = 0, && f(i, x, y) = ix, \\
k_+ \mathbb{E}((1-I)P) - k_- \mathbb{E}(IP) + k_2 \mathbb{E}(IM) - \gamma_2 \mathbb{E}(IP) = 0, && f(i, x, y) = iy,
\end{aligned}
$$

for the variables $E(M^2)$, $\mathbb{E}(P^2)$, $\mathbb{E}(MP)$, $\mathbb{E}(IM)$ and $\mathbb{E}(IP)$ gives the second moments of M and P. The proposition is proved. $\qquad\square$

As it will be seen in a more general framework, when the gene is always active, i.e. $k_- = 0$, the distribution of M is Poisson. This explains in particular the identity $\mathrm{Var}(M) = \mathbb{E}(M)$ of Proposition 2.2 in this case. See Proposition 4.4 of Section 4. Note that it is not the case for P since

$$\frac{\mathrm{Var}(P)}{\mathbb{E}(P)} > 1,$$

by Proposition 2.2. This is usually interpreted in the biological literature as the fact that the distribution of P is more variable than a Poisson distribution.

The above results, the first and second moments of the number of proteins at equilibrium, are essentially the main mathematical results of the current biological literature in this domain. As an example of application of the methods presented in this paper, we state several results which give some additional insight on the properties of the equilibrium distribution of the number of proteins. They are simple applications of general results which are proved in Section 3 and 6.

An explicit representation of the generating function of P is in fact available.

Proposition 2.3. *If the gene is always active, the generating function of the equilibrium distribution of the number of proteins, for $z \in (0, 1)$, is given by*

$$
\mathbb{E}\left(z^P\right) = \exp\left(-k_1 \int_{\mathbb{R}_+} \left[1 - e^{-(1-z)k_2\left(1-e^{-\gamma_2 u}\right)/\gamma_2}\right] e^{-\gamma_1 u}\, \mathrm{d}u \right.
$$
$$
\left. -k_1 \int_{\mathbb{R}_+^2} \left[1 - e^{-(1-z)k_2 e^{-\gamma_2 u}\left(1-e^{-\gamma_2 w}\right)/\gamma_2}\right] \gamma_1 e^{-\gamma_1 w}\, \mathrm{d}u\, \mathrm{d}w \right). \quad (2.2)
$$

Bokes et al. [3] gives, via an analytic approach with the master equation, an alternative representation of the generating function in terms of an hypergeometric function.

The classical relation (2.1) for the variance shows that when the gene is always active, i.e. $\delta_+ = 1$, and the average lifetime $1/\gamma_2$ of proteins gets large, then

$$
\frac{\mathrm{Var}P}{\mathbb{E}(P)} = \mathbb{E}\left(\left(\frac{P - \mathbb{E}(P)}{\sqrt{\mathbb{E}(P)}}\right)^2\right) \sim 1 + \frac{k_2}{\gamma_1}.
$$

The following proposition gives a more precise result, when lifetimes of proteins goes to infinity. It is a consequence of a general convergence theorem, Theorem 6.1 of Section 6.

Proposition 2.4. *If the gene is always active, for the convergence in distribution, the relation*

$$
\lim_{\gamma_2 \to 0} \frac{P - \mathbb{E}(P)}{\sqrt{\mathbb{E}(P)}} = \mathcal{N}(\sigma),
$$

holds, where $\mathcal{N}(\sigma)$ is a centered Gaussian random variable whose standard deviation is

$$
\sigma = \sqrt{1 + k_2/\gamma_1}.
$$

This proposition gives a Gaussian approximation result for the distribution of P as $\mathbb{E}(P) + \sqrt{\mathbb{E}(P)}\mathcal{N}(\sigma)$. The result will hold for protein types with a large number of copies in the cell.

The following proposition establishes a convergence in distribution when the death rate of mRNAs is going to infinity with the constraint that the transcription rate k_2 is sufficiently large so that each mRNA produces a fixed average number of proteins.

Proposition 2.5. *If the gene is always active, then, for $a>0$, the relation*

$$\lim_{\substack{\gamma_1 \to +\infty \\ k_2/\gamma_1 = a}} P = N(a),$$

holds for the convergence in distribution, where $N(a)$ is random variable whose distribution is given by, for $n \in \mathbb{N}$,

$$\mathbb{P}(N(a){=}n) = \frac{1}{(a{+}1)^{k_1/\gamma_2}} \frac{\Gamma(k_1/\gamma_2{+}n)}{n!\,\Gamma(k_1/\gamma_2)} \left(\frac{a}{1{+}a}\right)^n,$$

where Γ is the classical Gamma function.

The distribution of $N(a)$ is the *negative binomial* distribution with parameters k_1/γ_2 and a. See Chapter 5 of Johnson et al. [16]. This distribution is frequently mentioned in the approximations of the literature, Bokes et al. [3], Friedman et al. [12] and Shahrezaei and Swain [35] for example. It is also sometimes used to fit measurements of concentration of proteins. Its advantage compared to a simple Poisson distribution may be due to the fact that it is more flexible with its two parameters.

Proof. This is a straightforward application of Proposition 6.3 of Section 6. The elongation time of proteins is null, i.e. $E_2 \equiv 0$, so that the generating function of the limit $N(a)$ is given

$$\mathbb{E}\left(z^{N(a)}\right) = \exp\left(-k_1 \int_0^{+\infty} \frac{(1{-}z)\,ae^{-\gamma_2 u}}{1{+}(1{-}z)\,ae^{-\gamma_2 u}}\,\mathrm{d}u\right) = \left(\frac{1}{1{+}a{-}az}\right)^\kappa,$$

with $\kappa{=}k_1/\gamma_2$. Relations (B.3) and (B.4) of the Appendix give the representation of this distribution. □

3. A general stochastic model

We introduce in this section the stochastic processes which will be studied for the analysis of the time evolution of the number of proteins.

The distributions of lifetimes of mRNAs and proteins are general, in particular they are not assumed to be exponentially distributed as in the Markovian model of Section 2. The general distributions of the duration of elongations of mRNAs and proteins are also incorporated in the stochastic model. Once the polymerase/ribosome is bound to the gene/mRNA, the elementary components, nucleotides for mRNAs and amino-acids for proteins, are progressively added to build these components. As it can be seen in Figure 5 in the Appendix, the number of these elements can vary from $\sim$20 elements up to several thousands depending on the gene considered. This, of course, incurs delays which cannot be really neglected. This presentation is an extension of the framework of Fromion et al. [13]. See also Leoncini [21].

As it will be seen, a representation of the *equilibrium* of the protein production process in terms of marked Poisson point processes is established.

Contrary to a Markovian analysis, the resolution of the linear system of balance equations for the equilibrium, is not necessary in this case. More general assumptions on the distributions of some of the steps of the process can therefore be considered as mentioned before. It should be noted however, that even in the Markovian context of the classical three-step model, this representation simplifies much the analysis of the equilibrium. For example, asymptotic results concerning the *equilibrium distribution* of the number of proteins are simpler to obtain than the approach using analytic tools with hypergeometric functions. See Bokes et al. [3] and Section 6.2 for example.

3.1. Gene activation

The state of the gene for the type of protein is either active or inactive. It is activated at rate k_+ and inactivated at rate k_-. Time duration of these states are assumed to be exponentially distributed.

The process of activation of the gene is a Markov process $(I(t), t \in \mathbb{R})$ with values in $\{0, 1\}$ and whose Q-matrix R_I is given by $r_I(0, 1) = k_+$ and $r_I(1, 0) = k_-$. Without loss of generality, it will be assumed to be stationary. For this reason $(I(t))$ is defined on the whole real line, in some sense, the activation/inactivation process has started at $t = -\infty$. As it will be seen, this is a convenient formulation to describe properly the equilibrium of the protein production process.

This process can be represented as a marked point process $((t_i, X_i), i \in \mathbb{Z})$ where (t_i) is the increasing sequence of instants of change of state of the gene, with the convention that $t_0 \leq 0 < t_1$. The marks (X_i) in the set $\{0, 1\}$ indicate the state of the gene at these instants. In particular, for $i \in \mathbb{Z}$, conditionally on the event $\{X_i = 0\}$, (resp., on the event $\{X_i = 1\}$), the random variable $t_{i+1} - t_i$ is exponentially distributed with rate k_+, (resp., with rate k_-). Because of our assumption on stationarity of the activation/deactivation process, the sequence of activation instants of the gene, $(t_i \mathbb{1}_{\{X_i=1\}})$, is a stationary renewal point process. See Section 11.5 of Robert [33]. We denote

$$\delta_+ = \frac{k_+}{\Lambda} \text{ and } \Lambda \stackrel{\text{def}}{=} k_+ + k_-,$$

in particular $\mathbb{P}(I(t) = 1) = \delta_+$, for all $t \in \mathbb{R}$.

3.2. Transcription and translation: a fundamental Poisson process

When a gene is active, a *polymerase* can be bound to it according to a Poisson process with rate $k_1 > 0$. The *transcription phase* can start. An mRNA is then being built up by a process of aggregation of a specific sequence of nucleotides which are present in the cytoplasm. This is *the elongation phase* of an mRNA. Its duration has a distribution $E_1(\mathrm{d}x)$ on $\mathbb{R}_+$. Similarly a *ribosome* is bound to a given mRNA according to a Poisson process with parameter k_2. The *translation phase* starts. A chain of *amino-acids* is created as the ribosome progresses on the mRNA. This is the *elongation phase* of a protein. Its duration has a distribution

$E_2(\mathrm{d}x)$ on $\mathbb{R}_+$. Both mRNAs and proteins have a finite lifetime. The distribution of the lifetime of an mRNA (resp., a protein) is $L_1(\mathrm{d}y)$ (resp., $L_2(\mathrm{d}y)$) on $\mathbb{R}_+$. Throughout the paper, it is assumed that all these distribution have a finite first moment.

The fundamental Poisson process.
These two steps are represented by a marked Poisson point process. See Section B of the Appendix.

$$\mathcal{P} \stackrel{\mathrm{def}}{=} (u_n, E_{1,n}, L_{1,n}, \mathcal{N}_{2,n}, n\in\mathbb{Z}), \tag{3.1}$$

on the state space

$$\mathcal{S} \stackrel{\mathrm{def}}{=} \mathbb{R}\times\mathbb{R}_+^2\times\mathcal{M}_p(\mathbb{R}\times\mathbb{R}_+^2),$$

where $\mathcal{M}_p(\mathbb{R}\times\mathbb{R}_+^2)$ is the space of Radon point measures on $\mathbb{R}\times\mathbb{R}_+^2$, see Dawson [9] for example, and

1. (u_n) is a Poisson process on $\mathbb{R}$ with parameter k_1, this is the sequence of possible instants when a polymerase can be bound to the gene. For $n\in\mathbb{Z}$, u_n is indeed an instant of binding only if the gene is active at time u_n. The sequence $(u_n, n\in\mathbb{Z})$ is assumed to be non-decreasing and indexed with the convention $u_0 \leq 0 < u_1$.
2. $(E_{1,n}, n\in\mathbb{Z})$ is the sequence of the duration of the elongation for these mRNAs, it is an i.i.d. sequence whose common distribution is E_1.
3. $(L_{1,n})$ is the i.i.d. sequence of associated lifetimes of these mRNAs, its common law is L_1.
4. $(\mathcal{N}_{2,p})$ is an i.i.d. sequence with the same distribution as $\mathcal{N}_2$, a marked Poisson point process on $\mathbb{R}\times\mathbb{R}_+^2$ with intensity $k_2 \, \mathrm{d}x\otimes E_2\otimes L_2$.
   For $n\in\mathbb{Z}$, $\mathcal{N}_{2,n}=(x_{2,j}^n, E_{2,j}^n, L_{2,j}^n, j\in\mathbb{Z})$ is the process associated with the protein production for the mRNA with index n.

   (a) $(x_{2,j}^n, j\in\mathbb{Z})$ is a Poisson process on $\mathbb{R}$ with parameter k_2, this is the sequence of possible instants when a ribosome can be bound to the mRNA with index n. Only the instants occurring during the lifetime of this mRNA matter.

   (b) $(E_{2,j}^n)$ is the sequence of the duration of the elongation of proteins by this mRNA, it is an i.i.d. sequence whose common distribution is E_2.

   (c) $(L_{2,j}^n)$ is the i.i.d. sequence of associated lifetimes of associated proteins, its common law is L_2.

The coordinates of $\mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)$ are associated to the potential instants u of the binding of a polymerase on the gene, the variable v is the elongation time of the mRNA and the variable w is its lifetime. The last component m is the marked point process associated to the protein production process of this mRNA. The coordinates of $m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ are, x is associated to binding instants of ribosomes on this mRNA, y is the elongation time of the corresponding protein and z its lifetime.
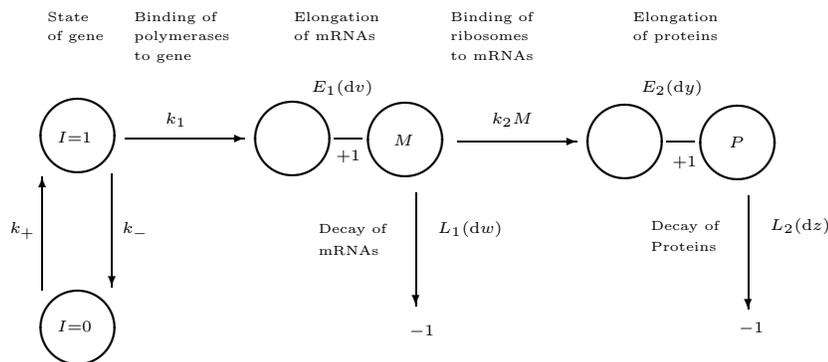
FIG 4. *A General Model for Protein Production.*

The intensity measure of $\mathcal{P}$ is

$$\nu_{\mathcal{P}} \stackrel{\text{def}}{=} k_1 \, \mathrm{d}u \otimes E_1 \otimes L_1 \otimes Q_2, \tag{3.2}$$

where Q_2, $Q_2(\mathrm{d}m) = \mathbb{P}(\mathcal{N}_2 \in \mathrm{d}m)$, is the distribution on $\mathcal{S}$ of the random variable $\mathcal{N}_2$, a Poisson point process on $\mathbb{R} \times \mathbb{R}_+^2$ with intensity $k_2 \, \mathrm{d}x \otimes E_2 \otimes L_2$.

Equivalently, if F is a non-negative Borelian function on $\mathcal{S}$, then

$$\int_{\mathcal{S}} F(u, v, w, m) \nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)$$

$$= \int_{\mathcal{S}} F(u, v, w, m) k_1 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w) Q_2(\mathrm{d}m)$$

$$= \int_{\mathbb{R}} \mathbb{E}(F(u, E_1, L_1, \mathcal{N}_2)) k_1 \, \mathrm{d}u,$$

where, with a slight abuse of notations, E_1, L_1 and $\mathcal{N}_2$ are independent random variables with respective distributions $E_1(\mathrm{d}v)$, $L_1(\mathrm{d}w)$ and $Q_2(\mathrm{d}m)$.

EXAMPLES. If the gene is always active and the system starts empty at time 0, the formula

$$\sum_{n \geq 1} \mathbb{1}_{\{u_n + E_{1,n} \leq t\}} = \int_{\mathbb{R}_+ \times \mathbb{R}_+^2} \mathbb{1}_{\{u+v \leq t\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w)$$

gives the total number of mRNAs created up to time t. The following notation has been used

$$\int_{\mathcal{S}} f(u, v, w) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w) = \int_{\mathcal{S}} f(u, v, w) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m),$$

if f is non-negative Borelian function on $\mathbb{R} \times \mathbb{R}_+^2$, i.e. $\mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w)$ is the marginal distribution of $\mathcal{P}$ with respect to the three first coordinates, i.e. it stands for $\mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+^2))$.

Similarly, still with a permanently active gene,

$$\sum_{n\geq 1}\mathbb{1}_{\{u_n+E_{1,n}\leq t<u_n+E_{1,n}+L_{1,n}\}} = \int_{\mathbb{R}_+\times\mathbb{R}_+^2}\mathbb{1}_{\{u+v\leq t<u+v+w\}}\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w)$$

is the number of mRNAs present at time $t\geq 0$.

If we include the gene dynamics into the formula, the number of messengers present at time t is

$$\sum_{n\geq 1}I(u_n)\mathbb{1}_{\{u_n+E_{1,n}\leq t<u_n+E_{1,n}+L_{1,n}\}}$$

$$= \int_{\mathbb{R}_+\times\mathbb{R}_+^2}I(u)\mathbb{1}_{\{u+v\leq t<u+v+w\}}\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w)$$

If the mRNA with index n is created at time $v=u_n+E_{2,n}$ and if its lifetime is w then

$$\mathcal{N}_{2,n}([v,v+w)\times\mathbb{R}_+^2) = \int_{[v,v+w)\times\mathbb{R}_+^2}\mathcal{N}_{2,n}(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z)$$

is the total number of proteins created by such an mRNA during its lifetime.

Comments on the model.
As a mathematical model, the above representation simplifies several aspects of the protein production process but the key steps of protein production are included. Furthermore, its main parameters have a clear biological interpretation.

1. *Expression Rates*. The parameter k_1 is the *affinity* of polymerases, or the *transcription rate* for the gene considered. The larger this rate is, the more likely polymerases will bind to this gene rather than to some other genes with a lower affinity. The same remark applies for k_2, the affinity of ribosomes, the *translation rate*, for the corresponding mRNAs.
   It should be stressed that if mRNAs and proteins are specific to a gene, polymerases and ribosomes are not. They can be used to create any type of mRNA and any type protein respectively. Polymerases and ribosomes can be seen as a kind of resources of the cell. The genes are, in some sense, competing to have access to polymerases. This approach with the parameters k_1 and k_2 considers that a specific gene receives a fixed portion of resources of the cell in terms of polymerases and ribosomes. In particular, it does not include the competition for access to resources between the different genes. Alternative stochastic models have to be considered to analyze these aspects.
2. *Elongation Times*. An mRNA being a sequence of N nucleotides, the elongation time E_1 of an mRNA is represented as a sum of N independent random variables, each of them corresponding to the duration of time required to get each of the nucleotides within the cytoplasm. A similar situation holds for E_2 for the number of amino-acids of the protein. See Figure 5 in the Appendix.

3. *Lifetimes*. The lifetimes of mRNAs are generally smaller than the lifetimes of proteins, mRNAs can be indeed degraded by other RNAs in the cytoplasm after a couple of minutes. This is one of the mechanisms used by the cell to regulate the protein production process. The situation is quite different for proteins, their lifetime may exceed the duration of time for cell division, of the order of 40mn for bacterium. At cell division, macromolecules are assumed to be allocated at random into one of the two daughter cells. In this case, if one follows a specific path of cells in the associated binary tree of successive cell divisions, proteins "vanish" simply because they have been allocated in the daughter cell. This is referred to as the *dilution* phenomenon.

4. *Multiple Copies of the Gene*. In a favorable environment, cells grow and divide very quickly. In particular a copy of the DNA is always on the way in the cell, and, therefore, another copy of the gene is present at some stage. In principle this may double the transcription rate of the gene when the corresponding part of the DNA has been duplicated. The rate is reduced at the division time. This aspect has been omitted in our model, mainly for the sake of simplicity. The main change would be that the process $(I(t))$ should have at least three values, depending on the number of copies of the gene which are active at a given time. See Paulsson [29] in a Markovian context.

The classical Markovian model of Section 2 corresponds to the case when the elongation times of mRNAs and proteins are null and that their respective lifetimes are exponentially distributed.

How to handle functionals of $\mathcal{P}$.
We will use repeatedly several possible representations in the calculations of expected values of functionals of $\mathcal{P}$. We give a quick sketch of the general approach. Let f be a non-negative Borelian function on $\mathbb{R}\times\mathbb{R}_+^2\times\mathbb{R}\times\mathbb{R}_+^2$ and let, for $(u,v,w,m)\in\mathcal{S}$,

$$F(u,v,w,m) \stackrel{\text{def}}{=} \int_{\mathbb{R}\times\mathbb{R}_+^2} f(u,v,w,x,y,z)m(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z).$$

When the gene is always active, most of random variables of interest can be expressed under the form

$$X = \langle\mathcal{P},F\rangle \stackrel{\text{def}}{=} \int_{\mathcal{S}} F(u,v,w,m)\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m).$$

As it will be seen, the variable activity of the gene adds some technical complications.

1. Calculation of the mean. Relation (B.1) of the Appendix gives that

$$\mathbb{E}(X) = \int_{\mathcal{S}} F(u,v,w,m)\nu_{\mathcal{P}}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m)$$

$$= \int_{\mathcal{S}} F(u,v,w,m) k_1 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w) Q_2(\mathrm{d}m)$$

$$= k_1 \int_{\mathbb{R}} \mathbb{E}(F(u, E_1, L_1, \mathcal{N}_2)) \, \mathrm{d}u,$$

with the same slight abuse of notations as before. For $(u,v,w) \in \mathbb{R}_+^3$, since the Poisson process $\mathcal{N}_2$ has intensity measure $k_2 \, \mathrm{d}x \otimes E_2 \otimes L_2$ on $\mathbb{R} \times \mathbb{R}_+^2$,

$$\mathbb{E}(F(u,v,w,\mathcal{N}_2)) = \mathbb{E}\left(\int_{\mathbb{R} \times \mathbb{R}_+^2} f(u,v,w,x,y,z) \mathcal{N}_2(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z) \right)$$

$$= \int_{\mathbb{R} \times \mathbb{R}_+^2} f(u,v,w,x,y,z) k_2 \, \mathrm{d}x E_2(\mathrm{d}y) L_2(\mathrm{d}z)$$

$$= k_2 \int_{\mathbb{R}} \mathbb{E}(f(u,v,w,x,E_2,L_2)) \, \mathrm{d}x,$$

hence

$$\mathbb{E}(X) = k_1 k_2 \int_{\mathbb{R}^2} \mathbb{E}(f(u, E_1, L_1, x, E_2, L_2)) \, \mathrm{d}u \, \mathrm{d}x.$$

For calculation of variances, the approach is similar, via Relation (B.2) of the Appendix and an additional trick. See the proof of Relation (4.6) for example.

2. Exponential Moments of X. For more precise results on the distribution of X, one has to express $\mathbb{E}(\exp(-\xi X))$, for some $\xi \geq 0$, the *Laplace transform* of X at ξ. Proposition B.1 of the Appendix gives the relation

$$\mathbb{E}\left(e^{-\xi \langle \mathcal{P}, F \rangle} \right) = \exp\left(-\int_{\mathcal{S}} \left(1 - e^{-\xi F(u,v,w,m)} \right) \nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m) \right)$$

$$= \exp\left(-\xi \int_{\mathcal{S}} \left(1 - e^{-\xi F(u,v,w,m)} \right) k_1 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w) Q_2(\mathrm{d}m) \right)$$

$$= \exp\left(-k_1 \int_{\mathbb{R} \times \mathbb{R}_+^2} \left(1 - \mathbb{E}\left(e^{-\xi F(u,v,w,\mathcal{N}_2)} \right) \right) \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w) \right).$$

By using again Proposition B.1 for the Poisson process $\mathcal{N}_2$, with the same arguments as before, we get, for $(u,v,w) \in \mathbb{R} \times \mathbb{R}_+^2$,

$$\mathbb{E}\left(e^{-\xi F(u,v,w,\mathcal{N}_2)} \right)$$

$$= \exp\left(-\int_{\mathbb{R}} \left(1 - \mathbb{E}\left(e^{-\xi f(u,v,w,x,y,z)} \right) \right) k_2 \, \mathrm{d}x E_2(\mathrm{d}y) L_2(\mathrm{d}z) \right)$$

$$= \exp\left(-k_2 \int_{\mathbb{R}} \left(1 - \mathbb{E}\left(e^{-\xi f(u,v,w,x,E_2,L_2)} \right) \right) \mathrm{d}x \right).$$

This expression for $\mathbb{E}(\exp(-X))$ though explicit is not simple to handle. It is nevertheless usable to get several limit results in Section 6. The trick of

using independent random variables E_a, L_a, $a \in \{1, 2\}$ or their distributions $E_a(\mathrm{d}s)$, $L_a(\mathrm{d}t)$ is used throughout this paper. Its main advantages are of simplifying the calculations *and* the expressions of the formulas obtained.

We state a simple result on some invariance properties of the Poisson process $\mathcal{P}$. For $m \in \mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+^2)$ and f is a Borelian function on $\mathbb{R} \times \mathbb{R}_+^2$, the measure $\check{m}$ is defined by

$$\langle \check{m}, f \rangle = \int_{\mathbb{R} \times \mathbb{R}_+^2} f(x, y, z) \check{m}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) = \int_{\mathbb{R} \times \mathbb{R}_+^2} f(-x, y, z) m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z).$$

Proposition 3.1. *If $\mathcal{P}$ is the marked Poisson point process on $\mathcal{S}$ defined by Relation (3.1) and F is some non-negative Borelian function on $\{0, 1\} \times \mathcal{S}$ then*

$$\int_{\mathcal{S}} F(I(u), u, v, w, m) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)$$

$$\stackrel{\text{dist.}}{=} \int_{\mathcal{S}} F(I(-u), -u, v, w, \check{m}) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)$$

Proof. This is a simple consequence of

— the reversibility of $(I(t))$. It is easily checked that $(I(-t))$ is a Markov process with the same Q-matrix, and with same distribution at $t=0$ as $(I(t))$. See Kelly [18] for example.
— If (t_n) is a Poisson process on $\mathbb{R}$ with rate $\lambda > 0$, then $(-t_n)$ is a Poisson process with the same rate.
— The independence of $(I(t))$ and $\mathcal{P}$.

Consequently, we have the identity

$$(I(u_n), u_n, L_{1,n}, \mathcal{N}_{2,n}) \stackrel{\text{dist}}{=} (I(-u_n), -u_n, L_{1,n}, \check{\mathcal{N}}_{2,n}),$$

and, therefore, our proposition. $\square$

3.3. Convergence to equilibrium

As before we denote by $M(t)$, (resp., $P(t)$) the number of mRNAs, (resp., of proteins) at time $t \geq 0$. It is assumed that the initial state is as follows: there are

— $M(0)$ mRNAS with respective remaining lifetimes $L_{1,k}^0$, $k=1, \ldots, M(0)$;
— $P(0)$ proteins with respective remaining lifetimes $L_{2,k}^0$, $k=1, \ldots, P(0)$. The marked Poisson process associated to the creation of proteins by the kth mRNA present at time 0 is denoted by $\mathcal{N}_{2,k}^0$.

For simplicity it is assumed that, initially, there are no mRNAs or proteins in their elongation phase. The proof of convergence in distribution does not change if the initial state includes components in their elongation phase.

The following theorem is the key result concerning the equilibrium distribution of the number of mRNAs and proteins.

1. It shows the convergence to equilibrium without having a Markovian framework which is, usually, the classical approach to prove such a convergence. As a consequence, the distribution of elongation times, lifetimes of mRNAs and proteins can be general.
2. The equilibrium of the number of mRNAs and of proteins are expressed in terms of functionals of the marked Poisson process $\mathcal{P}$. The closed form expressions of averages and variances are obtained with this representation.

Theorem 3.1 (Number of mRNAs and Proteins at Equilibrium). *As t goes to infinity, the random variable $(M(t), P(t))$ converges in distribution to (M, P) defined by*

$$M = \left(\int_{\mathcal{S}} I(u) \mathbb{1}_{\{v \leq u < v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w) \right) \tag{3.3}$$

and the variable P is defined as

$$\left(\int_{\mathcal{S}} I(u) \left(\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \begin{smallmatrix} x+v \leq u < x+v+w \\ y \leq x < y+z \end{smallmatrix} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m) \right), \tag{3.4}$$

where $\mathcal{P}$ is the marked Poisson point process on $\mathcal{S}$ defined by Relation (3.1).

The proof of the convergence relies on *coupling from the past arguments*. The idea consists in starting the process at time $-t$ and to study its state at time 0, which will have the same distribution as the original process at time t. If the process has convenient properties, of monotonicity for example, it may happen that the state at time 0 of the shifted process converges *almost surely* as t goes to infinity. This gives then the convergence in distribution of $(I(t), M(t), P(t))$ when t goes to infinity. This method applies in our case. One of the earliest works using this method seems to be Loynes [24] (1962). Its use has been popularized later by Propp and Wilson [30] to study the Ising Model. See Levin at al. [23] for a survey.

Proof. We first express the variables $M(t)$ and $P(t)$ in terms of the point process $\mathcal{P}$.

The variable $M(t)$ is the sum of the number of initial mRNAs still alive at time t and of the number of mRNAs born before time t and still alive at time t. This gives the following formula,

$$M(t) = \sum_{k=1}^{M(0)} \mathbb{1}_{\left\{ L_{2,k}^0 > t \right\}} + \int_{\mathcal{S}} I(u) \mathbb{1}_{\{0 < u+v \leq t < u+v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w).$$

Similarly, $P(t)$ is the sum of three quantities corresponding to the number of

1. the number of proteins present at time 0 and still alive at time t,

$$\sum_{k=1}^{P(0)} \mathbb{1}_{\left\{ L_{2,k}^0 > t \right\}};$$

2. the number of proteins created before time t by one of the mRNAs present at time 0 and still alive at time t,

$$\sum_{k=1}^{M(0)} \int_{\mathbb{R} \times \mathbb{R}_+^2} I(u) \mathbb{1}_{\left\{0 < u < L_{1,k}^0, u+v \le t < u+v+w\right\}} \mathcal{N}_{2,k}^0(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w);$$

3. proteins created before time t by new mRNAs and still alive at time t,

$$P_1(t) \overset{\mathrm{def}}{=} \int_{\mathcal{S}} I(u) \mathbb{1}_{\{u \in (0,t)\}}$$
$$\left(\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \begin{smallmatrix} u+v \le x < u+v+w \\ x+y \le t < x+y+z \end{smallmatrix} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m).$$

For simplicity of presentation, we will prove the convergence in distribution of $(P(t))$. The convergence in distribution of $(M(t), P(t))$ is similar. Clearly, the two first terms converge almost surely to 0 when t goes to infinity. We have thus only to take care of $P_1(t)$.

The stationarity of the process $\mathcal{P}$ with respect to translation by $-t$, Proposition 3.1, the identity $(I(s-t)) \overset{\mathrm{dist}}{=} (I(s))$, and the fact that $(I(s))$ is independent of $\mathcal{P}$ give that $P_1(t)$ has the same distribution as $P_2(t)$ with

$$P_2(t) \overset{\mathrm{def}}{=} \int_{\mathcal{S}} I(u) \mathbb{1}_{\{u \in (-t,0)\}}$$
$$\left(\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \begin{smallmatrix} u+v \le x < u+v+w \\ x+y \le 0 < x+y+z \end{smallmatrix} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m).$$

The quantity $P_2(t)$ can be seen the number of proteins at time 0 when the process starts empty at time $-t$, i.e. without mRNAs or proteins. This is a non-decreasing function of t converging almost surely to $P_2(\infty)$ defined by

$$\int_{\mathcal{S}} I(u) \mathbb{1}_{\{u < 0\}} \left(\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \begin{smallmatrix} u+v \le x < u+v+w \\ x+y \le 0 < x+y+z \end{smallmatrix} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m).$$

Proposition 3.1 gives that the quantity $P_2(\infty)$ has the same distribution as

$$\int I(u) \mathbb{1}_{\{u > 0\}} \left(\int \mathbb{1}_{\left\{ \begin{smallmatrix} x+v \le u < x+v+w \\ y \le x < y+z \end{smallmatrix} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m),$$

it is easy to see that this last term is the second coordinate of (3.4). The theorem is proved. □

4. Transcription

Relation (3.3) gives that the distribution of the number of mRNAs at equilibrium is given by the law of the random variable M defined by

$$M = \int_{\mathbb{R} \times \mathbb{R}_+^2} I(u) G_1(u, v, w) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w), \tag{4.1}$$

with

$$G_1(u, v, w) \overset{\text{def}}{=} \mathbb{1}_{\{v \leq u < v + w\}}. \tag{4.2}$$

The process of activation/deactivation of the gene complicates significantly the derivation of the mathematical expressions for the mean and variance of M. Formulas are much more simple when the gene is always active. We will need some technical results on this process. We denote by $\mathcal{F}^I$ the σ-field generated by the stochastic process $(I(t))$,

$$\mathcal{F}^I = \sigma \langle I(t), t \in \mathbb{R} \rangle.$$

A representation of the distribution of M in terms of a marked Poisson point process conditioned on the σ-field $\mathcal{F}^I$ is the main tool used. For the calculation of the variance an additional work has then to be done to "remove" this conditioning. It turns out that the correlation structure of the process $(I(t))$ plays a role at this stage.

Lemma 4.1 (Correlation Function of $(I(t))$). *For $t \geq 0$, $\mathbb{E}(I(t)) = \delta_+$ and*

$$\mathbb{P}(I(t) = 1 | I(0) = 1) = \delta_+ + (1 - \delta_+) e^{-\Lambda t},$$

where $\Lambda = k_+ + k_-$ and $\delta_+ = k_+ / \Lambda$.

Proof. Let $p(t) = \mathbb{P}(I(t) = 1 | I(0) = 1)$, then Kolmogorov's forward equations give the ODE

$$p'(t) = k_+(1 - p(t)) - k_- p(t)$$

which is easily solved. $\qquad\square$

Define $\mathcal{P}^I$ the marked point process by, for a non-negative Borelian function on the space $\mathcal{S} \overset{\text{def}}{=} \mathbb{R} \times \mathbb{R}_+^2 \times \mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+^2)$,

$$\langle \mathcal{P}^I, f \rangle = \int_{\mathcal{S}} I(u) f(u, v, m) \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}m). \tag{4.3}$$

The following proposition gives the intuitive, but important, result that, conditionally on $\mathcal{F}^I$, the point process $\mathcal{P}^I$ is a marked Poisson point process with intensity measure

$$\nu_{\mathcal{P}}^I(\mathrm{d}u, \mathrm{d}v, \mathrm{d}m) = I(u) k_1 \, \mathrm{d}u \, L_1(\mathrm{d}v) Q_2(\mathrm{d}m). \tag{4.4}$$

See Relation (3.2) for the unconditional case.

Proposition 4.1. *For any non-negative Borelian function f on the state space $\mathcal{S}$, the relation*

$$\mathbb{E}\left(\exp\left(-\langle \mathcal{P}^I, f \rangle\right) \big| \mathcal{F}^I\right)$$
$$= \exp\left(-\int_{\mathcal{S}} I(u) \left(1 - e^{-f(u,v,w,m)}\right) \nu_P(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)\right)$$

$$= \exp\left(-k_1 \int_{\mathcal{S}} I(u)\left(1-e^{-f(u,v,w,m)}\right) \mathrm{d}u E_1(\mathrm{d}v)L_1(\mathrm{d}w)Q_2(\mathrm{d}m)\right)$$

holds almost surely.

Proof. For the sake of rigor, despite of its intuitive content, a proof is given. The (formal) difficulty is of expressing rigorously the conditioning with respect to $\mathcal{F}^I$. We use the notations of Section 3.1 where the process $(I(t))$ is defined by the doubly infinite sequence (t_i, X_i). Denote, for $N \geq 1$,

$$\mathcal{F}_N^I \overset{\text{def}}{=} \sigma\left\langle(t_i, X_i), -N \leq i \leq N\right\rangle,$$

since the distribution of the finite marginals of the sequence $((t_i, X_i), i \in \mathbb{Z})$ determine its distribution, we have

$$\lim_{n \to +\infty} \uparrow \mathcal{F}_N^I = \mathcal{F}^I.$$

Let f a non-negative Borelian function on $\mathcal{S}$ with compact support in the following sense: there exists some $K > 0$ such that $f(z) = 0$ if $z = (u, v, w, m) \in \mathcal{S}$ when $u \notin [-K, K]$. On the event $\{t_{-N} \leq -K, t_N \geq K\}$,

$$\left\langle \mathcal{P}^I, f \right\rangle = \sum_{i=-N}^{N} \mathbb{1}_{\{X_i=1\}} \int_{\mathcal{S}} \mathbb{1}_{\{u \in [t_i, t_{i+1})\}} f(u, v, w, m)\, \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m).$$

By independence of (t_i, X_i) and $\mathcal{P}$ and by using Proposition B.1, it is easily checked that, almost surely on this event,

$$\mathbb{E}\left(\exp\left(-\left\langle \mathcal{P}^I, f \right\rangle\right) \middle| \mathcal{F}_N^I\right)$$
$$= \exp\left(-\sum_{i=-N}^{N} \mathbb{1}_{\{X_i=1\}} \int_{\mathcal{S}} \mathbb{1}_{\{u \in [t_i, t_{i+1})\}} \left(1-e^{-f(u,v,w,m)}\right) \nu_P(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)\right).$$

Indeed, one has to multiply both sides of this relation by some non-negative Borelian function of $((t_i, X_i), -N \leq i \leq N)$, take the expected value of these expressions. By using the independence of $((t_i, X_i), -N \leq i \leq N)$ and $\mathcal{P}$, and the expression of the Laplace transform of a Poisson process, see Proposition B.1, it is easy to check the equality of these two terms. Consequently, for N sufficiently large, almost surely,

$$\mathbb{E}\left(\exp\left(-\left\langle \mathcal{P}^I, f \right\rangle\right) | \mathcal{F}_N^I\right)$$
$$= \exp\left(-\int_{\mathcal{S}} I(u)\left(1-e^{-f(u,v,w,m)}\right) \nu_P(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)\right).$$

A classical result from martingale theory gives the almost sure convergence

$$\lim_{N \to +\infty} \mathbb{E}\left(\exp\left(-\left\langle \mathcal{P}^I, f \right\rangle\right) \middle| \mathcal{F}_N^I\right) = \mathbb{E}\left(\exp\left(-\left\langle \mathcal{P}^I, f \right\rangle\right) \middle| \mathcal{F}^I\right).$$

See Williams [44]. Hence the proposition holds for these class of functions f with compact support in the sense defined above. We conclude with the fact that any positive Borelian function can be expressed as a (monotone) limit of such functions. $\square$

4.1. Mean and variance of the number of mRNAs at equilibrium

Proposition 4.2 (The first moments of the number of mRNAs). *If M is the number of mRNAs at equilibrium, then*

$$\mathbb{E}(M) = \frac{k_+}{\Lambda} k_1 \mathbb{E}(L_1), \tag{4.5}$$

and

$$\frac{\mathrm{Var}(M)}{\mathbb{E}(M)} = 1 + \mathbb{E}(M) \frac{k_-}{k_+} \mathbb{E}\left(e^{-\Lambda|F_{L_1,1} - F_{L_1,2} + E_{1,1} - E_{1,2}|}\right), \tag{4.6}$$

where $\Lambda = k_+ + k_-$, and $E_{1,1}$ and $E_{1,2}$ are independent random variables with distribution E_1, and $F_{L_1,1}$ and $F_{L_1,2}$ are independent random variables with density $\mathbb{P}(L_1 \geq u)/\mathbb{E}(L_1)$ on $\mathbb{R}_+$.

Recall that the distribution of the elongation time of an mRNA is E_1 and the duration of its lifetime is L_1. Note that when the lifetimes L_1 of an mRNA is exponentially distributed with parameters γ_1, then $F_{L_1} \overset{\mathrm{dist}}{=} L_1$, and if the elongation time of an mRNA is null, i.e. $E_1 \equiv 0$, the above relation gives

$$\frac{\mathrm{Var}(M)}{\mathbb{E}(M)} - 1 = \mathbb{E}(M) \frac{k_-}{k_+} \mathbb{E}\left(e^{-\Lambda|L_{1,1} - L_{1,2}|}\right) = \frac{k_-}{\Lambda} \frac{k_1}{\gamma_1} \mathbb{E}\left(e^{-\Lambda L_1}\right) = \frac{k_-}{\Lambda} \frac{k_1}{\Lambda + \gamma_1},$$

since $|L_{1,1} - L_{1,2}| \overset{\mathrm{dist}}{=} L_1$, due to the exponential assumption. This is the relation for the variance of M in Proposition 2.2.

Proof. Proposition 4.1 gives that, almost surely,

$$\mathbb{E}\left(M|\mathcal{F}^I\right) = \int_{\mathcal{S}} I(u) \mathbb{1}_{\{v \leq u < v+w\}} \nu_{\mathcal{P}}^I(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m) \tag{4.7}$$

$$= \int_{\mathbb{R} \times \mathbb{R}_+^2} I(u) \mathbb{1}_{\{v \leq u < v+w\}} k_1\, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w)$$

$$= k_1 \int_{\mathbb{R}} I(u) \mathbb{P}(E_1 \leq u < E_1 + L_1)\, \mathrm{d}u,$$

by integrating this identity, we get from Lemma 4.1,

$$\mathbb{E}(M) = \delta_+ \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\{v \leq u < v+w\}} k_1\, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w) = \delta_+ k_1 \mathbb{E}(L_1). \tag{4.8}$$

From Representation (4.1) of M and Relation (B.2) of Corollary B.1, we obtain

$$\mathbb{E}\left(M^2|\mathcal{F}^I\right) = \left(\mathbb{E}\left(M|\mathcal{F}^I\right)\right)^2 + \int_{\mathbb{R} \times \mathbb{R}_+^2} I(u) \mathbb{1}_{\{v \leq u < v+w\}} k_1\, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w),$$

and, after integration,

$$\mathbb{E}\left(M^2\right) = \mathbb{E}\left(\mathbb{E}\left(M|\mathcal{F}^I\right)^2\right) + \mathbb{E}(M). \tag{4.9}$$

Define

$$\Delta^I(M) \overset{\text{def}}{=} \mathbb{E}\left(\mathbb{E}(M|\mathcal{F}^I)^2\right) - \mathbb{E}(M)^2,$$

so that $\text{Var}(M) = \Delta^I(M) + \mathbb{E}(M)$. Relations (4.7) and (4.8) give

$$\Delta^I(M) = \mathbb{E}\left(\left(\mathbb{E}(M|\mathcal{F}^I) - \mathbb{E}(M)\right)^2\right)$$
$$= \mathbb{E}\left[\left(k_1 \int_{\mathbb{R}} (I(u) - \delta_+)\mathbb{P}(E_1 \le u < E_1 + L_1)\, \mathrm{d}u\right)^2\right].$$

We can rewrite the square of the integral as a double integral in the following way,

$$\left(\int_{\mathbb{R}} (I(u) - \delta_+)\mathbb{P}(E_1 \le u < E_1 + L_1)\, \mathrm{d}u\right)^2$$
$$= \int_{\mathbb{R}^2} (I(u) - \delta_+)(I(u') - \delta_+)\mathbb{P}(E_{1,1} \le u < E_{1,1} + L_{1,1})\mathbb{P}(E_{1,2} \le u' < E_{1,2} + L_{1,2})\, \mathrm{d}u\, \mathrm{d}u',$$

where the variables $E_{1,i}$, $i=1,2$, and $L_{1,i}$, $i=1,2$, are independent. The quantity $\Delta^I(M)$ can be expressed as

$$\Delta^I(M) = k_1^2 \int_{\mathbb{R}^2} \mathbb{E}\left[(I(u) - \delta_+)(I(u') - \delta_+)\right]$$
$$\times \mathbb{P}(E_{1,1} \le u < E_{1,1} + L_1, E_{1,2} \le u' < E_{1,2} + L_{1,2})\, \mathrm{d}u\, \mathrm{d}u'.$$

From Lemma 4.1 we get, for u, $u' \in \mathbb{R}_+$,

$$\mathbb{E}[(I(u) - \delta_+)(I(u') - \delta_+)] = \mathbb{E}(I(u)I(u')) - \delta_+^2 = \delta_+(1 - \delta_+)e^{-\Lambda|u-u'|}, \quad (4.10)$$

hence the quantity $\Delta^I(M)/(k_1^2 \delta_+(1 - \delta_+))$ is given by

$$\int_{\mathbb{R}^2} e^{-\Lambda|u-u'|}\mathbb{P}(E_{1,1} < u < E_{1,1} + L_{1,1}, E_{1,2} \le u' < E_{1,2} + L_{1,2})\, \mathrm{d}u\, \mathrm{d}u'$$
$$= \mathbb{E}\left(\int_{\mathbb{R}_+^2} e^{-\Lambda|u+E_{1,1}-(u'+E_{1,2})|}\mathbb{P}(u \le L_{1,1}, u' < L_{1,2})\, \mathrm{d}u\, \mathrm{d}u'\right)$$
$$= \mathbb{E}(L_1^2)\mathbb{E}\left(e^{-\Lambda|F_{L_1,1}+E_{1,1}-(F_{L_1,2}+E_{1,2})|}\right).$$

If we gather these results into Relation (4.9), we obtain the identity

$$\frac{\text{Var}(M)}{\mathbb{E}(M)} = 1 + \mathbb{E}(M)\frac{1 - \delta_+}{\delta_+}\mathbb{E}\left(e^{-\Lambda|F_{L_1,1}+E_{1,1}-(F_{L_1,2}+E_{1,2})|}\right),$$

which is the desired formula. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ $\square$

We now turn to a more detailed analysis of the invariant distribution of $(M(t))$ in the case when the gene is always active.

4.2. The equilibrium distribution of the state of mRNAs

It is now assumed that the process $(I(t))$ is constant and equal to 1. In particular $\delta_+{=}1$. We introduce a random measure that describes precisely the state of the mRNAs,

$$\Lambda_1(t) \overset{\text{def}}{=} \sum_{k=1}^{M(t)} \delta_{R_i(t)} \text{ and}$$

where $R_i(t)$ is the residual lifetime of the ith mRNA present at time t. The measure $\Lambda_1(t)$ is the *empirical distribution* associated to the residual lifetimes.

Proposition 4.3 (Convergence to Equilibrium of the number of mRNAs). *When the gene is always active, the process $(\Lambda_1(t))$ converges in distribution to a random measure (Λ_1^*) defined by, for f a continuous function with compact support on $\mathbb{R}_+$,*

$$\langle \Lambda_1^*, f \rangle = \int_{\mathcal{S}} f(v{+}w{-}u) \mathbb{1}_{\{v \leq u < v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w).$$

Proof. The proof is similar to the proof of Theorem 3.1. Theorem 3.26 of Dawson [9] is used, it is enough to show that for the convergence in distribution

$$\lim_{t \to +\infty} \langle \Lambda_1(t), f \rangle = \langle \Lambda_1^*, f \rangle,$$

for every non-negative continuous function f with compact support on $\mathbb{R}_+$. By definition of the vector $(R_i(t))$,

$$\langle \Lambda_1(t), f \rangle = \sum_n f(u_n{+}E_{1,n}{+}L_{1,n}{-}t) \mathbb{1}_{\{0<u_n, u_n+E_{1,n} \leq t < u_n+E_{1,n}+L_{1,n}\}}$$

$$= \int_{\mathcal{S}} f(u{+}v{+}w{-}t) \mathbb{1}_{\{0<u, u+v \leq t < u+v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w)$$

$$\overset{\text{dist}}{=} \int_{\mathcal{S}} f(u{+}v{+}w) \mathbb{1}_{\{-t<u, u+v \leq 0 < u+v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w),$$

by invariance of the Poisson point process with respect to translation by $-t$, see Proposition 3.1. We conclude that $(\Lambda_1(t))$ converges in distribution to Λ defined by

$$\langle \Lambda, f \rangle = \int_{\mathcal{S}} f(u{+}v{+}w) \mathbb{1}_{\{u+v \leq 0 < u+v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w)$$

$$\overset{\text{dist}}{=} \int_{\mathcal{S}} f(v{+}w{-}u) \mathbb{1}_{\{-u+v \leq 0 < -u+v+w\}} \mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w) = \langle \Lambda_1^*, f \rangle,$$

by invariance with respect to the mapping $(u, v, w) \mapsto (-u, v, w)$, again by using Proposition 3.1. The proposition is proved. $\qquad\square$

This proposition states that at equilibrium the number of mRNAs is Poisson with parameter $k_1 \mathbb{E}(L_1)$ and the residual lifetimes of the mRNAs are independent and distributed as F_{σ_1}.

Proposition 4.4 (Equilibrium of the state of mRNAs)**.** *When the gene is always active,*

1. *the distribution of the variable M, the number of mRNAs at equilibrium, is a Poisson distribution with parameter $\mathbb{E}(M)$ given by*

$$\mathbb{E}(M) = k_1 \mathbb{E}(L_1).$$

2. *The residual lifetimes of the mRNAs are i.i.d. with common distribution*

$$F_{L_1}(\mathrm{d}x) = \frac{\mathbb{P}(L_1 > x)}{\mathbb{E}(L_1)}\,\mathrm{d}x.$$

Proof. We calculate the Laplace transform of Λ_1^*, by using Proposition B.1 of the Appendix and the fact that $\nu_{\mathcal{P}}$ defined by Relation (3.2) is the intensity measure of the Poisson process $\mathcal{P}$, we have

$$
\mathbb{E}\left(e^{-\langle \Lambda_1^*, f\rangle}\right) = \exp\left[-\int_{\mathcal{S}}\left(1 - e^{-f(v+w-u)}\right)\nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w)\right]
$$

$$
= \exp\left[-k_1\int_{\mathcal{S}}\left(1 - e^{-f(v+w-u)}\right)\mathbb{1}_{\{v \le u < v+w\}}\,\mathrm{d}u E_1(\mathrm{d}v)L_1(\mathrm{d}w)\right]
$$

$$
= \exp\left[-k_1\mathbb{E}(L_1)\left(1 - \mathbb{E}\left(e^{-f(F_{L_1})}\right)\right)\right]
$$

By taking $f \equiv -\log(z)$ for some fixed $z \in (0,1)$, we obtain that

$$\mathbb{E}\left(z^M\right) = \exp\left[-k_1\mathbb{E}(L_1)\left(1 - z\right)\right],$$

the distribution of M is Poisson with parameter $\mathbb{E}(M) = k_1\mathbb{E}(L_1)$. Additionally, since

$$\mathbb{E}\left(e^{-\langle \Lambda_1^*, f\rangle}\right) = \sum_{n=0}^{+\infty}\frac{\mathbb{E}(M)^n}{n!}e^{-\mathbb{E}(M)}\left(\mathbb{E}\left(e^{-f(F_{L_1})}\right)\right)^n,$$

we have that Λ_1^* has the same Laplace transform, and therefore the same distribution, as the random measure

$$\sum_{i=1}^{M}\delta_{F_{L_1,i}},$$

where $(F_{L_1,i})$ is an i.i.d. sequence of random variables with distribution F_{L_1} independent of M. The proposition is proved. $\qquad\square$

5. Translation

Theorem 3.1 shows that the distribution of the number of proteins at equilibrium has the same distribution as the random variable

$$P = \int_{\mathcal{S}} I(u)G_2(u, v, w, m)\mathcal{P}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m), \tag{5.1}$$

recall that $\mathcal{S}=\mathbb{R}\times\mathbb{R}_+^2\times\mathcal{M}_p(\mathbb{R}\times\mathbb{R}_+^2)$, and

$$G_2(u,v,w,m) \stackrel{\text{def}}{=} \int_{\mathbb{R}\times\mathbb{R}_+^2} \mathbb{1}_{\left\{\begin{subarray}{l} x+v\leq u<x+v+w \\ y\leq x<y+z \end{subarray}\right\}} m(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z) \tag{5.2}$$

where $\mathcal{P}$ is the marked Poisson point process on $\mathcal{S}$ defined by Relation (3.1).

5.1. Mean and variance of the number of proteins at equilibrium

We have an explicit representation of the variance of the number of proteins at equilibrium given by the following proposition. This is an extension of the results of Fromion et al. [13], see also Leoncini [21].

Proposition 5.1 (The two first moments of the number of proteins)**.** *If P is the number of proteins at equilibrium, then*

$$\mathbb{E}(P) = \frac{k_+}{\Lambda}k_1\mathbb{E}(L_1)k_2\mathbb{E}(L_2), \tag{5.3}$$

and

$$\frac{\mathrm{Var}(P)}{\mathbb{E}(P)} = 1+k_2\mathbb{E}(L_2)\mathbb{P}\left(F_{L_1}\geq|E_{2,1}-E_{2,2}+F_{L_2,1}-F_{L_2,2}|\right)$$
$$+ \frac{k_-}{k_+}\mathbb{E}(P)\mathbb{E}\left(e^{-\Lambda|F_{L_1,1}-F_{L_1,2}+F_{L_2,1}-F_{L_2,2}+E_{1,1}-E_{1,2}+E_{2,1}-E_{2,2}|}\right), \tag{5.4}$$

where $\Lambda=k_++k_-$ and, for $a=1$ and 2, the random variables $E_{a,1}$ and $E_{a,2}$ are independent with distribution E_a, and $F_{L_a,1}$ and $F_{L_a,2}$ are independent random variables with density $\mathbb{P}(L_a\geq u)/\mathbb{E}(L_a)$ on $\mathbb{R}_+$.

The proof is given in Section C of Appendix. It follows the same arguments as in the proof of Proposition 4.2 but with a more technical framework due to the mark m, in the space $\mathcal{M}_p(\mathbb{R}\times\mathbb{R}_+^2)$, of the Poisson process $\mathcal{P}$.

If the average of P does not depend on the distributions of elongation times of mRNAs and of proteins, the second moment does depend on these distributions, and also on the distributions of lifetimes. Note however that if Λ is large, i.e. at least one of the states of the gene is changing rapidly, Relation (5.4) shows that the dependence of the distribution of the elongation time of an mRNA on the variance is small in this case.

When the elongation times are null, i.e. $E_1\equiv0$ and $E_2\equiv0$ and the lifetimes L_1 and L_2 are exponentially distributed with respective parameters γ_1 and γ_2, in this case, for $a\in\{1,2\}$, the variable F_{L_a} is also exponential since $F_{L_a}\stackrel{\text{dist}}{=}L_a$, a simple calculation gives the classical formula (2.1) of the Markovian model.

Outside the Fano parameter, the biological literature defines the *noise* associated to the production of proteins as the variance of $P/\mathbb{E}(P)$, the quantity $\mathrm{Var}(P)/\mathbb{E}(P)^2$. The above formulas and Relation (4.5) give that it can be represented as

$$\frac{\mathrm{Var}(P)}{\mathbb{E}(P)^2} = \frac{1}{\mathbb{E}(P)} + \frac{1}{\mathbb{E}(M)} \mathbb{P}\left(F_{L_1} \geq |E_{2,1} - E_{2,2} + F_{L_2,1} - F_{L_2,2}|\right)$$

$$+ \frac{k_-}{k_+} \mathbb{E}\left(e^{-\Lambda|F_{L_1,1} - F_{L_1,2} + F_{L_2,1} - F_{L_2,2} + E_{1,1} - E_{1,2} + E_{2,1} - E_{2,2}|}\right).$$

5.2. The equilibrium distribution of the number of proteins

When the gene is always active, an explicit expression of the generating function of P can be obtained. As it will be seen, its form depends on the whole distribution of the lifetimes and elongation times. Nevertheless, by using appropriate scalings, it can be used to get limit results for its distribution and therefore some insight. This is the purpose of Section 6.

Note that for classical Markovian models, without elongation times in particular, Relation (26) of Bokes et al. [3] gives, via an analytic approach, an explicit expression of the joint generating function of M in terms of an hypergeometric function.

Proposition 5.2 (Generating Function of P). *If the gene is always active, i.e. the process $(I(t))$ is constant and equal to 1, and P is the random variable defined by Relation* (3.4), *the distribution of P is given by, for $z \in [0,1]$,*

$$\mathbb{E}\left(z^P\right) = \exp\left(-\int_{\mathbb{R}_+^2} \left[1 - e^{-(1-z)k_2 \mathbb{E}(L_2)\mathbb{P}\left(F_{L_2} + E_2 \in (u-w,u)\right)}\right] k_1 \,\mathrm{d}u L_1(\mathrm{d}w)\right),$$

where F_{L_2} and L_2 are two independent random variables and the distribution of F_{L_2} has density $\mathbb{P}(L_2 \geq u)/\mathbb{E}(L_2)$ on $\mathbb{R}_+$.

Proof. Relations (5.1) and (5.2) defining P and (B.1) of Appendix for the Laplace transform of Poisson point processes give

$$\mathbb{E}\left(\exp\left(-\xi P\right)\right) = \mathbb{E}\left(\exp\left(-\xi \int_{\mathcal{S}} G_2(u,v,w,m)\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m)\right)\right)$$

$$= \exp\left(-\int_{\mathcal{S}} \left[1 - e^{-\xi G_2(u,v,w,m)}\right]\nu_{\mathcal{P}}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m)\right)$$

$$= \int_{\mathcal{S}}\left(1 - \mathbb{E}\left(e^{-\xi G_2(u,v,w,\mathcal{N}_2)}\right)\right) k_1 \,\mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w),$$

where, as before, $\mathcal{N}_2$ is a Poisson process whose distribution is Q_2.

For $(u,v,w) \in \mathbb{R} \times \mathbb{R}_+^2$, by using again Relation (B.1) for the Laplace transform of the Poisson process $\mathcal{N}_2$,

$$\mathbb{E}\left(e^{-\xi G_2(u,v,w,\mathcal{N}_2)}\right) \tag{5.5}$$

$$= \mathbb{E}\left(\exp\left(-\xi \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{\substack{x+v \leq u < x+v+w \\ y \leq x < y+z}\right\}} \mathcal{N}_2(\mathrm{d}x,\mathrm{d}y,\mathrm{d}z)\right)\right)$$

$$= \exp\left(-\left(1 - e^{-\xi}\right)\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{\substack{x+v \leq u < x+v+w \\ y \leq x < y+z}\right\}} k_2 \,\mathrm{d}x E_2(\mathrm{d}y) L_2(\mathrm{d}z)\right)$$

$$= \exp\left(-\left(1-e^{-\xi}\right)\int_{\mathbb{R}_+}\mathbb{P}\left(\begin{smallmatrix}x+E_2+v<u<x+E_2+v+w\\0\le x<L_2\end{smallmatrix}\right)k_2\,\mathrm{d}x\right)$$

$$= \exp\left(-\left(1-e^{-\xi}\right)k_2\mathbb{E}(L_2)\mathbb{P}\left(F_{L_2}+E_2+v\le u<F_{L_2}+E_2+v+w\right)\right),$$

with the same arguments as in the proof of Relation (4.6). Hence,

$$\int_{\mathcal{S}}\left[1-e^{-\xi G_2(u,v,w,m)}\right]\nu_{\mathcal{P}}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m)$$

$$= \int_{\mathcal{S}}\left[1-\exp\left(-\left(1-e^{-\xi}\right)k_2\mathbb{E}(L_2)\mathbb{P}\left(u-v-w\le F_{L_2}+E_2<u-v\right)\right)\right]$$

$$k_1\,\mathrm{d}u E_1(\mathrm{d}v)L_1(\mathrm{d}w)$$

$$= \int_{\mathcal{S}}\left[1-\exp\left(-\left(1-e^{-\xi}\right)k_2\mathbb{E}(L_2)\mathbb{P}\left(u-w\le F_{L_2}+E_2<u\right)\right)\right]k_1\,\mathrm{d}u L_1(\mathrm{d}w),$$

which gives the desired formula for the generating function. $\qquad\square$

The explicit expression of the generating function of P for Markovian models is detailed in Proposition 2.3 of Section 2.

5.3. The joint distribution of the numbers of mRNAs and proteins

This topic is investigated in few references of the literature such as Bokes et al. [3] and Taniguchi et al. [37]. Its motivation lies in the fact that it could be useful if some information on the number of mRNAs could be extracted from knowledge on the number of proteins.

We begin to study the covariance function of M and P. Relations (4.1) and (5.1) give the representations

$$\begin{cases} M = \displaystyle\int_{\mathcal{S}} I(u)G_1(u,v,w)\,\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w), \\[2em] P = \displaystyle\int_{\mathcal{S}} I(u)G_2(u,v,w,m)\,\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m), \end{cases} \tag{5.6}$$

where the functions G_1 and G_2 are defined by Relations (4.2) and (5.2). By using Proposition 4.1 as before on the conditional distribution of (M,P) with respect to $\mathcal{F}^I$, the σ-field associated to the state of the gene. See Section 4. Relation (B.2) for the covariance of functionals of Poisson processes gives the formula

$$C^I \stackrel{\mathrm{def}}{=} \mathbb{E}\left(MP|\mathcal{F}^I\right) - \mathbb{E}\left(M|\mathcal{F}^I\right)\mathbb{E}\left(P|\mathcal{F}^I\right)$$

$$= \int_{\mathcal{S}} I(u)G_1(u,v,w,m)G_2(u,v,w,m)\nu_{\mathcal{P}}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m).$$

By integrating this identity and Relations (4.2) and (5.2) defining G_1 and G_2, we obtain that $\mathbb{E}(C^I)$ is equal to

$$\int_{\mathcal{S}} \mathbb{E}(I(u))\mathbb{1}_{\{v \leq u < v+w\}} \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{\substack{x+v \leq u < x+v+w \\ y \leq x < y+z}\right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)\nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)$$

$$=\delta_+ \int_{\mathcal{S}} \mathbb{1}_{\{v \leq u < v+w\}} \mathbb{E}\left(\int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{\substack{x+v \leq u < x+v+w \\ y \leq x < y+z}\right\}} \mathcal{N}_2(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)\right)$$
$$k_1 \, \mathrm{d}u E_1(\mathrm{d}v)L_1(\mathrm{d}w)$$

$$= \delta_+ k_1 k_2 \mathbb{E}\left(\int \mathbb{1}_{\{0 \leq u < L_1\}} \int_{\mathbb{R}^2} \mathbb{1}_{\left\{\substack{x \leq u < x+L_1 \\ E_2 \leq x < E_2+L_2}\right\}} \, \mathrm{d}x \, \mathrm{d}u\right)$$

$$= \delta_+ k_1 k_2 \mathbb{E}(L_2)\mathbb{E}\left(\left(L_1 - F_{L_2} - E_2\right)^+\right),$$

hence, with Relation (1.1),

$$\mathbb{E}(C^I) = \mathbb{E}(P)\mathbb{P}\left(F_{L_1} \geq F_{L_2} + E_2\right).$$

Concerning the product of conditional expectations in the definition of C^I, the following identities have established in the proof of Propositions 4.2 and 5.1,

$$\mathbb{E}\left(M|\mathcal{F}^I\right) = k_1 \int_{\mathbb{R}} I(u)\mathbb{P}(E_1 < u < E_1 + L_1) \, \mathrm{d}u,$$

$$\mathbb{E}\left(P|\mathcal{F}^I\right) = k_1 k_2 \int_{\mathbb{R}} I(u)\mathbb{E}\left(\int_{\mathbb{R}} \mathbb{1}_{\left\{\substack{x+E_1 \leq u < x+E_1+L_1 \\ E_2 \leq x < E_2+L_2}\right\}} \, \mathrm{d}x\right) \mathrm{d}u.$$

Using the same argument as before to express a product of integrals as a double integral, this gives the identity

$$\mathbb{E}\left(\mathbb{E}\left(M|\mathcal{F}^I\right)\mathbb{E}\left(P|\mathcal{F}^I\right)\right)/(k_1^2 k_2)$$
$$= \int_{\mathbb{R}^3} \mathbb{E}\left(I(u)I(u')\right)\mathbb{E}\left(\mathbb{1}_{\{E_{1,1} \leq u < E_{1,1}+L_{1,1}\}}\mathbb{1}_{\left\{\substack{x+E_{1,2} \leq u' < x+E_{1,2}+L_{1,2} \\ E_2 \leq x < E_2+L_2}\right\}}\right) \mathrm{d}x \, \mathrm{d}u' \, \mathrm{d}u.$$

By using Corollary 4.1 again, $\mathbb{E}[I(u)I(u')] = \delta_+(\delta_+ + (1-\delta_+)\exp(\Lambda|u-u'|)$, for u, $u' \geq 0$, the last expression is the sum of two terms,

$$\delta_+^2 \int_{\mathbb{R}^3} \mathbb{E}\left(\mathbb{1}_{\{0 \leq u < L_{1,1}\}}\mathbb{1}_{\left\{\substack{0 \leq u' < L_{1,2} \\ 0 \leq x < L_2}\right\}}\right) \mathrm{d}x \, \mathrm{d}u' \, \mathrm{d}u = \delta_+^2 \mathbb{E}(L_1)^2 \mathbb{E}(L_2),$$

and

$$\delta_+(1-\delta_+)\int_{\mathbb{R}^3} e^{-\Lambda|u-u'|}$$

$$\mathbb{E}\left(\mathbb{1}_{\{E_{1,1} \leq u < E_{1,1}+L_{1,1}\}}\mathbb{1}_{\left\{\substack{x+E_{1,2} \leq u' < x+E_{1,2}+L_{1,2} \\ E_2 \leq x < E_2+L_2}\right\}}\right) \mathrm{d}x \, \mathrm{d}u' \, \mathrm{d}u,$$

which is, after the same manipulations as in the proof of Proposition 5.1,

$$\delta_+(1-\delta_+)\mathbb{E}(L_1)^2\mathbb{E}(L_2)\mathbb{E}\left(e^{-\Lambda|F_{L_1,2}-F_{L_1,1}+F_{L_2}+E_2+E_{1,2}-E_{1,1})|}\right).$$

We have therefore proved the following proposition.

Proposition 5.3 (Covariance of M and P). *If M and P are the random variables defined by Relation* (3.3) *and* (3.4)*, then the covariance of M and P is given by*

$$\frac{\text{Cov}(M,P)}{\mathbb{E}(P)} = \mathbb{P}\left(F_{L_1}\geq F_{L_2}+E_2\right)+\mathbb{E}(M)$$
$$+\frac{k_-}{k_+}\mathbb{E}(M)\mathbb{E}\left(e^{-\Lambda|F_{L_1,2}-F_{L_1,1}+E_{1,2}-E_{1,1}+F_{L_2}+E_2|}\right),\quad (5.7)$$

for $E_{1,1}$ and $E_{1,2}$, are independent random variables with distribution E_1, and $F_{L_1,1}$ and $F_{L_1,2}$ are independent random variables with density $\mathbb{P}(L_1\geq u)/\mathbb{E}(L_1)$ on $\mathbb{R}_+$, F_{L_2} is a random variable with density $\mathbb{P}(L_2\geq u)/\mathbb{E}(L_2)$ on $\mathbb{R}_+$.

The lifetime L_1 of the mRNAs is usually much smaller than the lifetime L_2 of proteins, this implies that the same property also holds for the variables F_{L_1} and F_{L_2}. If, additionally, the average number of mRNAs $\mathbb{E}(M)$ is small, the right-hand-side of Relation (5.7) is small, so that $\text{Cov}(M,P)$ should be close to 0. This property has already been noticed in some measurements, see Taniguchi et al. [37] for example. We conclude this section with an explicit representation of the generating function of the random variable (M,P) when the gene is always active.

Proposition 5.4 (Joint Distribution of M and P). *When the gene is always active, for z_1 and $z_2\in(0,1)$,*

$$\mathbb{E}\left(z_1^M z_2^P\right)$$
$$= \exp\left[-k_1\mathbb{E}(L_1)\int_{\mathbb{R}_+}\left(1-z_1\exp\left[-(1-z_2)\,k_2\mathbb{E}(L_2)\mathbb{P}\left(F_{L_2}+E_2\leq u\right)\right]\right)F_{L_1}(\mathrm{d}u)\right.$$
$$\left.+k_1\int_{\mathbb{R}_+^2}\left(1-\exp\left[-(1-z_2)\,k_2\mathbb{E}(L_2)\mathbb{P}\left(F_{L_2}+E_2\in(u,u+w)\right)\right]\right)\mathrm{d}uL_1(\mathrm{d}w)\right],\quad (5.8)$$

where F_{L_1}, F_{L_2} and E_2 are independent random variables and, for $a\in\{1,2\}$, the random variable F_{L_a} has the density $\mathbb{P}(L_a\geq u)/\mathbb{E}(L_a)$ on $\mathbb{R}_+$.

Proof. For a_1, $a_2>0$, Relation (5.6) gives that

$$a_1M+a_2P = \int_{\mathcal{S}} F(u,v,w,m)\,\mathcal{P}(\mathrm{d}u,\mathrm{d}v,\mathrm{d}w,\mathrm{d}m)$$

where

$$F(u,v,w,m) \stackrel{\text{def}}{=} a_1 G_1(u,v,w) + a_2 G_2(u,v,w,m)$$

$$= a_1 \mathbb{1}_{\{v \le u < v+w\}} + a_2 \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \substack{x+v \le u < x+v+w \\ y \le x < y+z} \right\}} m(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z).$$

We have, by Relation (B.1) of the Appendix,

$$\mathbb{E}\left(e^{-a_1 M - a_2 P}\right) = \exp\left(-\int_{\mathcal{S}} \left(1 - e^{-F(u,v,w,m)}\right) \nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m)\right),$$

hence,

$$-\log\left(\mathbb{E}\left(e^{-a_1 M - a_2 P}\right)\right)$$
$$= \int_{\mathbb{R} \times \mathbb{R}_+^2} \left(1 - e^{-a_1 G_1(u,v,w)} \mathbb{E}\left(e^{-a_2 G_2(u,v,w,\mathcal{N}_2)}\right)\right) k_1 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w),$$

and, if we use Relation (5.5),

$$\mathbb{E}\left(\left(e^{-a_2 G_2(u,v,w,\mathcal{N}_2)}\right)\right) = \exp\left(-\left(1 - e^{-a_2}\right) k_2 \mathbb{E}(L_2) \mathbb{P}\left(F_{L_2} + E_2 \in (u-v-w, u-v)\right)\right),$$

hence

$$-\log\left(\mathbb{E}\left(e^{-a_1 M - a_2 P}\right)\right) = \int_{\mathbb{R} \times \mathbb{R}_+^2} \left(1 - \exp\left[-a_1 \mathbb{1}_{\{v \le u < v+w\}}\right.\right.$$
$$\left.\left. - \left(1 - e^{-a_2}\right) k_2 \mathbb{E}(L_2) \mathbb{P}\left(F_{L_2} + E_2 \in (u-v-w, u-v)\right)\right]\right) k_1 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w).$$

We conclude the proof with standard calculations. □

6. Convergence results

The purpose of this section is of revisiting, via rigorous convergence theorems, several classical research topics of the literature on the stochasticity of the gene expression. Most of studies in this domain take place in the Markovian setting of Section 2 without the elongation of mRNAs and proteins. Kolmogorov's equations associated to the Markov process, the "master equation" as it is usually presented, are the starting point of these studies.

As the equilibrium equations cannot be solved explicitly, several scenarios are investigated, via scalings, to get insight on the equilibrium distribution of the number of proteins: Fast Switching rates between active and inactive states, mRNAs with short lifetimes, proteins with long lifetimes, gene with large expression rates, ... See Friedman et al. [12], Raj et al. [31], Shahrezaei and Swain [35], Swain et al. [36] and Thattai and van Oudenaarden [38]. See Bokes et al. [3] for a quite extensive analysis of the Markovian model. We will look at three different scaling situations in the light of the results derived in the previous sections.

1. Long-Lived Proteins.
   The distribution of lifetimes of proteins is an exponential distribution with parameter γ_2, and its mean is converging to infinity. A central limit theorem is proved for the equilibrium of the number P of proteins. It is shown that, in distribution, $P \sim \mathbb{E}(P) + \sqrt{\mathbb{E}(P)}\mathcal{N}$, where $\mathcal{N}$ is a centered Gaussian random variable.
2. Short-Lived mRNAs and Long-Lived Proteins.
   The distribution of lifetimes of mRNAs and proteins are exponential with respective parameters γ_1 and γ_2, the mean lifetime $1/\gamma_1$ of mRNAs is converging to 0 and the death rate γ_2 of proteins is fixed so that $k_1 k_2/(\gamma_1 \gamma_2)$, the mean number of proteins at equilibrium, is fixed.
3. Short-Lived mRNAs and High Expression Rate of Proteins.
   The distribution of lifetimes of mRNAs and proteins are exponential with respective parameters γ_1 and γ_2. The mean lifetime of mRNAs is converging to 0 and the translation rate k_2 of proteins is fixed so that the quantity k_2/γ_1, the average number of proteins produced by an mRNA, is fixed.

For several of these regimes, in a Markovian setting, convergence results of this section can be seen in the context of a stochastic averaging principle: the system is driven by a fast process, associated with the mRNAs (for example), and by a slow process describing the time evolution of proteins. In the limit, the time evolution on finite time intervals of proteins can be described as if the fast process is at equilibrium at any time. A typical result states that this picture also holds for the equilibrium of the slow process, but this is a more difficult result to prove in general. See Kurtz [20] for example. A similar observation could also be done for the central limit theorem 6.1, see Chapter 4 of Anderson and Kurtz [1]. A direct approach is used here to prove convergence results by taking advantage of the explicit representation of equilibrium of Theorem 3.1.

In most cases, the gene will be assumed to be always active. The reason is that, if some results could be obtained on the distribution of P, we have not been able to get usable expressions to get some insight in the scaling regimes analyzed here. The next result shows that, in practice, a model with a permanently active gene is reasonable, provided some parameters are adjusted. This is a folk result of the biological literature: if the state of the gene switches more and more rapidly, then, in the limit, the model is equivalent to a model with a permanently active gene but with a reduced transcription rate.

6.1. Fast switching rates of gene between active and inactive states

Proposition 4.3 shows that when the gene is always active then the equilibrium distribution of the number of mRNAs is Poisson. This is minimal from the point of view of the Fano factor in this class of models. If the gene switches quickly between the two states active/inactive, one can expect an averaging effect. The following proposition establishes this intuitive result.

Proposition 6.1. *If P_N is the random variable defined by Relation* (3.4) *with the activation/deactivation rates are respectively given by $k_+ N$ and $k_- N$ for*

some scaling parameter $N \geq 1$, then, for the convergence in distribution,

$$\lim_{N \to +\infty} P_N = P,$$

where the distribution of P is the equilibrium distribution of the number of proteins for a model of gene expression for which the gene is always active and with the same parameters except for k_1 which is replaced by $k_1 k_+/(k_+ + k_-)$.

Proof. The stationary activation/deactivation process $(I_N(t))$ can clearly be expressed as $(I(Nt))$ where $(I(t))$ is a stationary process with switching rates k_+ and k_-

Let f be an integrable function on $\mathbb{R}_+$ and

$$X_N \stackrel{\text{def}}{=} \int_{\mathbb{R}_+} I(Nu) f(u) \ du,$$

we now prove the convergence in distribution

$$\lim_{N \to +\infty} X_N = \delta_+ \int f(u) \ du,$$

with $\delta_+ \stackrel{\text{def}}{=} k_+/(k_+ + k_-)$. Note that the convergence is certainly true for the first moment of X_N, since $\mathbb{E}(I(t)) = \delta_+$ for all $t \geq 0$.

By writing the square of X_N as a double integral, with Proposition 2.2 we obtain the relation

$$\mathbb{E}(X_N^2) = \mathbb{E}\left(\int_{\mathbb{R}_+^2} I(Nu) I(Nu') f(u) f(u') \ du \, du' \right)$$
$$= \left(\delta_+ \int f(u) \ du \right)^2 + \delta_+(1 - \delta_+) \int_{\mathbb{R}_+^2} e^{-N(k_+ + k_-)|u - u'|} f(u) f(u') \ du \, du'.$$

Therefore,

$$\lim_{N \to +\infty} \mathbb{E}\left((X_N - \mathbb{E}(X_N))^2 \right) = \lim_{N \to +\infty} \mathbb{E}\left(X_N^2 \right) - (\mathbb{E}(X_N))^2 = 0,$$

in particular, (X_N) converges in distribution to the limit of the sequence $(\mathbb{E}(X_N))$.

Denote by $\mathcal{F}^{I_N}$ the filtration associated to the process $(I(Nt))$. Relations (5.1) and (5.2) for P_N, and (B.1) of the Appendix for the Laplace transform of Poisson point processes give

$$\mathbb{E}\left(\exp\left(-\xi P_N\right) \big| \mathcal{F}^{I_N} \right)$$
$$= \mathbb{E}\left(\exp\left(-\xi \int_{\mathcal{S}} I(Nu) G_2(u, v, w, m) \mathcal{P}(du, dv, dw, dm) \right) \right)$$
$$= \exp\left(-\int_{\mathcal{S}} I(Nu) \left[1 - e^{-\xi G_2(u,v,w,m)} \right] \nu_{\mathcal{P}}(du, dv, dw, dm) \right)$$

$$= \exp\left(-\int_{\mathbb{R}} I(Nu)\left[1 - \mathbb{E}\left(e^{-\xi G_2(u, E_1, L_1, \mathcal{N}_2)}\right)\right] k_1 \, \mathrm{d}u\right),$$

where $\mathcal{N}_2$ is a Poisson process whose distribution is Q_2 defined by Relation (3.2).

By using the convergence result, we obtain that

$$\lim_{N \to +\infty} \mathbb{E}\left(\exp\left(-\xi P_N\right)\right) = \exp\left(-\int_{\mathbb{R}}\left[1 - \mathbb{E}\left(e^{-\xi G_2(u, E_1, L_1, \mathcal{N}_2)}\right)\right] \delta_+ k_1 \, \mathrm{d}u\right),$$

which is the Laplace transform of the variable P associated to a Poisson process with the same characteristics as $\mathcal{P}$ except that the transcription rate is $\delta_+ k_1$ and that the gene is always active. □

6.2. Asymptotic behavior of the equilibrium distribution

In this section, the invariant distribution of the number of proteins when the gene is always active is analyzed under some scaling conditions. Several of them rely on the fact that the average lifetime of a protein is much larger than the lifetime of an mRNA, which is of the order of 2 mn for an mRNA, where, for a protein, it is at least 30 mn. See the numbers of Section A of the Appendix. This approach is used in the literature to get a further insight on the distribution of P, i.e. with more information than the first two moments that have an explicit expression. It is usually done via approximations on the equation satisfied by the generating function of P. See Bokes et al. [3] or Shahrezaei and Swain [35] where this is done via an expansion of an hypergeometric function. The representation of Proposition 5.2 of the generating function of P will allow to get convergence results for the distribution of P in a quite general case and without too much technicality.

A limiting Gaussian distribution for the number of proteins.
The scaling considered here assumes that the average lifetime of proteins goes to infinity and the other parameters are fixed. In particular these proteins should be numerous within the cell. This setting is well suited for protein types having a large number of copies, of the order at least of 10,000 for example. The following result shows that a central limit theorem holds in this context.

Theorem 6.1 (Central Limit Theorem). *Under the conditions*

1. *the gene is always active;*
2. *the distribution L_2 of lifetimes of proteins is exponential with parameter γ_2;*
3. *the parameters k_1, k_2 are fixed as well as the distribution L_1 of the lifetimes of mRNAs and the respective distributions E_1 and E_2 of the elongation times of mRNAs and proteins. The distribution L_1 of lifetimes of mRNAs has a finite third moment, $\mathbb{E}(L_1^3) < +\infty$,*

if P is the random variable defined by Relation (3.4), then for the convergence in distribution,

$$\lim_{\gamma_2 \to 0} \sqrt{\gamma_2}\left(P - \mathbb{E}(P)\right) = \mathcal{N}\left(\sigma\right),$$

where

$$\sigma = \sqrt{k_1 \mathbb{E}(L_1) k_2} \left(1 + k_2 \int_{\mathbb{R}_+} \mathbb{P}\left(F_{L_1} \geq |E_{2,1} - E_{2,2} + x| \right) \, \mathrm{d}x \right)^{1/2},$$

and, for $a>0$, $\mathcal{N}(a)$ is a centered Gaussian random variable with standard deviation a.

Despite this framework is quite natural, curiously it does not seem to have been investigated in the biological literature, even for the classical Markovian three-step model. It gives in fact a Gaussian approximation for the number of proteins at equilibrium when the average lifetime $1/\gamma_2$ of proteins is large. It should be noted that, in this scaling regime, the second order does depend on the distribution of the elongation times of proteins.

Proof. By using Proposition 5.2, the relation $F_{L_2} \overset{\mathrm{dist}}{=} L_2$ due to the exponential assumption for L_2, and the calculation of the proof of Relation (5.3) of Proposition 5.1, we get that, for $\xi>0$,

$$\log\left[\mathbb{E}\left(\exp\left(-\sqrt{\gamma_2}\xi(P - \mathbb{E}(P)) \right) \right) \right] = -\int_{\mathbb{R}_+^2} I_{\gamma_2}(u, w) k_1 \, \mathrm{d}u L_1(\mathrm{d}w),$$

with, for $(u, w) \in \mathbb{R}_+^2$,

$$I_{\gamma_2}(u, w) \overset{\mathrm{def}}{=} 1 - e^{-\left(1 - e^{-\sqrt{\gamma_2}\xi}\right) \frac{k_2}{\gamma_2} \mathbb{P}(L_2 + E_2 \in (u-w, u))} - \xi \frac{k_2}{\sqrt{\gamma_2}} \mathbb{P}\left(L_2 + E_2 \in (u-w, u) \right).$$

The elementary inequality

$$\left| 1 - e^{-x} - x + \frac{x^2}{2} \right| \leq \frac{x^3}{6}, \quad \text{for } x = \left(1 - e^{-\xi\sqrt{\gamma_2}} \right) \frac{k_2}{\gamma_2} \mathbb{P}\left(L_2 + E_2 \in (u-w, u) \right) \quad (6.1)$$

gives the relation

$$\left| I_{\gamma_2}(u, v) - \left(1 - e^{-\xi\sqrt{\gamma_2}} - \xi\sqrt{\gamma_2} + \frac{1}{2}\xi^2\gamma_2 \right) \frac{k_2}{\gamma_2} \mathbb{P}\left(L_2 + E_2 \in (u-w, u) \right) \right.$$

$$\left. + \frac{1}{2}\xi^2\gamma_2 \frac{k_2}{\gamma_2} \mathbb{P}\left(L_2 + E_2 \in (u-w, u) \right) + \frac{1}{2}\left(1 - e^{-\xi\sqrt{\gamma_2}} \right)^2 \left(\frac{k_2}{\gamma_2} \right)^2 \mathbb{P}(L_2 + E_2 \in (u-w, u))^2 \right|$$

$$\leq \frac{1}{6}\left(1 - e^{-\xi\sqrt{\gamma_2}} \right)^3 \left(\frac{k_2}{\gamma_2} \right)^3 \mathbb{P}\left(L_2 + E_2 \in (u-w, u) \right)^3. \quad (6.2)$$

After integration with respect to the measure $k_1 \, \mathrm{d}u L_1(\mathrm{d}w)$ on $\mathbb{R}_+^2$ and by using the relation

$$\left| 1 - e^{-\xi\sqrt{\gamma_2}} - \xi\sqrt{\gamma_2} + \frac{1}{2}\xi^2\gamma_2 \right| \leq \frac{1}{6}\xi^3\gamma_2^{3/2},$$

we get the inequality

$$\left| \log \left[\mathbb{E} \left(\exp \left(- \sqrt{\gamma_2} \xi (P - \mathbb{E}(P)) \right) \right) \right] - A_1 - A_2(\gamma_2) \right|$$
$$\leq \frac{k_1}{6} \left(k_2^3 B_1(\gamma_2) + k_2 \xi^3 B_2(\gamma_2) \right), \quad (6.3)$$

with

$$A_1 \overset{\text{def}}{=} \frac{1}{2} k_1 k_2 \xi^2 \int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + E_2 \in (u - w, u) \right) \mathrm{d}u L_1(\mathrm{d}w),$$

$$A_2(\gamma_2) \overset{\text{def}}{=} \frac{1}{2} k_1 \left(1 - e^{-\xi \sqrt{\gamma_2}} \right)^2 \left(\frac{k_2}{\gamma_2} \right)^2 \int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + E_2 \in (u - w, u) \right)^2 \mathrm{d}u L_1(\mathrm{d}w),$$

$$B_1(\gamma_2) \overset{\text{def}}{=} \left(1 - e^{-\xi \sqrt{\gamma_2}} \right)^3 \frac{1}{\gamma_2^3} \int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + B_2 \in (u - w, u) \right)^3 \mathrm{d}u L_1(\mathrm{d}w),$$

$$B_2(\gamma_2) \overset{\text{def}}{=} \sqrt{\gamma_2} \int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + E_2 \in (u - w, u) \right) \mathrm{d}u L_1(\mathrm{d}w).$$

We now study the behavior of each of these four terms when γ_2 goes to 0. For the first term A_1, this is is straightforward since

$$A_1 = \frac{1}{2} k_1 k_2 \xi^2 \int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + E_2 \leq u \leq L_1 + L_2 + E_2 \right) \mathrm{d}u = \frac{1}{2} k_1 k_2 \xi^2 \mathbb{E}(L_1). \quad (6.4)$$

The second term $A_2(\gamma_2)$ is handled in the same way as in the derivation of Relation (5.4) of Proposition 5.1. First note that

$$\int_{\mathbb{R}_+^2} \mathbb{P} \left(L_2 + E_2 \in (u - w, u) \right)^2 \mathrm{d}u L_1(\mathrm{d}w)$$
$$= \int_{\mathbb{R}_+} \mathbb{E} \left((w - |L_{2,1} - L_{2,1} + E_{2,2} - L_{2,2}|)^+ \right) \mathrm{d}u L_1(\mathrm{d}w)$$
$$= \mathbb{E}(L_1) \mathbb{P} \left(F_{L_1} \geq |L_{2,1} - L_{2,1} + E_{2,2} - L_{2,2}| \right)$$
$$= \mathbb{E}(L_1) \int_0^{+\infty} \mathbb{P} \left(F_{L_1} \geq |x + E_{2,1} - E_{2,2}| \right) \gamma_2 e^{-\gamma_2 x} \, \mathrm{d}x,$$

with the help of the symmetry $E_{2,1} - E_{2,2} \overset{\text{dist}}{=} E_{2,2} - E_{2,1}$. Note that

$$\int_0^{+\infty} \mathbb{P} \left(F_{L_1} \geq |x + E_{2,1} - E_{2,2}| \right) \mathrm{d}x \leq \int_0^{+\infty} \mathbb{P} \left(x \leq F_{L_1} + |E_{2,1} - E_{2,2}| \right) \mathrm{d}x$$
$$\leq \mathbb{E} \left(F_{L_1} \right) + 2 \mathbb{E}(E_2) = \frac{\mathbb{E}(L_1^2)}{2 \mathbb{E}(L_1)} + 2 \mathbb{E}(E_2) < +\infty,$$

Lebesgue's Theorem gives therefore the relation

$$\lim_{\gamma_2 \to 0} A_2(\gamma_2) = \frac{1}{2}\xi^2 k_1 \mathbb{E}(L_1)k_2^2 \lim_{\gamma_2 \to 0} \int_0^{+\infty} \mathbb{P}\left(F_{L_1} \geq |x + E_{2,1} - E_{2,2}|\right) e^{-\gamma_2 x} \, \mathrm{d}x$$

$$= \frac{1}{2}\xi^2 k_1 \mathbb{E}(L_1)k_2^2 \int_0^{+\infty} \mathbb{P}\left(F_{L_1} \geq |x + E_{2,1} - E_{2,2}|\right) \mathrm{d}x. \qquad (6.5)$$

We now analyze the upper bound of Relation (6.3) expressed with the terms $B_1(\gamma_2)$ and $B_2(\gamma_2)$. The last one is easy to handle

$$B_2(\gamma_2) = \sqrt{\gamma_2} \int_{\mathbb{R}^2_+} \mathbb{P}\left(L_2 + E_2 \in (u - w, u)\right) \mathrm{d}u L_1(\mathrm{d}w) = \mathbb{E}(L_1)\sqrt{\gamma_2},$$

it converges to 0 as γ_2 goes to 0.

For the term $B_1(\gamma_2)$, Jensen's Inequality and Fubini's Theorem give

$$\int_{\mathbb{R}^2_+} \mathbb{P}\left(L_2 + E_2 \in (u - w, u)\right)^3 \mathrm{d}u L_1(\mathrm{d}w)$$

$$= \int_{\mathbb{R}^2_+} \left(\mathbb{E}\left(e^{-\gamma_2(u - w - E_2)^+} - e^{-\gamma_2(u - E_2)^+}\right)\right)^3 \mathrm{d}u L_1(\mathrm{d}w)$$

$$\leq \int_{\mathbb{R}^2_+} \mathbb{E}\left(\left(e^{-\gamma_2(u - w - E_2)^+} - e^{-\gamma_2(u - E_2)^+}\right)^3\right) \mathrm{d}u L_1(\mathrm{d}w)$$

$$= \int_{\mathbb{R}^2_+} \left(e^{-\gamma_2(u - w)^+} - e^{-\gamma_2 u}\right)^3 \mathrm{d}u L_1(\mathrm{d}w).$$

With simple calculations, we get that, for $w > 0$, the quantity

$$\int_{\mathbb{R}_+} \left(e^{-\gamma_2(u - w)^+} - e^{-\gamma_2 u}\right)^3 \mathrm{d}u$$

is the sum of two terms, C_1 and C_2, with

$$C_1 \stackrel{\mathrm{def}}{=} \int_w^\infty \left(e^{-\gamma_2(u - w)} - e^{-\gamma_2 u}\right)^3 \mathrm{d}u = \frac{(1 - e^{-w\gamma_2})^3}{3\gamma_2},$$

$$C_2 \stackrel{\mathrm{def}}{=} \int_0^w \left(1 - e^{-\gamma_2 u}\right)^3 \mathrm{d}u = \frac{2e^{-3w\gamma_2} - 9e^{-2wa} + 6w\gamma_2 + 18e^{-w\gamma_2} - 11}{6\gamma_2}.$$

The elementary inequality (6.1) gives the relations

$$C_1 \leq \gamma_2^2 \frac{w^3}{3} \text{ and } |C_2| \leq 4\gamma_2^2 w^3.$$

By using the fact that L_1 has a finite third moment, we deduce the relation

$$\lim_{\gamma_2 \to 0} \left(1 - e^{-\xi\sqrt{\gamma_2}}\right)^3 \left(\frac{k_2}{\gamma_2}\right)^3 \int_{\mathbb{R}^2_+} \mathbb{P}\left(L_2 + E_2 \in (u - w, u)\right)^3 \mathrm{d}u L_1(\mathrm{d}w) = 0.$$

Relations(6.3), (6.4) and (6.5) give therefore the convergence

$$\lim_{\gamma_2 \to 0} \log\left[\mathbb{E}\left(\exp(-\sqrt{\gamma_2}\xi(P-\mathbb{E}(P)))\right)\right]$$

$$= \frac{\xi^2}{2}k_1k_2\mathbb{E}(L_1)\left(1+k_2\int_0^{+\infty}\mathbb{P}\left(F_{L_1}\geq|x+E_{2,1}-E_{2,2}|\right)\mathrm{d}x\right).$$

The theorem is proved.   □

Short-lived mRNAs and long-lived proteins.
For the scaling considered in this section the death rate γ_2 of proteins goes to 0 and with the constraint that the death rate γ_1 of mRNAs is such that the average number of proteins is kept fixed and equal to $a>0$,

$$k_1\mathbb{E}(L_1)k_2\mathbb{E}(L_2) = \frac{k_1}{\gamma_1}\frac{k_2}{\gamma_2} = a.$$

In particular the average lifetime $1/\gamma_1$ of an mRNA goes to 0.

Proposition 6.2. *Under the conditions*

1. *the gene is always active;*
2. *the distribution of lifetimes of mRNAs and proteins are exponential with respective parameters γ_1 and γ_2;*
3. *the transcription and translation parameters k_1 and $k_2>0$ are fixed as well as the distributions of the elongation times,*

if P is the random variable defined by Relation (3.4), *then for $a>0$, the convergence in distribution,*

$$\lim_{\substack{\gamma_1 \to +\infty \\ \gamma_1\gamma_2=k_1k_2/a}} P = \mathrm{Pois}(a),$$

where $\mathrm{Pois}(a)$ *is the Poisson distribution with parameter a.*

Note that Proposition 2.2 gives readily that

$$\lim_{\substack{\gamma_1 \to +\infty \\ \gamma_1\gamma_2=k_1k_2/a}}\frac{\mathrm{Var}(P)}{\mathbb{E}(P)} = 1.$$

From the point of view of the Fano factor, the fluctuations of the number of proteins are minimal. A (rough) picture of that is that short-lived mRNAs and almost permanent proteins minimize fluctuations of gene expression. The proposition is more precise since it states that the number of proteins at equilibrium is in fact asymptotically Poisson.

Proof. In view of Proposition 5.2 and, since L_2 is exponentially distributed, $L_2 \overset{\mathrm{dist}}{=} F_{L_2}$, all we have to do is that the quantity

$$\Delta \overset{\mathrm{def}}{=} \int_{\mathbb{R}_+^2}\left[1-\exp\left(-(1-z)\,k_2\mathbb{E}(L_2)\mathbb{P}\left(L_2+E_2\in((u-w)^+,u)\right)\right)\right]k_1\,\mathrm{d}uL_1(\mathrm{d}w)$$

converges to $(1-z)a$, as γ_1 goes to infinity, with the constraint $\gamma_1\gamma_2=k_1k_2/a$.

With the elementary inequality (6.1), we get

$$|\Delta - (1-z)k_1 k_2 \Delta_1| \le \frac{(1-z)^2 k_1 k_2^2}{2}\Delta_2,$$

with

$$\Delta_1 \stackrel{\text{def}}{=} \int_{\mathbb{R}_+^2} \frac{1}{\gamma_2} \mathbb{P}\left(L_2 + E_2 \in ((u-w)^+, u)\right) \, du\gamma_1 e^{-\gamma_1 w} \, dw$$

and

$$\Delta_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}_+^2} \left(\frac{\mathbb{P}\left(L_2 + E_2 \in ((u-w)^+, u)\right)}{\gamma_2}\right)^2 \, du\gamma_1 e^{-\gamma_1 w} \, dw.$$

We now take care of these two quantities. By invariance by translation, we have

$$\Delta_1 = \frac{1}{\gamma_2}\mathbb{E}\left(\int \left(e^{-\gamma_2(u-E_2-w)^+} - e^{-\gamma_2(u-E_2)^+}\right) du\gamma_1 e^{-\gamma_1 w} \, dw\right)$$
$$= \frac{1}{\gamma_2}\mathbb{E}\left(\int \left(e^{-\gamma_2(u-w)^+} - e^{-\gamma_2 u}\right) du\gamma_1 e^{-\gamma_1 w} \, dw\right) = \frac{1}{\gamma_1\gamma_2}.$$

The Cauchy-Shwartz Inequality gives for the second term

$$\Delta_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}_+^2} \left[\frac{1}{\gamma_2}\mathbb{E}\left(\left(e^{-\gamma_2(u-E_2-w)^+} - e^{-\gamma_2(u-E_2)^+}\right)\right)\right]^2 du\gamma_1 e^{-\gamma_1 w} \, dw$$
$$\le \mathbb{E}\left(\int_{\mathbb{R}_+^2} \frac{1}{\gamma_2^2}\left(e^{-\gamma_2(u-w)^+} - e^{-\gamma_2 u}\right)^2 du\gamma_1 e^{-\gamma_1 w} \, dw\right) = \frac{1}{\gamma_1\gamma_2(\gamma_1+\gamma_2)}$$

hence it converges to 0 as γ_1 goes to infinity with $\gamma_1\gamma_2$ being constant. The proposition is proved by gathering these results. $\square$

The above result can be explained quite simply as follows. We assume that the elongation times are null for simplicity. With the "coupling from the past" approach of Section 3.3, if (u_n) denotes the non-decreasing sequence of instants of binding instants before time 0, since the distribution of the sequence of life-times $(L_{1,n})$ of mRNAs is converging to 0, with high probability there should at most one protein created by the mRNA with index n, so that

$$P \sim R \stackrel{\text{def}}{=} \sum_{n<0} \mathbb{1}_{\{\mathcal{N}_2([u_n+E_{1,n}, u_n+E_{1,n}+L_{1,n}]) \ne 0, u_n+E_{1,n}+L_{2,n}>0\}},$$

and a simple calculation gives

$$\mathbb{E}\left(z^R\right) = \exp\left(-(1-z)\frac{k_2 k_1}{\gamma_2(\gamma_1+k_2)}\right) \sim \exp\left(-(1-z)\frac{k_2 k_1}{\gamma_2\gamma_1}\right),$$

as γ_2 goes to 0 with $\gamma_1\gamma_2$ constant. The distribution of the variable R is asymptotically Poisson with parameter $k_2 k_1/(\gamma_2\gamma_1)$. It is not difficult to see that, with convenient estimates, a rigorous proof could be obtained by using this observation.

Short-lived mRNAs and high expression rate of proteins.
In this section the death rate γ_1 of proteins goes to infinity but it is assumed in addition that the quantity k_2/γ_1, the average number of proteins translated by an mRNA, is constant. In particular k_2, the expression rate of proteins, is large and the average number of proteins is fixed.

Proposition 6.3. *Under the conditions*

1. *the gene is always active;*
2. *the distribution of lifetimes of mRNAs and proteins are exponential with respective parameters γ_1 and γ_2;*
3. *the parameters k_1 and $\gamma_2 > 0$ are fixed as well as the distributions of the elongation times,*

if P is the random variable defined by Relation (3.4), *then for $a > 0$, the relation*

$$\lim_{\substack{\gamma_1 \to +\infty \\ k_2/\gamma_1 = a}} P = N(a),$$

holds for the convergence in distribution, where $N(a)$ is a discrete random variable whose generating function is given by, for $z \in (0,1)$,

$$\mathbb{E}\left(z^{N(a)}\right) = \exp\left(-k_1 \int_0^{+\infty} \frac{(1-z)\, a\mathbb{E}\left(e^{-\gamma_2(u-E_2)^+} \mathbb{1}_{\{u > E_2\}}\right)}{1 + (1-z)\, a\mathbb{E}\left(e^{-\gamma_2(u-E_2)^+} \mathbb{1}_{\{u > E_2\}}\right)}\, du\right).$$

The distribution of the random variable $N(a)$ belong to the class of generalized *Neyman Type A distributions*. See Section 9 of Chapter 9 of Johnson et al. [16] for example. In the Markovian case of Section 2, this is a negative binomial distribution. See Proposition 2.5.

Proof. For $z \in (0,1)$, the quantity $-\log(\mathbb{E}(z^P))/k_1$ is, since $F_{L_2} \overset{\text{dist}}{=} L_2$,

$$H(z) \overset{\text{def}}{=} \int_{\mathbb{R}_+^2} \left[1 - e^{-(1-z)k_2\mathbb{E}(L_2)\mathbb{P}(L_2 + E_2 \in (u-w,u))}\right] du\gamma_1 e^{-\gamma_1 w}\, dw$$

$$= \int_{\mathbb{R}_+^2} \left[1 - e^{-(1-z)k_2\mathbb{E}(L_2)\mathbb{E}\left(e^{-\gamma_2(u-w/\gamma_1-E_2)^+} - e^{-\gamma_2(u-E_2)^+}\right)}\right] due^{-w}\, dw. \quad (6.6)$$

With similar arguments as in the proof of Proposition 6.2 and Lebesgue's Theorem, we obtain the relation

$$\lim_{\substack{\gamma_1 \to +\infty \\ k_2/\gamma_1 = a}} H(z) = \int_{\mathbb{R}_+^2} \left[1 - e^{-(1-z)a\gamma_2\mathbb{E}(L_2)w\mathbb{E}\left(e^{-\gamma_2(u-E_2)^+} \mathbb{1}_{\{u > E_2\}}\right)}\right] due^{-w}\, dw.$$

The proposition is proved. $\square$

6.3. The impact of elongation on variability

In this section we discuss the impact of the elongation of proteins on the variability of the protein production process. This aspect has been rarely considered in the mathematical models of the literature. Note that the first moments of M and P do not depend on the distributions of the elongation times. The distribution of the random variable M and the second moment of P do not depend on the distribution of the elongation phase of mRNAs when the gene is always active.

To simplify the presentation, we assume in this section that the lifetimes of mRNAs and proteins are exponentially distributed with respective parameters γ_1 and γ_2 and that the gene is always active.

If the protein is composed of N amino-acids, it is natural to assume that the average of the elongation time is proportional to N, i.e. given by N/α for some $\alpha > 0$. An index N is added to the variables E_2 and P defined by Relation (3.4). Relation (5.4) of Proposition 5.1 gives that the variance of the number of proteins at equilibrium is given by,

$$\mathrm{Var}(P^N) - \frac{k_1 k_2}{\gamma_1 \gamma_2} = \frac{k_1 k_2^2}{\gamma_1 \gamma_2^2} \mathbb{E}\left(e^{-\gamma_1 \left| S_2^N + Y_2 \right|}\right), \qquad (6.7)$$

by the equality in distribution $F_{L_a} \stackrel{\mathrm{dist}}{=} L_a$ for $a \in \{1, 2\}$ due to the exponential distribution assumption, with $S_2^N = E_{2,1}^N - E_{2,2}^N$ and $Y_2 = L_{2,1} - L_{2,2}$.

Two assumptions on the distribution of the elongation time of a protein are now considered.

1. MARKOV MODEL. The variable E_2^N is an exponential random variable given by $E_{2,M}^N$ with

$$E_{2,M}^N \stackrel{\mathrm{def}}{=} N V_\alpha,$$

   where V_α is an exponential random variable with parameter α.
   A similar assumption can be done for the elongation of mRNAs, i.e. that its distribution is also exponential. If $M^0(t)$, (resp., $P^0(t)$), denotes the number of mRNAs, (resp., proteins), being built at time $t \geq 0$, then the process

$$(I(t), M^0(t), M(t), P^0(t), P(t))$$

   has clearly the Markov property. This is a natural Markovian extension of the classical Markovian model including elongation.

2. ADDITIVE MODEL. The variable E_2^N is given by $E_{2,A}^N$ with

$$E_{2,A}^N \stackrel{\mathrm{def}}{=} \sum_{k=1}^{N} V_{\alpha,i},$$

   where $(V_{\alpha,i})$ is an i.i.d. sequence of exponential random variables with parameter α. This is also a natural assumption since it considers that it takes a random amount of time with an exponential distribution to find each amino-acid of the protein within the cytoplasm.

The variable P defined by Relation (3.4) is denoted as P_M^N for the Markov model and P_A^N for the additive model. Similarly for the variable S_2^N of Relation (6.7), we define the corresponding variables $S_{2,M}^N$ and $S_{2,A}^N$.

A non-intuitive phenomenon.
A simple calculation gives

$$\mathrm{Var}(E_{2,M}^N) = \frac{N^2}{\alpha^2} \ \text{ and } \ \mathrm{Var}(E_{2,A}^N) = \frac{N}{\alpha^2}.$$

As it can be seen the variability of the elongation time is larger for the Markov model. Intuitively, this could suggest that the same property holds for the number of proteins, that the variance of P_M^N is larger than the variance of P_A^N. But a glance at Relation (6.7) suggests in fact the contrary. Indeed, the variable $S_2^N + Y_2$ will be more variable for the Markov model and, therefore, due to this relation, the variance should be smaller. The next proposition gives a more formal formulation of this observation. See also Leoncini [21] for related numerical results of this kind.

Proposition 6.4 (Variance for a Large Number of Amino-Acids). *Under the conditions*

1. *the gene is always active;*
2. *the number of amino-acids of the protein is N, the mean elongation time of a protein is N/α for $\alpha > 0$;*
3. *the distribution of lifetimes of mRNAs and proteins are exponential with respective parameters γ_1 and γ_2;*

the variable P defined by Relation (3.4) is denoted as P_M^N for the Markov model and P_A^N for the additive model, then

$$\begin{cases} \mathbb{E}\left(P_M^N\right) = \mathbb{E}\left(P_A^N\right) = \dfrac{k_1 k_2}{\gamma_1 \gamma_2} \stackrel{\text{def}}{=} a, \\ \displaystyle\lim_{N \to +\infty} \mathrm{Var}\left(P_M^N\right) = \lim_{N \to +\infty} \mathrm{Var}\left(P_A^N\right) = a. \end{cases}$$

Furthermore, if $\gamma_1 < \gamma_2$, there exists $\alpha_0 > 0$ such that if $\alpha > \alpha_0$ then there exist constants $C > c > 0$ such that the relations

$$\begin{cases} c \le N\left(\mathrm{Var}(P_M^N) - a\right) \le C, \\ c \le \sqrt{N}\left(\mathrm{Var}(P_A^N) - a\right) \le C, \end{cases}$$

hold for any $N \ge 1$.

This proposition shows that, if the second moment of P depend on the distribution of the elongation times, its dependence is somewhat limited for both models, Markov and additive. It converges to some constant as the average (and the variance) of the elongation time goes to infinity. Recall that the average number of proteins does not depend at all on the distribution of the elongation time.

The inequalities of the proposition show that the convergence rate of the variance to a, as N gets large, is of the order of $1/\sqrt{N}$ for the additive model, and $1/N$ for the Markov model.

Proof. With the notations of Relation (6.7), for $B \in \{A, M\}$, the elementary relations

$$\mathbb{E}\left(e^{-\gamma_1|Y_2|}\right)\mathbb{E}\left(e^{-\gamma_1\left|S_{2,B}^N\right|}\right) \leq \mathbb{E}\left(e^{-\gamma_1\left|S_{2,B}^N+Y_2\right|}\right) \leq \mathbb{E}\left(e^{-\gamma_1\left|S_{2,B}^N\right|}\right)\mathbb{E}\left(e^{\gamma_1|Y_2|}\right)$$

hold. Since $\gamma_1 < \gamma_2$, one has

$$\mathbb{E}\left(e^{\gamma_1|Y_2|}\right) < +\infty,$$

hence, one has only to take care of the limiting behavior of $\mathbb{E}\left(\exp\left(-\gamma_1|S_2^N|\right)\right)$ as N goes to infinity.

For the Markov model, a simple calculation gives that $|S_{2,M}^N|$ is exponentially distributed with parameter α/N,

$$\mathbb{E}\left(e^{-\gamma_1\left|S_{2,M}^N\right|}\right) = \frac{\alpha}{\alpha+\gamma_1 N}, \text{ so that } \lim_{N\to+\infty} N\mathbb{E}\left(e^{-\gamma_1\left|S_{2,M}^N\right|}\right) = \frac{\alpha}{\gamma_1},$$

For the additive model, $S_{2,A}^N$ is a sum of N i.i.d. centered random variables

$$S_{2,A}^N = \sum_{k=1}^{N}(V_{\alpha,i}^1 - V_{\alpha,i}^2),$$

where $(V_{\alpha,i}^1)$ and $(V_{\alpha,i}^2)$ are independent i.i.d. sequences of exponential random variables with parameter α. Let

$$\sigma \stackrel{\text{def}}{=} \sqrt{\text{Var}(V_{\alpha,1}^1 - V_{\alpha,1}^2)} = \frac{\sqrt{2}}{\alpha} \text{ and } \rho \stackrel{\text{def}}{=} \mathbb{E}\left(\left|V_{\alpha,1}^1 - V_{\alpha,1}^2\right|^3\right) = \frac{6}{\alpha^3},$$

Berry-Esseen's theorem, See Feller [11] for example, gives that, for all $N \in \mathbb{N}$,

$$\sup_{x\in\mathbb{R}_+}\left|\mathbb{P}\left(\frac{|S_{2,A}^N|}{\sigma\sqrt{N}} \leq x\right) - \mathbb{P}(|\mathcal{N}(1)|\leq x)\right| \leq \frac{6\rho}{\sigma^3\sqrt{N}} = \frac{18}{\sqrt{2N}}, \qquad (6.8)$$

where $\mathcal{N}(1)$ is a centered Gaussian random variable with standard deviation 1. From Fubini's Theorem, we have

$$\mathbb{E}\left(e^{-\gamma_1\left|S_{2,A}^N\right|}\right) = \gamma_1 \int_0^{+\infty} \mathbb{P}\left(\left|S_{2,A}^N\right| \leq x\right) e^{-\gamma_1 x} \, \mathrm{d}x,$$

and, with Relation (6.8), this gives

$$\left|\mathbb{E}\left(e^{-|S_2^N|}\right) - \gamma_1 \int_0^{+\infty} \mathbb{P}\left(|\mathcal{N}(1)|\leq\frac{x}{\sigma\sqrt{N}}\right) e^{-\gamma_1 x} \, \mathrm{d}x\right| \leq \frac{18}{\sqrt{2N}}. \qquad (6.9)$$

Let $a_N = \gamma_1 \sigma \sqrt{N}$, then

$$\gamma_1 \int_0^{+\infty} \mathbb{P}\left(|\mathcal{N}(1)| \leq \frac{x}{\sigma\sqrt{N}}\right) e^{-\gamma_1 x} \, \mathrm{d}x = \mathbb{E}\left(\exp\left(-a_N |\mathcal{N}(1)|\right)\right)$$

$$= \frac{1}{a_N}\sqrt{\frac{2}{\pi}} \int_0^{+\infty} e^{-x^2/(2a_N^2)} e^{-x} \, \mathrm{d}x.$$

With Lebesgue's Theorem, we obtain

$$\lim_{N\to+\infty} \sqrt{N}\gamma_1 \int_0^{+\infty} \mathbb{P}\left(|\mathcal{N}(1)| \leq \frac{x}{\sigma\sqrt{N}}\right) e^{-\gamma_1 x} \, \mathrm{d}x = \frac{1}{\gamma_1\sigma}\sqrt{\frac{2}{\pi}} = \frac{\alpha}{\gamma_1\sqrt{\pi}}.$$

If α is such that

$$\alpha > 18\gamma_1\sqrt{\frac{\pi}{2}},$$

Inequality (6.9) gives the relations

$$0 < \liminf_{N\to+\infty} \sqrt{N}\mathbb{E}\left(e^{-|S_2^N|}\right) \leq \limsup_{N\to+\infty} \sqrt{N}\mathbb{E}\left(e^{-|S_2^N|}\right) < +\infty.$$

This concludes the proof of the proposition. $\qquad\square$

Appendix A: Some numbers for *Escherichia coli* bacterial cells

We give some numbers and some statistics for a single cell, to give an idea of the orders of magnitude of the different elements of this production process. It is also used to justify some of the assumptions in the mathematical models introduced. The sources of these numbers can be found at the address www.bioNumbers.org. The references Bremer and Dennis [4] and Chen et al. [8] have also been used.

1. GLOBAL PARAMETERS.
   — Diameter of bacteria: $\sim 2 \cdot 10^{-6}$ m.
   — Water content of a cell: 50–70%.
   — Dry content of a cell: $\sim$55% proteins, 20% mRNAs, 10% lipids.

2. POLYMERASES AND RIBOSOMES.
   A ribosome is a complex macro-molecule composed of ribosomal proteins, of the order of 50 units, and ribosomal RNAs corresponding to a total of $\sim$3000 nucleotides. Polymerases and ribosomes can be seen, for the production process, as resources of the cell. They are used by all types of genes of the cell, in some way, genes (resp., the mRNAs) of all types are competing to have access to polymerases (resp., ribosomes) to produce mRNAs (resp., proteins).
   — Number of polymerases: 1500–11400.
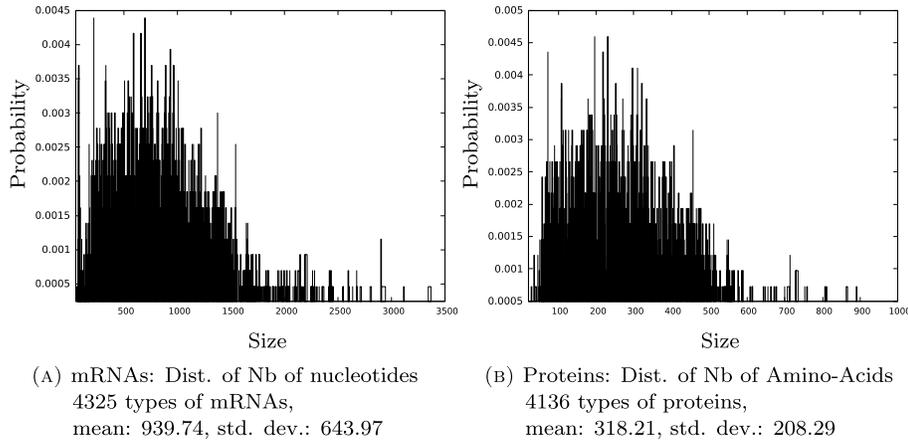   — Number of ribosomes: 6800–72000.

(A) mRNAs: Dist. of Nb of nucleotides
4325 types of mRNAs,
mean: 939.74, std. dev.: 643.97

(B) Proteins: Dist. of Nb of Amino-Acids
4136 types of proteins,
mean: 318.21, std. dev.: 208.29

FIG 5. *Distributions of Sizes for* Escherichia coli. *Sources of Data: Karp et al. [17] and Valgepea et al. [39].*

— Rate of transcription by an RNA polymerase $\sim$85 nucleotides/sec.

— Rate of translation by a ribosome $\sim$20 amino-acids/sec.

Note that the range of variations of these numbers depends on the growth rate of the cells, i.e. of the environment.

3. MRNAS AND PROTEINS.

— Number of mRNAs in a cell: $\sim$7800.

— Average lifetime of an mRNA: $\sim$2 mn.

— Size of protein: $\sim$2$\cdot$10^{-9} m.

— Number of proteins in a cell: $\sim$2.36$\cdot$10^6.

— Diffusion coefficient of a protein: $\sim$10^{-2} sec to cross cell.

Appendix B: Poisson point processes

The main results concerning Poisson processes used in this chapter are briefly recalled, in the following H is a separable locally compact metric space. See Kingman [19], Neveu [27], or Robert [33].

Definition B.1. *If $\lambda>0$, μ is a probability distribution on H, a marked Poisson process on $\mathbb{R}\times H$ with intensity $\lambda\,\mathrm{d}x\otimes\mu$ is a sequence $\mathcal{N}_\lambda=((t_n,X_n),n\in\mathbb{Z})$ of elements of $\mathbb{R}\times H$ where*

— *$(t_n,n\in\mathbb{Z})$ is a (classical) Poisson process on $\mathbb{R}$ with rate λ, with the convention that $t_n\leq t_{n+1}$ for any $n\in\mathbb{Z}$ and $t_0\leq0<t_1$.*

— *$(X_n,n\in\mathbb{Z})$ is an i.i.d. sequence of random variables with values in H and whose distribution is $X(\mathrm{d}x)$.*

The sequence $\mathcal{N}_\lambda$ can also be seen as a marked point process on $\mathbb{R} \times H$, if F is a non-negative Borelian function on $\mathbb{R} \times H$, then

$$\langle \mathcal{N}_\lambda, F \rangle = \int_{\mathbb{R} \times H} F(u,x) \mathcal{N}_\lambda(\mathrm{d}u, \mathrm{d}x) = \sum_{n \in \mathbb{Z}} F(t_n, X_n).$$

The following important proposition characterizes marked Poisson point processes.

Proposition B.1 (Laplace Transform of a Poisson Process). *The marked point process $\mathcal{N}_\lambda = (t_n, X_n)$ is a Poisson point process with intensity $\lambda \, \mathrm{d}u \otimes X(\mathrm{d}x)$ if and only if the relation*

$$\mathbb{E}\left(\exp\left(- \langle \mathcal{N}_\lambda, F \rangle \right) \right) = \exp\left(-\lambda \int_{\mathbb{R} \times H} \left(1 - e^{-F(s,x)} \right) \mathrm{d}s \, X(\mathrm{d}x) \right)$$

holds for any non-negative continuous function F on $\mathbb{R} \times H$.

Corollary B.1. *If F and G are non-negative functions in $L_2 \cap L_1(\mathrm{d}u \otimes X)$, then*

$$\mathbb{E}\left(\int_{\mathbb{R} \times H} F(s,x) \mathcal{N}_\lambda(\mathrm{d}s, \mathrm{d}x) \right) = \int_{\mathbb{R} \times H} F(s,x) \lambda \, \mathrm{d}s X(\mathrm{d}x) \qquad (\text{B.1})$$

and

$$\mathrm{Cov}\left(\int_{\mathbb{R} \times H} F(s,x) \mathcal{N}_\lambda(\mathrm{d}s, \mathrm{d}x), \int_{\mathbb{R} \times H} G(s,x) \mathcal{N}_\lambda(\mathrm{d}s, \mathrm{d}x) \right)$$
$$= \int_{\mathbb{R} \times H} F(s,x) G(s,x) \lambda \, \mathrm{d}s X(\mathrm{d}x) \quad (\text{B.2})$$

Negative binomial distribution.
For a, $b > 0$, let G be a random variable whose distribution has the density

$$\frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

on $\mathbb{R}_+$, where Γ is the classical Gamma function

$$\Gamma(a) = \int_0^{+\infty} u^{a-1} e^{-u} \, \mathrm{d}u.$$

See Whittaker and Watson [43] for example. Let $\mathcal{N}_1$ be a Poisson process on $\mathbb{R}_+$ with rate 1, then, for $z \in [0,1]$,

$$\mathbb{E}\left(z^{\mathcal{N}([0,G])} \right) = \mathbb{E}\left(e^{-G(1-z)} \right)$$
$$= \int_0^{+\infty} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-(b+1-z)x} \, \mathrm{d}x = \left(\frac{b}{1+b-z} \right)^a, \quad (\text{B.3})$$

additionally, for $n \in \mathbb{N}$,

$$
\begin{aligned}
\mathbb{P}(\mathcal{N}([0, G]) = n) &= \frac{b^a}{n! \Gamma(a)} \int_0^{+\infty} x^{a+n-1} e^{-(b+1)x} \, \mathrm{d}x \\
&= \left(\frac{b}{b+1} \right)^a \frac{\Gamma(a+n)}{n! \Gamma(a)} \frac{1}{(1+b)^n}. \quad \text{(B.4)}
\end{aligned}
$$

Appendix C: Proof of Proposition 5.1

As in section 4, $\mathcal{F}^I$ denotes the σ-field generated by the stochastic process $(I(t))$, by using again Proposition 4.1, we have

$$
\begin{aligned}
\mathbb{E}\left(P \mid \mathcal{F}^I\right) &= \int_{\mathcal{S}} G_2(u, v, w, m) I(u) \nu_{\mathcal{P}}(\mathrm{d}u \, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m), \\
&= k_1 \int_{\mathcal{S}} I(u) G_2(u, v, w, m) \, \mathrm{d}u \, L_1(\mathrm{d}v) Q_2(\mathrm{d}m) \\
&= k_1 \int_{\mathbb{R}} I(u) \mathbb{E}\left(G_2(u, E_1, L_1, \mathcal{N}_2)\right) \mathrm{d}u,
\end{aligned}
$$

by definition of Q_2, see Relation (3.2), and G_2 is defined by Relation (5.1). The integration of this identity gives

$$
\begin{aligned}
\mathbb{E}(P) &= \delta_+ k_1 \mathbb{E}\left(\int_{\mathbb{R}} \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \substack{x+E_1 \leq u < x+E_1+L_1 \\ y \leq x < y+z} \right\}} \mathcal{N}_2(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \, \mathrm{d}u \right) \\
&= \delta_+ k_1 \mathbb{E}\left(\int_{\mathbb{R}^2} \mathbb{1}_{\left\{ \substack{x+E_1 \leq u < x+E_1+L_1 \\ E_2 \leq x < E_2+L_2} \right\}} k_2 \, \mathrm{d}x \, \mathrm{d}u \right) \\
&= \delta_+ k_1 \mathbb{E}(L_1) k_2 \mathbb{E}(L_2),
\end{aligned}
$$

where $\delta_+ = k_+ / \Lambda$. The first identity of the proposition is proved.

From Representation (5.1) of P, Proposition 4.1 and Relation (B.2) of Corollary B.1, we obtain

$$
\begin{aligned}
\mathbb{E}\left(P^2 \mid \mathcal{F}^I\right) &= \left(\mathbb{E}\left(P \mid \mathcal{F}^I\right)\right)^2 + \int_{\mathcal{S}} I(u) G_2(u, v, w, m)^2 \nu_{\mathcal{P}}(\mathrm{d}u, \mathrm{d}v, \mathrm{d}w, \mathrm{d}m) \quad \text{(C.1)} \\
&= \left(\mathbb{E}\left(P \mid \mathcal{F}^I\right)\right)^2 + k_1 \int_{\mathbb{R}} I(u) \mathbb{E}\left[G_2\left(u, E_1, L_1, \mathcal{N}_2\right)^2\right] \mathrm{d}u.
\end{aligned}
$$

We now take care successively of the two terms of the right hand side of the last equality. First we study $\mathbb{E}\left[G_2\left(u, E_1, L_1, \mathcal{N}_2\right)^2\right]$.

For u, v, $w \in \mathbb{R}_+$, the distribution of the random variable

$$
G_2\left(u, v, w, \mathcal{N}_2\right) = \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \substack{x+v \leq u < x+v+w \\ y \leq x < y+z} \right\}} \mathcal{N}_2(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)
$$

is Poisson with parameter

$$\kappa(u, v, w) \stackrel{\text{def}}{=} k_2 \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{1}_{\left\{ \substack{x+v \le u < x+v+w \\ y \le x < y+z} \right\}} \, \mathrm{d}x E_2(\mathrm{d}y) L_2(\mathrm{d}z)$$

$$= k_2 \mathbb{E} \left(\int_{\mathbb{R}} \mathbb{1}_{\left\{ \substack{x+v \le u < x+v+w \\ E_2 \le x < E_2+L_2} \right\}} \, \mathrm{d}x \right).$$

By using again Relation (B.2) of Corollary B.1, we obtain that

$$k_1 \mathbb{E} \left(\int_{\mathbb{R}_+} I(u) \mathbb{E} \left[G_2 \left(u, E_1, L_1, \mathcal{N}_2 \right)^2 \right] \mathrm{d}u \right)$$

$$= \delta_+ k_1 \int_{\mathbb{R}} \left[\mathbb{E} \left(\kappa(u, E_1, L_1) \right) + \mathbb{E} \left(\kappa(u, E_1, L_1)^2 \right) \right] \mathrm{d}u$$

$$= \mathbb{E}(P) + \delta_+ k_1 \int_{\mathbb{R} \times \mathbb{R}_+^2} \kappa(u, v, w)^2 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w).$$

For $i=1$, 2, as before $(E_{2,i})$ and $(L_{2,i}$ are independent random variables with the same distribution as E_2 and L_2 respectively. By rewriting $\kappa(u, v, w)^2$ as a double integral, we obtain

$$\int_{\mathbb{R} \times \mathbb{R}_+^2} \kappa(u, v, w)^2 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w)$$

$$= k_2{}^2 \int_{\mathbb{R} \times \mathbb{R}_+^2} \mathbb{E} \left(\int_{\mathbb{R}^2} \mathbb{1}_{\left\{ \substack{x+v \le u < x+v+w \\ E_{2,1} \le x < E_{2,1}+L_{2,1}}, \substack{x'+v \le u < x'+v+w \\ E_{2,2} \le x < E_{2,2}+L_{2,2}} \right\}} \, \mathrm{d}x \, \mathrm{d}x' \right) \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w)$$

$$= k_2{}^2 \int_{\mathbb{R}_+^2} \mathbb{E} \left(\int_{\mathbb{R}^2} \left(w + x \wedge x' - x \vee x' \right)^+ \mathbb{1}_{\left\{ \substack{E_{2,1} \le x < E_{2,1}+L_{2,1} \\ E_{2,2} \le x' < E_{2,2}+L_{2,2}} \right\}} \, \mathrm{d}x \, \mathrm{d}x' \right) E_1(\mathrm{d}v) L_1(\mathrm{d}w),$$

after integration with respect to u, the translation invariance of Lebesgue's measure on $\mathbb{R}$ and Relation (1.1) give the relation

$$\int_{\mathbb{R} \times \mathbb{R}_+^2} \kappa(u, v, w)^2 \, \mathrm{d}u E_1(\mathrm{d}v) L_1(\mathrm{d}w)$$

$$= k_2{}^2 \mathbb{E}(L_1) \mathbb{E} \left(\int_{\mathbb{R}_+^2} \mathbb{P}(F_{L_1} \ge |E_{2,1} + x - (E_{2,2} + x')|) \mathbb{P}(L_2 \ge x) \mathbb{P}(L_2 \ge x') \, \mathrm{d}x \, \mathrm{d}x' \right)$$

$$= \mathbb{E}(L_1)(k_2 \mathbb{E}(L_2))^2 \mathbb{P}(F_{L_1} \ge |E_{2,1} + F_{L_2,1} - (E_{2,2} + F_{L_2,2})|),$$

hence

$$\frac{1}{\mathbb{E}(P)} k_1 \int_{\mathbb{R}} I(u) \mathbb{E} \left[G_2 \left(u, E_1, L_1, \mathcal{N}_2 \right)^2 \right] \mathrm{d}u$$

$$= 1 + k_2 \mathbb{E}(L_2) \mathbb{P} \left(F_{L_1} \ge |E_{2,1} + F_{L_2,1} - (E_{2,2} + F_{L_2,2})| \right) \quad \text{(C.2)}$$

We now investigate the term $\mathbb{E}\left(\mathbb{E}(P|\mathcal{F}^I)^2\right)$ of Relation (C.1). As for the variable M of the last section, observe that

$$\Delta^I(P) \stackrel{\text{def}}{=} \mathbb{E}\left(\mathbb{E}(P|\mathcal{F}^I)^2\right) - \mathbb{E}(P)^2 = \mathbb{E}\left(\left(\mathbb{E}(P|\mathcal{F}^I) - \mathbb{E}(P)\right)^2\right)$$

$$= \mathbb{E}\left(\left(k_1 \int_{\mathbb{R}} (I(u)-\delta_+)\mathbb{E}\left(G_2(u,E_1,L_1,\mathcal{N}_2)\right) du\right)^2\right).$$

We can rewrite the square of the integral as

$$\left(\int_{\mathbb{R}} (I(u)-\delta_+)\mathbb{E}\left(G_2(u,E_1,L_1,\mathcal{N}_2)\right) du\right)^2$$

$$= \int_{\mathbb{R}^2} (I(u)-\delta_+)(I(u')-\delta_+)\mathbb{E}\left(G_2(u,E_1,L_1,\mathcal{N}_2)\right)\mathbb{E}\left(G_2(u',E_1,L_1,\mathcal{N}_2)\right) du\,du'$$

$$= \int_{\mathbb{R}^2} (I(u)-\delta_+)(I(u')-\delta_+)$$

$$\times \int_{\mathbb{R}^2} \mathbb{P}\left(\begin{smallmatrix} x+E_{1,1}\leq u<x+E_{1,1}+L_{1,1},\ x'+E_{1,2}\leq u'<x'+E_{1,2}+L_{1,2} \\ E_{2,1}\leq x<E_{2,1}+L_{2,1}, \qquad E_{2,2}\leq x'<E_{2,2}+L_{2,2} \end{smallmatrix}\right) k_2{}^2\,dx\,dx'\,du\,du'$$

By using Relation (4.10) and the translation invariance of Lebesgue's measure on $\mathbb{R}$, we get

$$\mathbb{E}\left(\Delta^I(P)\right) = \delta_+(1-\delta_+)k_1{}^2 \int_{\mathbb{R}^4} e^{-\Lambda|u-u'|}$$

$$\times \mathbb{P}\left(\begin{smallmatrix} x+E_{1,1}\leq u<x+E_{1,1}+L_{1,1},\ x'+E_{1,2}\leq u'<x'+E_{1,2}+L_{1,2} \\ E_{2,1}\leq x<E_{2,1}+L_{2,1}, \qquad E_{2,2}\leq x'<E_{2,2}+L_{2,2} \end{smallmatrix}\right) k_2{}^2\,dx\,dx'\,du\,du'$$

$$= \delta_+(1-\delta_+)k_1{}^2 k_2{}^2 \int_{\mathbb{R}^4} e^{-\Lambda|u+x+E_{1,1}+E_{2,1}-(u'+x'+E_{1,2}+E_{2,2})|}$$

$$\times \mathbb{P}\left(\begin{smallmatrix} 0\leq u<L_{1,1},\ 0\leq u'<L_{1,2} \\ 0\leq x<L_{2,1},\ 0\leq x'<L_{2,2} \end{smallmatrix}\right) dx\,dx'\,du\,du',$$

which gives the identity

$$\mathbb{E}\left(\Delta^I(P)\right) = \frac{1-\delta_+}{\delta_+}\mathbb{E}(P)^2$$

$$\times \mathbb{E}\left(e^{-\Lambda|F_{L_1,1}+F_{L_2,1}+E_{1,1}+E_{2,1}-(F_{L_1,2}+F_{L_2,2}+E_{1,2}+E_{2,2})|}\right) \quad \text{(C.3)}$$

By plugging the identities (C.2) and (C.3) into Relation (C.1), we obtain the desired result.

Acknowledgments

References

[1] ANDERSON, D. F. AND KURTZ, T. G. (2015). *Stochastic analysis of biochemical systems*. Mathematical Biosciences Institute Lecture Series. Stochastics in Biological Systems, Vol. **1**. Springer, Cham; MBI Mathematical Biosciences Institute, Ohio State University, Columbus, OH.

[2] BERG, O. G. (1978). A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of Theoretical Biology* **71**, 4 (Apr.), 587–603.

[3] BOKES, P., KING, J. R., WOOD, A. T. A., AND LOOSE, M. (2012). Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of Mathematical Biology* **64**, 5 (Apr.), 829–854.

[4] BREMER, H. AND DENNIS, P. (2008). Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*.

[5] BRENNER, S., JACOB, F., AND MESELSON, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**, 4776, 576–581.

[6] BRESSLOFF, P. C. (2014). *Stochastic processes in cell biology*. Interdisciplinary Applied Mathematics. Springer, Cham [u.a.].

[7] BURRILL, D. R. AND SILVER, P. A. (2010). Making cellular memories. *Cell* **140**, 1, 13–18.

[8] CHEN, H., SHIROGUCHI, K., GE, H., AND XIE, X. (2015). Genome-wide study of mRNA degradation and transcript elongation in escherichia coli. *Mol. Syst. Biol.* **11**, 1 (Jan.), 781. Published correction appears in *Mol. Syst. Biol.* **11**, (May), 808 (2015).

[9] DAWSON, D. A. (1993). Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—1991*. Lecture Notes in Math., Vol. **1541**. Springer, Berlin, 1–260.

[10] ELDAR, A. AND ELOWITZ, M. B. (2010). Functional roles for noise in genetic circuits. *Nature* **467**, 7312, 167.

[11] FELLER, W. (1971). *An introduction to probability theory and its applications*, 2nd ed. Vol. **II**. John Wiley & Sons Ltd, New York.

[12] FRIEDMAN, N., CAI, L., AND XIE, X. S. (2006). Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters* **97**, 16, 168302.

[13] FROMION, V., LEONCINI, E., AND ROBERT, P. (2013). Stochastic gene expression in cells: A point process approach. *SIAM Journal on Applied Mathematics* **73**, 1 (Jan.), 195–211. arXiv:1206.0362.

[14] GUPTA, A., BRIAT, C., AND KHAMMASH, M. (2014). A scalable computational framework for establishing long-term behavior of stochastic reaction networks. *PLoS Computational Biology* **10**, 6, e1003669.

[15] JACOB, F. AND MONOD, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* **3**, 3, 318–356.

[16] JOHNSON, N. L., KEMP, A. W., AND KOTZ, S. (2005). *Univariate discrete*

distributions, Third ed. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[17] KARP, P. D., KOTHARI, A., FULCHER, C. A., KESELER, I. M., LATENDRESSE, M., KRUMMENACKER, M., SUBHRAVETI, P., ONG, Q., BILLINGTON, R., CASPI, R., PALEY, S. M., ONG, W. K., AND MIDFORD, P. E. (2017). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*.

[18] KELLY, F. P. (1979). *Reversibility and stochastic networks*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

[19] KINGMAN, J. F. C. (1993). *Poisson processes*. Oxford Studies in Probability.

[20] KURTZ, T. (1992). Averaging for martingale problems and stochastic approximation. In *Applied Stochastic Analysis, US-French Workshop*. Lecture Notes in Control and Information Sciences, Vol. **177**. Springer Verlag, 186–209.

[21] LEONCINI, E. (2013). Ph.D. document. Ph.D. thesis, École Polytechnique. https://pastel.archives-ouvertes.fr/pastel-00924232.

[22] LESTAS, I., VINNICOMBE, G., AND PAULSSON, J. (2010). Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**, 7312, 174.

[23] LEVIN, D. A., PERES, Y., AND WILMER, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society, Providence, RI.

[24] LOYNES, R. (1962). The stability of queues with non independent interarrival and service times. *Proc. Cambridge Ph. Soc. 58*, 497–520.

[25] MACKEY, M. C., SANTILLÁN, M., TYRAN-KAMIŃSKA, M., AND ZERON, E. S. (2016). *Simple mathematical models of gene regulatory dynamics*. Lecture Notes on Mathematical Modelling in the Life Sciences. Springer, Cham.

[26] MURRAY, J. D. (2002). *Mathematical biology. I*, Third ed. Interdisciplinary Applied Mathematics, Vol. **17**. Springer-Verlag, New York. An introduction.

[27] NEVEU, J. (1977). Processus ponctuels. In *École d'été de Probabilités de Saint-Flour*, P.-L. Hennequin, Ed. Lecture Notes in Mathematics, Vol. **598**. Springer-Verlag, Berlin, 249–445.

[28] NORMAN, T. M., LORD, N. D., PAULSSON, J., AND LOSICK, R. (2013). Memory and modularity in cell-fate decision making. *Nature* **503**, 7477, 481.

[29] PAULSSON, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews* **2**, 2 (June), 157–175.

[30] PROPP, J. G. AND WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms* **9**, 1-2, 223–252.

[31] RAJ, A., PESKIN, C. S., TRANCHINA, D., VARGAS, D. Y., AND TYAGI, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology* **4**, 10, e309.

[32] RIGNEY, D. AND SCHIEVE, W. (1977). Stochastic model of linear, con-

tinuous protein synthesis in bacterial populations. *Journal of Theoretical Biology* **69**, 4 (Dec.), 761–766.

[33] ROBERT, P. (2003). *Stochastic networks and queues.* Stochastic Modelling and Applied Probability Series. Springer-Verlag, New York.

[34] RUSSELL, J. B. AND COOK, G. M. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol. Mol. Biol. Rev.* **59**, 1, 48–62.

[35] SHAHREZAEI, V. AND SWAIN, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* **105**, 45, 17256–17261.

[36] SWAIN, P. S., ELOWITZ, M. B., AND SIGGIA, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences* **99**, 20, 12795–12800.

[37] TANIGUCHI, Y., CHOI, P. J., LI, G.-W., CHEN, H., BABU, M., HEARN, J., EMILI, A., AND XIE, X. S. (2010). Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 5991, 533–538.

[38] THATTAI, M. AND VAN OUDENAARDEN, A. (2001). Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences* **98**, 15, 8614–8619.

[39] VALGEPEA, K., ADAMBERG, K., SEIMAN, A., AND VILU, R. (2013). Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Molecular BioSystems* **9**, 9, 2344–2358.

[40] VAN KAMPEN, N. G. (1992). *Stochastic processes in physics and chemistry.* Vol. **1**. Elsevier.

[41] WATSON, J. D., BAKER, T. A., BELL, S. P., GANN, A., LEVINE, M., LOSICK, R., AND CSHLP, I. (2007). *Molecular biology of the gene*, 6th ed. Pearson/Benjamin Cummings; Cold Spring Harbor Laboratory Press, San Francisco; Cold Spring Harbor, N.Y.

[42] WATSON, J. D., CRICK, F. H., AND OTHERS. (1953). Molecular structure of nucleic acids. *Nature* **171**, 4356, 737–738.

[43] WHITTAKER, E. T. AND WATSON, G. N. (1996). *A course of modern analysis.* Cambridge Mathematical Library. Cambridge University Press, Cambridge. Reprint of the fourth (1927) edition.

[44] WILLIAMS, D. (1991). *Probability with martingales.* Cambridge University Press, Cambridge.