Stochastic models for protein production : the impact of autoregulation, cell cycle and protein production interactions on gene expression

Renaud Dessalles

NNT : 2017SACLX005

Thèse de doctorat de l'Université Paris-Saclay préparée à l'École Polytechnique

Ecole doctorale n°574
École doctorale de mathématiques Hadamard
Spécialité de doctorat: Mathématiques Appliquées

par

M. Renaud Dessalles

Stochastic models for protein production: the impact of autoregulation, cell cycle and protein production interactions on gene expression

Thèse présentée et soutenue à Paris, le 11 janvier 2017.

Composition du Jury :

M.	Olivier Martin	Directeur de Recherche INRA	(Président du jury)
Mme.	Amandine Véber	Chargée de recherche École Polytechnique	(Membre du jury)
M.	Jerome Robert	Maître de conférence UPMC	(Rapporteur)
M.	Pierre Vallois	Professeur Université de Lorraine	(Rapporteur)
M.	Fabien Campillo	Directeur de recherche INRIA	(Rapporteur)
M.	Philippe Robert	Directeur de recherche INRIA	(Directeur de thèse)
M.	Vincent Fromion	Directeur de recherche INRA	(Directeur de thèse)

Stochastic Models for Protein Production: the Impact of Autoregulation, Cell Cycle and Protein Production Interactions on Gene Expression

Renaud Dessalles

Résumé

Le mécanisme de production des protéines, qui monopolise la majorité des ressources d'une bactérie, est hautement stochastique : chaque réaction biochimique qui y participe est due à des collisions aléatoires entre molécules, potentiellement présentes en petites quantités. La bonne compréhension de l'expression génétique nécessite donc de recourir à des modèles stochastiques et discrets qui sont à même de caractériser les différentes origines de la variabilité dans la production ainsi que les dispositifs biologiques permettant éventuellement de la contrôler.

Dans ce contexte, nous avons analysé la variabilité d'une protéine produite avec un mécanisme d'autorégulation négatif : c'est-à-dire dans le cas où la protéine est un répresseur pour son propre gène. Le but est de clarifier l'effet de l'autorégulation sur la variance du nombre de protéines exprimées. Pour une même production moyenne de protéine, nous avons cherché à comparer la variance à l'équilibre d'une protéine produite avec le mécanisme d'autorégulation et celle produite en « boucle ouverte ». En étudiant un modèle limite, avec une mise à l'échelle (scaling), nous avons pu faire une telle comparaison de manière analytique. Il apparaît que l'autorégulation réduit effectivement la variance, mais cela reste néanmoins limité : un résultat asymptotique montre que la variance ne pourra pas être réduite de plus de 50%. L'effet sur la variance à l'équilibre étant modéré, nous avons cherché un autre effet possible de l'autorégulation : nous avons observé que la vitesse de convergence à l'équilibre est plus rapide dans le cadre d'un modèle avec autorégulation.

Les modèles classiques de production des protéines considèrent un volume constant, sans phénomènes de division ou de réplication du gène, avec des ARN-polymérases et les ribosomes en concentrations constantes. Pourtant, la variation au cours du cycle de chacune de ces quantités a été proposée dans la littérature comme participant à la variabilité des protéines. Nous proposons une série de modèles de complexité croissante qui vise à aboutir à une représentation réaliste de l'expression génétique. Dans un modèle avec un volume suivant le cycle cellulaire, nous intégrons successivement le mécanisme de production des protéines (transcription et traduction), la répartition aléatoire des composés à la division et la réplication du gène. Le dernier modèle intègre enfin l'ensemble des gènes de la cellule et considère leurs interactions dans la production des différentes protéines à travers un partage commun des ARN-polymérases et des ribosomes, présents en quantités limitées. Pour les modèles où cela était possible, la moyenne et la variance de la concentration de chacune des protéines ont été déterminées analytiquement en ayant eu recours au formalisme des Processus Ponctuels de Poisson Marqués. Pour les cas plus complexes, nous avons estimé la variance au moyen de simulations stochastiques. Il apparaît que, dans l'ensemble des mécanismes étudiés, la source principale de la variabilité provient du mécanisme de production des protéines lui-même (bruit dit « intrinsèque »). Ensuite, parmi les autres aspects « extrinsèques », seule la répartition aléatoire des composés semble avoir potentiellement un effet significatif sur la variance ; les autres ne montrent qu'un effet limité sur la concentration des protéines. Ces résultats ont été confrontés à certaines mesures expérimentales, et montrent un décalage encore inexpliqué entre la prédiction théorique et les données biologiques, ce qui appelle à de nouvelles hypothèses quant aux possibles sources de variabilité.

En conclusion, les processus étudiés ont permis une meilleure compréhension des phénomènes biologiques en explorant certaines hypothèses difficilement testables expérimentalement. Des modèles étudiés, nous avons pu dégager théoriquement certaines tendances, montrant que la modélisation stochastique est un outil important pour la bonne compréhension des mécanismes d'expression génétique.

Abstract

The mechanism of protein production, to which is dedicated the majority of resources of the bacteria, is highly stochastic: every biochemical reaction that is involved in this process is due to random collisions between molecules, potentially present in low quantities. The good understanding of gene expression requires therefore to resort to stochastic models that are able to characterise the different origins of protein production variability as well as the biological devices that potentially control it.

In this context, we have analysed the variability of a protein produced with a negative autoregulation mechanism: *i.e.* in the case where the protein is a repressor of its own gene. The goal is to clarify the effect of this feedback on the variance of the number of produced proteins. With the same average protein production, we sought to compare the equilibrium variance of a protein produced with the autoregulation mechanism and the one produced in "open loop". By studying the model under a scaling regime, we have been able to perform such comparison analytically. It appears that the autoregulation indeed reduces the variance; but it is nonetheless limited: an asymptotic result shows that the variance won't be reduced by more than 50%. The effect on the variance being moderate, we have searched for another possible effect for autoregulation: it havs been observed that the convergence to equilibrium is quicker in the case of a model with autoregulation.

Classical models of protein production usually consider a constant volume, without any division or gene replication and with constant concentrations of RNA-polymerases and ribosomes. Yet, it has been suggested in the literature that the variations of these quantities during the cell cycle may participate to protein variability. We propose a series of models of increasing complexity that aims to reach a realistic representation of gene expression. In a model with a changing volume that follows the cell cycle, we integrate successively the protein production mechanism (transcription and translation), the random segregation of compounds at division, and the gene replication. The last model integrates then all the genes of the cell and takes into account their interactions in the productions of different proteins through a common sharing of RNA-polymerases and ribosomes, available in limited quantities. For the models for which it was possible, the mean and the variance of the concentration of each proteins have been analytically determined using the Marked Poisson Point Processes. In the more complex cases, we have estimated the variance using computational simulations. It appears that, among all the studied mechanisms, the main source of variability comes from the protein production mechanism itself (referred as "intrinsic noise"). Then, among the other "extrinsic" aspects, only the random segregation of compounds at division seems to have potentially a significant impact on the variance; the other aspects show only a limited effect on protein concentration. These results have been confronted to some experimental measures, and show a still unexplained decay between the theoretical predictions and the biological data; it instigates the formulations of new hypotheses for other possible sources of variability.

To conclude, the processes studied have allowed a better understanding of biological phenomena by exploring some hypotheses that are difficult to test experimentally. In the studied models, we have been able to indicate theoretically some trends; hence showing that the stochastic modelling is an important tool for a good understanding of gene expression mechanisms.

Remerciements

À l'issue de cette thèse, je pense que je regarderai ces trois ans de travail avec une certaine nostalgie malgré un parcours qui n'a pas toujours été aussi simple que cela (Quelle thèse l'est ?). Les traditionnels remerciements sont de mise pour celles et ceux qui ont contribué directement ou indirectement à cet aboutissement.

Je tiens tout d'abord à remercier les membres de mon jury en les personnes d'Amandine Véber, Olivier Martin, Jérôme Robert, Pierre Vallois, Fabien Campillo. Je remercie tout particulièrement Jérôme Robert, Pierre Vallois, Fabien Campillo pour avoir lu ma thèse d'un bout à l'autre et pour leurs commentaires éclairants et constructifs. Je remercie bien sûr mes deux directeurs de thèse Vincent et Philippe avec qui j'ai étroitement travaillé pendant ces trois années : leur aide précieuse, leur disponibilité, leur pédagogie, leur patience parfois sont autant de choses indispensables pour m'avoir permis de mener à bout ce doctorat. Je tiens aussi à remercier mes tuteurs de thèse : Gregory Batt et Lydia Robert pour leurs conseils en cours de thèse.

Pour les autres remerciements, je me tourne en premier lieu vers mes parents : l'un comme l'autre merci de m'avoir cherché à me donner cette ouverture sur le monde, de m'avoir encouragé et pour m'avoir partagé de votre intérêt pour les sciences, et ce dès le plus jeune âge. Je pense aussi aux autres membres de ma famille : frères, cousin·e·s, (grandes) tantes et oncles ainsi que leur conjoint·e·s. Je fais aussi un gros bisou à Lisa, Matthieu et Éric. J'ai une pensée toute particulière pour mes grands-parents qui nous ont malheureusement quittés récemment.

Je remercie nombre de mes amis de longue date. Je remercie Claire et Camille pour les escapades, parfois londoniennes, parfois au Grand Palais. Je me suis beaucoup amusé avec vous et merci beaucoup pour le soutien dans les moments compliqués. Merci à JN qui a perdu ;). Merci à Nastassia pour m'avoir hébergé dans son appart de Philadelphie. Plus généralement, je remercie les ENSTAs Anne-Cha, Antoine C., Sadako, Antoine T., Ximun, et les autres qui me tiennent compagnie depuis quelques années maintenant.

Merci à Loïc pour les repas à se raconter ses malheurs et ses bonheurs le midi, merci aussi à Irène. Merci aux membres de cette belle aventure qu'est Kinea : Théo, Cécile, Annalí, Thomas, Guillaume. C'est un très beau projet qu'il faut continuer. Moltíssimes grácies al meu petit cor (de carxofa) que em fa tan feliç avui.

Je remercie la team des coureurs (et parfois nageurs) de l'INRIA : Pauline, Jonathan, Victorien, Jacques-Henri, Gabriel, Cédric, Thierry. À défaut d'un marathon, j'ai au moins pu décompresser en courant régulièrement en bonne compagnie. Je remercie les autres membres actuel et passé de RAP : Davit, Othmane, Wen, Christine, Nicolas, Virginie, Nelly, Jelena, Jim, Vianney et Pierre. Merci à Emanuele pour son aide quand je suis arrivé. Merci à Guilherme pour les bonnes rigolades et pour son aide toujours spontanée. Un grand merci à ma sœur de thèse, Sarah, dont les conversations, les soutiens mutuels m'ont considérablement rendu la thèse beaucoup plus agréable. Merci aussi à Aurora, Chloé et Noémie.

Avoir la chance d'avoir deux affiliations permet de démultiplier les souvenirs. Là l'équipe est beaucoup plus large, mais j'ai une pensée toute particulière pour divers membres de MaIAGE : Cyprien, Stéphan, Vincent H., Marc, Laurent, Arnaud, Anne, Louise et Mahendra. J'ai beaucoup aimé la soirée Minecraft (heureusement que c'était après ma rédaction). Merci à Estelle pour m'avoir donné deux fois l'occasion de parler au HeadBang. Merci aussi à Juliette pour m'avoir aidé dans les tracasseries administratives pour que je puisse être payé de mes cours (je n'y croyais pus).

Toutes mes excuses à la personne que j'aurai oubliée (c'est une de mes spécialités). Encore un grand merci à tous.

Contents

Chapter 1

Introduction

Friedrich Wöhler's famous experiment in 1828 is considered as a pioneer work of organic chemistry: for the first time an organic molecule, urea, was synthesised using non-organic materials. This study brought the idea in the scientific community that the composition of and the reactions inside living cells are not different in nature to the composition and reactions within non-organic bodies. Fundamental principles of physics and chemistry should then be the basis to explain all biological phenomena.

Under this perspective, the ambition is *in fine* to explain how macroscopic phenomena can result from elementary physical and chemical mechanisms. For a bacteria for example, a complete understanding would require to describe for different scales how the different mechanisms occur. At the end we would have a step-by-step vertical integration that starts from the molecular level, up to the cell as a whole. This local/global integration is a current important challenge for modern biology.

At a molecular level, one important aspect of the cell is its stochastic nature. Locally, the cell is composed of individual molecules in constant interaction and every elementary reaction is the result of a random collision between two of these molecules. The countless molecular reactions in the cell seem then completely unorganised and subject to large variability as it is the direct result of chance. The local behaviour of bacteria appears therefore largely disordered.

Yet, when looking at the whole cell, global cellular mechanisms seem much more stable and robust. Bacteria follow quite straightforward directions in their cell cycle: they manage to double all their contents, they replicate their entire DNA and segregate the copies in two parts of the cell and then trigger division. In favourable conditions, all of this process is done regularly in cycles of less than one hour. Therefore, robust and relatively predictable phenomena seem to globally prevail in bacteria.

The opposition between the local stochastic nature and the global stable behaviour of the cell is not easy to explain. In order to fulfil the program initiated by Wöhler's experiment, that all processes of living cells are *in fine* explained by physical and chemical phenomena, one needs to understand how individual interactions of *disordered components* can self-organise into complex systems.

Classical chemistry deals with this problem by considering that the number of every reactant is so high (in the order of magnitude of the Avogadro constant) that a deterministic approach is suitable to analyse the evolution of the reaction. Thanks to large number properties, amounts of entities can then be considered as continuous and deterministic: the behaviours of molecules average away. But the same argument cannot

be applied in a biological context like bacteria. The number of compounds intervening in a cellular chemical reaction can be very low (a sticking example is the DNA molecule that is potentially present in a single copy) so that the by-product of the reaction suffers from a lot of variability (see the first chapter of Schrödinger (1944)). Cells display countless possible intertwined reactions, with reactants of many different nature, but possibly in very few amounts. Then, randomness seems at first an obstacle to the realisation of complex biological processes: all global mechanisms of the cell are the result of potentially highly erratic reactions. Moreover the low number of entities emphasizes the discrete nature of the reactants. Understanding how this randomness is organised, or at least self-controlled, in order to be able to produce complex resilient structures, is a major issue for the global understanding of living cells.

Variability in gene expression

The question of heterogeneity in the molecular processes is especially true for the main mechanism that occurs in bacteria: *gene expression*. Gene expression is the process by which the genetic information is used to produce functional products: the proteins. It is the main process in the cell as it is estimated that *Escherichia coli* dedicates most of its energy to this usage. The production of each type of protein involves small number of entities such as DNA molecules and messenger-RNAs, and needs commonly shared macromolecules like RNA-polymerases and ribosomes. This process of protein creation is therefore subject to high viability. Moreover, some proteins are known to be involved in important cellular mechanisms (like DNA-replication initiation, cell division, responds to external threat, etc.); thereby fluctuations in the expression of these individual proteins can be reflected in the whole cell dynamics.

Fluctuations in gene expression have indeed been highlighted since the beginning of molecular biology (for example by Novick and Weiner (1957)); but it is only since the early 2000s that modern techniques of fluorescent microscopy allow a new experimental highlight on this topic (pioneered by Elowitz et al. (2002), Swain et al. (2002)). They permit quantitative measurement of the noise for particular types of proteins. Since then, variability in gene expression has become an important topic in experimental biology. They aim to estimate the variability for different types of proteins and try to determine the different origins of this heterogeneity, the potential strategies of noise reduction for some important genes, and so on.

Yet many different hypotheses are not easily testable experimentally. For example, the impact of some cellular aspects on protein variability are still not well understood: some mechanisms of protein production like auto-regulation have been proposed to reduce the variability; global cellular processes like DNA-replication and division are supposed to have an important impact on protein heterogeneity; and fluctuations of commons resources (like RNA-polymerases and ribosomes) in protein productions are said to be the prime source for variability of highly expressed genes. All these hypotheses are yet to be investigated.

Theoretical Modelling

The goal of this work is to provide new perspectives on these challenging unsolved biological questions. To do so, I have relied on theoretical models that represent the different steps of the production of proteins. Their analysis leads to a better understanding of these biological mechanisms.

The aim of this modelling process is to offer a simple but relevant representation of a particular aspect of protein production, that can be analysed using mathematical language and computational tools. The notion of "model" here completely differs from the concept of "model organism" usually used in biology: a "model organism" is an example (for instance, *Escherichia coli* and *Bacillus subtilis* are used to represent the whole realm of bacteria; *Saccharomyces cerevisiae* is the model of yeast and more generally to every eukaryotic cells, etc.). Theoretical models, on the other hand, are rather simplifications. They do not represent the whole reality, but only particular aspects of it.

Each of the models presented in this work is designed to address one particular biological question about protein production. They represent only the part of the reality that is thought to be directly involved in the studied phenomenon (for example, a model adapted to study the effect of auto-regulation). Therefore, the model may simplify many aspects of the real world, and even discard many others (*e.g.* by considering the production of only one type of protein independently from the rest of the cell for instance).

Figure 1.1 depicts the modelling methodology I have followed. At first, the proposed model needs to represent what is known from biology. From the observed biological phenomenon, one has to specify the concepts, the definitions and the questions using mathematical formalism (for instance: "What quantity to consider in order to represent 'protein noise'?"; "What do the notions of internal and external variability refer to exactly?"; etc.).

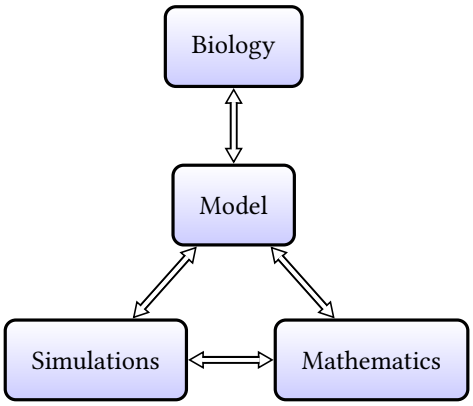


Figure 1.1: Diagram of the modelling process

We have tried to consider at first models that are parsimonious: we want to predict the biological feature, with a minimal amount of hypotheses. Sometimes these simple models are enough to essentially predict the experimentally observed effects. If some real aspects remain unexplained by the model, one may consider adding new features one by one to the model in order to better reflect the reality (this will be our approach in Chapter 3 and Chapter 4). Thanks to this series of models of increasing complexity, one can understand the relative impact of each of these added features on the simulated phenomenon. A model with many features from the outset may well fit the observations as there are more degrees of freedom, but it may also hinder the understanding of the relative impact that each of the different modelling hypotheses has on the phenomenon.

Two possible ways have been used to analyse models: through mathematics or through computational simulations. Both methods are complementary. Mathematical analyses are able to give analytical results (for example explicit formulas of protein variance). These kinds of results can indicate the theoretical possibility of a phenomenon, or provide a region of parameter values that corresponds with a particular behaviour of the system. Simulations, on the other hand, can investigate more complex mechanisms, but they lack the generality of mathematical analysis: every possible situation has to be simulated with a particular set of parameters, and the cost in terms of computational time can be non-negligible.

Both the computational and mathematics analyses enrich each other, as the simulations can indicate some interesting property worth analysing mathematically and as mathematical results on simple models can provide directions to follow when simulating more complex models (as it is the case in Chapter 2 of the present work). Finally, the results obtained by these methods have to be confronted with the biological results and, in the best case, propose new orientations for experimentation. These constant interactions between biology, the model, the mathematical and computational analyses determine the relevance of the model and enrich our understanding of the biological phenomena.

Plan of the Chapter The remaining of the chapter is an introduction to the topic of heterogeneity in gene expression. It introduces the main biological notions and modelling tools that are used in the different chapters of the manuscript. In Section 1.1 are presented the main notions, terms and mechanisms relative to the process of gene expression that will be used in the whole manuscript; it will be useful to the reader unfamiliar with the biological aspects. We then present experimental studies performed on the topic of gene expression and

what theoretical question they raise (Section 1.2). Then, we present the classical three-stage model of protein production that is largely used in the literature and that will be our basis in the different models developed here (Section 1.3); the section also give a concrete example of the kind of results that can be expected from a theoretical model.

1.1 Biological Aspects of Gene Expression

We present in this section the main biological mechanisms concerning gene expression. The goal is to present to the reader unfamiliar with this topic the basic concepts and terms relative to this subject. Many aspects of protein production are not exhaustively described here. The aim is to explain the main notions that we will refer to in the manuscript. We mainly focus on bacterial mechanisms as all the models presented within the manuscript specifically take place in prokaryotic cells.

Proteins are the main functional molecules of any cell from eukaryotic to prokaryotic cells. Their function can be to transport other molecules, to catalysis reactions, to make up the structure of the cell, or to regulate other proteins. For example, in an *Escherichia coli* bacteria, there are about 3.6×10^6 proteins of approximately 2000 different types with a great variability in concentration, depending on their types: from a few dozen up to 10^5. In total, it represents more than half of the dry mass of the bacteria (see Neidhardt and Umbarger (1996)). The time of the cell cycle (between the birth of the cell and its division) varies from 20 min (in the richest medium) to more than 150 min (for the poorest medium). During this time frame, the cell manages to approximately double its content; especially there is about twice more proteins in the cell just before division as there was at birth. As a consequence, it is estimated that *E. coli* devotes more than 67% of its energy to this usage (Russell and Cook, 1995), which make gene expression the main process in the bacteria.

Transcription and Translation

Particular chunks of genetic information on the DNA, the genes, can be interpreted to produce various types of proteins. The process by which the information of the gene (a sequence composed of four possible nucleobases) is transformed into proteins (a sequence of twenty possible amino acids) is called *gene expression*. This process is performed in two main steps: transcription and translation.

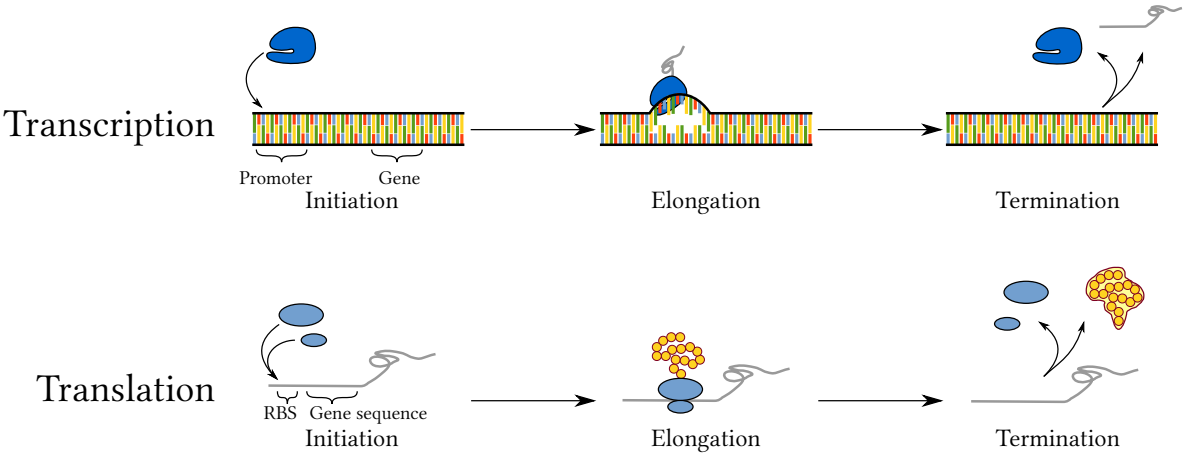


Figure 1.2: Schematic steps of transcription and translation

Transcription The transcription is the process by which the information contained on the DNA is copied into an intermediate short sequence of nucleotides: the messenger RNA (mRNA or transcript). The catalyst of this reaction is a macromolecule called RNA-polymerase.

At start, an RNA-polymerase, binds in a specific position upstream of the gene called the promoter. The binding tendency of the RNA-polymerase depends on several factors: the degree of affinity with the promoter sequence, the possible presence of transcription factors (proteins that bind close to the promoter and that can repress or promote the binding of an RNA-polymerase, see below), as well as the local structure of the chromosome that might change the accessibility of the RNA-polymerase to the promoter. Then the double strand of the DNA can be opened around the RNA-polymerase in order to create the transcription bubble: this is the initiation.

After the initiation stage, the elongation of the mRNA begins (at this stage, the process is irreversible): the RNA-polymerase "reads" one strand of the DNA in order to create the mRNA. To each nucleobase of the DNA sequence is associated a complementary nucleobase on the mRNA (the only difference between DNA and RNA is the substitution of the nucleobase Thymine in DNA, by Uracile in RNA). The RNA-polymerase creates the mRNA while advancing on the DNA nucleobase by nucleobase.

The transcription usually terminates when the RNA-polymerase elongates a specific sequence: the local mRNA conformation (determined by its sequence) provokes the disruption of the elongation complex. After the termination, the newly formed mRNA is released to the medium, as well as the RNA-polymerase that is anew available for transcription of another part of the DNA.

Translation The translation is the process by which the information on the mRNA is converted into a sequence of amino acids that constitutes a particular type of protein. The reaction is performed by another macro molecule, the ribosome. This process shares some similarity with the transcription process, as it involves the three main stages of initiation, elongation and termination as well.

The initiation sees the formation of the ribosome complex from different subunits on the mRNA. It assembles on a sequence just upstream from the beginning of the gene, called ribosomal binding site (RBS). The probability of formation of the ribosome and its ability to start elongation depends in particular on the sequence around the RBS that determines its affinity for the ribosome.

Once the initiation part is completed, the ribosome begins elongation. This process consists in associating a triplet of nucleobases (codons) to one of the twenty possible amino acids. To each codon possibly corresponds one amino acid. As for the transcription, the ribosome moves forward on the mRNA codon by codon, elongating the protein one amino acid after another.

Specific codons on the mRNA are responsible for the termination of translation: they are called STOP codons. Once the ribosome has reached one of them, the ribosome is disassembled into its different subunits and the protein is released in the medium.

It is noticeable that, contrary to eukaryotes, mRNAs are directly elongated in the cytoplasm, where ribosomes can bind on it while they are still elongated: a translation can begin on an mRNA molecule whose transcription is still ongoing.

Gene regulation

The cell has to orchestrate its protein production to be able to trigger all cellular mechanisms (like division) or to respond to environmental change. It mainly does so through transcriptional regulation: each gene sees its transcription controlled as it is prevented from or promoted to produce mRNAs during a certain period of time.

Transcriptional regulation can occur in many ways, but it is usually induced by transcription factors, i.e. proteins responsible for gene elongation. In *E. coli*, there are up to 300 different types of different transcription factors (Madan Babu and Teichmann, 2003) (which represent less than 10% of the different types of protein produced). Transcription factors bind on designated sequences on the DNA (usually close to or overlapping with the promoter of a specific gene). Once bound, they can promote (in this case, it is called *activator*) or prevent transcription (it is then called *repressor*) by modifying the affinity of the promoter for RNA-polymerases or by changing the local structure of the chromosome. The RNA-polymerase binding ability is affected as long as the transcription factor is present on it. A repressor can completely disable transcription as long as it is bound on the DNA; in this case, it can be in two states: it is either activated and is able to transcribe mRNAs, or inactivated and the RNA-polymerase cannot initiate transcription (see Figure 1.3).

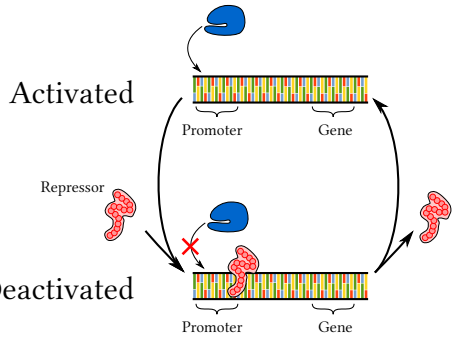


Figure 1.3: Gene activation and deactivation through a repressor.

The activity of the gene depends therefore on the nature of its transcription factors present in the cell. Transcription factors may change function depending on the environment: they may associate with other compounds (for instance nutriments or other proteins) that change their conformation and therefore affect their repressing ability, or even changing them into activators.

As transcription factors are ordinary proteins, their target promoter can control the expression of their own gene. In that case the protein can influence its own production: this auto-regulation is called the *autogenous feedback*. This mechanism is the central phenomenon studied in Chapter 2.

Messenger-RNA degradation

In bacteria, mRNAs have lifetimes of few minutes: Taniguchi et al. (2010) measured mRNA half-life of around 4 minutes. It is much shorter than their counterparts in eukaryotic cells and shorter than the doubling time of the cell. The rate of degradation depends on the type of mRNA: their sequence and their spatial conformation can influence their degradation speed. This rapid decay allows a quick turnover in the transcripts repartition, that is needed in the adaptation to sudden environmental changes.

During degradation, the mRNA is disassembled into individual mononucleotides which can be recycled in another translation or in the DNA replication. It is an active reaction as several types of enzymes intervene in the process: the *ribonucleases (RNase)* . In *E. coli*, this process usually requires two kinds of reactions (see Figure 1.4):

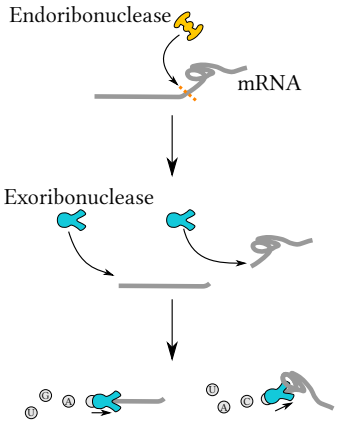


Figure 1.4: mRNA degradation process in *E. coli*.

1. First, the mRNA is cleaved by a kind of ribonuclease: the endoribonuclease that intervene in the middle of the mRNA chain. The most common endoribonuclease in *E. coli* is RNase E which binds on regions rich in Adenine-Uracil. Once bound to mRNA, the endoribonuclease performs cleavage and the mRNA is cut into two pieces. Once it happens, the mRNA is likely to loose its translation ability. Several cleavages can occur in quick succession so that the messenger is split into multiple small mRNA fragments.

2. Once the cleavages operated, other types of ribonuclease intervene to finish mRNA degradation: the exoribonucleases (in *E. coli*, the most common are PNPase, RNase II and RNase R). They are able to attack mRNA fragments by one extremity and degrade them one nucleotide at a time.

One can refer to Deutscher (2006) for an introduction on the subject.

Proteolysis

The protein degradation process, called *proteolysis*, also exists in bacteria. It has two main objectives: to degrade proteins that are misfolded, damaged or not functional; and it participates in the regulation of some functional proteins. The process shares many similarities with mRNA degradation. In particular, it relies on a type of enzymes called *protease*, which subdivides into two families: the endoproteases that cleave the protein from the middle of the chain, and the exopeptidases that catalyse the degradation from the extremity of the chain. One can refer to Miller (1996) for a complete description.

Most of the proteins are quite stable and have a much longer lifetime than mRNAs. It often exceeds several cell cycles (Koch and Levy, 1955). The exceptions are usually proteins that are regulated by proteolysis: for instance the protein SulA that is involved in the response to DNA damage is degraded by Lon protease in around 1 min (Miller, 1996).

1.2 Variability in Gene Expression

Since the beginning of molecular biology, insights about the variability in gene expression have been found. For instance Novick and Weiner (1957) describe the fluctuation in a population of genetically identical bacteria *E. coli* in the expression of β-galactosidase. Experimental techniques at that time did not allow a close examination of single-cell protein production, and this lack was an obstacle to further analysis.

Eventually, this topic blossomed during the 2000s, thanks to the emergence of fluorescent microscopy (whose principles were developed by chemistry Nobel prise laureates Betzig, Morner and Hell). This wide range of techniques enables the observation through microscopy of the expression of specific genes in a given cell: the protein of interest produces fluorescence that can be detected with optical microscopes (see Figure 1.5). As the fluorescence in the cell depends on the given protein abundance, it is possible to estimate the number of protein with a precision of one molecule unit.

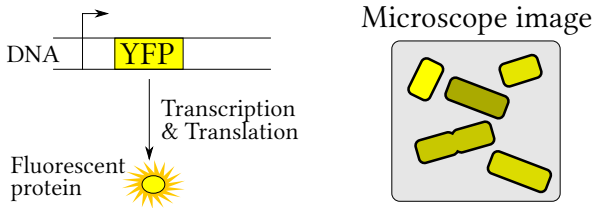


Figure 1.5: Quantitative fluorescent microscopy experiment example: the sequence of a gene (called reporter gene) is introduced into the bacterial genome; this reporter gene usually codes for a fluorescent protein (here the Yellow Fluorescent Protein) detectable with a fluorescent microscope. Using microscopy imaging, we can deduce the quantity of proteins for each cell by the fluorescence observed, and thus cell-to-cell variability in the reporter gene can be estimated.

The first two articles on this topic were Ozbudak et al. (2002) on *Bacillus subtilis* and Elowitz et al. (2002) on *Escherichia coli* (both are prokaryotic cells). It was quickly followed by experiments on eukaryotes such as yeast (Raser and O'Shea, 2004, Bar-Even et al., 2006, Cai et al., 2006, Newman et al., 2006).

We present below the main experimental results, topics and questions raised in the biological literature on the subject (one can also refer to the reviews of Raj and van Oudenaarden (2008) and Kærn et al. (2005) for additional information).

Importance of Noise in Bacteria

All these studies show a substantial amount of variability in gene expression: in isogenic population (where all individuals are genetically identical) and in a similar and constant environment, the production of a given type of protein shows large cell-to-cell variability. For two different cells in the population, a given protein may be produced in different concentrations; and inside the same cell, its temporal production may show large variations.

One reason for this variability is the small number of molecules that intervene directly or indirectly in the protein production. For instance, there is usually one or two copies of each gene; transcription factors can be present in small numbers is the cell (a dozen for the lac repressor (Kalisky et al., 2007)), and there is usually less than one mRNA for each gene at the same time (one can refer to Table 1.A.1 at the end of this chapter). Each chemical reaction is due to random encounters between molecules in the medium through random diffusion. These small numbers of entities induce variability in the protein production.

Cell scale events are naturally source of heterogeneity. For instance, the division separates the cytoplasm and its content in two parts; every compound can be in either one of two of the daughter cells. If the molecule is present in very few copies, this can have a significant effect on the variance of the distribution. DNA replication is another cause of variability: as the replication fork reaches a promoter, it can unbind the transcription factors on it (thus inducing a parasite transcription for a highly repressed gene for instance); or, as the gene is replicated, its transcription rate gets doubled.

Transcriptional and Translational Bursts

When using fluorescent microscopy to measure gene expression, it clearly appears in many cases that proteins are produced during short periods of times followed by long periods without any translation: proteins are produced in intermittent bursts. There are two possible explanations for such profiles: translational and transcriptional bursts.

Ozbudak et al. (2002) conducted a series of experiments, where the expression of a reporter gene (*gfp*) was measured. The idea was to control the transcription rate by using an inducible promoter (so the transcription rate can be controlled with environmental conditions determined in the experiment), and to control the translation rate by changing the ribosome-RBS affinity by point mutation on the RBS. With these elegant methods, both the rates of transcription and translation varied among the experiments and the authors were able to determine the respective impact on the protein expression of each step of gene expression. The results showed that the protein relative variance strongly depends on translation efficiency: it increases linearly with the average protein abundance with stronger ribosome-RBS affinity. On the other hand, the influence of the transcription efficiency on protein noise was much less apparent. The most probable explanation for these results is related to the *translational bursts*: a low number of mRNAs (possibly unique) are highly transcribed, so that the number of proteins highly depends on the small discrete number of mRNAs. Similar studies were performed in eukaryotes, showing similar results (Blake et al., 2003).

But Golding et al. (2005) proposed another possible mechanism that explains the profile of protein production: the transcriptional bursts. In addition to the measure of protein production, and contrary to Ozbudak et al. (2002), they were able to monitor mRNA production. Using the MS2-GFP method, they managed to quantify the transcript number with a single-molecule precision: some fluorescent proteins have a high tendency to bind on a specific messenger, so this messenger can be easily monitored through fluorescent microscopy.

They discovered that in their case, the mRNA is not synthesised uniformly in time, through uncorrelated random events, but during burst episodes, in which several mRNAs are produced in a short period of time. These burst periods were then followed by long periods of transcriptional inactivity. The natural interpretation is to consider that these long inactivation periods are due to gene regulation. Strong repressors bind on the gene promoter for long periods of time, giving only short time windows for transcriptions. During transcription episodes, the created mRNAs are translated, thus increasing the protein abundance for a short period of time. The bursts observed in the protein profile are here explained by an underlying transcriptional burst.

It is noticeable that these two concepts are not incompatible as they can be both specific to different types of genes. The translational burst can occur in a constitutive gene (without regulation) with rare mRNA transcriptions, or with a very rapid gene regulation; transcriptional burst rather occurs when the gene is inactivated for long periods of time and gets strongly transcribed when it is active. In both cases, the protein translation rate needs to be high in order to exhibit bursts. Moreover the protein production signal is not significantly different between the two kinds of burst: its abundance is still suddenly increasing. So that it is not easy to differentiate between these two sources of noise without directly looking at mRNA production (only Singh et al. (2012) proposed a protocol to distinguish these different sources of variability just by looking at the protein expression).

Intrinsic and Extrinsic Noise

Both transcriptional and translational noises originate from the stochastic biochemical reactions inherent to the protein production mechanism: gene activation and deactivation, mRNA production and degradation, protein elongation, etc. It is commonly referred as *intrinsic noise*. But many other external aspects have been proposed to add variability in protein production: division, gene-replication, resource availability, etc. All these additional sources of heterogeneity are usually denoted as the *extrinsic noise*. Some articles were able to propose means to quantify these two origins of variability.

One of the first article on stochasticity in gene expression, Elowitz et al. (2002), introduced the *dual reporter* technique. The idea is to compare two similar genes: they are simultaneously expressed in the same cell, and they both possess an identical promoter and are hence identically regulated (see Figure 1.6). By observing correlations in the signals of the two proteins in a given cell, the authors were able to separate two possible origins of noise. The intrinsic noise is supposed to be gene-specific, it is supposed to affect both genes independently; while the extrinsic noise, being a cell-scale fluctuation, has an identical impact on both genes; as a consequence, the extrinsic noise has a correlated impact on both expressions. The authors showed with their experiments that the extrinsic contribution is predominant in gene expression and that the proportion of each noise significantly depends on the promoter activity. In yeast, Raser and O'Shea (2004), implemented the same dual-reporter technique by introducing two almost identical genes on the same locus on homologous chromosomes in the same cell. They also obtained a high extrinsic noise resulting from a high degree of correlation between genes.

Hilfinger and Paulsson (2011) analysed the underlying theoretical idea behind the separation of extrinsic and intrinsic noise using dual reporter techniques. They interpreted the dual-reporter decomposing method as an estimation of an *environmental state decomposition*: by using the notation of the article, if X represents the number of proteins of a given cell at a given time, and if Z represents the known state of the cell (its RNA-polymerase number, its volume etc.), then it is theoretically possible to decompose the variance of X into a part which is explained by Z and another part which is completely uncorrelated (this decomposition is sometimes called the law of total variance):

$$\mathbb{V}\mathrm{ar}\left[X\right] = \underbrace{\mathbb{E}\left[\mathbb{V}\mathrm{ar}\left[X|Z\right]\right]}_{\text{unexplained by } Z} + \underbrace{\mathbb{V}\mathrm{ar}\left[\mathbb{E}\left[X|Z\right]\right]}_{\text{explained by } Z}.$$

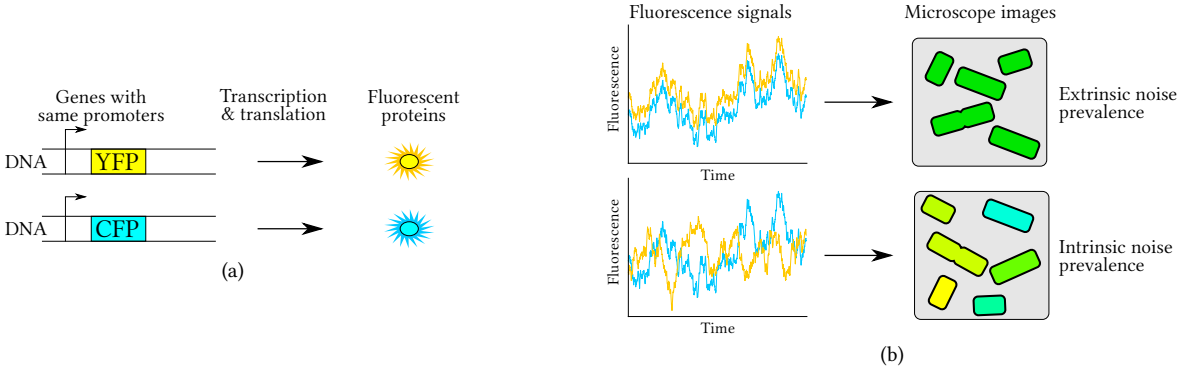


Figure 1.6: Dual reporter technique principle. **(a)**: on the same chromosome, there are two genes with the same promoter, each coding for a fluorescent protein emitting a different wave length. **(b)**: uncorrelated production give cells that express more of one of the fluorescent molecules. Above: a hypothetical situation showing fully correlated protein production; below: uncorrelated protein production (inspired by Elowitz et al. (2002)).

In dual-reporter experiment, in a specific cell, each of the two gene expressions can be interpreted as the realisation of X in a common Z environment (as they are in the same cell). In that sense, the decomposition of the signals of the two genes in the dual-reporter experiment allows to estimate the environmental state decomposition. As the authors observe, this decomposition is only possible in constant (or near constant) conditions. But in experimental work, the environment is not always constant as the promoter of the gene of interest is often induced by external changes. They proposed a slightly new version of the decomposition procedure that takes into account not only the current state Z, but also the history of the state in the case of non-constant conditions.

Several years since the first introduction of the concept by Elowitz et al. (2002), the concepts of intrinsic and extrinsic noise still remain incompletely understood at least for two reasons:

1. By considering intrinsic noise as only resulting from transcription and translation variability, the largely used dual reporter technique is still sensitive to some extrinsic contribution such as imperfect timing in replication and intracellular heterogeneity (Kærn et al., 2005). As we will see in Chapter 3, this technique is for example not suited to detecting the effect of division on protein variance, as division is commonly considered as part of the extrinsic noise in this context.

2. The expression "extrinsic noise" is often a way of denoting the unexplained part of the noise (it is exactly what is meant by environmental state decomposition in Hilfinger and Paulsson (2011)). Many mechanisms have been proposed to explain this additional noise (partition at cell division, gene replication, fluctuations in the availability of RNA-polymerases and ribosomes, uncertainty on the division etc.), but it is not easy to understand the real importance of each of these factors on protein production heterogeneity.

A complete theoretical decomposition of the different possible origins of noise will be the subject of Chapter 3 and Chapter 4 in the current manuscript.

Genome-wide Variability

In recent years, it has become possible to consider measuring the activity of a large number of genes, possibly of the whole genome.

Newman et al. (2006), by using flow cytometry method monitored the expression of more than 2500 GFP-tagged genes in yeasts (*Saccharomyces cerevisiae*) strained in a rich or a minimal media. The flow cytometry technique allows, for each individual cell, to quantify the GFP-labelled protein and at the same time to have a measure of some features of the cell such as the cell size and its granularity. For each type of protein, having determined the protein abundance in each cell, the authors were able to compute the mean and the variance of the number of proteins in the population. One of the main results concerned the importance of population heterogeneity in the extrinsic noise: within a population, yeasts display a wide range of sizes and cell cycle states. This simple fact is sufficient to add extra variability. Using the dual reporter technique, and considering only cells of a certain size and granularity, thus having approximately cells at the same stage in the cell cycle, they observed that the extrinsic noise was considerably reduced. Once filtered, protein variability clearly depended primarily on protein abundance.

Taniguchi et al. (2010) performed an analogous analysis on bacteria (*E. coli*) by using fluorescent microscopy. In each experiment, a strain of cells was considered in which a type of protein was fused with the fluorescent YFP molecule. This technique allows the direct quantification of the fused protein. They were able to measure about 1000 different proteins, and for each type of protein, the measured protein abundances range from 10^{-1} up to 10^{4} copies. On top of that, they detect simultaneously mRNAs abundance using *Fluorescence in situ hybridisation* (FISH) (fluorescent probes that are able to bind on a complementary specific sequence of nucleotides). They discovered two regimes for the protein noise: for low expressed proteins, the variability depends on protein abundance; while for highly expressed proteins, protein noise becomes independent of its abundance. They interpreted this second regime as dominated by the extrinsic noise since the noise was not gene specific and cannot be due to the gene expression mechanism. They compared their results with the yeast experiment of Newman et al. (2006): they showed that a similar noise plateau due to extrinsic factors is present in both cases, but that the extrinsic noise seems larger in *E. coli*.

In these genome-wide studies, the "extrinsic noise" is measured at the scale of the cell. Its global impact on all the proteins seems to follow specific trends. Nonetheless its possible origins are still unclear: different hypotheses have been given in these articles but without decisive arguments. But these experiments give us also measures for a large variety of proteins of the cell: proteins with different levels, essential and nonessential genes, etc. In particular, the simultaneous measures of mRNA and protein production in Taniguchi et al. (2010) could make the comparison with classical theoretical predictions possible, and thus for a majority of genes of the cell. Even more important, with these measures, it is possible to study the impact of interactions between the productions of different genes: we can consider the genes altogether in a single model rather than independently and check the model against these genome-wide experiments.

Detrimental and Advantageous Effects of the Noise

The consequence of the variability in gene expression can be noxious for the cell as it can corrupt the quality of protein signals. For instance, the fluctuation of a transcription factor can spread over entire gene networks (Pedraza and Oudenaarden, 2005, McAdams and Arkin, 1997); important choice making processes are dependent on the relative concentration of particular types of proteins (Balázsi et al., 2011, Süel et al., 2006); some highly produced proteins (like the subunits of ribosomes) can have a high cost of production in terms of energy, and fluctuations in their production could induce wasteful consumption. In the case of multicellular organisms, the development stages rely on precise spatial and temporal gene expression, in which case noise control is vital (Arias and Hayward, 2006).

But in some cases, heterogeneity among cells is proposed as a possible way of adapting to changing environments (Balaban et al., 2004, Acar et al., 2008). This strategy is called *bet-hedging*. Thanks to heterogeneity in bacterial population phenotype, some cells may be, by chance, fit to resist some external threat. Heterogeneity has been invoked to explain the selective resistance of some bacteria to antibiotics in isogenic population, or to explain the competence (the ability to take up DNA from the environment) of only a fraction of *Bacillus subtilis* populations growing in the same environment (see Raj and van Oudenaarden (2008) for further examples).

Strategies of Noise Reduction

As noise in bacteria seems to be often disadvantageous, some cellular mechanisms have been proposed to be a way for the bacteria to reduce the variability of at least some specific genes.

Fraser et al. (2004) showed that in yeast the expression of essential genes (i.e. which are critical for its survival) and genes coding for complex subunits have their protein production optimised in such a way that it minimises their production noise. The authors started from the idea of translational burst observed experimentally by Ozbudak et al. (2002) and Blake et al. (2003): a transcription burst presupposes that a large part of the variability in protein production is due to the low number of mRNAs. Genes with lower mRNA-protein average abundance ratio should have less noisy protein production. They ensured that the essential genes and the genes participating to the formation complexes, have a global tendency to be more transcribed than their nonessential counterparts. Of course, this noise optimisation comes with the cost of an extra mRNA production. But it stresses the idea that noise is an important aspect in the cell and that it is subject to natural selection.

Another possible way proposed to reduce the protein variability is negative feedback. As previously said, a protein can be a transcription factor of its own production. The hypothesis that this mechanism might be a way to reduce the variability of a protein has been emitted by theoretical models of Savageau (1974) or Thattai and van Oudenaarden (2004): a protein production that fluctuates above its mean is driven down as it decreases its own gene activity, and fluctuations below the mean would activate the gene. This hypothesis was tested experimentally by Dublanche et al. (2006) and Austin et al. (2006). In these articles, several protein production circuits on plasmids (a small DNA molecule within a cell distinct from the main chromosomal DNA) have been analysed: circuits that are autoregulated, and the others that are in "open loop". They show a decrease in the noise of the autoregulated proteins which tends to go in favour of the hypothesis. Nonetheless, it has been objected that this noise diminution mainly affects the variability induced by external changes in the number of plasmids (Paulsson, 2004). This would suggest that autoregulation has an impact only on the extrinsic noise that might come from plasmid variation, and is inefficient in reducing intrinsic noise.

The fact the autoregulation is used as a convenient way for the cell to reduce the variability of some of its proteins is still debated. Other authors like Camas et al. (2006) and Rosenfeld et al. (2002) emit the hypothesis that the autogenous feedback is used mainly in genes that need to quickly change their expression in case of environmental changes. The theoretical analysis of negative feedback autoregulation is the subject of Chapter 2.

1.3 Mathematical Modelling for Protein Production

The previous experimental works have raised several questions about the different origins of noise in protein production, the different possible ways of reducing the variability on particular proteins, or how the cell globally manages to deal with fluctuations to fulfil its genetic program. Tackling with these questions only experimentally is difficult: experimental techniques are not sufficiently advanced to allow a real time observation of every particular mechanisms in the cell, and knowing all local interactions does not directly explains how the global system behaves. The use of theoretical models has been a natural complementary means to

investigate these questions. We propose in this section to describe one important classical model of protein production: the three stage model (Subsection 3.2.1). Then we will present other theoretical works of the literature often derived from the three-stage model (Subsection 1.3.2). Finally, we present different limitations of these classical models to address several important biological questions (Subsection 1.3.3).

1.3.1 A Canonical Example: the Three-Stage Model

Let's have an insight in one canonical model of gene expression: the three-stage model, also referred as Paulsson's model as it was fully analytically described in Paulsson (2005). It is important, as it is widely used when it comes to interpreting biological results (for instance Blake et al. (2003), Raser and O'Shea (2004), Golding et al. (2005), Bar-Even et al. (2006), Taniguchi et al. (2010)). Moreover, it displays analytical formulas for the mean and the variance of protein expression, and thus can be used to decompose different causes of variability. Another interest is that its basic features have inspired many other modelling works (for instance Innocentini and Hornos (2007), Shahrezaei and Swain (2008), Fromion et al. (2013), Jansen (2014), Fromion et al. (2015)), including those of the next chapters of this manuscript.

Presentation of the Model

The three-stage model relies on several hypotheses that are commonly shared with other stochastic models of protein production. In particular, it is a "gene-centred model" as it aims to represent the production of a particular type of protein without considering interactions with the expressions of the other genes of the cell. Therefore, there is only one type of protein in the system, produced by one type of mRNA, translated from one copy of the gene[1].

Like the pioneering works of Berg (1978) and Rigney and Schieve (1977), and like many other stochastic models of gene expression since then, it relies on a common hypothesis: all the events (protein production, mRNA degradation, etc.) are represented as occurring at times that are exponentially distributed. The rates of these random variables may depend on the current state of the system. This naturally leads to a Markovian description: the model is "memoryless", the future of the system only depends on its current state, and not on its history.

Even if it is often not made explicit in the literature, as the system represents finite quantities of mRNAs and proteins, one needs to consider some sort of spatial limitation. One natural way to do so is to consider that the model only represents the number of mRNAs and proteins in an arbitrary fixed volume around the considered gene. Usually, this volume may be considered as being of the order of magnitude of the cell size so that the number of compounds in the system would approximately represent the number of compounds in a cell.

The model aims at representing all the steps that intervene in gene expression (see Section 1.1). To do so it describes the evolution of three entities: I, M and P that respectively represent the state of the gene (active of inactive), the number of mRNAs and the number of proteins. The biological mechanisms of gene expression are represented: gene regulation, transcription and translation (see Figure 1.7).

Gene regulation We consider that there are only two possible states for the gene: it can either be active (represented by $I = 1$), in which case the translation is possible; or inactive (represented by $I = 0$), in which case the translation is disabled. The repressor that inactivates the gene is independent of the system and binds on the promoter at rate λ_1^-. The repressors only leaves the promoter after an

[1] Paulsson (2005) considers a case with possibly multiple copies of the same gene, but for the sake of simplicity, we consider the case of one gene copy.

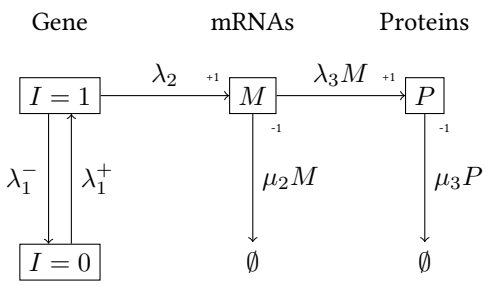


Figure 1.7: Three-stage model presentation

exponentially randomly distributed time of rate λ_1^+. Both λ_1^+ and λ_1^- are fixed: it implicitly represents a case where the concentration of the repressor in the medium is constant; and that the repressor-promoter dissociation is a spontaneous event (for instance due to thermal agitation).

Transcription While the gene is active, it can be transcribed and it produces an mRNA molecule at rate λ_2; the number M of mRNAs is then increased by 1. Every mRNA is then degraded at rate μ_2, so the global mRNA degradation rate is $M\mu_2$.

Translation Each mRNA can be translated at rate λ_3 thus creating a protein (which gives a global translation rate of $\lambda_3 M$). A protein is considered as part of the system until its decay, which occurs with rate μ_3. The total rate of protein decay is hence $\mu_3 P$.

Decay Versus Degradation

The "decay" rates μ_2 for mRNAs and μ_3 for proteins are often understood as degradation rates. As explained in Section 1.1, both the mRNAs and the proteins are broken down through active catalysed reactions. However another mechanism can be interpreted as a possible source of compound decay. During the cell cycle, the cell grows making additional space for the compounds inside the cell and as it divides, around half of the entities leaves the volume: it is dilution. As previously said, in the current case, the model takes place in a volume of about the size of a cell. As a consequence, any compound, if not degraded before, may leave the volume of interest in a time that is about the time of the cell cycle (also called doubling time). More precisely, the dilution decay has a halftime that is equal to the cell cycle.

From these two perspectives, the decay rate of mRNAs and of proteins is the combined effect of degradation and of dilution. But, as generally observed, the mRNA lifetime is much smaller than the doubling time and, on the contrary, most of proteins are stable enough to subsist several cell cycles. This leads to the following distinction about the nature of rates μ_1 and μ_2:

- The mRNA decay rate μ_1 represents a degradation rate of mRNA of the order of few minutes.
- For proteins, the decay rate μ_2 is similar to all (stable) proteins, and represents the dilution effect. It is given by the doubling time of the cell.

Analytical Expressions for the Mean and the Variance of Proteins

This system is described by $(I(t), M(t), P(t))$ which is a Markovian process with a unique invariant distribution. The moments of the equilibrium distributions of P can be calculated recursively using equilibrium equations. In particular, we get explicit solutions for its mean and its variance:

$$\mathbb{E}[P] = \delta \frac{\lambda_2}{\mu_2} \frac{\lambda_3}{\mu_3} \tag{1.1}$$

$$\mathbb{V}\mathrm{ar}[P] = \mathbb{E}[P] \left(1 + \frac{\lambda_3}{\mu_2 + \mu_3} + \frac{\lambda_2 \lambda_3 (1-\delta)(\Lambda + \mu_2 + \mu_3)}{(\mu_2 + \mu_3)(\Lambda + \mu_2)(\Lambda + \mu_3)} \right) \tag{1.2}$$

with the notations $\delta := \lambda_1^+ / (\lambda_1^+ + \lambda_1^-)$ and $\Lambda := \lambda_1^+ + \lambda_1^-$. One can refer to Paulsson (2005) or Fromion et al. (2013) to know in detail how to establish these expressions.

One can at first remark the following property: in any case, the protein variance $\mathbb{V}\mathrm{ar}[P]$ is always larger than protein average $\mathbb{E}[P]$, and so whatever the parameter choice. It indicates a theoretical lower bound for any protein variability: the protein signal cannot be precise beyond a certain limit. But as the three-stage model does not represent complex mechanisms like autogenous feedback, the question of this "Poisson lower bound" as being an actual biological limit is still unsolved (it will be the subject of Chapter 2).

Application to Transcriptional and Translational Bursts

We can use the formula as a way to describe the translation/transcription burst phenomena. Translational and transcriptional bursts have similar effects (burst in protein expression) but they are caused by two different mechanisms. The transcriptional burst effect shows bursts in the mRNA expression due to the activation of the gene during short periods of time. For the translational burst, mRNA are regularly produced during the cell cycle in very few copies and then the high translation rates provoke a sudden protein creation. The three-stage model previously described gives the decomposition of these two origins of noise (see two examples in Table 1.1).

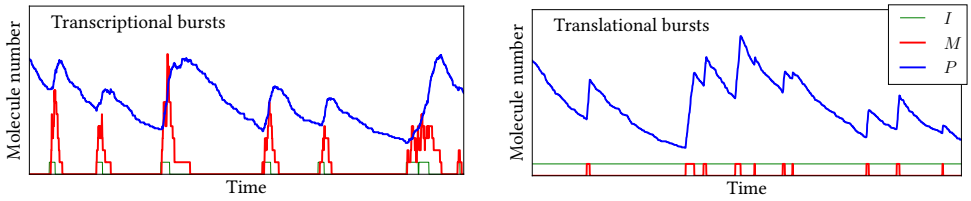
	Transcriptional burst only	Translational burst only
Noise origin	Gene activation/deactivation	Spontaneous mRNA variations
Conditions on Equation (1.3)	3rd term predominant	2nd term predominant
Example		

Table 1.1: Transcriptional and translational bursts

We may wonder, for genes with the same average production ($\mathbb{E}[P]$ is the same), for which sets of parameters transcriptional or translational bursts. For that, let's consider the relative variance deduced from Equation (1.2):

$$\frac{\mathbb{V}\mathrm{ar}[P]}{\mathbb{E}[P]} = 1 + \frac{\lambda_3}{\mu_2 + \mu_3} + \mathbb{E}[P] \times \frac{\mu_2 \mu_3}{\mu_2 + \mu_3} \cdot \frac{1-\delta}{\delta} \cdot \frac{\Lambda + \mu_2 + \mu_3}{(\Lambda + \mu_2)(\Lambda + \mu_3)}. \tag{1.3}$$

It appears that only the third term has parameters involved in gene regulation: Λ and δ. Therefore it is the only term whose contribution to the protein variance is due to the gene activation/deactivation process. For a protein with high transcriptional burst contribution, this term is predominant. In particular, low activation/deactivation phases compared to mRNA and protein decay (i.e. $\Lambda \ll \mu_2, \mu_3$) and short times of activation (i.e. $\delta \ll 1$) give a large value to this third term.

On the contrary, the second term of Equation (1.3) is independent from parameters linked to gene activity. This term represents the contribution of mRNA spontaneous fluctuations to protein variability. A protein with a high translational burst contribution has this second term predominant. For instance, one can consider a protein whose gene activation/deactivation contribution to the variance is small (for instance by having $\delta = 1$, then the gene is always activated; or $\Lambda \gg \mu_2, \mu_3$, then the gene activation/deactivation is on a quick timescale). In that case, if the mRNA activity is high as compared to the mRNA and protein degradation rates (i.e. $\lambda_3 \gg \mu_2, \mu_3$), then we get translational bursts.

The three-stage model and its variants are broadly used in the literature as it represents the basic steps of protein production, and they display analytical results to express the variability of the model.

1.3.2 Other Models

Analytical Distributions for Stochastic Gene Expression

Shahrezaei and Swain (2008) proposed analytical solutions for the distribution of proteins taking advantage of the quick mRNA decay as compared to proteins (i.e. $\mu_2 \gg \mu_3$ by using the notations of the previous section).

Basing their analysis on the three-stage model previously described, they considered at first the case without gene regulation (with $\lambda_1^- = 0$). They examined the model with the hypothesises of very short-lived mRNAs with high translational activity. In this case, the protein equilibrium distribution is shown to follow a negative binomial distribution of parameter a and b:

$$\mathbb{P}[P = n] = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)} \left(\frac{b}{1+b}\right)^n \left(\frac{1}{1+b}\right)^a \tag{1.4}$$

with P denoting the number of proteins. They also gave biological interpretations for parameters a and b: parameter a represents the average number of mRNA created in a protein lifetime, and parameter b represents the average number of proteins create by one copy of mRNA before its degradation (using the notation of the previous section, respectively $a = \lambda_2/\mu_3$ and $b = \lambda_3/\mu_2$).

The authors performed the same analysis on the complete three-stage model with gene activation/deactivation. Doing the same analysis, they give an analytical distribution for protein distributions in this case as well:

$$\begin{aligned}\mathbb{P}[P = n] = \frac{\Gamma(\alpha+n)}{\Gamma(n+1)\Gamma(\alpha)} \cdot \frac{\Gamma(\beta+n)}{\Gamma(\beta)} \cdot \frac{\Gamma(\lambda_1^+ + \lambda_1^-)}{\Gamma(\lambda_1^+ + \lambda_1^- + n)} \times \left(\frac{b}{1+b}\right)^n \left(\frac{1}{1+b}\right)^a \\ \times\ {}_2F_1\left(\alpha+n, \lambda_1^+ + \lambda_1^- - \beta, \lambda_1^+ + \lambda_1^- + n; \frac{b}{1+b}\right)\end{aligned}$$

with α and β being two parameters depending on the parameters of the system and ${}_2F_1$ a hyper-geometric function. This more complicated distribution can display a bimodal density function in the case of very long periods of gene activation/deactivation. It also converges to the previous negative binomial distribution when $\lambda_1^- \to 0$ (gene always active), or $\lambda_1^+ + \lambda_1^- \to \infty$ (activation/deactivation of the gene on a quick timescale).

These models have often been used in the literature. In particular the negative binomial distribution Equation (1.4), often approximated "continuous version", the gamma distribution with the same parameters a and

b, has been used by Friedman et al. (2006), Cai et al. (2006), Yu et al. (2006), Taniguchi et al. (2010) to fit the experimental protein distribution. The advantage of this model is that it gives an explicit formula for the whole protein distribution and for its easily biologically interpretable parameters. Yet, this model is not as general as the three-stage model: even if they are not the majority some genes may have long mRNA lifetime, or may produce few proteins.

Changing the "Exponential Assumption"

The "exponential assumption" is an important hypothesis made in the three-stage model (and many other models): each event is supposed to happen at times that follow exponentially distributed random variables. Yet, if it is a reasonable assumption to represent random collisions between two individual molecules, it may not be the case for more complex aspects. For instance, the elongation times of RNA (or protein) chains are the result of 100-300 individual steps where nucleotides (or amino-acids) are added one by one; the exponential assumption does not seem to fit in this case. Therefore, some works reinterpret the three-stage model where some mechanisms were represented by more realistic distributions.

Fromion et al. (2013) present a more general case for the mRNA and protein decay that does not consider the exponential assumption. Using the framework of Marked Point Poisson Possesses, they derive general formula for the three-stage model with arbitrary distributions for the decay distribution. In particular, they compare the protein variance in the case of a deterministic protein decay. They showed that a deterministic protein decay increases protein degradation.

In Leoncini (2013, Chapter 3), this work was continued with a "four-stage model". The translation process is separated in two distinct steps: first an initiation step, followed by an elongation step. The initiation is still represented as occurring at times that follow exponential distributions. But the elongation step is supposed to follow an Erlang distribution as it seems to be a more biologically realistic assumption: elongation results from hundreds of steps (the successive addition of amino acids). By supposing that each amino-acid requires the same amount of time to be processed, it can be represented as a finite sum of exponentially distributed times, that is to say, an Erlang distribution. They compare the protein variance in this case and in the "classical" case where the protein elongation is represented as an exponentially distributed random variable. It appears that the Erlang distribution increases the variability, but qualitative comparisons show that this impact is small in the case of biologically relevant stable proteins.

1.3.3 Limitations of Classical Models

The three-stage model and its variants presented above all share a common basis: they suppose that the direct environment of the gene is not changing through time. For instance, the transcription and translation rates (respectively λ_2 and λ_3) are constants; as in real cells, these rates depend on the availability of RNA-polymerases and ribosomes. It means that the model implicitly supposes that these entities are in constant concentration in the cell and do not fluctuate through time. Similarly, the constant rate λ_1^- supposes that the concentration of repressors is not changing throughout the cell cycle. So, any fluctuations in these quantities are not represented by the classical models.

More generally, different events in the cell cycle may have an impact on protein variability. At some point in the cell cycle, the DNA replication doubles the gene copy number, therefore instantaneously doubling the rate of transcription. At division, each protein either goes to one of the daughter cells or the other. These two events induce additional periodic fluctuations in the cell cycle that are not taken into account by classical models.

Moreover, the production of some proteins can be more complicated than what is simply described in the three-stage model: for instance, some proteins needs an extra step of maturation with the intervention of

chaperones (proteins that assist in the good conformation of proteins). Also, as previously said, some proteins intervene in their own production: some are their own repressor and bind to the promoter of their gene, thus deactivating their production. In order to determine the impact that each of these mechanisms has on protein variance, we need to consider more complex models.

In this manuscript we tackle several of these limitations, trying to offer an exhaustive description of the impact that different cellular mechanisms have on the protein heterogeneity.

1.4 Outline

The next three chapters present the results of my research activity during the three last years. As Chapter 2 addresses problems that are quite distinct from the remainder of the manuscript, it can be read independently from the rest. As for Chapter 3 and 4, they develop the same series of models of increasing complexity and should be read in that order.

Chapter 2: Model of Protein Production with Feedback. In this first chapter, we examine the variability of protein production when it is under the control of an autoregulation mechanism. The autoregulation considered here relies on a negative feedback: the considered protein is a repressor of its own gene. We propose to adapt the three-stage model to represent this mechanism and we clarify the impact of such regulation. The goal is to compare, for a same average protein production, the model with autoregulation and the classical three-stage model.

Even with a Markovian model that simply represents the feedback mechanism, there is no simple way to obtain analytical solutions for the mean and the variance of proteins at equilibrium. We therefore consider a scaling regime under which the classical three-stage model and the feedback model can be compared. In this regime, compared to the protein dynamics, the gene activation-deactivation and the mRNAs dynamics are considered to be on a quick time-scale; they both reach quickly a local equilibrium that depends on the current number of protein. We prove that the process describing the number of proteins converges then to a birth and death process where the birth rate follows a Hill repression with a hyperbolic control. In this regime, we have an explicit expression of the protein distribution. It appears in particular that the feedback indeed decreases the protein variability. But this effect is limited: an asymptotic result shows that the variance cannot be reduced of more than 50% compared to the model without feedback. With have performed simulations with parameters close to real genes, and show that in this case the variance descrease is even less import.

The limited reduction of the equilibrium variance by the autoregulation has lead us to search for other possible roles for the feedback in the cell. With additional simulations, we observe that the convergence to a new equilibrium is quicker in the case of the feedback. This feature gives a possible new role for the autoregulation: the quick adaptation of the protein production to environmental changes.

Chapter 3: Models with Cell Cycle. Usually, classical models do not explicitly represent several aspects of the cell cycle: the volume variations, the division and the gene replication. Yet these aspects have been proposed in literature to impact the protein production. In this chapter, we therefore propose a series of "gene-centred" models (that concentrates on the production of only one type of protein) that integrates successively all the aspects of the cell cycle. The goal is to obtain a realistic representation of the expression of one particular gene during the cell cycle. When it was possible, we analytically determined the mean and the variance of the protein concentration using Marked Poisson Point Process framework.

We based our analysis on a simple model where the volume changes across the cell cycle, and where only the mechanisms of protein production (transcription and translation) are represented. The variability

predicted by this model is usually assimilated to the "intrinsic noise". We then add the random segregation of compounds at division to see its effect on protein variability: at division, every mRNA and every protein has an equal chance to go to either of the two daughter cells. It appears that this division sampling of compounds can add a significant variability to protein concentration. This effect directly depends on the relative variance (Fano factor) of the protein concentration: this effect is stronger as the relative variance is low. The dependence on the relative variance can be explained by considering a simplified model. With parameters deduced from real experimental measures, we estimate that the random segregation of compounds can double the variability of the genes with the lowest relative variance.

Finally, we integrate the gene replication to the model: at some point in the cell cycle, the gene is replicated, hence doubling the transcription rate. We are able to give analytical expressions for the mean and the variance of protein concentration at any moment of the cell cycle; it allows to directly compare the variance with the previous model of the chapter with division. We show that gene replication has little impact on the protein variability: an environmental state decomposition shows that the part of the variance due to gene replication represents only at most 2% of the total variability predicted by the model.

In the end, these results are compared to the real experimental measure of protein variability. It appears that the models of this chapter tend to underestimate the protein variability especially for highly expressed proteins.

Chapter 4: Multi-protein Model. In continuation of Chapter 3, we propose a model that still considers the division and the gene replication but which also integrates the sharing of common resources: the different genes are in competition for the limited quantity of RNA-polymerases and ribosomes in order to produce the mRNAs and proteins. The goal is to examine if fluctuations in the availability of these macromolecules have an important impact on the protein variability, as it has been suggested in literature. As the model considers the interaction between the different protein productions, one needs to represent all the genes of the bacteria altogether: it is therefore a multi-protein model.

As this model is too complex to be studied analytically, we develop a procedure to estimate the parameters so that they correspond to real experimental measures. We then perform simulations in order to determine the variance of each protein and compare them with the one predicted by the models of the previous chapter. It appears that the common sharing of RNA-polymerases and ribosomes has a limited impact on the protein production: for most of proteins the variance increases of at most 10%.

In a last part, we investigate other possible sources of variability by presenting other simulations that integrate some specific aspects: variability in the production of RNA-polymerases and ribosomes, uncertainty in the division and DNA replication decisions, etc. None of the considered aspects seems to have a significant impact on the protein variability.

In the last two chapters, we then have studied many of aspects that are usually suggested as possible sources of protein concentration variability. It appears that the main contribution to the protein heterogeneity is the "intrinsic noise" due to the production mechanism itself. The only important "extrinsic" contribution is due to the random sampling of mRNAs and proteins at division. All other mechanisms studied have a limited impact. New hypotheses need to be proposed in order to explain the difference of the variability predicted by the models and the one observed experimentally.

In conclusion, this work explores many hypotheses that are difficult to test experimentally. We have been able to explore unknown features of biology such the effect of the binomial division compared to the exact division. We have been able to explore important biological hypothesis such as effect of the sharing of the RNA-polymerases and ribosomes on the variance. We also give some clear theoretical limitations of some

mechanisms, such the effect of the autogenous feedback on the variance. It shows that stochastic modelling is an important tool for the good understanding of gene expression mechanisms.

1.A Appendix: Useful Numbers

The Table 1.A.1 regroups some biological useful numbers. They are not meant to represent precise quantities but to give to the reader orders of magnitude for the different characteristics of the cell. We consider figures that corresponds to *E. coli* bacteria that are in slow growth (as it will be the case in Chapter 3 and Chapter 4). Then using numbers provided by the literature, we are able to get some insight in the rate of different events in the cell and other quantities; it is presented in Table 1.A.2.

Name	Symbol	Value	Source
Number of coding gene	K	4000	Blattner et al. (1997)
Total number of proteins	P	2.6×10^6	Neidhardt and Umbarger (1996), Table1
Total number of mRNAs	M	1.4×10^3	
Total number of RNA-polymerases	N_Y	1.5×10^3	Bremer and Dennis (1996), Table 3, for a time of division of 100 min (the closest to the doubling time of Taniguchi et al. (2010))
Total number of ribosomes	N_R	6.8×10^3	
mRNA elongation speed	c_Y	39 Nucl/s	
Protein elongation speed	c_R	12 aa/s	
mRNA average lifetime	τ_m	4 min	Taniguchi et al. (2010)
Doubling time	τ_D	150 min	

Table 1.A.1: Useful numbers in *E. coli*.

Name	Expression	Value
Transcriptions per second	M/τ_m	$6\,\mathrm{s}^{-1}$
Translations per second	P/τ_D	$3 \times 10^2\,\mathrm{s}^{-1}$
Average mRNA number per genes	M/K	0.32
Average proteins number per genes	P/K	6.0×10^2
mRNA number produced in one cell cycle	$M \cdot \tau_D/\tau_m$	5.2×10^4

Table 1.A.2: Rate of events and other quantities deduced from Table 1.A.1.

Chapter 2

Model of Protein Production with Feedback

This chapter analyses, in the context of a prokaryotic cell, the stochastic variability of the number of proteins when there is a control of gene expression by an autoregulation scheme. The goal of this work is to estimate the efficiency of the regulation to limit the fluctuations of the number of copies of a given protein. The autoregulation considered in this chapter relies mainly on a negative feedback: the proteins are repressors of their own gene expression. The efficiency of a production process without feedback control is compared to a production process with an autoregulation of the gene expression assuming that both of them produce the same average number of proteins. The main characteristic used for the comparison is the standard deviation of the number of proteins at equilibrium. With a Markovian representation and a simple model of repression, we prove that, under a scaling regime, the repression mechanism follows a Hill repression scheme with an hyperbolic control. An explicit asymptotic expression of the variance of the number of proteins under this regulation mechanism is obtained. Simulations are used to study other aspects of autoregulation such as the rate of convergence to equilibrium of the production process and the case where the control of the production process of proteins is achieved via the inhibition of mRNAs.

2.1 Introduction

2.1.1 Biological Context

The *gene expression* is the process by which genetic information is used to produce functional products of gene expression: proteins and non-coding RNAs. This chapter concerns itself with the production of proteins. The information flow from DNA genes to proteins is a fundamental process. It is composed of three main steps: *Gene Activation*, *transcription* and *translation*.

1. The initiation of transcription is strongly regulated. Schematically the gene is said to be in "inactive state" if a repressor is bound on the gene's promoter preventing the RNA polymerase from binding and is in "active state" otherwise.

2. When the gene is in active state, the RNA polymerase binds and initiates transcription that leads to the creation of a mRNA, a copy of a specific DNA sequence.

3. The translation of the messenger into a protein is achieved by a large complex molecule: the *ribosome*. A ribosome binds to an active mRNA, initiates the translation and proceeds to protein elongation. Once the elongation terminates, the protein is released in the medium and the ribosome is anew available for any another translation.

The production of proteins is the most important cellular activity, both for the functional role and the high associated cost in terms of resources. In a *E. Coli* bacterium for example there are about 3.6×10^6 proteins of approximately 2000 different types with a large variability in concentration, depending on their types: from a few dozen up to 10^5. The gene expression is additionally a highly stochastic process and results from the realization of a very large number of elementary stochastic processes of different nature. The three main steps are the results of a large number of encounters of macromolecules following random motions, due in particular to thermal excitation, in the viscous fluid of the cytoplasm. One of the key problems is to understand the basic mechanisms which allow a cell to produce a large number of proteins with very different concentrations and in a random context. This can be seen as a problem of minimization of the variance of the number of proteins of each type.

To study this problem, one can take a simple stochastic model, with a limited set S of parameters preferably, describing the three steps of the production of a given type of protein. Once a closed form expression of the variance of the number of proteins is obtained, it is natural to find the parameters of the set S which minimizes the variance with the constraint that the mean number of proteins is fixed. See the survey Paulsson (2005).

A more effective way to regulate the number of proteins can be of using a direct feedback control, an *autoregulation* mechanism, so that the production of proteins is either sped up or slowed down depending on the current number of proteins. It should be noted that the feedback control loop can involve other intermediate proteins to achieve this goal, like the classical lac operon, but it is not considered here. See Yildirim and Mackey (2003) for example.

The protein can regulate the gene activation simply, for example by being a repressor and tend to bind on his own gene's promoter. This is the *autogenous regulation* scheme. See Goldberger (1979) and Maloy and Stewart (1993). See also Thattai and van Oudenaarden (2004). Other autoregulation mechanisms are possible in cells, such as an autoregulation on the mRNAs where a protein inhibits its own translation initiation by binding to the translation initiation region of its own mRNAs. It occurs for example in the production of ribosomal proteins, see Kaczanowska and Rydén-Aulin (2007). The idea being that a feedback mechanism may reduce significantly the number of large excursions from the mean. In this chapter, the mathematical analysis will mainly focus on a negative autogenous feedback, when the rate of inactivation of the gene expression grows with the number of proteins.

2.1.2 Literature

The classical results concerning the mathematical analysis of the variance of the number of proteins has been investigated in Berg (1978) and Rigney and Schieve (1977) and reviewed more recently by Paulsson (2005), see also Raj and van Oudenaarden (2008) for the biological aspects. These references use the three stage model, the state of the system is given by three variables: the state of the promoter, the number of mRNAs and the number of proteins. Mathematically, the techniques used rely on the Fokker-Planck equations of the associated three dimensional Markov process and the observation that at equilibrium, a recurrence on the moments of the number of proteins holds. Fromion et al. (2013) investigates a more general model (elongation times are not necessarily exponentially distributed in particular) and an alternative technique to a Markovian approach is introduced.

Concerning the evaluation of autoregulation, most of mathematical models use a continuous state space, the rate of production of proteins depends linearly on the number of mRNAs and the rate of production of mRNAs is a nonlinear function $k(p)$ exhibiting a non-linear dependence on the current number p of proteins. In Rosenfeld et al. (2002) and Becskei and Serrano (2000), based on experiments the constant $k(p)$ is taken a *Hill repression function*, i.e. $k(p) = a/(b + p^n)$ for some constants a and b and $n \geq 1$ is the Hill coefficient. See also Thattai and van Oudenaarden (2004). Related models in a similar framework with further results are presented in Bokes et al. (2011) and Yvinec et al. (2013). For most of these models the state of the promoter, active or inactive, which is a source of variability is not taken into account, it is in some way encapsulated in the constant $k(p)$ whose representation is rarely discussed. In Hornos et al. (2005) the state of the gene expression, on or off, is taken into account but not the number of mRNAs and therefore the fluctuations generated by transcription. The parameter of activation $k(p)$ is of course crucial in our case since autogenous regulation rely on the state of the promoter which can be inactivated by proteins. Our model includes it. See also Fournier et al. (2007) for some simulations of these stochastic models of autoregulation as well as some experiments.

2.1.3 Results of the Chapter

The main goal of this chapter is to estimate the possible benefit of the autogenous regulation to control the fluctuations of the number of copies of a given protein. The efficiency of a production process without feedback control is compared to a production process with an autoregulation of the gene expression, assuming that both of them produce the same average number proteins. The main characteristic used for the comparison is the standard deviation of the number of proteins at equilibrium. For this purpose, two approaches are used.

Mathematical Analysis One first studies the distribution of the number of proteins via a stochastic model. When there is no regulation, the corresponding classical mathematical model has been investigated in detail for some time now. In particular, the standard deviation of the number of proteins at equilibrium has a closed form expression in terms of the basic parameters of the production process. See for example the survey Paulsson (2005), and also Fromion et al. (2013).
To represent the negative feedback of the autogenous regulation, a simple model is used: each protein can be bound, at some rate and for some random duration of time, on its own gene expression. In this situation the gene expression is inactive and the transcription is not possible during that time. This amounts to say that the gene expression is deactivated at a rate proportional to the number of proteins. The activation rate is constant.

As will be seen, the mathematical model of the autogenous regulation is more complicated, in particular there is no recurrence relationship between the moments of the number of proteins at equilibrium as in the classical model of protein production process. For this reason, a limiting procedure is used, it amounts to assume that the dynamics of the activation of the gene expression and of the evolution of mRNAs occur on a much faster time scale than the dynamics of the proteins. The values of the key parameters are presented in Subsection 2.5.1. The scaling parameter is the multiplicative factor describing the difference of speed of these two time scales. The main convergence result is Theorem 2.2. The assumption of a fast time scale for gene expression activation and mRNAs is quite common in the literature, see Bokes et al. (2011) and Yvinec et al. (2013). The techniques used in these references rely on singular perturbation methods to deal with the two time scales. In our setting, a probabilistic approach is used, as will be seen, it gives precise results on the asymptotic stochastic evolution of the number of proteins.

Under this limiting regime it is shown that, asymptotically, the protein production process can be described as a birth and death process. See Keilson (1974) for example. In state $x \in \mathbb{N}$, the birth rate is $a/(b+x)$ for some constants a and b. This is a contribution of the chapter that, with a simple model of the autoregulation, one can show that the repression mechanism follows indeed a Hill repression scheme with an hyperbolic control,

i.e. with Hill coefficient 1. The death rate is not changed by the limiting procedure, it is proportional to x. Consequently, one can get an asymptotic closed form expression of the standard deviation of the number of proteins by using the explicit representation of the equilibrium of this birth and death process. See Corollary 2.1. It is shown that, in this limiting regime, the standard deviation is reduced by 30%. The corresponding results are presented in Section 2.3 and Section 2.4 and in Section 2.A. The mathematical results are obtained via convergence theorems for sequence of Markov process, the proof of a stochastic averaging principle and a saddle point approximation result.

Simulation We also analyze, via simulations, autogenous regulation but also other aspects related to the regulation of protein production. This is presented in Section 2.5. Simulations are used mainly because of the complexity of the mathematical models of some aspects of the autogenous regulation. By using plausible biological parameters, one gets an improvement of 15% for the standard deviation of the number of proteins can be expected. This is significantly less than the performances of the limiting mathematical model studied in Section 2.3. The main reason seems to be that that the scaling parameter is not, in some cases, sufficiently large to have a reasonable accuracy with the limit given by the convergence result of Theorem 2.2.
Via simulations, one also investigates the case when the regulation is not on the gene expression but on the corresponding mRNAs: a protein can block an mRNA for some time. In this situation, it could be expected that the production process is modulated more smoothly by playing on the inactivation of a fraction of the mRNAs and not on the rough on-off control of the gene expression. It is shown that the improvement is real but not that big (less than 10%). It is nevertheless remarkable that if the average life time of mRNAs is significantly increased, our experiments show that the benefit of such regulation can be of the order of more than 30% on the standard deviation of the number of proteins.

Coming back to regulation on the gene expression. Our experiments show that, despite the impact of autogenous regulation on fluctuations of the number of proteins can be limited, it has nevertheless a very interesting property. Starting with a number of proteins significantly less (or greater) than the average number of proteins at equilibrium, the autogenous regulation returns to the "correct" number of proteins much faster than the classical production process without regulation. This is a clear advantage of this mechanisms to adapt quickly when biological conditions change due to an external stress for example. See Subsection 2.5.6. This phenomenon has been observed, via experiments, in Rosenfeld et al. (2002). See also Camas et al. (2006). Finally Subsection 2.5.5 investigates the comparison of production processes with and without a feedback on the gene expression through the estimation of their respective power spectral density.

2.2 Stochastic Models of Protein Production

We present the stochastic models used to investigate the protein production process. We will use the three stepsou model describing the activation-deactivation of the gene, the transcription phase and the translation phase. Like in most of the literature, it is assumed that the various events, like the encounter of two macromolecules, occurring within the cell have a duration with an exponential distribution. We start with the classical model used in this domain since the late 70's by Berg (1978) and Rigney and Schieve (1977). See also Thattai and van Oudenaarden (2004) and Paulsson (2005).

2.2.1 The Classical Model of Protein Production

1. The inactive gene is activated at rate λ_1^+ and deactivated at rate λ_1^- otherwise.

2. If the gene is active, an mRNA is produced at rate λ_2. An mRNA is degraded at rate μ_2.

3. Given M mRNAs at some moment, a protein is produced at rate $\lambda_3 M$. Each protein is degraded at rate μ_3.

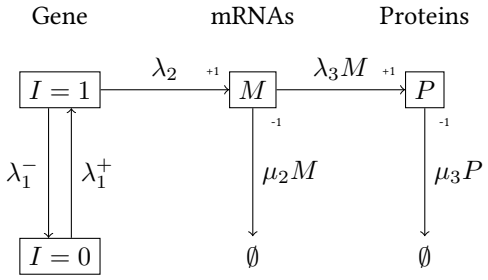


Figure 2.1: Classical Three Stage Model for Protein Production.

The stochastic processes describing the protein production process are: $I(t)$ the state of the gene at time t which is 0 if it is inactive and 1 otherwise. The number of mRNA at time t is $M(t)$ and $P(t)$ denotes the number of proteins at that moment. The process $(I(t), M(t), P(t))$ is Markovian with state space

$$\mathcal{S} \overset{\text{def.}}{=} \{0, 1\} \times \mathbb{N}^2,$$

its transition rates are given by, if $(I(t), M(t), P(t)) = (i, m, p) \in \mathcal{S}$,

$$\begin{cases} (0,m,p) \to (1,m,p) & \text{at rate } \lambda_1^+, \quad (1,m,p) \to (0,m,p) & \text{at rate } \lambda_1^-, \\ (i,m,p) \to (i,m+1,p) & \lambda_2 i, \quad (i,m,p) \to (i,m-1,p) & \mu_2 m, \\ (i,m,p) \to (i,m,p+1) & \lambda_3 m, \quad (i,m,p) \to (i,m,p-1) & \mu_3 p. \end{cases}$$

See Figure 2.1.

Lemma 2.1. *The previous Markov process has a unique invariant distribution.*

Proof. We can construct the coupling $(\widetilde{M}(t), \widetilde{P}(t))$ such as $M(t) \leq \widetilde{M}(t)$ and $P(t) \leq \widetilde{P}(t)$ such as which corresponds to the case where the gene is always active. It is enough to prove that this process is ergodic to show the result. Using the Liapunov function $f(m,p) = m + ap$ with a a positive number smaller than μ_2/λ_3. In that case we have

$$Qf(m,p) = \lambda_2 + (a\lambda_3 - \mu_2)m - a\mu_3 p,$$

with Q the Q-matrix of the process $(\widetilde{M}(t), \widetilde{P}(t))$. By choosing

$$K > \max\left(\lambda_2/(a\lambda_3 - \mu_2), \lambda_2/(a\mu_3)\right)$$

we have that for any (m,p) such as $f(m,p) > K$, it follows that $Qf(m,p) < -\varepsilon$ with $\varepsilon > 0$. Then, using the Proposition 8.14 of Robert (2010), it follows the result. □

An explicit expression of the distribution of P at equilibrium is not known but, due to the linear transition rates, the moments of P can be calculated recursively. In the following (I, M, P) will denote random variables whose law is invariant for $(I(t), M(t), P(t))$.

Proposition 2.1. *At equilibrium, the two first moments of P can be expressed by*

$$\mathbb{E}[P] = \frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^-} \frac{\lambda_2}{\mu_2} \frac{\lambda_3}{\mu_3} \tag{2.1}$$

$$\mathbb{V}\mathrm{ar}[P] = \mathbb{E}[P] \left(1 + \frac{\lambda_3}{\mu_2 + \mu_3} + \frac{\lambda_1^- \lambda_2 \lambda_3 \left(\lambda_1^+ + \lambda_1^- + \mu_2 + \mu_3 \right)}{\left(\lambda_1^+ + \lambda_1^- \right) \left(\mu_2 + \mu_3 \right) \left(\lambda_1^+ + \lambda_1^- + \mu_2 \right) \left(\lambda_1^+ + \lambda_1^- + \mu_3 \right)} \right). \tag{2.2}$$

See Paulsson (2005), Shahrezaei and Swain (2008), Swain et al. (2002) and Fromion et al. (2013) for example. Explicit closed expressions for the moments are not that common to obtain for stochastic models of gene expression, in the continuation, we will see that it is for instance not the case for our model with autogenous regulation.

2.2.2 A Stochastic Model of Protein Production with Autogenous Regulation

The regulation is done via proteins which can inactivate the gene corresponding to the protein. If there are P proteins at some moment then the gene is activated at a rate proportional to P. Compared to the above model, only the first step changes.

1. The inactive gene is activated at rate λ_1^+ and inactivated at rate $\lambda_1^- P$ otherwise.

See Figure 2.2. For the sake of simplicity, we use the same notations λ_1^+ and λ_1^- as for the classical model of protein production instead of $\lambda_{F,1}^+$ and $\lambda_{F,1}^-$ for example. It should be noted that in our comparisons in Section 2.5, these quantities are not necessarily the same for these two models.

The corresponding Markov process is denoted as $(I_F(t), M_F(t), P_F(t))$, its transitions have the same rate as $(I(t), M(t), P(t))$ except for those concerning the first coordinate.

$$\Big\{ (0,m,p) \to (1,m,p) \quad \text{at rate } \lambda_1^+, \quad (1,m,p) \to (0,m,p) \quad \text{at rate } \lambda_1^- p.$$

As before, (I_F, M_F, P_F) will denote random variables whose law is the invariant distribution of the Markov process $(I_F(t), M_F(t), P_F(t))$. The following proposition is the analogue of Proposition 2.1 for the feedback model but with unknown quantities related to the the activity of the gene, $\mathbb{E}[I_F]$, and the correlation of the activity of the gene and the number of mRNAs, $\mathbb{E}[I_F M_F]$.

Proposition 2.2. *At equilibrium, the first moment of P_F can be expressed by*

$$\mathbb{E}[P_F] = \mathbb{E}[I_F] \frac{\lambda_2}{\mu_2} \frac{\lambda_3}{\mu_3}. \tag{2.3}$$

Proof. We can prove that the process has a unique unique invariant distribution similarly as in Lemma 2.1. By equality of input and output for $(M(t))$ and $(P(t))$ at equilibrium, one gets the relations

$$\lambda_2 \mathbb{E}[I_F] = \mu_2 \mathbb{E}[M_F], \quad \lambda_3 \mathbb{E}[M_F] = \mu_3 \mathbb{E}[P_F],$$

and therefore (Equation (2.3)). □

It does not seem that an expression for $\mathbb{E}[I_F]$ can be obtained, the relation $\lambda_1^- \mathbb{E}[I_F P_F] = \lambda_1^+ (1 - \mathbb{E}[I_F])$ of equality of flows for activation/deactivation process introduces the correlation between I_F and P_F. This is in fact the main obstacle to get more insight on the fluctuations of the number of proteins. The next section investigates a scaling where the activation/deactivation phase is much more rapid than the production process of proteins.

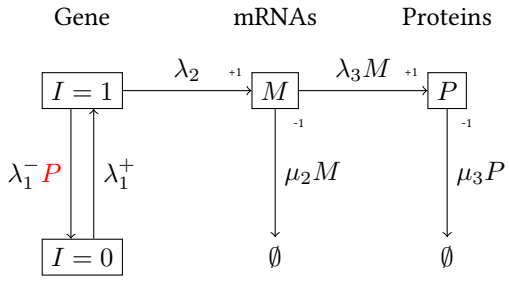


Figure 2.2: Three Stage Model for Protein Production with Autogenous Regulation

2.3 A Scaling Analysis

It has been seen in the previous section that, for the feedback mechanism, an explicit representation of the variance of the number of proteins at equilibrium seems to be difficult to derive. In this section we use the fact that the time scale of the first two steps, activation/deactivation of the gene and production of mRNAs is more rapid than the time scale of protein production. This is illustrated by the fact that the lifetime of an mRNA is of the order of 2 min. whereas the doubling time of a bacteria is around 40 min giving a lifetime of a protein of the order of one hour. See Taniguchi et al. (2010), Li and Elf (2009) and Hammar et al. (2012). As will be seen, this assumption simplifies the analysis of the feedback mechanism. We will be able to get an asymptotic explicit expression for the distribution of the number of proteins at equilibrium.

A (large) scaling parameter N is used to stress the difference of time scale. When there is a feedback control, an upper index N is added to the variables so that the corresponding Markov process is denoted as $(X_F^N(t)) = (I_F^N(t), M_F^N(t), P_F^N(t))$ on the state space $\mathcal{S} = \{0,1\} \times \mathbb{N}^2$. The transition rates of the Markov process are given by

$$\begin{cases} (0,m,p) \to (1,m,p) & \text{at rate } \lambda_1^+ N, & (1,m,p) \to (0,m,p) & \text{at rate } \lambda_1^- Np, \\ (i,m,p) \to (i,m+1,p) & i\lambda_2 N, & (i,m,p) \to (i,m-1,p) & \mu_2 mN, \\ (i,m,p) \to (i,m,p+1) & \lambda_3 m, & (i,m,p) \to (i,m,p-1) & \mu_3 p. \end{cases} \tag{2.4}$$

The initial state is constant with N given by $X_F^N(0) = (i_0, m_0, p_0) \in \mathcal{S}$.

The aim of this section is of proving that the non-Markovian process $(P_F^N(t))$ converges in distribution to a limiting Markov process $(\overline{P}_F(t))$. As will be seen, an averaging principle, proved in the appendix, holds: locally the "fast" process $(I_F^N(t), M_F^N(t))$ reaches very quickly some equilibrium depending on the current value of the "slow" variable $P_F^N(t)$. It turns out that the equilibrium of this limiting process $(\overline{P}_F(t))$ can be analyzed in detail. The proof of the averaging principle relies on stochastic calculus applied to Markov processes in the same spirit as in Papanicolaou et al. (1977) in a Brownian setting, see also Kurtz (1992).

Notations

Throughout the rest of this chapter, we will use the following notations $\rho_1 = \lambda_1^+/\lambda_1^-$ and, for $i = 2, 3$, $\rho_i = \lambda_i/\mu_i$.

2.3.1 Scaling of the Classical Model of Protein Production

One first states a scaling result for the classical model of protein production. The result being much simpler to prove than the corresponding result, Theorem 2.2, for the feedback process, its proof is skipped. One denotes by $(X^N(t)) = (I^N(t), M^N(t), P^N(t))$ the corresponding Markov process, its transition rates are the same as for feedback in (Equation (2.4)) except for deactivation:

$$(1, m, p) \to (0, m, p) \text{ at rate } \lambda_1^- N.$$

The following result shows that, in the limit, the evolution of the number of proteins converges to the time evolution of an $M/M/\infty$ queue. See Chapter 6 of Robert (2010) for example.

Theorem 2.1. *If $X^N(0) = (i_0, m_0, p_0) \in \mathcal{S}$, the sequence of processes $(P^N(t))$ converges in distribution on the Skorohod space to a birth and death process $(\overline{P}(t))$ on $\mathbb{N}$ whose respective birth and death rates (β_x) and (δ_x) are given by*

$$\beta_x = \frac{\lambda_3 \rho_2 \rho_1}{\rho_1 + 1} \text{ and } \delta_x = \mu_3 x.$$

The equilibrium distribution of $(\overline{P}(t))$ is a Poisson distribution with parameter $\rho_1\rho_2\rho_3/(1+\rho_1)$.

Proof. The intuition of this result can be described quickly as follows. The processes $(I^N(t), M^N(t))$ live on a much faster time scale than $(P^N(t))$ and therefore reach quickly the equilibrium. When N gets large, the process $(M^N(t))$ is an $M/M/\infty$ queue with arrival rate $\lambda_2\lambda_1^+/(\lambda_1^+ + \lambda_1^-)$ and service rate μ_2. See Chapter 6 of Robert (2010) for example. Its equilibrium distribution is therefore Poisson with parameter $\rho_2\rho_1/(1+\rho_1)$. The process $(P^N(t))$ can then be seen as an $M/M/\infty$ queue with arrival rate $\lambda_3\rho_2\rho_1/(1+\rho_1)$ and service rate μ_3, i.e. a birth and death process with the transition rates of the theorem. Its equilibrium is Poisson with parameter $\rho_1\rho_2\rho_3/(1+\rho_1)$.

The proof of a corresponding result in a more complicated setting, for the production process with feedback, is done below. For this reason the proof of this result is skipped. □

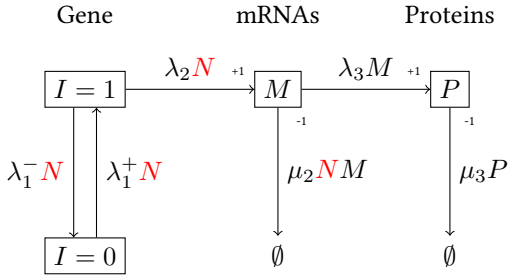


Figure 2.1: Feedback Model with Scaling Parameter N

2.3.2 Scaling of the Production Process with Feedback

The following theorem is the main result of this section. As in the case of the classical model of protein production, it relies on the fact that, due to the scaling, the activation/deactivation of the gene and the production of mRNAs occurs on a fast time scale so that an averaging principle holds. See below. Some of the technical results used to establish the following theorem are presented in the Appendix.

Theorem 2.2 (Hill Repression Scheme). *If $X_F^N(0) = (i_0, m_0, p_0) \in \mathcal{S}$, the sequence of processes $(P_F^N(t))$ converges in distribution to a birth and death process $(\overline{P}_F(t))$ on $\mathbb{N}$ whose respective birth and death rates (β_x) and (δ_x) are given by*

$$\beta_x = \frac{\lambda_3 \rho_2 \rho_1}{\rho_1 + x} \text{ and } \delta_x = \mu_3 x,$$

with $\rho_1 = \lambda_1^+/\lambda_1^-$ and $\rho_2 = \lambda_2/\mu_2$.

Proof. If f is a function on $\mathbb{N}$ with finite support then

$$V_f^N(t) \overset{\text{def.}}{=} f(P_F^N(t)) - f(p_0) - \int_0^t \lambda_3 M_F^N(u) \Delta^+(f)(P_F^N(u))\, du - \int_0^t \mu_3 P_F^N(u) \Delta^-(f)(P_F^N(u))\, du,$$

is a local martingale. See Rogers et al. (1987) for example. The operators Δ^+ and Δ^- are defined as follows, for a real-valued function f on $\mathbb{N}$,

$$\Delta^+(f)(x) = f(x+1) - f(x) \text{ and } \Delta^-(f)(x) = f(x-1) - f(x), \quad x \in \mathbb{N}.$$

With a similar method as in the proof of Assertion 1) of Lemma 2.2 in the appendix and by using the criterion of the modulus of continuity, see Theorem 7.2 page 81 of Billingsley (1999), it is easy to show that the two processes

$$\left(\int_0^t \lambda_3 M_F^N(u) \Delta^+(f)(P_F^N(u))\, du \right) \text{ and } \left(\int_0^t \mu_3 P_F^N(u) \Delta^-(f)(P_F^N(u))\, du \right)$$

are tight. Because of the tightness of $(P_F^N(t))$ of Proposition 2.5 of the appendix, one can take (N_k) a subsequence such that the process

$$\left(P_F^{N_k}(t), \int_0^t \lambda_3 M_F^{N_k}(u) \Delta^+(f)(P_F^{N_k}(u)) du, \int_0^t \mu_3 P_F^{N_k}(u) \Delta^-(f)(P_F^{N_k}(u)) du \right)$$

converges in distribution.

Let $(\overline{P}_F(t))$ be a possible limit of $(P_F^{N_k}(t))$, then by continuity of the mapping

$$(z(t)) \mapsto \left(\int_0^t z(u) \Delta^-(f)(z(u))\, du \right)$$

on $\mathcal{D}([0,T])$ endowed with the Skorohod topology then, for the convergence in distribution.

$$\lim_{k \to +\infty} \left(P_F^{N_k}(t), \int_0^t P_F^{N_k}(u) \Delta^-(f)(P_F^{N_k}(u))\, du \right) = \left(\overline{P}_F(t), \int_0^t \overline{P}_F(u) \Delta^-(f)(\overline{P}_F(u))\, du \right).$$

For $t \leq T$, by using the definition of Λ^N and of $\mathcal{E}_T$ in Subsection 2.A.2 of the Appendix, one has the relation

$$\int_0^t M_F^{N_k}(u) \Delta^+(f)(P_F^{N_k}(u))\, du = \int_{\mathcal{E}_T} m \Delta^+(f)(p) \mathbb{1}_{[0,t]}(u)\, \Lambda^{N_k}(dz),$$

hence, by Proposition 2.6 of Appendix, for the convergence in distribution

$$\begin{aligned} \lim_{k \to +\infty} \int_{\mathcal{E}_T} m \Delta^+(f)(p) \mathbb{1}_{[0,t]}(u)\, \Lambda^{N_k}(dz) &= \int_0^t \sum_{p \in \mathbb{N}} \Delta^+(f)(p) \sum_{(i,m) \in \{0,1\} \times \mathbb{N}} m \ell(i,m,p)\, du \\ &= \int_0^t \sum_{p \in \mathbb{N}} \Delta^+(f)(p) \frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^- p} \frac{\lambda_2}{\mu_2} \nu_u(p)\, du \end{aligned}$$

by Equation (2.17) of Proposition 2.6 of the Appendix. By convergence of the sequence (Λ^{N_k}) this last expression can be expressed as

$$\left(\int_0^t \sum_{p\in\mathbb{N}} \Delta^+(f)(p)\frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^- p}\nu_u(p)\,du\right) = \lim_{k\to+\infty}\left(\int_0^t \Delta^+(f)(P_F^{N_k}(u))\frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^- P_F^{N_k}(u)}\,du\right)$$
$$\overset{\text{dist.}}{=} \left(\int_0^t \Delta^+(f)(\overline{P}_F(u))\frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^-\overline{P}_F(u)}\,du\right)$$

for the convergence in distribution.

For $0 \le s \le t$, the characterisation of a Markov process as the solution of a martingale problem gives the relation

$$\mathbb{E}\left[f(P_F^N(t)) - f(P_F^N(s)) - \int_s^t \lambda_3 M_F^N(u)\Delta^+(f)(P_F^N(u))\,du - \int_s^t \mu_3 P_F^N(u)\Delta^-(f)(P_F^N(u))\,du\,\middle|\,\mathcal{F}_s\right] = 0,$$

from which we deduce the identity

$$\mathbb{E}\left[f(\overline{P}_F(t)) - f(\overline{P}_F(s)) - \int_s^t \lambda_3\frac{\rho_1\rho_2}{\rho_1+\overline{P}_F(u)}\Delta^+(f)(\overline{P}_F(u)) - \int_s^t \mu_3\overline{P}_F(u)\Delta^-(f)(\overline{P}_F(u))\,du\,\middle|\,\mathcal{F}_s\right] = 0.$$

See Theorem II.2.42 of Jacod and Shiryaev (1987). Consequently, a possible limit is the solution of the martingale problem associated to the birth and death process with birth rate (β_x) and death rate (δ_x) and with initial state in p_0. One gets therefore the desired convergence in distribution of $(P_F^N(t))$. The theorem is proved. □

There exist cases where the autoregulation is not achieved by the regulated protein but by a complex of this protein, e.g by a dimer (2 copies of the protein) or a tetramer (4 copies) to cite few examples. In order to handle such cases, it is necessary to add to the gene expression model, a preliminary step describing the reaction scheme of the complex formation based on the law of mass action (as it is done in Rosenfeld et al. (2002), Bokes et al. (2011)). In general, the dynamics involved in the reaction scheme are (very) rapid compared to the other processes of the gene expression and leads, by a singular perturbation like argument, to represent in case of deterministic model the rate of production of mRNAs as a non-linear function of protein concentration. Furthermore, when the reaction scheme possesses suitable properties, a Hill like repression function could also be obtained. See Weiss (1997) for details. In the stochastic context, that leads to introduce a suitable scaling factor in the dynamics of the complex formation and to extend the previous derivation in the previous theorem to Hill functions, $x \mapsto a/(b+x^n)$, with order n greater than 1.

The next section analyses, in this limiting regime, the fluctuations of the number of proteins at equilibrium.

2.4 Fluctuations of the Number of Proteins

This section is devoted to the analysis of the equilibrium of the asymptotic process $(\overline{P}_F(t))$ of Theorem 2.2 describing the evolution of the number of proteins with feedback. We start with a classical result for birth and death processes.

Proposition 2.3. *The invariant distribution π_F of the birth and death process $(\overline{P}_F(t))$ of Theorem 2.2 is given by*

$$\pi_F(x) = \frac{1}{Z}\frac{(\rho_2\rho_3)^x}{x!}\prod_{i=0}^{x-1}\frac{\rho_1}{\rho_1+i}, \qquad x\in\mathbb{N},$$

where $\rho_1 = \lambda_1^+/\lambda_1^-$, $\rho_i = \lambda_i/\mu_i$ *for* $i = 2,\ 3$ *and* Z *is the normalization constant.*

The expression of π_f is explicit but with a normalization constant which is not simple. The constant Z can be expressed in terms of hypergeometric functions. See Abramowitz and Stegun (1964) for example. Even if we can get a numerical evaluation of the average and of the variance of π_F, it is much more awkward to get some insight on the dependence of these quantities with respect to some of the parameters like ρ_2 or ρ_3 for example. In the following we give an asymptotic description of the ratio of the variance and the mean of the number of proteins at equilibrium when the value of the quantity $\rho_1\rho_2\rho_3$ is large. In a biological context the numerical value of this parameter is not always large but this limit results sheds some light on the qualitative behaviour of the auto-regulation mechanism. See Corollary Corollary 2.1 for example. A Laplace method is in particular used to investigate the asymptotic behaviour of the first two moments of π_F.

Theorem 2.1 shows that the distribution of the process $(P(t))$ at equilibrium is Poisson with parameter $\mathbb{E}[P(t)] = x_\rho = \rho_1\rho_2\rho_3/(1+\rho_1)$. In particular, one has the relation $\mathbb{V}\mathrm{ar}[P(t)] = \mathbb{E}[P(t)]$. In the rest of this section, we will be interested in the corresponding quantity for the feedback process.

For $\eta > 0$ and $\rho > 0$, denote by ν_ρ the probability distribution on $\mathbb{N}$ defined by

$$\nu_{\rho,\eta}(k) = \frac{1}{Z_\rho}\frac{\rho^k}{k!}\prod_{i=1}^{k}\frac{1}{\eta+i} = \frac{1}{Z_\rho}\exp\left(\sum_{i=1}^{k}\log\left(\frac{x}{i(\eta+i)}\right)\right), \tag{2.5}$$

where Z_ρ is the normalization constant. It is easily seen that π_F is $\nu_{\rho,\eta}$ with $\rho = \rho_1\rho_2\rho_3$ and $\eta = \rho_1 - 1$.

Proposition 2.4. *If, for* $\rho > 0$ *and* $\eta > -1$, A_ρ *is a random variable with distribution* $\nu_{\rho,\eta}$ *defined by Equation (2.5), then for the convergence in distribution*

$$\lim_{\rho\to+\infty}\frac{A_\rho - a_\rho}{\sqrt{a_\rho}} = \mathcal{N}\left(0, 1/\sqrt{2}\right),$$

where $a_\rho = \left(\sqrt{\eta^2+4\rho} - \eta\right)/2$ *and* $\mathcal{N}\left(0,1/\sqrt{2}\right)$, *is a centered Gaussian random variable with variance* $1/2$. *In particular, for the convergence in distribution,*

$$\lim_{\rho\to+\infty}\frac{A_\rho}{\sqrt{\rho}} = 1.$$

Proof. If ϕ is a bounded function on $\mathbb{R}$, denote

$$\Delta_\rho(\phi) \overset{\text{def.}}{=} \frac{1}{\sqrt{a_\rho}}\sum_{k=0}^{+\infty}\phi\left(\frac{k-\lceil a_\rho\rceil}{\sqrt{a_\rho}}\right)\exp\left(\sum_{i=\lceil a_\rho\rceil}^{k}\log\left(\frac{\rho}{i(\eta+i)}\right)\right),$$

with the following convention, to take care of the order of summation in discrete sums, if $(a_n, n\in\mathbb{Z})$ is a sequence of real numbers, for $\ell, m \in \mathbb{Z}$, then

$$\sum_{i=m}^{\ell} a_i \overset{\text{def.}}{=} -\sum_{i=\ell}^{m-1} a_i.$$

The definition of $\nu_{\rho,\eta}$ gives that

$$\mathbb{E}\left[\phi\left(\frac{A_\rho - \lceil a_\rho\rceil}{\sqrt{a_\rho}}\right)\right] = \frac{\Delta_\rho(\phi)}{\Delta_\rho(1)} \tag{2.6}$$

Fix ϕ some continuous function with compact support on $[-K_0, K_0]$ for some $K_0 > 0$. Since a_ρ is the solution of the equation $a_\rho(\eta + a_\rho) = \rho$, a change of variable gives the relation

$$\Delta_\rho(\phi) = \frac{1}{\sqrt{a_\rho}} \sum_{k=-\lfloor K_0\sqrt{a_\rho}\rfloor}^{\lceil K_0\sqrt{a_\rho}\rceil} \phi\left(\frac{k}{\sqrt{a_\rho}}\right) \exp\left(\sum_{i=0}^{k} \log\left(\frac{a_\rho(\eta+a_\rho)}{(i+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+i)}\right)\right).$$

The uniform estimation

$$\sum_{i=0}^{k} \log\left(\frac{a_\rho(\eta+a_\rho)}{(i+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+i)}\right) = \int_0^k \left(\log\left(\frac{a_\rho(\eta+a_\rho)}{(u+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+u)}\right) du + O\left(\frac{1}{\sqrt{a_\rho}}\right)\right)$$

for all $k \in \mathbb{Z}$ with $|k| \leq K_0\sqrt{a_\rho}$ and the fact that ϕ has a compact support give that the quantity $\Delta_\rho(\phi)$ is equivalent to

$$\begin{aligned}
&\frac{1}{\sqrt{a_\rho}} \sum_{k=-\lfloor K_0\sqrt{a_\rho}\rfloor}^{\lceil K_0\sqrt{a_\rho}\rceil} \phi\left(\frac{k}{\sqrt{a_\rho}}\right) \exp\left(\int_0^k \log\left(\frac{a_\rho(\eta+a_\rho)}{(u+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+u)}\right) du\right) \\
&= \frac{1}{\sqrt{a_\rho}} \sum_{k=-\lfloor K_0\sqrt{a_\rho}\rfloor}^{\lceil K_0\sqrt{a_\rho}\rceil} \phi\left(\frac{k}{\sqrt{a_\rho}}\right) \\
&\qquad \times \exp\left(\int_0^{k/\sqrt{a_\rho}} \sqrt{a_\rho} \log\left(\frac{a_\rho(\eta+a_\rho)}{(u\sqrt{a_\rho}+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+u\sqrt{a_\rho})}\right) du\right).
\end{aligned}$$

Again, with the uniform estimation

$$\sqrt{a_\rho} \log\left(\frac{a_\rho(\eta+a_\rho)}{(u\sqrt{a_\rho}+\lceil a_\rho\rceil)(\eta+\lceil a_\rho\rceil+u\sqrt{a_\rho})}\right) = -2u + O\left(\frac{1}{\sqrt{a_\rho}}\right),$$

for u in some fixed finite interval, one gets that

$$\Delta_\rho(\phi) \sim \frac{1}{\sqrt{a_\rho}} \sum_{k=-\lfloor K_0\sqrt{a_\rho}\rfloor}^{\lceil K_0\sqrt{a_\rho}\rceil} \phi\left(\frac{k}{\sqrt{a_\rho}}\right) \exp\left(-2\int_0^{k/\sqrt{a_\rho}} u\, du\right) \sim \int_{-\infty}^{+\infty} \phi(v)\, e^{-v^2}\, dv.$$

With similar estimations for $\Delta_\rho(1)$ (which imply in fact the tightness of the random variables $(A_\rho - \lfloor a_\rho\rfloor)/\sqrt{a_\rho}$) and Equation (2.6), the proposition is proved. □

Corollary 2.1 (Asymptotic Number of Proteins with Regulation). *If $\overline{P}_F$ is a random variable with distribution π_F then, for the convergence in distribution*

$$\lim_{\rho_2\rho_3\to+\infty} \frac{\mathbb{E}\left[\overline{P}_F\right]}{\sqrt{\rho_1\rho_2\rho_3}} = 1 \text{ and } \lim_{\rho_2\rho_3\to+\infty} \frac{\mathbb{V}\text{ar}\left[\overline{P}_F\right]}{\mathbb{E}\left[\overline{P}_F\right]} = \frac{1}{2}. \tag{2.7}$$

Furthermore, for the convergence in distribution,

$$\lim_{\rho_2\rho_3\to+\infty} \frac{\overline{P}_F - a_\rho}{\sqrt{a_\rho}} = \mathcal{N}\left(0, 1/\sqrt{2}\right),$$

where $a_\rho = \left(\sqrt{(\rho_1-1)^2 + 4\rho_1\rho_2\rho_3} - \rho_1 + 1\right)/2$.

The equivalent of Equation (2.7) for the scaling of the classical model of protein production is

$$\mathbb{E}[P] = \frac{\rho_1}{1+\rho_1}\rho_2\rho_3 \text{ and } \frac{\mathbb{V}\text{ar}[P]}{\mathbb{E}[P]} = 1.$$

by Theorem 2.1. it shows that a feedback mechanism reduces the variance of the number of proteins in this limiting regime by a factor 2 for the ratio of the second moment and the first moment.

2.5 Discussion

In this section, other aspects of regulation of protein production are discussed via simulations in a plausible biological context whose parameters are going to be defined. These simulations are performed using the Gillespie (1977) algorithm. Simulation follows the models in Subsection 2.2.2 and simulates the variables I_F, M_F, and P_F, not their scaling limits.

2.5.1 Numerical Values of Biological Parameters

For the model with feedback, there are six parameters to determine. By using the literature one can estimate the common orders of magnitude of these parameters in a biological context. We therefore propose a set of parameters corresponding to an "ordinary" gene.

1. Gene regulation. The parameter λ_1^- gives the rate at which a given protein reaches its own promoter. It has been shown that this motion combines a three-dimensional diffusion in the cytoplasm and one-dimensional sliding along the DNA, see Halford (2009).

   Experiments on the lac repressor, using live-cell single-molecule imaging techniques, show that this time is of the order of 5 min, see Li and Elf (2009) and Hammar et al. (2012). For this reason we will take $\lambda_1^- = 3.3 \times 10^{-3}\, s^{-1}$.

   The parameter λ_1^+ can be quite variable, depending on the affinity of the protein to the DNA sequence, we set $\lambda_1^+ = 1\, s^{-1}$.

2. mRNAs. The lifetime of an mRNA is $\mu_2^{-1} \simeq 4$ min, see Taniguchi et al. (2010). When the gene expression is always active (corresponding to the case where our variable I remains equals to 1), there is an average of 2 messengers, that is to say $\lambda_2^{-1} = \mu_2^{-1}/m = 120\, s$ which gives $\lambda_2 = 8.3 \times 10^{-3}\, s^{-1}$.

3. Proteins. A doubling time for the cell of $t_{1/2} \simeq 40$ min gives a protein decay of around one hour. For this reason one takes $\mu_3 = \log 2/t_{1/2} = 2.8 \times 10^{-4}\, s^{-1}$ for the rate of protein decay. It is assumed that a give type of protein that is produced in $p = 300$ copies when the gene expression is always active. From one messenger, a protein should be produced in a duration of time of the order of $\lambda_3^{-1} = m \times \mu_3^{-1}/p$ which gives $\lambda_3 = 4 \times 10^{-2}\, s^{-1}$.

These parameters may correspond to an "ordinary bacterial" gene: in a E. Coli genome of 4300 genes, there are around 3.6×10^6 proteins and 1.4×10^3 mRNAs per gene, see Table 1 of Chapter 3 of Neidhardt and Umbarger (1996), the number of messengers and proteins is of the order of magnitude of our numerical estimation of the parameters.

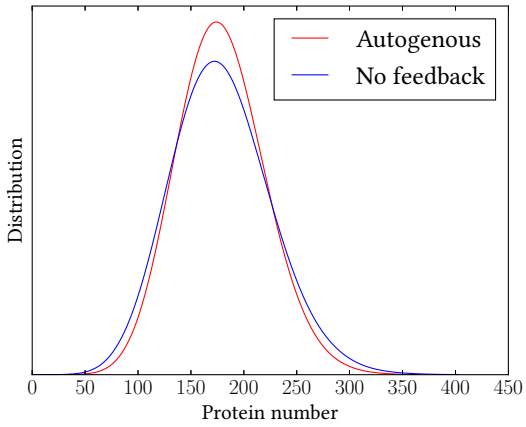


Figure 2.1: Simulations: Protein distribution with and without autogenous regulation with a fixed mean number of proteins of 178.

2.5.2 Impact of Autogenous Regulation on Gene Expression

We have compared two mechanisms: the classical model without regulation and the autogenous regulation process. The mean number of proteins is the same as well as the mean number of mRNAs produced $\mathbb{E}[M] = \mathbb{E}[M_F]$. Parameters λ_1^+ and λ_1^- are adapted in the classical model to fulfil these conditions. The other parameters are as defined in the previous section.

The comparison is shown in Figure 2.1. The mean number of proteins is 178, as can be seen that the curve for the autogenous regulation is slightly more concentrated around the mean but not that much. The values of the corresponding standard deviations are not really different $\sqrt{\mathbb{V}\text{ar}[P]} = 42.2$ and $\sqrt{\mathbb{V}\text{ar}[P_F]} = 35.8$. The impact of the autogenous regulation on the variability of the number of proteins is non-trivial but not really spectacular for the set of parameters associated to a "typical" gene. This is significantly less than the performances of the limiting mathematical model studied in Section 2.3. The main reason seems to be that the scaling parameter is not, in some cases, sufficiently large to have a reasonable accuracy with the limit given by the convergence result of Theorem 2.2.

2.5.3 The Limiting Scaling Regime as a Lower Bound

Roughly speaking, Theorem 2.1 and Corollary 2.1 give that for N and $\rho_2\rho_3$ large, then the ratio $\mathbb{V}\text{ar}\left[P_F^N\right]/\mathbb{E}\left[P_F^N\right]$ converges to $1/2$. In Figure 2.2, one considered a simulation with fixed product $\rho_2\rho_3 = 71.43$ with N varying. The interesting feature is that the ratio is decreasing with N, this suggests that the variance of the limit of the scaling procedure should provide a lower bound for the variance of the real model. We have not been able to show rigorously this phenomenon. For $N = 250$, the value of the ratio $\mathbb{V}\text{ar}\left[P_F^N\right]/\mathbb{E}\left[P_F^N\right] = .7964$ which is quite far from its limiting value $1/2$ given by Corollary 2.1. This can be explained by the fact that the quantities N and $\rho_2\rho_3$ are not very large.

2.5.4 Regulation of the Production Process on mRNAs

The regulation on the gene has the effect of an ON/OFF mechanism. When the gene is active, it is producing mRNAs at full speed and no mRNA is produced when it is inactive. This suggests that the production of proteins follows roughly the same pattern: steady production rate at some instants and little is produced

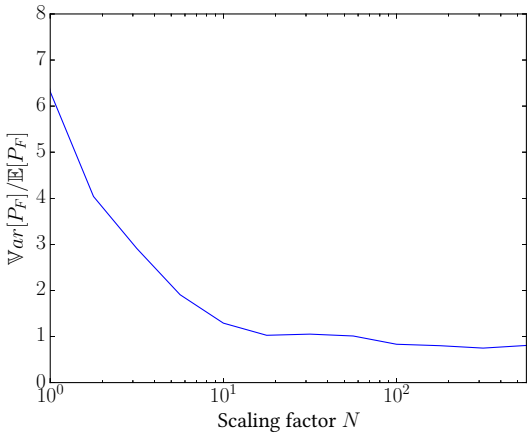


Figure 2.2: Simulations: Evolution of the ratio $\mathbb{V}\text{ar}\left[P_F^N\right]/\mathbb{E}\left[P_F^N\right]$ as a function of N.

otherwise. This scheme can consequently increase the variability of the production process of proteins. A possible idea to reduce the variance due to the activation/inactivation of the gene is to transfer the activation/inactivation process at the level the mRNAs. This possibility is investigated in this section. Each mRNA can be inactivated by a protein at rate λ_2^-, in this state it cannot produce proteins. An inactivated mRNAs becomes active at rate λ_2^+. In this way the production process can, hopefully, be modulated more smoothly by playing on the inactivation of a fraction of the mRNAs. In this way at time t, if the number of active [resp. inactive] mRNAs is $M(t)$ [resp. $M^*(t)$], the process $(M(t), M^*(t), P(t))$ is Markov with transition rates, for $(m, m^*, p) \in \mathbb{N}^3$,

$$\begin{cases} (m, m^*, p) \to (m+1, m^*, p) \text{ at rate } \lambda_2, \\ (m, m^*, p) \to (m-1, m^*+1, p) \text{ at rate } \lambda_2^- mp, \\ (m, m^*, p) \to (m+1, m^*-1, p) \text{ at rate } \lambda_2^+ m^*, \end{cases}$$

the other transitions are as before, active of inactive mRNAs die at rate μ_2 and proteins are produced at rate $\lambda_3 m$ and die at rate μ_3.

To compare the two regulation processes, either on the gene or on mRNAs, simulations have been done with the following constraints: the average number of proteins is fixed around 1400.[1] To have a fair comparison, we add the constraint that the number of mRNAs produced should be the same in all simulations. The numerical values have been estimated by using similar methods as in Section 2.3 but for this setting. Experiment (3) considers the case of an average lifetime of an mRNA of 40 min, if this is far from a "normal" biological setting, as it will be seen, this scenario has the advantage of stressing the importance of this parameter in this configuration.

Numerical Values of Parameters

Regulation on the gene.

[1] For the model with regulation on the gene, we determined the parameters by using Equation (2.3) and by fixing $\mathbb{E}[P] = 1400$. We also make the approximation that $\mathbb{E}[I] \simeq \lambda_1^+/(\lambda_1^+ + \lambda_1^- \mathbb{E}[P])$. The resulting simulations show a relatively precise (the mean around 1403). A similar strategy to determine the parameters of the model with regulation on mRNAs.

λ_1^+	λ_1^-	λ_2	μ_2	λ_3	μ_3
0.21”	5’	12”	4’	25”	1h.

Regulation on mRNAs (I). For this experiment, the expected lifetime of an mRNA is twice the corresponding value of case (1).

λ_2	λ_2^+	λ_2^-	μ_2	λ_3	μ_3
23”	2”	45’	8’	25”	1h.

Regulation on mRNAs (II). For this second experiment on the regulation of mRNAs, the expected lifetime of an mRNA is 10 times than in case (1).

λ_2	λ_2^+	λ_2^-	μ_2	λ_3	μ_3
23.8”	2”	45’	40’	25”	1h.

Results of the Experiments

Table Table 2.1 shows that the mean number of mRNAs produced per unit of time is essentially the same in all experiments as well as the mean number of active mRNAs. It should be noted the impact of regulation on mRNAs for the standard deviation of the number of proteins when the mean life time is 8 min is not really significant (10% gain) than the regulation on the gene. When the mean lifetime is 40 min the improvement, 36%, of the standard deviation becomes significant, showing that in this case the production process is “smoothed” by this mechanism. The three distributions of the number of proteins of these experiments are presented in Figure 2.3.

Regulation on	Gene	mRNAs/8 min	mRNAs/40 min
Mean number of mRNAs	10.33	19.74	99.04
Mean number of Active mRNAs	10.33	9.77	9.81
Mean number of Proteins	1403.63	1400.29	1403.36
Standard Deviation of number of Proteins	92.66	84.22	59.04

Table 2.1: Comparison of Regulation Processes on Gene or on mRNAs with Different Lifetimes

2.5.5 Impact of Feedback on Frequency

In this section, we study the nature of the fluctuations of the number of proteins at equilibrium from the point of view of signal processing or automatic control. The aim of a feedback is often of changing the nature of the signal, attenuating disturbances by reducing, for instance, high frequencies. In these cases, spectral analysis gives a characterisation of the nature of changes.

By analogy, we consider our model as a system that has to achieve a command (the production of a given mean number of proteins) and where the resulting signal $P(t)$ or $P_F(t)$ is altered by some noise. In this framework, one can study if the effect of the feedback has an impact on the signal, by rejection of some frequency ranges.

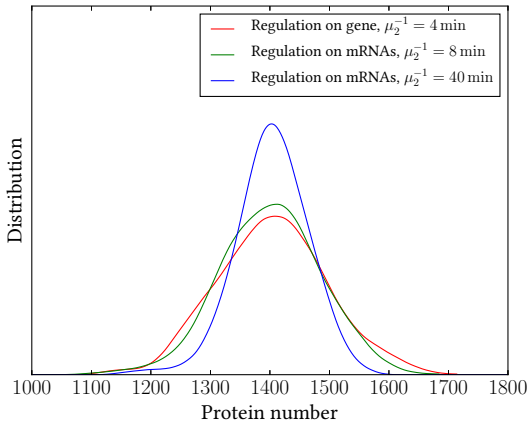


Figure 2.3: Simulations: Probability Distribution of the Number of Proteins with Regulation on Gene or on mRNAs, μ_2^{-1} is the average lifetime of an mRNA. The average number of proteins is 1400.

To do so, consider the signals $(P(t))$ and $(P_F(t))$ of two simulations with or without autogenous regulation. The analysis of these signals is done by estimating the power spectral density, that describes the spectral characteristics of stochastic process. We estimate the power spectral density for each signal, using classical estimator of smoothed periodogram. See George et al. (1978) and Chapter 10 of Miller and Childers (2012) for example.

The result is shown in Figure 2.4. Both spectra seem to represent a low-pass filter with a cut off frequency in the order of magnitude of the dilution factor $\mu_3 = 2.8 \times 10^{-4}\, s^{-1}$. The two power spectral densities do not seem to exhibit significant differences. The feedback has therefore no noticeable effect in terms of reduction of frequency disturbances.

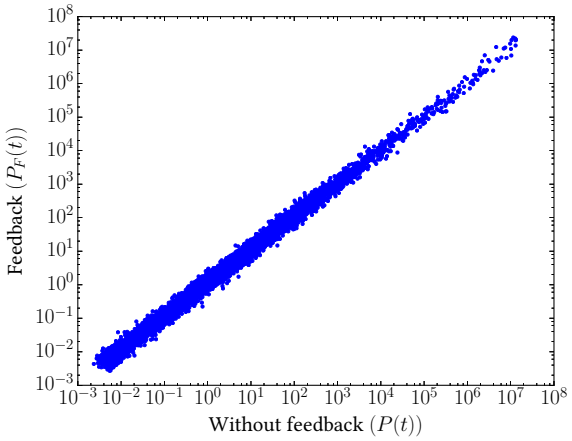


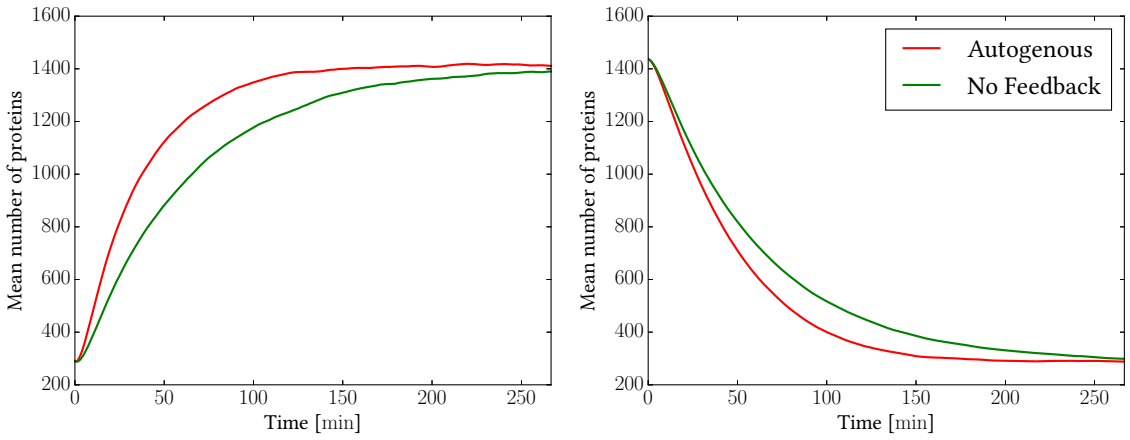
Figure 2.4: Power spectral density estimation of signals with and without regulation

2.5.6 Versatility of the Protein Production Process

This section is devoted to the impact of autogenous regulation on another aspect of protein production. Up to now, we have considered the production process of proteins at equilibrium, by assuming that the production rate of a given protein has to be fixed. It may happen nevertheless that, due to an external stress, such as antibiotics, DNA damage by UV, see Camas et al. (2006), or nutriment absorption, see Schleif (2000), the cell has to change rapidly its production rate to quickly produce a large amount of proteins for example. The affinity of the transcription factor for the promoter of the gene can be adapted for that purpose. Conversely, when the external stress disappears, the production of the protein has to be quickly reduced to minimize the consumption of resources.

We consider the situation when the two production processes, with and without autogenous regulation, give the same average output of proteins at equilibrium. Two cases are investigated: when the initial number of proteins is below the value equilibrium, see Figure 2.5a, or above this value, see Figure 2.5b. As it can be seen, the autogenous production process converges more rapidly to equilibrium in both cases. Our simulations show that when the initial value is 290, the autogenous production process is 40% faster than the process without feedback to reach the level 1300 (the equilibrium is at 1400 in this case). A similar result holds in the other case.

These interesting properties are related to the modulation of the gene activity. In the experiment of Figure 2.5a, for the autogenous process the rate of activity of the gene is of the order of 50% at the beginning and it is only of the order of 0.1 later at equilibrium. Without regulation this rate is constant throughout the simulation. This explains the "fast start" of the autogenous process. An analogous explanation holds for the experiment of Figure 2.5a, in the autogenous process. The gene is rapidly switched off due to the large number of proteins, thereby decreasing rapidly the number of proteins. This is consistent with experiments described in Camas et al. (2006) and especially Rosenfeld et al. (2002) where the improvement has been estimated at 80% in some cases.



(a) Initial Point at 290, equilibrium at 1400.

(b) Initial Point at 1400, equilibrium at 290.

Figure 2.5: Simulations: Evolution of the Mean Number of Proteins

2.A Appendix: Convergence Results

We first introduce some notations that will be used throughout this section.

2.A.1 Evolution Equations

We will use the Skorohod's topology for convergence in distribution in the space $\mathcal{D}([0,T],\mathbb{R}_+)$ of càdlàg processes. See Chapter 3 of Billingsley (1999) for example. To simplify the presentation, all our processes will be defined on the same probability space in the following way.

Let $\mathcal{N}_i^+$, $\mathcal{N}_i^-$, $i=1,2,3$ be independent Poisson processes on $\mathbb{R}_+^2$ with rate 1 defined on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. If $A\in\mathcal{B}(\mathbb{R}_+^2)$ is a Borelian subset of $\mathbb{R}_+^2$ and $(i,c)\in\{1,2,3\}\times\{+,-\}$, $\mathcal{N}_i^c(A)$ denotes the number of points of the process $\mathcal{N}_i^c$ in the subset A. For $t\geq 0$, one denotes by $\mathcal{F}_t$ the σ-field generated by the random variables

$$\mathcal{N}_i^c(B\times[0,t]) \text{ for } B\in\mathcal{B}(\mathbb{R}_+) \text{ and } (i,c)\in\{1,2,3\}\times\{+,-\}.$$

It is easily seen that the process $(X_F^N(t))$ has the same distribution as the solution of the following stochastic differential equations (SDE)

$$\begin{aligned} dI_F^N(t) &= \mathbb{1}_{\{I_F^N(t-)=0\}}\mathcal{N}_1^+([0,\lambda_1^+N]\times[dt]) \\ &\qquad - \mathbb{1}_{\{I_F^N(t-)=1\}}\mathcal{N}_1^-([0,\lambda_1^-NP_F^N(t-)]\times[dt]) \qquad (2.8)\\ dM_F^N(t) &= \mathbb{1}_{\{I_F^N(t-)=1\}}\mathcal{N}_2^+([0,\lambda_2N]\times[dt]) - \mathcal{N}_2^-([0,\mu_2NM_F^N(t-)]\times[dt]) \qquad (2.9)\\ dP_F^N(t) &= \mathcal{N}_3^+([0,\lambda_3M_F^N(t-)]\times[dt]) - \mathcal{N}_3^-([0,\mu_3P_F^N(t-)]\times[dt]) \qquad (2.10)\end{aligned}$$

with the same initial condition. For any $N\geq 1$, $(X_F^N(t))$ is a $\mathcal{F}_t$-Markov process adapted to the filtration $\mathcal{F}_t$ defined as $\sigma\left(X_F^N(0);\mathcal{N}_i^c(A\times[0,s[),s\leq t,A\in\mathcal{B}(\mathbb{R}_+^2),(i,c)\in\{1,2,3\}\times\{+,-\}\right)$. These SDE can be rewritten as, for some function f with finite support on $\mathcal{S}$,

$$\begin{aligned} f(X_F^N(t)) = f(X_F^N(0)) &+ \int_0^t \lambda_1^+N(1-I_F^N(u))\Delta_1(f)(X_F^N(u))\,du \\ &+ \int_0^t \lambda_1^-NP_F^N(u)I_F^N(u)\Delta_1(f)(X_F^N(u))\,du \\ &+ \int_0^t \lambda_2NI_F^N(u)\Delta_2^+(f)(X_F^N(u))\,du + \int_0^t \mu_2NM_F^N(u)\Delta_2^-(f)(X_F^N(u))\,du \\ &+ \int_0^t \lambda_3M_F^N(u)\Delta_3^+(f)(X_F^N(u))\,du + \int_0^t \mu_3P_F^N(u)\Delta_3^-(f)(X_F^N(u))\,du + W_f^N(t) \qquad (2.11)\end{aligned}$$

where, for $x=(i,m,p)\in\mathcal{S}$, the operators $\Delta^{+/-}$ are defined by

$$\begin{cases} \Delta_1(f)(x) = f(1-i,m,p)-f(x) & \\ \Delta_2^+(f)(x) = f(i,m+1,p)-f(x), & \Delta_2^-(f)(x) = f(i,m-1,p)-f(x) \\ \Delta_3^+(f)(x) = f(i,m,p+1)-f(x), & \Delta_3^-(f)(x) = f(i,m,p-1)-f(x), \end{cases}$$

and $(W_f^N(t))$ is a local martingale whose previsible increasing process is given by

$$\begin{aligned}\left\langle W_f^N \right\rangle(t) = &\int_0^t \left[\lambda_1^+ N(1 - I_F^N(u)) + \lambda_1^- N P_F^N(u) I_F^N(u)\right] \left[\Delta_1(f)(X_F^N(u))\right]^2 du \\ &+ \int_0^t \lambda_2 N I_F^N(u) \left[\Delta_2^+(f)(X_F^N(u))\right]^2 du + \int_0^t \mu_2 N M_F^N(u) \left[\Delta_2^-(f)(X_F^N(u))\right]^2 du \\ &+ \int_0^t \lambda_3 M_F^N(u) \left[\Delta_3^+(f)(X_F^N(u))\right]^2 du + \int_0^t \mu_3 P_F^N(u) \left[\Delta_3^-(f)(X_F^N(u))\right]^2 du. \quad (2.12)\end{aligned}$$

See Rogers et al. (1987) for example.

Definition 2.1. Let $(\overline{M}^N(t), \overline{P}^N(t))$ be the Markov process with transition rates given by

$$\begin{cases} (m,p) \to (m+1,p) \text{ at rate } \lambda_2 N, & (m,p) \to (m-1,p) \quad " \quad \mu_2 m N, \\ (m,p) \to (m,p+1) \quad " \quad \lambda_3 m, & (m,p) \to (m,p-1) \quad " \quad \mu_3 p \end{cases} \quad (2.13)$$

and initial state $(\overline{M}^N(0), \overline{P}^N(0)) = (m_0, p_0)$.

The process $(\overline{M}^N(t), \overline{P}^N(t))$ is simply the analogue of our process $(M_F^N(t), P_F^N(t))$ when the gene is always active.

Lemma 2.2. *1. For the convergence in distribution for the uniform norm on compact sets*

$$\lim_{N \to +\infty} \left(\int_0^t \overline{M}^N(u)\, du \right) = (\rho_2 t).$$

2. For $T > 0$,

$$\sup_{N \geq 1} \mathbb{E}\left[\sup_{0 \leq t \leq T} \overline{P}^N(t) \right] < +\infty.$$

Proof. From Equation (2.13), it is easily seen that the process $(\overline{M}^N(t))$ can be expressed $(L_1(Nt))$ where $(L_1(t))$ is an $M/M/\infty$ queue with arrival rate λ_2 and service rate μ_2 with $L_1(0) = m_0$. See Chapter 6 of Robert (2010) for example. Elementary stochastic calculus gives, for $t > 0$,

$$L_1(Nt) = m_0 + \lambda_2 N t - \mu_2 \int_0^{Nt} L_1(u)\, du + \mathcal{M}_1^N(t), \quad (2.14)$$

where $(\mathcal{M}_1^N(t))$ is a local martingale whose previsible increasing process is given by

$$\left\langle \mathcal{M}_1^N \right\rangle(t) = \lambda_2 N t + \mu_2 \int_0^{Nt} L_1(u)\, du.$$

It is possible to show that $(\mathcal{M}_1^N(t))$ is a martingale: we have that for every $t > 0$

$$\left|\mathcal{M}_1^N(t)\right| < m_0 + \mathcal{N}_2[0, Nt[+ \lambda_2 N t + \sum_{i=0}^{\infty} \int_0^{Nt} \mathbb{1}_{i \leq L_1(u)} \mathcal{N}_{\mu 2}^i(ds) + \mu_2 \int_0^{Nt} L_1(u)\, du,$$

with $\left(\mathcal{N}_{\mu 2}^{i}\right)$ independent point Poisson processes of rate μ_2. It follows that

$$\mathbb{E}\left[\sup_{s\leq t}\left|\mathcal{M}_1^N(t)\right|\right] \leq m_0 + 2\lambda_2 Nt + 2\lambda\mu N^2 t^2,$$

so with Theorem A.7 of Robert (2010), it comes that $(\mathcal{M}_1^N(t))$ is a martingale. By applying Doobs' inequality, it shows that the process $(\mathcal{M}_1^N(t)/N)$ vanishes for the convergence in distribution as N gets large.

For $\varepsilon > 0$ and $x \in \mathbb{N}$, if

$$T_x = \inf\{t \geq 0 : L_1(u) \geq x\},$$

Proposition 6.10 of Robert (2010) shows the convergence in distribution

$$\lim_{x\to+\infty} \frac{\rho_2^x}{(x-1)!} T_x = E_0$$

where E_0 is an exponential random variable with parameter $\mu_2 \exp(-\rho_2)$. This shows in particular the process $(L_1(Nt)/N)$ converges in distribution to 0 for the uniform convergence on compact intervals since

$$\mathbb{P}\left[\left(\sum_{0\leq t\leq T} \frac{L_1(Nt)}{N} \geq \varepsilon\right)\right] \leq \mathbb{P}\left[\left(T_{\lfloor \varepsilon N\rfloor} \leq NT\right)\right].$$

From Equation (2.14), one gets

$$\int_0^t \overline{M}^N(u)\,du = \frac{1}{N}\int_0^{Nt} L_1(u)\,du = \rho_2 t + \frac{1}{\mu_2}\left(\frac{m_0}{N} - \frac{L_1(Nt)}{N} + \frac{\mathcal{M}_1^N(t)}{N}\right)$$

and therefore assertion 1) of the lemma.

For the last assertion, the method is similar: one first write the evolution equation

$$\overline{P}^N(t) = p_0 + \lambda_3 \int_0^t \overline{M}^N(u)\,du - \mu_3 \int_0^t \overline{P}^N(u)\,du + \mathcal{M}_2^N(t),$$

where $(\mathcal{M}_2^N(t))$ is a local martingale whose previsible increasing process is given by

$$\left\langle \mathcal{M}_2^N \right\rangle(t) = \lambda_3 \int_0^t \overline{M}^N(u)\,du + \mu_3 \int_0^t \overline{P}^N(u)\,du.$$

As for $(\mathcal{M}_1^N(t))$, it is possible to show that the local martingale $(\mathcal{M}_2^N(t))$ is indeed a martingale by showing that for every t, $\mathbb{E}\left[\sup_{s\leq t}\left|\mathcal{M}_2^N(t)\right|\right]$ is finite.

Define $\overline{P}_*^N(t) = \sup\{\overline{P}^N(u) : 0 \leq u \leq t\}$, then for $0 \leq t \leq T$

$$\mathbb{E}\left[\overline{P}_*^N(t)\right] \leq p_0 + \lambda_3 \mathbb{E}\left[\int_0^T \overline{M}^N(u)\,du\right] + \mathbb{E}\left[\sup_{0\leq u\leq t}\left|\mathcal{M}_2^N(u)\right|\right] + \mu_3 \int_0^t \mathbb{E}\left[\overline{P}_*^N(u)\right]\,du. \tag{2.15}$$

Doob's Inequality gives, for $t \leq T$,

$$\mathbb{E}\left[\sup_{0\leq u\leq t}\left|\mathcal{M}_2^N(u)\right|\right] \leq 2\lambda_3 \int_0^T \mathbb{E}\left[\overline{M}^N(u)\right]\,du + 2\mu_3 \int_0^t \mathbb{E}\left[\overline{P}_*^N(u)\right]\,du,$$

and from the ergodic theorem for $(L_1(t))$ (recall that $\overline{M}^N(t) = L_1(Nt)$) one gets

$$\lim_{N\to+\infty} \mathbb{E}\left[\int_0^T \overline{M}^N(u)\,du\right] = \rho_3 T.$$

One concludes by using Gronwall's Lemma. □

Proposition 2.5. *The sequence $(P_F^N(t))$ is tight for the convergence in distribution of càdlàg processes.*

Proof. Aldous' criterion for tightness is used. See Theorem 4.5 page 320 of Jacod and Shiryaev (1987) for example. For $T > 0$, one denotes by $\mathcal{T}_T$ the set of stopping times associated to the filtration $(\mathcal{F}_t)$ which are bounded by T. For $\eta > 0$, let $\tau_1, \tau_2 \in \mathcal{T}_T$ be such that $\tau_1 \le \tau_2 \le \tau_1 + \eta$. The respective probabilities that, on the time interval $[\tau_1, \tau_2]$, no protein is made or that no protein is degraded are respectively given by

$$\mathbb{E}\left[\exp\left(-\lambda_3 \int_{\tau_1}^{\tau_2} M_F^N(u)\,du\right)\right] \text{ and } \mathbb{E}\left[\exp\left(-\mu_3 \int_{\tau_1}^{\tau_2} P_F^N(u)\,du\right)\right]$$

By using the strong Markov property, one gets the relation

$$\mathbb{P}\left[\left(|P_F^N(\tau_1) - P_F^N(\tau_2)| \ge 1\right)\right] \le 1 - \mathbb{E}\left[\left(\exp\left(-\lambda_3 \int_{\tau_1}^{\tau_2} M_F^N(u)\,du\right)\right)\right] + 1 - \mathbb{E}\left[\left(\exp\left(-\mu_3 \int_{\tau_1}^{\tau_2} P_F^N(u)\,du\right)\right)\right].$$

With a simple coupling using the same Poisson processes $\mathcal{N}_{2/3}^{+/-}$ of Equation (2.9) and Equation (2.10) gives a process as in Definition 2.1 on the same probability space such that the relations $M_F^N(t) \le \overline{M}^N(t)$ and $P_F^N(t) \le \overline{P}^N(t)$ hold almost surely for all $t \ge 0$. From the last relation, one gets the inequality

$$\begin{aligned}
\mathbb{P}\left[\left(|P_F^N(\tau_1) - P_F^N(\tau_2)| \ge 1\right)\right] &\le 1 - \mathbb{E}\left[\exp\left(-\lambda_3 \int_{\tau_1}^{\tau_1+\eta} \overline{M}^N(u)\,du\right)\right] \\
&\qquad + 1 - \mathbb{E}\left[\exp\left(-\mu_3 \eta \sup_{0\le t\le T} \overline{P}^N(t)\right)\right] \\
&\le 1 - \mathbb{E}\left[\exp\left(-\lambda_3 \sup_{0\le t\le T} \int_t^{t+\eta} \overline{M}^N(u)\,du\right)\right] \\
&\qquad + 1 - \mathbb{E}\left[\exp\left(-\mu_3 \eta \sup_{0\le t\le T} \overline{P}^N(t)\right)\right].
\end{aligned}$$

Lemma 2.2 gives the relation

$$\lim_{N\to+\infty} \sup_{\tau_1\in\mathcal{T}_T} \mathbb{E}\left[\exp\left(-\lambda_3 \int_{\tau_1}^{\tau_1+\eta} \overline{M}^N(u)\,du\right)\right] = e^{-\lambda_3\rho_2\eta}$$

and, for $\varepsilon > 0$, the existence of $K > 0$ such that

$$\sup_{N\ge1} \mathbb{P}\left[\left(\sup_{0\le t\le T} \overline{P}^N(t) \ge K\right)\right] \le \varepsilon.$$

Consequently

$$\lim_{\eta\to0} \lim_{N\to+\infty} \sup_{\substack{\tau_1,\tau_2\in\mathcal{T}_T \\ \tau_1\le\le\tau_2\le\tau_1+\eta}} \mathbb{P}\left[\left(|P_F^N(\tau_1) - P_F^N(\tau_2)| \ge 1\right)\right] = 0,$$

hence, by Aldous' criterion, the tightness of the sequence $(P_F^N(t))$ is established. The proposition is proved. □

2.A.2 Convergence of Occupation Measures

For $N \geq 1$ and $T > 0$, one defines the random measure Λ^N on $\mathcal{E}_T \overset{\text{def.}}{=} \{0,1\} \times \mathbb{N}^2 \times [0,T]$ as follows, for a non-negative Borelian function G on $\mathcal{E}_T$,

$$\Lambda^N(G) = \int_0^T G(X_F^N(u), u)\, du.$$

If A is a Borelian subset of $\mathcal{E}_T$, $\Lambda^N(A)$ denotes $\Lambda^N(\mathbb{1}_A)$.

Proposition 2.6. *The sequence Λ^N of random measures is tight and any of its limiting points Λ can be written as*

$$\Lambda(F) = \sum_{(i,m,p)\in\mathcal{S}} \int_0^T G(i,m,p,u)\pi_p(i,m)\nu_u(p)\, du.$$

where, for any $u \leq T$, ν_u is a positive measure on $\mathbb{N}$ such that, almost surely,

$$\int_0^t \nu_u(\mathbb{N})\, du = t, \quad \forall t \leq T,$$

and, for $p \in \mathbb{N}$, π_p is the invariant distribution of the Markov process on $\{0,1\} \times \mathbb{N}$ whose transition rates are given by, for $(i,m) \in \{0,1\} \times \mathbb{N}$,

$$\begin{cases} (i,m) \to (1-i,m) & \text{at rate } \lambda_1^+ i + \lambda_1^- p(1-i), \\ (i,m) \to (i,m+1) & \lambda_2 i, \\ (i,m) \to (i,m-1) & \mu_2 m. \end{cases} \tag{2.16}$$

Additionally, one has

$$\sum_{(i,m)\in\{0,1\}\times\mathbb{N}} m\pi_p(i,m) = \frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^- p}\frac{\lambda_2}{\mu_2}. \tag{2.17}$$

Proof. For $K > 0$, if $\mathcal{K}_K$ is the compact subset $\{0,1\} \times [0,K]^2 \times [0,T]$ of $\mathcal{E}_T$, then

$$\mathbb{E}\left[\Lambda^N\left(\mathcal{E}_T \setminus \mathcal{K}_K\right)\right] \leq \int_0^T \mathbb{P}\left[\left(M_F^N(u) \geq K\right)\right] du + T\mathbb{P}\left[\left(\sup_{0\leq u\leq T} P_F^N(u) \leq K\right)\right].$$

By using the same coupling as in the proof of Proposition 2.5, one gets that

$$\mathbb{E}\left[\Lambda^N\left(\mathcal{E}_T \setminus \mathcal{K}_K\right)\right] \leq \int_0^T \mathbb{P}\left[\left(\overline{M}^N(u) \geq K\right)\right] du + T\mathbb{P}\left[\left(\sup_{0\leq u\leq T} \overline{P}^N(u) \leq K\right)\right].$$

By Lemma 2.2, for $\varepsilon > 0$, there exists some K such that

$$\sup_{N\geq 1} \mathbb{E}\left[\Lambda^N\left(\mathcal{E}_T \setminus \mathcal{K}_K\right)\right] \leq \varepsilon.$$

Consequently, the sequence (Λ^N) of random Radon measures on $\mathcal{E}_T$ is tight. See Dawson (1993, Lemma 3.28, page 44) for example.

Proof. Let Λ be a limiting point of some subsequence $(\Lambda^{N_k}(\cdot))$. By using Radon-Nikodym's Theorem, see Chapter 8 of Rudin (1986) for example, it is not difficult to see that there exists some non-negative random variables $(\ell_u(x)(\omega), (\omega, x, u) \in \Omega \times \mathcal{S} \times [0,T])$ such that $(\omega, x, u) \mapsto \ell_u(x)(\omega)$ is measurable and Λ can be expressed as

$$\Lambda(G) = \sum_{x \in \mathcal{S}} \int_0^T G(x,u)\ell_u(x)\, du.$$

From the domination relation of Lemma 2.2, one gets that, almost surely, there is no loss of mass, i.e.

$$\int_0^t \ell_u(\mathcal{S})\, du = t, \quad \forall t \leq T, \tag{2.18}$$

holds almost surely. Now take a function f with bounded support on $\mathcal{S}$ and let's use Equation (2.12). As previously, we can apply the Doob's Inequality to this process $(\langle W_f^N \rangle (t \wedge T))$ and show that the process $(\langle W_f^N \rangle (t))$ satisfies the relation

$$\lim_{N \to +\infty} \frac{1}{N^2} \mathbb{E}\left[\left\langle W_f^N \right\rangle (T)\right] = 0.$$

It implies that the martingale $(W_f^N(t)/N)$ converges in distribution to 0 for the uniform norm on $[0,T]$. □

By dividing Equation (2.11) by N, one gets that, for the convergence in distribution, the relation

$$\begin{aligned}\lim_{N \to +\infty} \Bigg(& \int_0^t \lambda_1^+ \left(1 - I_F^N(u)\right) \Delta_1(f)(X_F^N(u))\, du \\ & + \int_0^t \lambda_1 P_F^N(u) I_F^N(u) \Delta_1(f)(X_F^N(u))\, du \\ & + \int_0^t \lambda_2 I_F^N(u) \Delta_2^+(f)(X_F^N(u))\, du + \int_0^t \mu_2 M_F^N(u) \Delta_2^-(f)(X_F^N(u))\, du \Bigg) = 0.\end{aligned}$$

holds almost surely for all $0 \leq t \leq T$ and for all indicator functions of elements $\mathcal{S}$. Now, for $p \in \mathbb{N}$ and g a function with finite support on $\{0,1\} \times \mathbb{N}$, define $f(i,m,p) = g(i,m)$, the above relation gives

$$\begin{aligned}\sum_{x=(i,m,p) \in \mathcal{S}} & \ell_u(i,m,p)\lambda_1^+(1-i)\Delta_1(g)(i,m) + \ell_u(i,m,p)\lambda_1^- ip \Delta_1(g)(i,m) \\ & + \ell_u(i,m,p)\lambda_2 i \Delta_2^+(g)(i,m) + \ell_u(i,m,p)\mu_2 m \Delta_2^-(g)(i,m) = 0\end{aligned} \tag{2.19}$$

holds almost surely for all $u \in \mathcal{A} \subset [0,T]$ and $[0,T] - \mathcal{A}$ is negligible for Lebesgue measure. Equation (2.19) shows that for $u \in \mathcal{A}$, the vector $(\ell_u(i,m,p))$ is proportional to the invariant distribution π_p of the Markov process on $\{0,1\} \times \mathbb{N}$ whose transition rates are given by Equation (2.16).

One gets therefore the existence of a constant $\nu_u(p)$ such that $\ell_u(i,m,p) = \nu_u(p)\pi_p(i,m)$ for all $(i,m,p) \in \mathcal{S}$. Equation (2.18) gives the relation

$$\int_0^t \nu_u(\mathbb{N})\, du = t, \quad \forall t \leq T.$$

Hence one has $\nu_u(\mathbb{N}) = 1$ almost surely for all $u \in \mathcal{A}_1 \subset [0,T]$ and $[0,T] - \mathcal{A}_1$ is negligible for Lebesgue measure.

Straightforward calculations as in the proof of Proposition 2.2 complete the proof of the proposition to give Equation (2.17). □

Chapter 3

Models with Cell Cycle

Since the beginning of the 2000s, fluorescent microscopy experiments permit to quantitatively measure cell by cell gene expression (see for instance Elowitz et al. (2002), Taniguchi et al. (2010), Valgepea et al. (2013)). In particular, the article Taniguchi et al. (2010) presents a comprehensive study of messengers and proteins production in *E. coli*. It describes the behaviour of a large number of proteins, not only in terms of average expression but also in terms of variability: in populations of cells, the means and the variances of many types of mRNAs and proteins are measured. In total, about 1000 gene are considered.

These data can be confronted with stochastic models of production of proteins which exist since the 1970s: Berg (1978), Rigney and Schieve (1977) (see a review of Paulsson (2005)). The usual model presented in Paulsson (2005) is a three-stage model where gene regulation, transcription and translation are represented. All events occur at exponentially distributed times: the activation and deactivation of the gene, mRNA transcription and degradation, protein production and degradation. The rates of these events depend on the current state of the model.

Yet, these classical models are not considering many aspects that may yet have an impact on the protein variability. For instance, they do not integrate events of the cell cycle such as gene replication and division. Moreover each of them is based on the production of one particular type of protein: only one gene is considered and it produces only one type of protein. No interaction between the different protein production processes are considered (like the common sharing of common resources like RNA-polymerases and ribosomes to produce mRNA and proteins). The aim of the two following chapters is to determine the impact on the protein variability of these different aspects that are not considered by the classical models.

In this Chapter, we begin by considering successively three different origins of protein variability: the noise that directly originated from the transcription and translation mechanisms, then the effect of division and finally the impact that has gene replication. The next chapter will provide a new step toward the global understanding of the whole protein production as we will be interested in interactions between production of different classes of proteins.

Plan of the Chapter Section 3.1 will present in detail the experimental study of Taniguchi et al. (2010) on which will be based all the comparisons with our models. The techniques, the results and the interpretations of the article will be displayed. In Section 3.2, we will discuss the pertinence to use classical models to reproduce

the experimental measures. Because they do not represent explicitly the growth, we show that they are unfit for quantitative comparisons with Taniguchi et al. (2010) measures of protein variability. To address this problem, one will need new kinds of models that take into account this aspect. The Section 3.3 will present our first two models that integrate volume growth. The first model aims to represent aspects only due to the production of proteins, so that the noise predicted is only due the gene expression process. The second model includes the random segregation of mRNAs and proteins at division. We will show that even if both these aspects are important sources of variability, they are not sufficient to reproduce the results observed in Taniguchi et al. (2010). In Section 3.4 we continue our study by providing a model that also introduces gene replication at some point in the cell cycle. In this section, we propose the main theoretical results as will be able to give explicit solutions for the mean and the variance of every proteins. We will show that the noise induced by the replication of the gene is negligible compared the two previous sources of variability.

3.1 Taniguchi et al. Measures

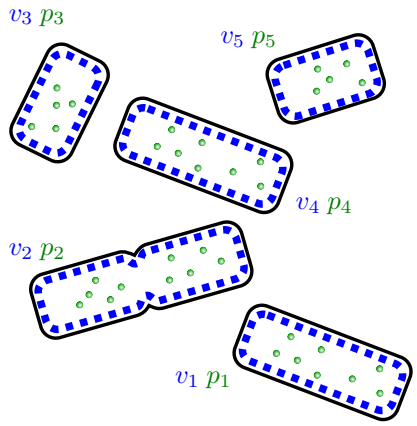
In the article Taniguchi et al. (2010), an overall experimental study of mRNA and protein expression of 1018 genes was performed using single-molecule fluorescence microscopy in bacteria (*Eschierichia coli*). A series of experiments was conducted; and each of them considers a strain of cells where a particular gene was fused with the sequence coding for the fluorescent YFP molecule. It gives a population of cells, for which the fluorescence abundance denotes the specific protein quantity. For each strain, the emitted fluorescence was hence measured in each cells. The obtained global fluorescence was normalised by the fluorescence emitted and by one single protein and divided by the cell size.

In each experiment, what result is the concentration of the considered protein in each cells. So its distribution among the population could then be deduced. In particular, for each type of proteins, the value of the mean μ_p and the standard-deviation σ_p of the concentration of proteins was deduced from these distributions (see Figure 3.1a). For each type of protein, the obtained concentration of proteins range from 10^{-1} and 10^4 copies per μm^3 ($1\,\mu m^3$ is in the order of magnitude of the volume of a cell).

On top of that, the article shows measures in the mRNA abundance performed by two techniques. First 137 mRNA types (the most expressed mRNAs) were detected, using *Fluorescence in situ hybridisation* (FISH) technique: in that case, mRNA expression is measured at the same time as the corresponding proteins and provides the mean μ_m and the standard-deviation σ_m of the concentration of each mRNA type. The other method is the mRNA-sequencing technique that allows the measure of 841 average mRNA concentration (but this method does not determine the cell to cell variability in the population). This analysis was completed with the measure of average mRNA lifetime τ_m.

The analysis of the article considers the coefficient of variation (sometimes called "noise") of each mRNA and protein concentration. The coefficient of variation (CV) is used in biology literature as a way to normalise the variance: it is defined as the the variance divided by the mean squared. For instance, Figure 3.1b and Figure 3.1c depict, for every gene, the CV of mRNA concentration (defined as σ_m^2/μ_m^2) and protein concentration (defined σ_p^2/μ_p^2) as a function of the average of mRNA and protein concentration. Among other things, these graphs allow, for every gene, to compare the distribution of mRNA and protein with a Poisson distribution: for a Poisson distribution, for an average expression μ, the noise would be $1/\mu$; the noise would scale inversely with the mean.

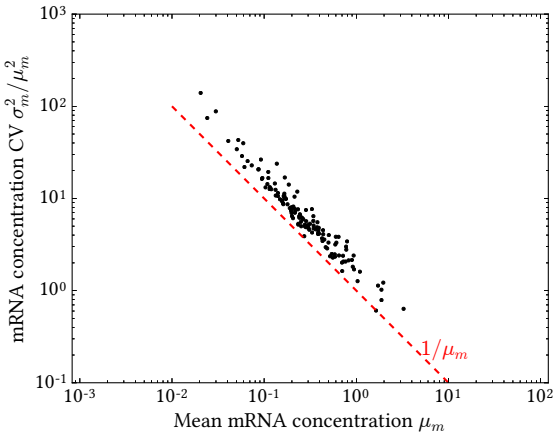
For the mRNAs, the noise in Figure 3.1b appears indeed to scale inversely with mRNA mean concentration. But it is higher than expected for Poisson distributions: for a mean μ_m the noise here appears to be around $1.7/\mu_m$ instead of $1/\mu_m$. The Figure 3.1c also shows the noise of proteins as a function of the average concentration of proteins. It clearly appears that there are two regimes for the protein CV, regimes that the article Taniguchi et al. (2010) denotes as "intrinsic noise" and "extrinsic noise" regimes. For low expressed



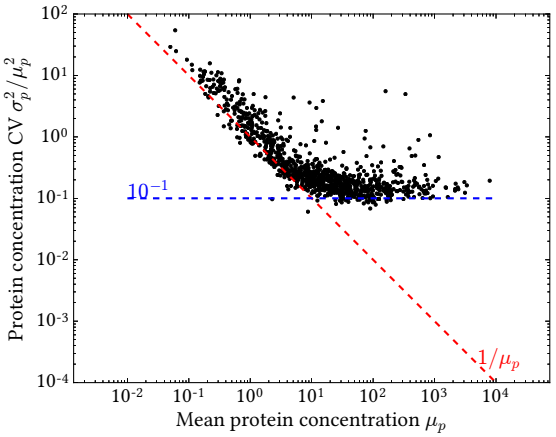
Empirical mean $\mu_p = \frac{1}{N}\sum_{i=1}^{N}\frac{p_i}{v_i}$

Empirical variance $\sigma_p^2 = \frac{1}{N}\sum_{i=0}^{N}\left(\frac{p_i}{v_i}\right)^2 - \mu_p^2$

(a) Experiment principle for the proteins



(b) mRNA coefficient of variation.



(c) Protein production coefficient of variation.

Figure 3.1: Results on mRNA and protein production in the article Taniguchi et al. (2010). **(a)**: In the experimental dataset, a large population of cells are considered; each cell i of the population has a specific number of proteins p_i and a specific volume v_i; these values are used to compute empirical mean μ_m and variance σ_m^2. **(b)**: mRNA production coefficient of variation (σ_m^2/μ_m^2) as a function of the average mRNA expression μ_m for every gene. The CV is inversely proportional to mRNA mean concentration, but it is higher than expected for Poisson distributions (red dashed lines). (Measures have been made using FISH technique on 137 mRNA types.) **(c)**: Protein production CV function of the average protein copy number for every gene. For low expressed proteins (mean protein number < 10), the CV is inversely proportional to the average protein production, this part is considered lowered by an "intrinsic noise limit" (red dashed line). For genes with higher protein expression, the CV becomes in independent of the protein expression level, protein expression is here denoted as dominated by the "extrinsic noise" (blue dashed line).

proteins (mean protein concentration < 10) the CV roughly scales inversely with the average concentration, the protein variability is dominated by the "intrinsic noise". For genes with higher protein production (mean protein concentration > 10), the CV becomes independent of the average protein production level, the plateau is around 10^{-1}; this regime of gene expression is denoted as dominated by the "extrinsic noise".

These terms of "intrinsic and extrinsic noise" were firstly introduced by Elowitz et al. (2002) and Swain et al. (2002) (see Section 1.1) to differentiate the noise coming from the protein production mechanism itself through translation and transcription (intrinsic noise) and the impact of global fluctuations of the cells on the whole gene expression efficiency (extrinsic noise). The authors of Taniguchi et al. (2010) use these terms because in the "intrinsic noise" regime, the fluctuations seem gene-specific, as the CV depends on the average protein production; as a consequence, the noise in this area seems to only depend on variables intrinsically specific to the considered gene, and not to any external other factors. In the second "extrinsic noise" regime on the contrary, fluctuations in the protein concentration are gene-independent and are therefore supposed to have an origin not directly linked to the protein production mechanism itself. They presumed that this external heterogeneity is the result of low fluctuations of "global" cellular components such as such as metabolites, ribosomes, and RNA-polymerases. The stochastic behaviour of these compounds is said to have a similar global impact on all protein productions; in particular, it is said to dominate the noise for highly expressed proteins.

Several arguments in the article are brought to justify the extrinsic nature of the noise observed in the lower plateau of protein noise of Figure 3.1c. In particular they show large heterogeneity in the protein production between cells of the same population, while dynamic fluctuations inside a cell are very low (the timescale is in the order or several cell cycles). Nevertheless, it is not fully dismissed that the origin of this noise can lie in the mechanism of protein production itself or some global cell events like division and gene replication. [1]

In this Chapter we want to estimate the relative contributions of the mechanism of protein production, the division and the gene replication to the protein variance, and check that they cannot take into account the observed two regimes of noise observed in Taniguchi et al. (2010). To do so we first show in the next section that the classical models are not fit to be quantitatively compared with the measures; we will show the need for new kinds of models that take into account the cell cycle.

3.2 Inadequacies of Classical Models

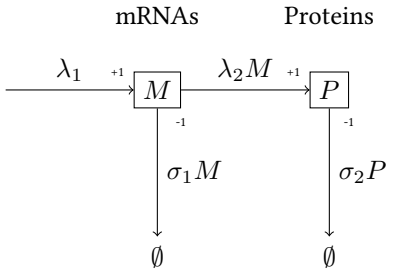


Figure 3.1: Classical model: Gene constitutive model

[1] Comparisons with standard models are indeed made in Taniguchi et al. (2010) , but to explain the poor correlation between mRNA and their protein number inside a cell (section 17 of S.I.), not to explain the lower plateau of the noise for highly produced proteins of figure Figure 3.1c.

In this section we propose to investigate the interpretation of experiment measures by classical models, by taking the example of gene constitutive model. After presenting the gene constitutive model (the other models of the chapter will be based on it), we show why it is not adequate to represent real experiments as they lack the notion of cell growth and division.

3.2.1 Constitutive Gene Model

Let's consider one of the simplest models that describes the production of one type of protein: the gene constitutive model, also referred as the two-stage model. It is a particular case of the three-stage model described in Subsection 1.3.1 (without the gene regulation part), but as it serves as a base for all models of this chapter, it is useful to recall its main mechanisms.

We consider the productions of each protein as being independent from each other: for instance, the production of the protein Adk has no influence on the production of YjiE. It is a "gene-centred" model. Every event (transcription, degradation, etc.) is supposed to happen at times that follow exponential distribution. Moreover it represents the number of mRNAs and proteins in an arbitrary fixed volume V around the considered gene. In literature, the value of V is often not explicitly given. But a reasonable value for it would be $1\,\mu m^3$ in order to directly interpret the quantity of compounds described by the model as the actual concentration expressed in copies per μm^3 (we will discuss the consequences of this choice in Subsection 3.2.2).

The gene constitutive model considers that each gene is continually available for transcription. Hence, for a given gene, two types of entities are considered (see Figure 3.1):

mRNAs mRNAs are transcribed at rate λ_1; the number M of mRNAs is then increased by 1. Each mRNA exists for a certain time determined by the rate σ_1 until it disappears; as there is M mRNAs, the total rate of mRNA disappearance is $\sigma_1 M$. When a disappearance occurs, M is decreased by 1.

Proteins Each mRNA can be translated into a protein at rate λ_2; since the number of mRNA is given by M, the total rate of protein production is $\lambda_2 M$. When a translation occurs, the number of proteins P is increased by 1. Analogously to messengers, each protein exists during a certain time until its decay which occurs at an exponential time with rate σ_2; the total rate of protein decay is hence $\sigma_2 P$.

The "decay" rates σ_1 for mRNAs and σ_2 for proteins are not representing the same effect:

- The mRNA decay is mainly due to the rapid degradation through an active catalysed reaction with enzymes. Through this mechanism, the cell ensures the quality control of the molecules which can be denatured through time. For the mRNAs, this degradation is of the order of few minutes; it is much quicker than the cell cycle (the median lifetime of mRNAs is about 5 min in Taniguchi et al. (2010), while the cell cycle is around 150 min). Therefore, in that case the mRNA decay rate σ_1 represents a degradation rate; it is specific to the type mRNA since each of them has different affinity with the degradation enzymes.

- For the proteins, the active degradation also exists (in this case, it is called *proteolysis*), but it usually occurs only in very long period of times, much higher than the duration of the cell cycle (see Koch and Levy (1955)). As the model takes place in a fixed volume V, the protein will certainly leave it before being degraded and is considered to never return inside: it is the decay by dilution. In the case of proteins, the decay rate σ_2 hence represents the dilution effect; this rate is similar for all proteins and represents the time for a compound to leave the considered volume.

Different properties of this model are known (it is a particular case of the three-stage model presented in Subsection 1.3.1). In particular, the mRNA copy number at equilibrium is known to follow a Poisson distribution and explicit expressions for the mean and the variance of the number of proteins at equilibrium are

known and can possibly be used to fit the different measurements of Taniguchi et al. (2010) experiments: it will make predictions for the variances of mRNAs and proteins that can be compared to experimental measures. But as we show in the next subsection, this quantitative comparison may not be relevant as this model (and by extension all classical models) does not take into account the real volume of the cell.

3.2.2 Impact of the Considered Volume in Classical Models

As previously said, in classical models, a fixed volume V is considered surrounding the gene of interest. In this volume, only one copy of the gene is considered, and when one compound (either a protein or a mRNA) leaves the volume, it is assumed that it never returns inside (see Figure 3.2).

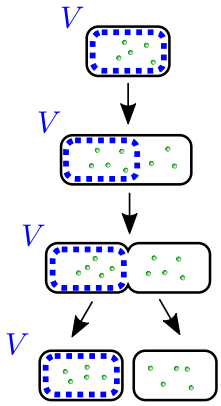


Figure 3.2: Volume in classical models are based on dilution. The volume V is fixed, once a compound (mRNA or proteins in green) leaves the volume, it does not return.

At a time t, from the number M_t of mRNAs and P_t of proteins inside the volume V, it is possible to consider their concentration as their number per unit of volume V: at time t, the concentration of mRNAs and proteins are respectively

$$\frac{M_t}{V} \text{ and } \frac{P_t}{V}.$$

With this definition of the concentration, the mean and the variance of the concentration of proteins is interpreted as:

$$\begin{aligned} \mathbb{E}\left[P_t/V\right] &= \mathbb{E}\left[P_t\right]/V, \\ \mathbb{V}\mathrm{ar}\left[P_t/V\right] &= \mathbb{V}\mathrm{ar}\left[P_t\right]/V^2, \end{aligned}$$

and similarly for the mean and the variance of mRNA concentration. The choice for V of $1\,\mu m^3$ permits directly to interpret the mean $\mathbb{E}\left[P_t\right]$ and the variance $\mathbb{V}\mathrm{ar}\left[P_t\right]$ of the number of proteins directly as the mean $\mathbb{E}\left[P_t/V\right]$ and the variance $\mathbb{V}\mathrm{ar}\left[P_t/V\right]$ of their concentration per μm^3.

Nonetheless, one can wonder if this particular choice of V has an impact on the obtained values of the mean and the variance of the concentration. The following example shows that it is indeed the case.

Example 3.1. Let's consider two volumes of size V. In each of these volumes two independent but with identical dynamics processes occur: $(P_{1,t})$ and $(P_{2,t})$ would be the processes that represent the number of proteins in respectively the first and in the second volume. Concentration being an extensive quantity, one expects that the behaviour of the concentration in one volume would be similar as in both volumes taken altogether. But the mean and the variance of the concentration in the large volume are:

$$\begin{aligned} \mathbb{E}\left[(P_{1,t}+P_{2,t})/(2V)\right] &= \mathbb{E}\left[P_{1,t}/V\right], \\ \mathbb{V}\mathrm{ar}\left[(P_{1,t}+P_{2,t})/(2V)\right] &= \mathbb{V}\mathrm{ar}\left[P_{1,t}/V\right]/2. \end{aligned}$$

The mean is indeed identical if the volume $2V$ is considered instead of V, but not the variance.

The previous example shows that, with a classical models, the concentration is not an extensive quantity in terms of variance, and the distribution of the concentration of each compounds depends on the considered volume V. This has important consequences and it raises difficulties when it comes to interpreting experimental results with those predicted with classical models.

The usual comparison between a classical model and the measures are done as following: in order to deduce the parameters of the model, the equilibrium mean $\mathbb{E}\left[P/V\right]$ of the model is interpreted with the

empirical mean μ_p of the measures; once all the parameters known, one compares the obtained variance $\mathbb{V}\mathrm{ar}\left[P/V\right]$ with the empirical variance σ_p^2 of the measures. Figure 3.1a explains schematically how empirical mean and variance are computed: in each cell of the population, concentrations were computed using the specific volume of the cell and not an abstract volume V. In the model, on the contrary, one can fix $\mathbb{E}\left[P/V\right]$ and still have different $\mathbb{V}\mathrm{ar}\left[P/V\right]$ depending on the volume V chosen. To sum up, the predicted variance $\mathbb{V}\mathrm{ar}\left[P/V\right]$ is volume V dependent, therefore it cannot be directly be interpreted as the empirical variance σ_p^2.

Comparisons between theoretical variances obtained with classical models and the empirical variances measured in real experiments are hence problematic. They do not represent the same thing: $\mathbb{V}\mathrm{ar}\left[P/V\right]$ denotes the variance of proteins inside an abstract volume V, as the empirical variance is computed using *real* volume of cells. In order to represent what are exactly the empirical mean and the empirical variance described in Figure 3.1a, one needs to propose a model with a quantity that depicts the actual volume of the cell, and this volume changes across time as the cell grows and divides.

3.3 Model with Cell Cycle

The model of this section is close to the gene constitutive classical model, but we introduce the notion of volume of the cell that changes across the cell cycle. This aspect will enable quantitative comparisons with experimental dataset of Taniguchi et al. (2010). This model only considers aspects due to the protein production; as a consequence, the variability predicted here arises from protein synthesis mechanism itself and not from external factors.

First, in Subsection 3.3.1, we present this new feature, as it will be common in all the remaining models of this chapter and also the model of Chapter 4. In Subsection 3.3.2 is the presented the model in detail; the Subsection 3.3.3 is dedicated to its theoretical analysis in order to fit parameters to the measures. We will analyse the simulations of the model and show that the protein coefficient of variation globally follows what the authors Taniguchi et al. (2010) denotes as the "intrinsic noise limit". In Subsection 3.3.6 we will interest in the effect of the random segregation of each cell compounds at division and show that it has a significant impact on some protein variability.

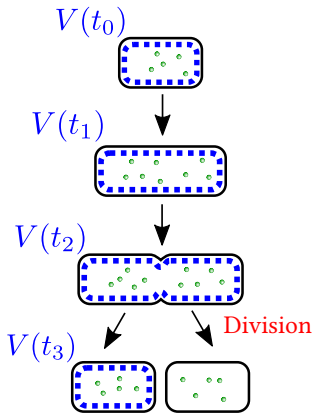


Figure 3.1: Volume in models with cell division. The volume V increase as the cell grows. At division, the model follows only one cell, segregation occurs on the compounds (mRNA or proteins in green) to know if there are in the considered daughter cell.

3.3.1 New Feature: A Time Dependent Volume

The baseline of all the models of this chapter and the next is the following two key features: the cell growth and the division. They are the minimum notions to introduce in order to somewhat take into account the cell cycle. So we describe in this subsection aspects that will remain true for all the models to come.

At any time t, the volume $V(t)$ considered is the entire volume of the cell which is increasing as the cell grows (see Figure 3.1). From this volume, concentrations can be considered: if M_t and P_t respectively denote the number of mRNAs and proteins at time t, it comes the concentrations

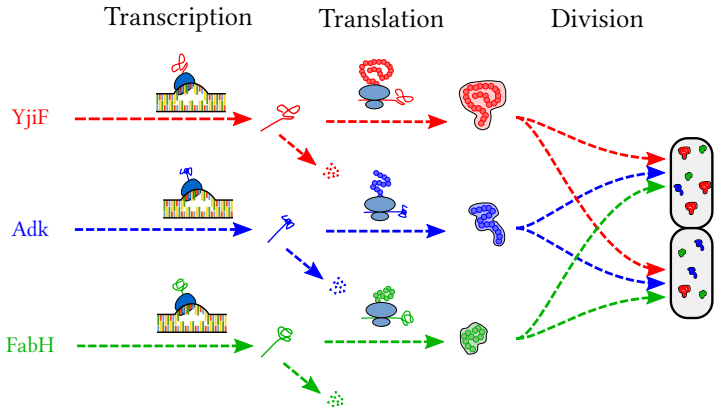
$$\frac{M_t}{V(t)} \text{ and } \frac{P_t}{V(t)}.$$

At periodic times, divisions occur; it corresponds to the step between t_2 and t_3 in Figure 3.1. Two daughter cells are created and the model only focuses on one of them; the volume considered then is the one of this

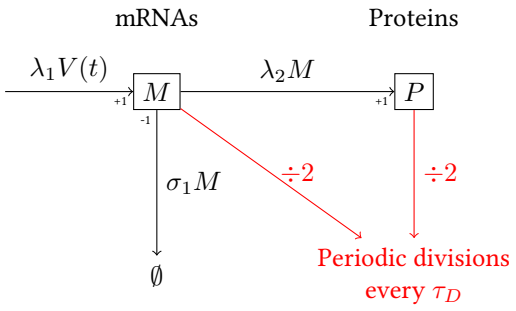
newborn cell. During the event of division, each mRNA and protein either goes in the next considered cell or not. In a first step, we consider that this segregation is exact, that is to say that exactly one half of mRNAs and proteins goes to the considered cell (in Subsection 3.3.6 we will consider the case where this segregation is random: each mRNA and protein has an equal chance to be in the considered daughter cell or not).

From this perspective, the notion of dilution of compounds introduced for classical models (the phenomenon by which a compound can spontaneously leave the volume of the model) is no longer used. It is replaced by the molecule segregation at division where the number of compounds is halved. The spontaneous "decay" of compounds will only be due to their hydrolysis, the active catalysed degradation of mRNAs or proteins.

3.3.2 Presentation of the Model



(a) Biological mechanisms for three typical genes.



(b) Production of one type of protein.

Figure 3.2: Model with cell cycle. **(a)** The model of this section considers genes independently from each other; the three mechanism of transcription, translation and division are represented. **(b)** For one particular type of protein, the number of mRNAs and proteins are respectively M and P (see main text for more details).

To take into account the growth and division of the cell, we propose the model shown in Figure 3.2. It is based on the constitutive gene model of Subsection 3.2.1, in particular the production and degradation of

mRNA are similar:

mRNAs In classical model Subsection 3.2.1, in a fixed volume of $1\,\mu m^3$, mRNAs are spontaneously created at a constant rate. In this model, we keep this concept of spontaneous creation per volume unit as the rate of mRNA creation is $\lambda_1 V(t)$, with $V(t)$ the volume of the cell at time t; hence per volume unit, the rate of production will remain constant. When an mRNA creation occurs, the number of mRNAs M is increased by one. As for the gene-constitutive model, each mRNA degrades at an exponential time given by the rate σ_1.

Proteins As in the model with constitutive gene, each mRNA can be translated into a protein at rate λ_2; the number of proteins P is then increased by 1. But here there is no disappearance rate: since the proteolysis occurs in a timescale much longer than the cell cycle, its decay is dominated by protein segregation that occurs at division.

Division Periodically, every time τ_D, a division occurs. At this instant, the number of mRNAs and the number of proteins undergo an exact division to keep only half of the molecules that are in the considered daughter cell.

On top of that, one also considers the volume growth of the cell independently. In real life experiments, bacteria volume grows exponentially (see Wang et al. (2010)) and approximately doubles its volume at the time of division τ_D. As a consequence, the model considers that, if s is the time spent since the last division, then the volume grows as

$$V(s) = V_0 2^{s/\tau_D}$$

with V_0 being the typical size of a cell at birth.

Remark 3.1. *This model only considers aspects that are specific to protein production, so that randomness induced by the model is only due to the transcription and translation mechanisms:*

- *mRNA rate of production is proportional to the volume so that there is no effect of gene replication. As a result, the average concentrations of mRNAs and proteins remain constant across the whole cell cycle (it will be proven in the next subsection).*
- *exact segregation at division minimises the effect of division: as the volume is strictly halved as well as the number of mRNAs and proteins, their concentrations remain unchanged during the division process.*

The goal is to have a basic model with the notion of the cell cycle that only consider sources of noise that are specific to the mechanism of protein production: this model is a way to estimate the "intrinsic noise" of the gene expression variability. It is our starting point for the analysis of the protein variability as, later in the chapter, external features will be added one by one to the model.

Now, our aim is to analyse the variability predict by this model. To do so, in the next subsection, we conduct at first a theoretical analysis of the average mRNA and protein production by the model. This analysis will permit to fit parameters to the experimental measures of Taniguchi et al. (2010).

3.3.3 Dynamic of the Number of mRNAs and Proteins

The content of this section is technical: the goal is to justify the Proposition 3.3 of the next subsection. The reader more interested in biological aspects can skip this and go directly to Subsection 3.3.5, while admitting its first proposition.

Messenger-RNA Dynamic

For any time $t \in \mathbb{R}_+$, M_t denotes the number of mRNAs at this instant. Let's depict the distribution of M_t for any $t \in \mathbb{R}_+$. We suppose that the initial time $t = 0$ is a time of division; in this case, at each time $i \cdot \tau_D$ with $i \in \mathbb{N}$ are moments of division. For any $i \in \mathbb{N}$, $M_{i\tau_D}$ denotes the number of mRNAs at the beginning of i-th cell cycle and $M_{i\tau_D-}$ the number of mRNAs in the $(i-1)$-th cell cycle just *before* division.

We suppose that a lot of cell divisions already occurred even before time $t = 0$, and hence the considered cell cycle occurs when the embedded Markov chain $(M_{i\tau_D})_i$ has already reached its equilibrium: it means that the distribution of $M_{i\tau_D}$ is the same as the distribution of $M_{(i+1)\tau_D}$. If the equilibrium is already reached at time 0, it implies that the distribution of any $M_{i\tau_D+s}$ for any $i \in \mathbb{N}$ and $s \in [0, \tau_D[$ is equal to the distribution of M_s. As a consequence, we can only consider the first cell cycle $s \in [0, \tau_D[$ to fully characterise the behaviour of M_s at any time $s \in \mathbb{R}_+$.

The dynamic of M_s for $s \in [0, \tau_D[$ resembles the classical constitutive gene model (Subsection 3.2.1); but in the classical model, equilibrium properties were used to describe its behaviour. Here we need to describe the evolution of (M_s) between times 0 and τ_D (during this period of time, the number of mRNA approximately doubles).

As previously said, in our case, the number of messengers M_s after a time s does not reach its equilibrium; we therefore need, not to describe the equilibrium, but the *dynamics* of M_s.

To do so, for any time s of the cell cycle, let's group mRNAs two categories:

- First group: mRNAs which were present at the birth of the cell. Each mRNA i of the first group is characterised by $E^i_{\sigma_1}$, its lifetime given by an exponential random variable of rate σ_1. The i-th mRNA still exists at time s if and only if $E^i_{\sigma_1} > s$. As a consequence, the number of mRNAs of this group still exists at time s is given by
$$\sum_{i=1}^{M_0} \mathbb{1}_{\{E^i_{\sigma_1} > s\}}. \tag{3.1}$$

- Second group: mRNAs which have been created since the birth of the cell. The description of the number of mRNA of this group is more complicated. It is necessary to resort to the framework of Marked Poisson Point Processes (MPPP). A MPPP is a two-dimensional process. It is based on a Poisson process where each of its random point is "marked" with another random variable; each point of a MPPP is a couple (x, y) where x is part of a Poisson point process and y is the mark distributed according to a certain distribution. One can refer to the first Chapter of Robert (2010) or Kingman (1993) for the main results concerning MPPP.

  We use this tool to characterise the number of mRNAs of the second group. In our case, the first variable x represents the time at which the mRNA is created and the second variable y represents the mRNA lifetime. Let's define $\mathcal{N}$ an MPPP of intensity
$$\nu(dx, dy) = \lambda_1 V(x)dx \otimes \sigma_1 e^{-\sigma_1 y} dy.$$

  It is noticeable that the underlying Poisson Process of this MPPP is not homogeneous. If the i-th mRNA of this group is born at time x_i and its lifetime is y_i, then it exists at time s if and only if $(x_i, y_i) \in \Delta_s$ with
$$\Delta_s = \left\{(x, y) \in \mathbb{R}^2_+,\ 0 < x < s,\ y > s - x\right\}.$$
  One can refer to Figure 3.3. Therefore the number of mRNA of this group still present at time s is given by
$$\mathcal{N}(\Delta_s) = \iint_{\mathbb{R}^2_+} \mathbb{1}_{(x,y)\in\Delta_s} \mathcal{N}(dx, dy). \tag{3.2}$$

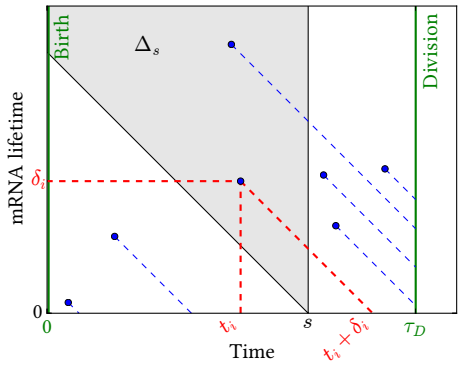


Figure 3.3: Illustration of the Marked Point Poisson Process describing the dynamic of mRNAs: each mRNA is characterised by the point (x_i, y_i) (with (x_i, y_i) following the MPPP $\mathcal{N}$, whose distribution is of intensity ν). The random variable x_i represents the time at mRNA creation and y_i its lifetime, hence this mRNA exists from volume x_i up to volume $x_i + y_i$; that is to say that mRNA is still present at time s, if and only if the point (x_i, y_i) is in the set with $\Delta_s = \left\{(x, y) \in \mathbb{R}_+^2,\, 0 < x < s,\, y > s - x\right\}$.

By summing the number of mRNAs for each group (Equation (3.1) and Equation (3.2)), it comes the total number of mRNAs present at time $s \in [0, \tau_D[$:

$$M_s = \sum_{i=1}^{M_0} \mathbb{1}_{\{E_{\sigma_1}^i > s\}} + \mathcal{N}(\Delta_s)\,. \tag{3.3}$$

This description of the dynamic of M_s, together with the equilibrium hypothesis which implies that $M_0 \overset{\mathcal{D}}{=} M_{\tau_D}$, allows to prove the next proposition.

Proposition 3.1. *At equilibrium, the mean number of mRNAs at time $s \in [0, \tau_D[$ of the cell cycle is*

$$\mathbb{E}[M_s] = V(s)\frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2}.$$

Proof. By taking the mean of Equation (3.3), it follows for any time s of the cell cycle:

$$\mathbb{E}[M_s] = \mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{1}_{\{E_{\sigma_1}^i > s\}}\right] + \mathbb{E}[\mathcal{N}(\Delta_s)]\,.$$

Since all $\left(E_{\sigma_1}^i\right)_i$ are i.i.d. and independent of M_0, the first term is given by

$$\mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{1}_{\{E_{\sigma_1}^i > s\}}\right] = \mathbb{E}[M_0]\, e^{-s\sigma_1}.$$

For the second term, one has to remark that as $\mathcal{N}$ is a MPPP, $\mathcal{N}(\Delta_{\tau_D -})$ is a Poisson random variable (Proposition 1.13.a of Robert (2010)). The parameter of this Poisson random variable is given by

$$\nu(\Delta_s) \quad = \quad \iint_{\Delta_s} \nu(dx, dy)\,.$$

Since, in the definition of Δ_s

$$\begin{aligned}\nu(\Delta_s) &= \int_0^s \int_{s-x}^{\infty} \lambda_1 V(x)\sigma_1 e^{-\sigma_1 y}\, dy\, dx \\ &= V_0\lambda_1\sigma_1 \int_0^s 2^{x/\tau_D} \int_{s-x}^{\infty} e^{-\sigma_1 y}\, dy\, dx \\ &= V_0\lambda_1 e^{-\sigma_1 s} \int_0^s \exp\left(x\left(\frac{\log 2}{\tau_D}+\sigma_1\right)\right) dx \\ &= V_0 \frac{\lambda_1\sigma_1}{\log 2 + \sigma_1\tau_D}\left(2^{s/\tau_D} - e^{-\sigma_1 s}\right).\end{aligned}$$

As a consequence, it comes that for any time s in the cell cycle:

$$\mathbb{E}[M_s] = \mathbb{E}[M_0]\, e^{-s\sigma_1} + V_0 \frac{\lambda_1\tau_D}{\tau_D\sigma_1 + \log 2}\cdot\left(2^{s/\tau_D} - e^{-s\sigma_1}\right).$$

We still have to specify the mean number of mRNAs at birth $\mathbb{E}[M_0]$. At the end of the cell cycle, for $s = \tau_D-$, the average number of mRNAs is given by

$$\mathbb{E}[M_{\tau_D-}] = \mathbb{E}[M_0]\, e^{-\tau_D\sigma_1} + V_0 \frac{\lambda_1\tau_D}{\tau_D\sigma_1 + \log 2}\cdot\left(2 - e^{-\tau_D\sigma_1}\right),$$

and since at equilibrium,

$$\mathbb{E}[M_{\tau_D}] = \mathbb{E}[M_0] = \frac{\mathbb{E}[M_{\tau_D-}]}{2}.$$

Hence

$$\mathbb{E}[M_0]\left(2 - e^{-\tau_D\sigma_1}\right) = V_0 \frac{\lambda_1\tau_D}{\tau_D\sigma_1 + \log 2}\cdot\left(2 - e^{-\tau_D\sigma_1}\right),$$

which gives the result. □

Protein Dynamic

The mean number of mRNAs is now determined for any moment of the cell cycle. Each of the mRNAs potentially produces proteins at rate λ_2. As for the mRNAs, we describe the number of proteins at time s by grouping them into two categories:

- The P_0 proteins that were present at birth and which remain in the bacteria during all the cell cycle (as said in Subsection 3.3.2 the proteolysis is not considered in this model).
- The proteins that have been created during the current cell cycle. The rate of production is depending on the current number of mRNAs. We consider $\mathcal{N}^i_{\lambda_2}$ (for $i \in \mathbb{N}$ and $i \geq 1$) independent Poisson Point Processes of intensity λ_2. If the i-th mRNA exists at time s (that is to say if $i \leq M_s$), then the number of proteins produced by this mRNA between s and $s + ds$ is $\mathcal{N}^i_{\lambda_2}(ds)$.

To sum up, the number of proteins at a time s of the cell cycle is:

$$P_s = P_0 + \sum_{i=1}^{\infty} \int_0^s \mathbb{1}_{i\leq M_u} \mathcal{N}^i_{\lambda_2}(du). \tag{3.4}$$

The first term is the number of proteins at birth, and the second take into account all the proteins created between times 0 and s. Based on that, one can determine the mean number of proteins at any time s of the cell cycle:

Proposition 3.2. *At equilibrium, the mean number of proteins at any time $s \in [0, \tau_D[$ of the cell cycle is*

$$\mathbb{E}[P_s] = V(s) \cdot \frac{\lambda_2 \tau_D}{\log 2} \cdot \frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2}.$$

Proof. Taking the average of Equation (3.4) gives

$$\begin{aligned}\mathbb{E}[P_s] &= \mathbb{E}[P_0] + \sum_{i=1}^{\infty} \mathbb{E}\left[\int_0^s \mathbb{1}_{i \leq M_u} \mathcal{N}^i_{\lambda_2}(du)\right] = \mathbb{E}[P_0] + \sum_{i=1}^{\infty} \mathbb{E}\left[\int_0^s \mathbb{1}_{i \leq M_u} \lambda_2 du\right] \\ &= \mathbb{E}[P_0] + \lambda_2 \int_0^s \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{i \leq M_u}\right] du = \mathbb{E}[P_0] + \lambda_2 \int_0^s \mathbb{E}[M_u]\, du.\end{aligned}$$

As we know the mean number of mRNAs $\mathbb{E}[M_u]$ at any time u of the cell cycle with Proposition 3.1:

$$\mathbb{E}[P_s] = \mathbb{E}[P_0] + \frac{\lambda_2 \tau_D}{\log 2} \cdot \frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2} \cdot (V(s) - V_0).$$

Since the system is at equilibrium, we have for time τ_D-, $\mathbb{E}[P_{\tau_D-}] = 2\mathbb{E}[P_0]$; so

$$\begin{aligned}\mathbb{E}[P_0] &= \frac{\lambda_2 \tau_D}{\log 2} \cdot \frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2} \cdot (V(\tau_D-) - V_0) \\ &= \frac{\lambda_2 \tau_D}{\log 2} \cdot \frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2} \cdot V_0.\end{aligned}$$

Consequently, for any time s of the cell cycle,

$$\mathbb{E}[P_s] = \lambda_2 \frac{\tau_D}{\log 2} \cdot \frac{\lambda_1 \tau_D}{\sigma_1 \tau_D + \log 2} \cdot V_0 \left(1 + 2^{s/\tau_D} - 1\right);$$

hence the result. □

Expressions for the variance of the number of mRNAs and proteins are not easily obtained for this model, we will hence determine them with simulations. (In Section 3.4, will be presented a model, that better represent the real dynamic of the cell, and from which we have then analytical expressions of the mean and the variance of mRNAs and proteins).

3.3.4 Parameter Computations

In this subsection, we determine the parameters λ_1, $\sigma_1$and λ_2 so that the average production of mRNAs and proteins correspond to the measure of Taniguchi et al. (2010). To do so, we will use the previous analytical results on the mean number of mRNAs and proteins.

3.3.4.1 Mean and Variance of Concentration over the Cell Cycle

This model represents mRNAs and proteins in cells with a volume that changes across time: we have described mRNA and protein number at any time of the cell cycle. But experimental measures are not done at a particular time in the cell cycle. So we need to characterise the mean and the variance of compound concentration, not at a given time s in the cell cycle, but over the cell cycle.

At any moment of the cell cycle, the concentration of any compound (either mRNAs or proteins) is defined as its current number divided by the current volume : for instance, in the case of mRNAs, the concentration at time $s \in [0, \tau_D[$ is $M_s/V(s)$ (with $V(s) = V_0 2^{s/\tau_D}$). One can consider the global average concentration over the cell cycle $\widehat{\mathbb{E}}[M/V]$, as simply the mean $\mathbb{E}[M_s/V(s)]$ averaged over the cell cycle for s from 0 to τ_D. $\widehat{\mathbb{E}}[M/V]$ is then defined as:

$$\widehat{\mathbb{E}}[M/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \mathbb{E}\left[\frac{M_s}{V(s)}\right] ds. \tag{3.5}$$

Let's then define the global variance of mRNA concentration $\widehat{\mathbb{V}}ar[M/V]$. On can look at how much the concentration $M_s/V(s)$ deviates from the global average $\widehat{\mathbb{E}}[M/V]$ at any time s of the cell cycle: $\mathbb{E}\left[(M_s/V(s))^2\right] - \widehat{\mathbb{E}}[M/V]^2$. Then let's denote the global variance $\widehat{\mathbb{V}}ar[M/V]$ as the average over the cell cycle of this deviation:

$$\widehat{\mathbb{V}}ar[M/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \left[\mathbb{E}\left[\left(\frac{M_s}{V(s)}\right)^2\right] - \widehat{\mathbb{E}}[M/V]^2\right] ds. \tag{3.6}$$

Equivalently for the proteins, let's define:

$$\widehat{\mathbb{E}}[P/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \mathbb{E}\left[\frac{P_s}{V(s)}\right] ds, \tag{3.7}$$

$$\widehat{\mathbb{V}}ar[P/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \left[\mathbb{E}\left[\left(\frac{P_s}{V(s)}\right)^2\right] - \widehat{\mathbb{E}}[P/V]^2\right] ds. \tag{3.8}$$

Remark 3.2. *Let's consider a population of independent cells, where each cells have specific mRNA and protein concentrations (as it is the case in real experiments). If the ages of cells of the population are uniformly distributed in the interval $[0, \tau_D[$, the mean and the variance of mRNAs and proteins concentration would be equivalent to Equations (3.5), (3.6), (3.7) and (3.8). In reality, the experimental population considered are in exponential growth, which means that the population age distribution is not uniformly distributed, but we will see in the next Section (subsubsection 3.4.5.4), that this has little impact on the population distribution.*

In the case of this model, we have determined in the previous subsection the average mRNA and protein number for any time s of the cell cycle in Proposition 3.1 and Proposition 3.2, in particular, the concentrations at any time s remain constant and are given by

$$\mathbb{E}\left[\frac{M_s}{V(s)}\right] = \frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2} \quad \text{and} \quad \mathbb{E}\left[\frac{P_s}{V(s)}\right] = \frac{\lambda_2\tau_D}{\log 2}\cdot\frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2}.$$

With the definition of $\widehat{\mathbb{E}}[M/V]$ and $\widehat{\mathbb{E}}[P/V]$ (Equation (3.9) and Equation (3.10)), it follows the next proposition.

Proposition 3.3. *The average mRNA and protein concentrations over the cell cycle are:*

$$\widehat{\mathbb{E}}[M/V] = \frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2} \tag{3.9}$$

and

$$\widehat{\mathbb{E}}[P/V] = \frac{\lambda_2\tau_D}{\log 2}\cdot\frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2} \tag{3.10}$$

3.3.4.2 Parameters Deduced from Experimental Dataset

For each gene, we want to identify the parameters λ_1, σ_1 and λ_2. We also need to determine the "global" quantities τ_D and V_0. To do so, we use measurements of Taniguchi et al. (2010).

First, let's fix the parameters common to all genes: the division time τ_D is said to be 150 min in the article and the volume at birth V_0 is taken equal to $1.3\,\mu m^3$.[2]

Then we have to determine for each gene the three gene-specific parameters λ_1, σ_1 and λ_2. We consider the genes of the article for which was measured the empirical mean of messengers μ_m and proteins μ_p concentrations, as well as the mRNA half-life time τ_m.

First we determine the rate of mRNA degradation for each gene with the measured mRNA half-life time τ_m: a half-life τ_m indicates that a mRNA has a probability $1/2$ to disappear within a duration τ_m, hence $e^{-\sigma_1 \tau_m} = 1/2$. From that, we can compute the rate σ_2 (specific for each type of mRNA):

$$\sigma_1 = \log 2/\tau_m.$$

Then we can identify the averages of mRNA and protein concentrations of the model (respectively $\widehat{\mathbb{E}}\,[M/V]$ and $\widehat{\mathbb{E}}\,[P/V]$) with the empirical averages μ_m of mRNA concentration and μ_p of protein concentration of the article. As a consequence, with Equation (3.9) and Equation (3.10), the parameters λ_1 and λ_2 are:

$$\lambda_1 = \mu_m \cdot \frac{\sigma_1 \tau_D + \log 2}{\tau_D},$$
$$\lambda_2 = \mu_p \cdot \frac{\log 2}{\tau_D} \cdot \frac{\sigma_1 \tau_D + \log 2}{\lambda_1 \tau_D}.$$

A summary of the different parameters can be seen in Table 3.1. Having determined all the parameters allows to perform simulations of the model using stochastic algorithm in order to assess the variability of every protein and compare them with those experimentally obtained in Taniguchi et al. (2010). When performing simulations, one needs to take care of the non-homogeneity of the Poisson processes describing mRNA creation times: the rate protein production $\lambda_1 V(s)$ is not a homogeneous rate as it changes with time. That does not allow a direct application of Gillespie (1977) algorithm and a more complex algorithm need to be used. One can refer to appendix Section 3.A for more information.

Param	Median	Mean	Maximum	Minimum
λ_1^{-1}	74.8	$2.32 \cdot 10^2$	$9.97 \cdot 10^3$	0.90
λ_2^{-1}	0.62	7.88	$1.70 \cdot 10^3$	$8.96 \cdot 10^{-3*}$
σ_1^{-1}	5.05	6.68	52.1	0.91

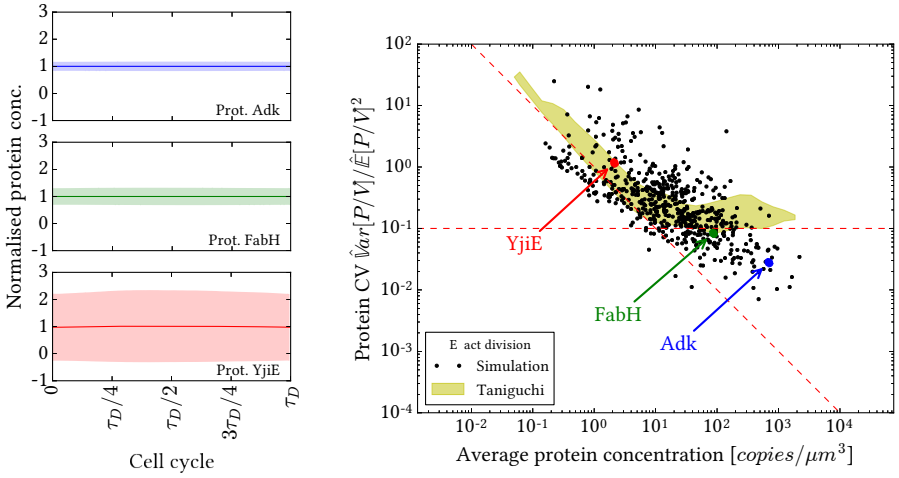
Table 3.1: Quantitative summary of the parameters in min (*: this value of the gene *yjiY* is biologically unrealistic ; maybe due to an error on the measure of its mRNA in Taniguchi et al. (2010))

3.3.5 Results of the Model with Cell Cycle

For each gene, we have performed a simulation using the parameters previously described. In each case, we regularly recorded the protein concentration at different moments of the cell cycle for thousands of generations. From these results, the behaviour of each protein concentration distribution during the whole cell cycle can be deduced.

The results of the protein variance are shown in Figure 3.4, where the variability of the protein concentration in the simulations of the model and of the experiments are shown. In the first figure is shown the profiles

[2] The value of V_0 is not explicitly given in Taniguchi et al. (2010). But it can be estimated via the area range of cells and the typical width given in the supplementary material of Taniguchi et al. (2010).



(a) Normalised protein concentration profile.

(b) Simulated CV of protein concentrations

Figure 3.4: Result of simulations of the model with cell cycle, compared with Taniguchi et al. (2010) experiments. **(a)**: Normalised protein concentration profile over the cell cycle for three representative proteins. The thick line represents the normalised mean concentration over the cell cycle $\mathbb{E}[P_s/V(s)]$, and the coloured areas represents the standard deviation. As predicted, the mean protein concentration remains constant over the cell cycle. Furthermore, we observe that the relative variance is higher for the less expressed proteins such as YjiE. **(b)**: coefficient of variation (CV) of the protein concentration (defined as $\widehat{\mathbb{V}}ar[P/V]/\widehat{\mathbb{E}}[P/V]^2$) as a function of the average protein concentration. It appears that the CV predicted by the simulation scales approximately inversely with the average protein production. Though, unlike Taniguchi et al. (2010) experiments, indicated by the yellow area (corresponding to the point cloud of Figure 3.1c), there is no lower plateau for highly expressed proteins.

of three different representative proteins (Adk, FabH and YjiE) across the cell cycle; these three proteins are respectively highly, moderately and lowly expressed. As predicted, the mean protein concentration remains constant during the cell cycle: there is no average periodic fluctuation due to cell cycle in this model. The figure also shows that the relative standard deviation decreases as the average production increases.

It is confirmed by Figure 3.4b. The figure shows the protein CV (defined as $\widehat{\mathbb{V}}ar[P/V]/\widehat{\mathbb{E}}[P/V]^2$) against the average protein concentration $\widehat{\mathbb{E}}[P/V]$. It appears that the noise approximately scales inversely the average protein concentration like in the first "intrinsic noise" regime of Taniguchi et al. (2010). But unlike in Taniguchi et al. (2010) experiment, there is no lower plateau for highly expressed proteins.

As previously said, the protein variance of this model is only due to the protein production mechanism. Therefore the results presented in Figure 3.4 confirm that the variability due to protein production itself (the intrinsic noise) cannot explain all the protein variability experimentally observed. In the next subsection we add the first contribution to the "extrinsic" noise: the effect of random distribution of compounds (either mRNA or proteins) in daughter cells during division.

3.3.6 Model with Cell Cycle and Binomial Division

In this subsection, we propose our first model extension. In the model presented in Figure 3.2b, the division performed is considered as exact: the numbers of mRNAs and proteins are halved. In reality, this division is not exact as the position of any compound in the dividing cell is random. By supposing that the size of the two daughter cells are equivalent, every compound has an equal chance to be in the next considered bacteria or not. Given the number of mRNAs M_{τ_D-} and proteins P_{τ_D-} just *before* the division, the total number of mRNAs and proteins after division is no longer deterministic, it is the result of a random sampling. We called this sample the *binomial sampling*: for instance, knowing M_{τ_D-}, the number of mRNAs M_{τ_D} after division follows a binomial distribution $\mathcal{B}(n,p)$ with parameters $n = M_{\tau_D-}$ (the number to distribute) and $p = 1/2$ (equal chance to be in the considered cell). Every other aspects remain identical to the previous model (see Figure 3.5a).

Even if the sampling at division is now random, on average we still have

$$\mathbb{E}[M_{\tau_D}] = \frac{1}{2}\mathbb{E}[M_{\tau_D-}] \qquad \text{and} \qquad \mathbb{E}[P_{\tau_D}] = \frac{1}{2}\mathbb{E}[P_{\tau_D-}]$$

and it does not change the results for the mean of mRNA and protein production: the proofs of Proposition 3.1 and Proposition 3.2 remain correct even with the binomial sampling. In particular, we still have

$$\widehat{\mathbb{E}}[M/V] = \frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2} \qquad \text{and} \qquad \widehat{\mathbb{E}}[P/V] = \frac{\lambda_2\tau_D}{\log 2}\cdot\frac{\lambda_1\tau_D}{\sigma_1\tau_D + \log 2},$$

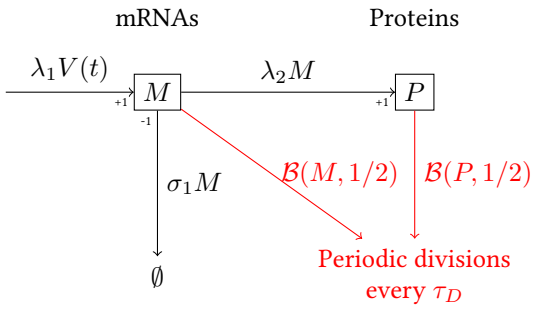
which allows to determine parameters λ_1 and λ_2 based on Taniguchi et al. (2010) dataset.

Comparative simulations of the two models have been performed: with exact division and with binomial division. On Figure 3.5b is shown the protein concentration CV in the model with exact division divided by the variance in the CV with binomial sample. This ratio allows us to know the proportion of noise that is due to the binomial sampling. As this ratio is below 1 for every gene, it shows that, as expected, the binomial sampling indeed adds variability. For most of proteins around 10% of the noise is due to binomial sampling. This proportion can increase up to 50%. It corresponds to proteins that have the lowest Fano factor (defined as $\widehat{\mathbb{V}}ar[P/V]/\widehat{\mathbb{E}}[P/V]$) such as OmpC. It is noticeable that, as low expressed genes tend to have a low Fano factor (not shown), these genes tend to be more sensitive to the binomial effect of division.

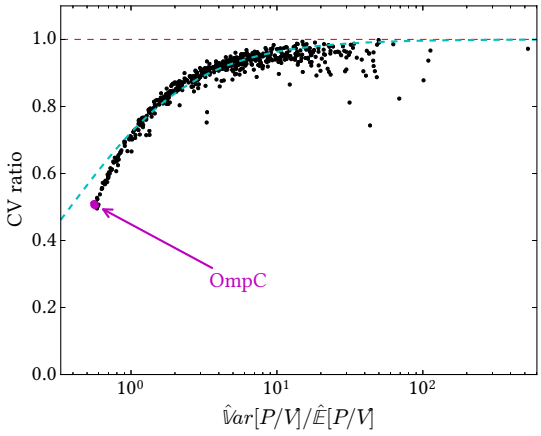
One can explain this clear dependence on the relative variance with a toy model (in dashed cyan line in the figure) which is explained with further details in Section 3.C. In Figure 3.5c is shown the profile of the protein OmpC, with a high contribution of the binomial sampling to the variance: a higher variability at the beginning of the cell cycle that tends to diminish during the cell cycle due to dilution.

Even if the additional noise can be important for some genes, the variability imposed by protein production still prevails. In particular, as it has an impact primarily on less expressed proteins, this effect is not able to explain the "extrinsic noise" lower plateau observed for highly expressed proteins in Taniguchi et al. (2010). The protein CV as a function of the average protein concentration shows behaviour not that different as in Figure 3.4b

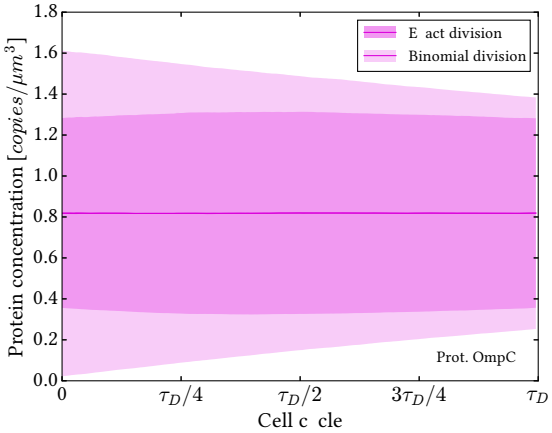
In the next section we propose a more complete model where the mRNAs are no longer created spontaneously in every unit of volume, but their creations depend on the number of gene copies in the cell.



(a) Model with binomial sampling.



(b) Protein noise ratio of the two models



(c) Protein concentration profile for one gene in the two models.

Figure 3.5: Effect of the binomial sample on the protein concentration noise. **(a)**: The model differs from the model with cell cycle (Figure 3.2b) only through division: at division the mRNAs and proteins undergo a binomial sampling. **(b)**: Fractions of the coefficient of variation (CV) of protein concentration in simulations with exact division over simulations with binomial division. The effect of the binomial division can represent up to 50% of the total CV (proteins with low Fano factor). A simplified model of the ratio of noise (cyan dashed line) reproduces this behaviour (see Section 3.C). **(c)**: profile example of the gene OmpC, the central line represents the mean protein concentration over the cell cycle, and the coloured areas represent the standard-deviation in both models.

3.4 Model with Cell Cycle and Gene Replication

This section presents the main model of the chapter. It takes into account all the basic features that can be expected for the production of a type of protein inside a cell cycle: the transcription, the translation, the gene replication and the division. Unlike the previous models, it also represents the gene as an entity that is replicated at some point in the cell cycle, hence doubling the transcription rate at some point in the cell cycle. The goal of the section will be to quantify its contribution to the protein noise. To do so, and contrary to the previous models, we will be able to give analytical expressions for the mean and the variance of the mRNAs and proteins; so that we will not need simulations to estimate the variability each gene expression.

In Subsection 3.4.1 we will present the model and its mechanisms in detail. The two subsections that follow contain the main theoretical results of this section: the explicit distribution for the number of mRNA is given by Theorem 3.3 in Subsection 3.4.2 and the mean and the variance of the number of proteins is given by Theorem 3.4 in Subsection 3.4.3. These two analytical results will be helpful in the last part Subsection 3.4.4 to determine the parameters, and in Subsection 3.4.5 to predict the protein variance of the parameters. In this last subsection, we will show that the gene replication has a low impact on the protein variability.

3.4.1 Presentation of the Model

In the models of the previous section, every gene sees its mRNAs spontaneously created in every volume unit in the cell. It would represent a case where, the gene quantity increases continuously with the volume thus keeping the gene concentration constant. In reality the gene quantity is a discrete number that doubles with DNA-replication. In this section, we propose an extension of the model: cell growth and binomial sampling are still considered but we add the notion of gene replication during the cell cycle.

As for the previous model, it still focuses on one particular gene, the cell volume $V(s)$ is still increasing exponentially during the cell cycle until time τ_D; at division, all compounds undergo a binomial sampling before beginning the new cycle. The difference here is in the rate of mRNA production: it is no longer proportional to the volume, but it remains constant until it doubles at the deterministic time of gene replication τ_R (with $0 < \tau_R < \tau_D$) and remains anew constant until the time of division.[3]

The model represents four aspects of that intervene in protein production (see Figure 3.1b):

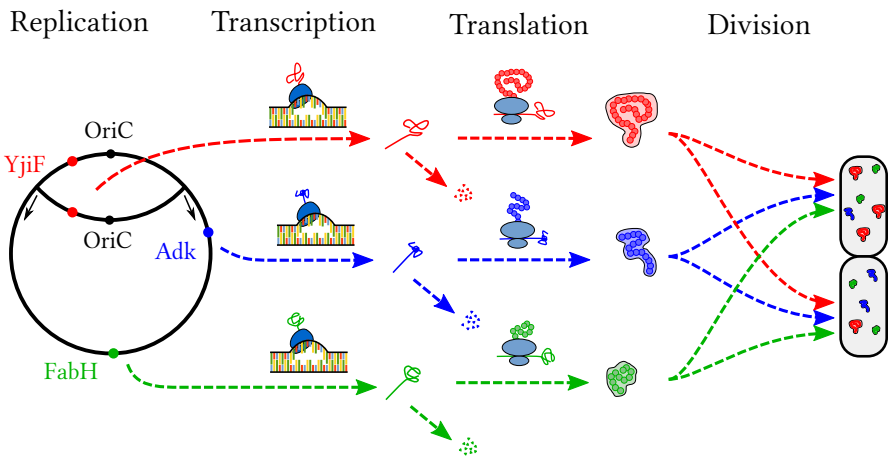
mRNAs Messenger-RNAs are transcribed at constant rate λ_1 before the replication and at constant rate $2\lambda_1$ after gene replication. When transcription happens, the number M of mRNAs then increased by 1. As in the previous model, each mRNA has a lifetime of rate σ_1 (so the global mRNA degradation rate is $\sigma_1 M$).

Proteins Each protein has the same dynamic as presented in the model with cell division gene (Subsection 3.3.2).

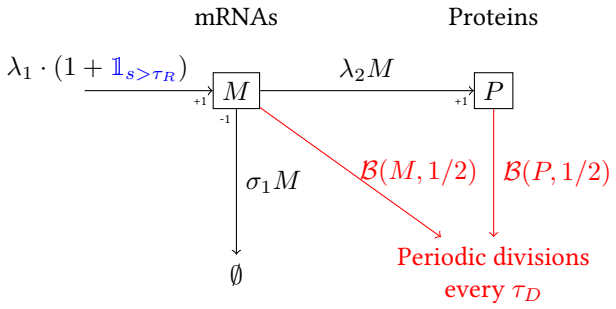
Gene replication At deterministic times τ_R after each division (with $\tau_R < \tau_D$), it occurs the gene replication. The gene responsible for the mRNA transcription is replicated, hence doubling the mRNA transcription rate until next division.

Division Divisions still occur periodically at deterministic times τ_D. The effect is a binomial sampling that only keeps molecules that are in the considered daughter cell. Moreover, as there is only one copy of the gene in the newborn cell, the mRNA transcription rate is anew set to λ_1 until the next gene replication.

[3] Only one replication during the cell cycle is considering here, as it is the case for the slowly growing bacteria of Taniguchi et al. (2010). But the work of this section can be generalised for more than one replication during the cell cycle.



(a) Biological mechanism



(b) Representation of the production of one protein

Figure 3.1: Model with cell cycle and gene replication. **(a)** Four biological mechanisms are represented: DNA-replication, transcription, translation and cell division. **(b)** For one particular type of protein, the number of mRNAs and proteins are respectively M and P; events occur at stochastic times that depend on parameters λ_1, σ_1, λ_2 and on the current state of the system (see main text for more details).

On top of that, the growth of the cell volume is still considered as deterministic: at any time s of the cell cycle the volume of the cell is

$$V(s) = V_0 2^{s/\tau_D}$$

with V_0 being the typical size of a cell at birth.

In this section we will be able to have explicit expressions to describe the mean and the variance for both mRNAs and proteins. It is helpful as it allows to determine the parameters that corresponds to the experimental genes and also that the variability of each protein can be computed directly, without resorting to simulations. The next two subsections gather technical proofs that are needed in order to have analytical results for the protein production. The reader more interested in biological interpretations of the model can directly go to Subsection 3.4.5 and admit the analytical expressions of Theorem 3.3 (depicts the mRNA distribution) and Theorem 3.4 (depicts the protein mean and variance).

3.4.2 Dynamics of mRNA number

The aim of this section is to prove Theorem 3.3 which states that at any time of the cell, the mRNA number follows a Poisson distribution. To do so, we first give a description of the number mRNAs at any time in the cell cycle using Marked Poisson Point Process. Using this description, we will be able to show Proposition 3.4, that the distribution of M_0 at the beginning of the cell cycle is a Poisson distribution. This proposition will allow to finally prove the main theorem of the subsection.

Let's consider that time $s = 0$ is the beginning of a new cell cycle and that the system is already at equilibrium in the same sense as the previous models (see Subsection 3.3.3). We consider that M_0, the number of mRNAs at birth is known. As in Subsection 3.3.3, we assort mRNAs in independent groups; here let's consider three categories:

- mRNAs which were present at the birth of the cell. Each of them is characterised by its lifetime given by an exponential time of rate σ_1. The i-th mRNA is still present at time s if and only if $E^i_{\sigma_1} > s$, with $\left(E^i_{\sigma_1}\right)$ being i.i.d. exponential random variables of parameter σ_1.
- mRNAs created since the birth of the cell by the first copy of the gene. The i-th mRNA of this group is characterised by the time of creation t_i given by a Poisson Process of rate λ_1 and its lifetime δ_i given by a exponential time of rate σ_1.
- mRNAs created since the gene replication by the second copy of the gene. As in the previous group, the i-th mRNA is characterised by the time of creation t_i given by a Poisson Process of rate λ_1 and its lifetime δ_i given by an exponential time of rate σ_1. But here, the Poisson Process of rate λ_1 begins at time τ_R, the time of replication of the gene.

As in Subsection 3.3.3, one can represent the number of mRNAs of the second and the third group as two independent MPPPs $\mathcal{N}$ and $\mathcal{N}'$. The first variable x of each of these MPPPs will represent the time. The intensity of each of the MPPP is the same:

$$\nu(dx, dy) = \lambda_1 dx \otimes \sigma_1 e^{-\sigma_1 y} dy.$$

The only difference between $\mathcal{N}$ and $\mathcal{N}'$ is the fact that they begin at time 0 for $\mathcal{N}$ and at time τ_R for $\mathcal{N}'$ (see Figure 3.2). As a consequence, if we consider an mRNA of either group, the conditions of its existence at time $s \in [0, \tau_D[$ are respectively:

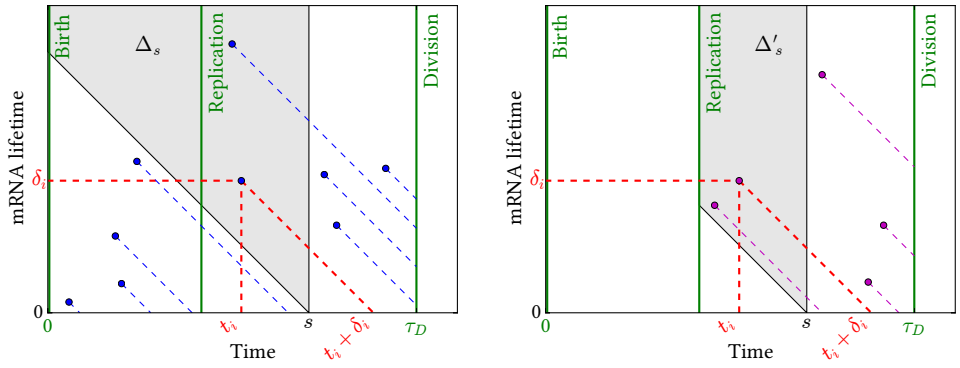
- if it is in the second group: $(t_i, \delta_i) \in \Delta_s$ with $\Delta_s = \{(x, y),\, 0 < x < s,\, y > s - x\}$,
- if it is in the third group: $(t_i, \delta_i) \in \Delta'_s$ with $\Delta'_s = \{(x, y),\, \tau_R < x < s,\, y > s - x\}$.

Hence, we can describe the number of mRNAs at any time $s \in [0, \tau_D[$ as follows:

$$M_s = \sum_{i=1}^{M_0} \mathbb{1}_{\{E^i_{\sigma_1} > s\}} + \mathcal{N}(\Delta_s) + \mathbb{1}_{s \geq \tau_R} \mathcal{N}'_{\lambda_1}(\Delta'_s). \tag{3.11}$$

Each term corresponds to each group of mRNAs previously described.

At first we want to characterise the distribution of M_0, the number of mRNAs at the birth of the cell. To do so, we use the equilibrium hypothesis that implies that $M_0 \overset{\mathcal{D}}{=} M_{\tau_D}$:



(a) Representation of $\mathcal{N}$ (b) Representation of $\mathcal{N}'$

Figure 3.2: The illustration of the Marked Point Poisson Processes that describe the dynamic of mRNAs: each mRNA is characterised by the point (t_i, δ_i), with (t_i, δ_i) following MPPP $\mathcal{N}$ or $\mathcal{N}'$. Both MPPPs are of intensity ν. The random variable t_i represents its birth time and δ_i its lifetime, hence this mRNA exists from time t_i up to time $t_i + \delta_i$. The only difference for the two processes is the starting time: the process $\mathcal{N}$ in **(a)** begins at birth (in particular, an mRNA is still present at time s, if and only if the point (t_i, δ_i) is in the set $\Delta_s = \{(x, y),\ 0 < x < s,\ y > s - x\}$); the process $\mathcal{N}'$, in **(b)**, begins at replication (in particular an mRNA is still present at time s, if and only if the point (t_i, δ_i) is in the set $\Delta'_s = \{(x, y),\ \tau_R < x < s,\ y > s - x\}$).

Proposition 3.4. *At equilibrium, the number of mRNAs at birth M_0 follows a Poisson distribution of parameter:*

$$x_0 = \frac{\lambda_1}{\sigma_1}\left[1 - \frac{e^{-(\tau_D - \tau_R)\sigma_1}}{2 - e^{-\tau_D \sigma_1}}\right].$$

Proof. When $s = \tau_D-$, there is by Equation (3.11):

$$M_{\tau_D-} = \sum_{i=1}^{M_0} \mathbb{1}_{\{E^i_{\sigma_1} > \tau_D -\}} + \mathcal{N}_{\lambda_1}\left(\Delta_{\tau_D-}\right) + \mathcal{N}'_{\lambda_1}\left(\Delta'_{\tau_D-}\right).$$

The first term corresponds to initial messengers that were not degraded after the time τ_D. Let's suppose that M_0 is distributed according to a Poisson distribution of parameter x_0. Then the random variable

$$\sum_{i=1}^{M_0} \mathbb{1}_{\{E^i_{\sigma_1} > \tau_D -\}}$$

also follows a Poisson distribution of parameter $x_0 e^{-\tau_D \sigma_1}$.

The second term corresponds to mRNAs that were created by the first copy of the gene and which are still present at division. Since $\mathcal{N}$ is a MPPP, $\mathcal{N}\left(\Delta_{\tau_D-}\right)$ is a Poisson random variable (Proposition 1.13 of Robert (2010)) with parameter

$$\nu\left(\Delta_{\tau_D-}\right) = \int_0^{\tau_D} \int_{\tau_D - x}^{\infty} \lambda_1 \sigma_1 e^{-\sigma_1 y}\, dy\, dx = \frac{\lambda_1}{\sigma_1}\left(1 - e^{-\tau_D \sigma_1}\right).$$

The third term corresponds to mRNAs that were created by the second copy of the gene (replicated at time τ_R) and which are still present at division. As before, $\mathcal{N}'\left(\Delta'_{\tau_D-}\right)$ is a Poisson random variable of parameter

$$\nu\left(\Delta'_{\tau_D-}\right) = \int_{\tau_R}^{\tau_D}\int_{\tau_D-x}^{\infty} \lambda_1\sigma_1 e^{-\sigma_1 y}\, dy\, dx = \frac{\lambda_1}{\sigma_1}\left(1-e^{-(\tau_D-\tau_R)\sigma_1}\right).$$

As M_{τ_D-} is the sum of three independent Poisson random variables, it comes that

$$\begin{aligned} M_{\tau_D-} &\sim \mathcal{P}\left(x_0 e^{-\sigma_1\tau_D} + \frac{\lambda_1}{\sigma_1}\left(1-e^{-\tau_D\sigma_1}\right) + \frac{\lambda_1}{\sigma_1}\left(1-e^{-(\tau_D-\tau_R)\sigma_1}\right)\right) \\ &\sim \mathcal{P}\left(x_0 e^{-\sigma_1\tau_D} + \frac{\lambda_1}{\sigma_1}\left(2-e^{-\tau_D\sigma_1}-e^{-(\tau_D-\tau_R)\sigma_1}\right)\right). \end{aligned}$$

Between τ_D- and τ_D, with the binomial sampling, each mRNA has an equal chance to stay or to disappears, therefore

$$M_{\tau_D} = \sum_{1=0}^{M_{\tau_D-}} B_{1/2,i}$$

with $\left(B_{1/2,i}\right)$ i.i.d. Bernoulli random variables of parameter $1/2$. The random variable $B_{1/2,i}$ determines if the i-th mRNA is in the next considered cell or not. The random variable M_{τ_D} hence follow a Poisson distribution such as:

$$M_{\tau_D} \sim \mathcal{P}\left(\left[x_0 e^{-\sigma_1\tau_D} + \frac{\lambda_1}{\sigma_1}\left(2-e^{-\tau_D\sigma_1}-e^{-(\tau_D-\tau_R)\sigma_1}\right)\right]/2\right).$$

As the system is at equilibrium, it comes that $M_0 \overset{\mathcal{D}}{=} M_{\tau_D}$, therefore

$$x_0 = \left(x_0 e^{-\sigma_1\tau_D} + \frac{\lambda_1}{\sigma_1}\left(2-e^{-\tau_D\sigma_1}-e^{-(\tau_D-\tau_R)\sigma_1}\right)\right)/2$$

which gives:

$$x_0 = \frac{\lambda_1}{\sigma_1}\left[1-\frac{e^{-(\tau_D-\tau_R)\sigma_1}}{2-e^{-\tau_D\sigma_1}}\right].$$

Since the equilibrium distribution is unique, the number of mRNAs at birth follows a Poisson distribution of parameter x_0 at equilibrium. □

We have determined the equilibrium distribution of the embedded Markov Chain $(M_{i\tau_D})_{i\in\mathbb{N}}$. Now, let's look at the distribution of mRNA number at any time s of the cell cycle:

Theorem 3.3. *At equilibrium, at a time s in the cell cycle, the mRNA number M_s follows a Poisson distribution of parameter*

$$x_s = \frac{\lambda_1}{\sigma_1}\left[1-\frac{e^{-(s+\tau_D-\tau_R)\sigma_1}}{2-e^{-\tau_D\sigma_1}} + \mathbb{1}_{s\geq\tau_R}\left(1-e^{-(s-\tau_R)\sigma_1}\right)\right].$$

In particular, the mean and the variance of mRNA number are known at any time s of the cell cycle:

$$\begin{aligned} \mathbb{E}\left[M_s\right] &= \frac{\lambda_1}{\sigma_1}\left[1-\frac{e^{-(s+\tau_D-\tau_R)\sigma_1}}{2-e^{-\tau_D\sigma_1}} + \mathbb{1}_{s\geq\tau_R}\left(1-e^{-(s-\tau_R)\sigma_1}\right)\right], \\ \mathbb{V}\mathrm{ar}\left[M_s\right] &= \frac{\lambda_1}{\sigma_1}\left[1-\frac{e^{-(s+\tau_D-\tau_R)\sigma_1}}{2-e^{-\tau_D\sigma_1}} + \mathbb{1}_{s\geq\tau_R}\left(1-e^{-(s-\tau_R)\sigma_1}\right)\right]. \end{aligned}$$

Proof. At a moment s of the cell cycle, let's consider the moment-generating function of M_s with $\xi < 0$:

$$\mathbb{E}\left[\exp\left(\xi M_s\right)\right] = \mathbb{E}\left[\exp\left(\xi\left(\sum_{i=1}^{M_0} \mathbb{1}_{\{E_{\sigma_1}^i > s\}} + \mathcal{N}\left(\Delta_s\right) + \mathbb{1}_{s\geq\tau_R}\mathcal{N}'_{\lambda_1}\left(\Delta'_s\right)\right)\right)\right].$$

Since M_0, $E_{\sigma_1}^i$, $\mathcal{N}_{\lambda_1}$ and $\mathcal{N}'_{\lambda_1}$ are all independent, it follows that

$$\mathbb{E}\left[\exp\left(\xi M_s\right)\right] = \mathbb{E}\left[\exp\left(\sum_{i=0}^{M_0} \xi\mathbb{1}_{\{E_{\sigma_1}^i > s\}}\right)\right] \cdot \mathbb{E}\left[\exp\left(\xi\mathcal{N}\left(\Delta_s\right)\right)\right] \cdot \mathbb{E}\left[\exp\left(\xi\mathbb{1}_{s\geq\tau_R}\mathcal{N}'\left(\Delta'_s\right)\right)\right].$$

For the first factor, since all the random variables $\mathbb{1}_{\{E_{\sigma_1}^i > s\}}$ are i.i.d. Bernoulli variables of parameter $e^{-s\sigma_1}$ and independent of M_0, it comes then

$$\mathbb{E}\left[\exp\left(\sum_{i=0}^{M_0} \xi\mathbb{1}_{\{E_{\sigma_1}^i > s\}}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\xi\mathbb{1}_{\{E_{\sigma_1}^1 > s\}}\right)|M_0\right]^{M_0}\right] = \mathbb{E}\left[\exp\left(1 + e^{-s\sigma_1}(e^\xi - 1)\right)^{M_0}\right].$$

With Proposition 3.4, M_0 is known to be a Poisson random variable of parameter x_0, hence, with the probability-generating function of a Poisson random variable, it holds:

$$\mathbb{E}\left[\exp\left(\sum_{i=0}^{M_0} \xi\mathbb{1}_{\{E_{\sigma_1}^i > s\}}\right)\right] = \mathbb{E}\left[\exp\left(x_0 e^{-s\sigma_1}\left(e^\xi - 1\right)\right)\right]$$

For the second factor, one can recall that $\mathcal{N}\left(\Delta_s\right)$ is a Poisson random variable. As in Proposition 3.4, its parameter can be calculated:

$$\nu\left(\Delta_s\right) = \int_0^s \int_{\tau_D - x}^\infty \lambda_1\sigma_1 e^{-\sigma_1 y}\, dy\, dx = \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right).$$

Identically for the third factor, $\mathcal{N}'\left(\Delta'_s\right)$ is a Poisson random variable of parameter:

$$\nu\left(\Delta'_s\right) = \int_{\tau_R}^s \int_{\tau_D - x}^\infty \lambda_1\sigma_1 e^{-\sigma_1 y}\, dy\, dx = \frac{\lambda_1}{\sigma_1}\left(1 - e^{-(s-\tau_R)\sigma_1}\right).$$

As a consequence, the moment-generating function of M_s is

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\xi M_s\right)\right] &= \mathbb{E}\left[\left(x_0 e^{-s\sigma_1} + \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right) + \mathbb{1}_{s>\tau_R}\frac{\lambda_1}{\sigma_1}\left(1 - e^{-(s-\tau_R)\sigma_1}\right)\right)\left(e^\xi - 1\right)\right] \\
&= \mathbb{E}\left[\left(x_0 e^{-s\sigma_1} + \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1} + \mathbb{1}_{s>\tau_R}\left(1 - e^{-(s-\tau_R)\sigma_1}\right)\right)\right)\left(e^\xi - 1\right)\right]
\end{aligned}$$

which is the moment-generating function of a Poisson random variable of parameter

$$x_0 e^{-s\sigma_1} + \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1} + \mathbb{1}_{s>\tau_R}\left(1 - e^{-(s-\tau_R)\sigma_1}\right)\right).$$

□

In this subsection, we have managed to obtain explicit expressions for the mean and the variance of mRNA number at any time s of the cell cycle at equilibrium (and in particular, its mean and its variance are known). In the next subsection, we are interested in obtaining the same kind of results for proteins.

3.4.3 Dynamics of protein number

As for the previous analysis of the mRNA number, we search an expression for the protein production through the cell cycle. This case is more complicated than the mRNA case: the protein distribution is not as simple as a Poisson distribution, and we will only calculate analytical expressions only for the first two moments of P_s.

Theorem 3.4 is the main theoretical result of this section: for any time s of the cell cycle, it gives explicit expressions for the mean $\mathbb{E}[P_s]$ and the variance $\mathbb{V}\text{ar}[P_s]$ of the protein number. This result is important as in the next section, it will be used to calculate directly the mean $\widehat{\mathbb{E}}[P/V]$ and variance $\widehat{\mathbb{V}}ar[P/V]$ of the protein concentration averaged across the cell cycle without using simulations: only with the parameters of the model (λ_1, σ_1, λ_2, τ_R and τ_D), we will be able to know the behaviour of the protein concentration in terms of variance.

In order to prove the Theorem 3.4, we will characterise $\mathbb{E}[P_s]$ and $\mathbb{V}\text{ar}[P_s]$ in the two following cases:

1. In a first step, we consider the case before replication ($s < \tau_R$). We begin by considering that the state of the cell at birth (M_0, P_0) is known and we calculate the first two moments of P_s for any time $s < \tau_R$ (Corollary 3.1). Then, we integrate over all the possible initial states (M_0, P_0) to determine expressions for $\mathbb{E}[P_s]$ and $\mathbb{V}\text{ar}[P_s]$ for any time $s < \tau_R$ (Proposition 3.6). These expressions are dependant of the first moments of (M_0, P_0): they depend on $\mathbb{E}[M_0]$, $\mathbb{E}[P_0]$, $\mathbb{V}\text{ar}[M_0]$, $\mathbb{V}\text{ar}[P_0]$ and $\mathbb{C}\text{ov}[M_0, P_0]$.

2. In a second step, we consider the case after replication ($s \geq \tau_R$). Similarly the first case, we will consider that the state of the cell at replication (M_{τ_R}, P_{τ_R}) is known and we calculate the first two moments of P_s for any time $\tau_R \leq s < \tau_D$ (Proposition 3.8). After integration, expressions for $\mathbb{E}[P_s]$ and $\mathbb{V}\text{ar}[P_s]$ for any time s after replication are determined, these expressions depend on $\mathbb{E}[M_{\tau_R}]$, $\mathbb{E}[P_{\tau_R}]$, $\mathbb{V}\text{ar}[M_{\tau_R}]$, $\mathbb{V}\text{ar}[P_{\tau_R}]$ and $\mathbb{C}\text{ov}[M_{\tau_R}, P_{\tau_R}]$ (Proposition 3.8).

In the end, in Theorem 3.4, are presented the mean and variance of protein number at any time s of the cell cycle, only depending on the first moments of (M_0, P_0) and (M_{τ_R}, P_{τ_R}). Additional results are presented in the appendix Section 3.B, which determine explicitly the first moments of (M_0, P_0) and (M_{τ_R}, P_{τ_R}) so that the mean and variance of protein number will be fully characterised.

Description of the Protein Number Process

Before beginning, let's at first have a description for the number of proteins P_s for any time s. We will use this description in the upcoming proofs. Similarly to mRNA case (Equation (3.11)) , we group them into two categories:

- The proteins that were there at birth and which remain in the cell during all the cell cycle (as said in Subsection 3.4.1 the proteolysis is not considered in this model).
- The proteins that were created during the cell cycle. The rate of production depends on the current number of mRNAs. For that we consider $\left(\mathcal{N}^i_{\lambda_2}\right)_{i\in\mathbb{N}}$, a sequence of i.i.d. Poisson Point Processes of intensity λ_2; if the i-th mRNA exists at time s (that is to say if $i \leq M_s$), then the number of proteins produced by this mRNA between s and $s+ds$ is $\mathcal{N}^i_{\lambda_2}(ds)$. Hence, the total number of proteins produced between s and $s + ds$ is then $\sum_{i=1}^{\infty} \mathbb{1}_{i\leq M_u}\mathcal{N}^i_{\lambda_2}(ds)$.

To sum up, the number of proteins at a time s of the cell cycle is:

$$P_s = P_0 + \sum_{i=1}^{\infty} \int_0^s \mathbb{1}_{i\leq M_u}\mathcal{N}^i_{\lambda_2}(du)\,. \tag{3.12}$$

The first term is the number of proteins at birth, and the second takes into account all proteins created between times 0 and s.

Protein Number Before Replication

Let's beginning with the case before replication, where $s < \tau_R$. The random variables M_0 and P_0 are supposed to be known; so that we use the notation $\mathbb{E}_{M_0,P_0}[.] := \mathbb{E}[.|(M_0,P_0)]$. At first, we want to characterise the first two moments of P_s for any time s of the cell cycle conditionally to (M_0,P_0). To do so, as for the mRNAs, we determine at first the moment-generating function of P_s.

Proposition 3.5. *For any $s \in [0,\tau_R[$, by supposing the birth state (M_0,P_0) known, it comes that the moment generating function of P_s is:*

$$\mathbb{E}_{M_0,P_0}[\exp(\xi P_s)] = \exp(\xi P_0)\cdot h_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right)$$

for any $\xi < 0$ and such as $h_{1,s}$ is the moment generating function of $\int_0^s M_u\,du$. The expression of $h_{1,s}$ is:

$$h_{1,s}(\xi) := \exp\left[M_0\log\left[\frac{\sigma_1-\xi e^{-(\sigma_1-\xi)s}}{\sigma_1-\xi}\right]+\lambda_1\frac{\xi}{\sigma_1-\xi}\left(s-\frac{1-e^{-(\sigma_1-\xi)s}}{\sigma_1-\xi}\right)\right].$$

Proof. With Equation (3.12), it is is easy to show that

$$\mathbb{E}_{M_0,P_0}[\exp(\xi P_s)] = \exp(\xi P_0)\cdot\mathbb{E}_{M_0,P_0}\left[\prod_{i=1}^{\infty}\mathbb{E}_{M_0,P_0}\left[\exp\left(\xi\int_0^s \mathbb{1}_{i\le M_u}\mathcal{N}^i_{\lambda_2}(du)\right)|(M_u)_{u\le s}\right]\right].$$

We then consider the Laplace functional for the Poisson process $\mathcal{N}^i_{\lambda_2}$:

$$\begin{aligned}\mathbb{E}_{M_0,P_0}\left[\exp\left(\xi\int_0^s \mathbb{1}_{i\le M_u}\mathcal{N}^i_{\lambda_2}(du)\right)|(M_u)_{u\le s}\right] &= \exp\left[\lambda_2\int_0^{\infty}\left(\exp\left(\xi\mathbb{1}_{i\le M_u}\mathbb{1}_{u\le s}\right)-1\right)du\right]\\ &= \exp\left[\lambda_2\left(e^{\xi}-1\right)\int_0^s \mathbb{1}_{i\le M_u}du\right].\end{aligned}$$

By making the product for i from 1 to infinity, it comes that:

$$\prod_{i=1}^{\infty}\mathbb{E}_{M_0,P_0}\left[\exp\left(\xi\int_0^s \mathbb{1}_{i\le M_u}\mathcal{N}^i_{\lambda_2}(du)\right)|(M_u)_{u\le s}\right] = \exp\left[\lambda_2\left(e^{\xi}-1\right)\int_0^s M_u du\right].$$

As a consequence, it indeed follows that

$$\mathbb{E}_{M_0,P_0}[\exp(\xi P_s)] = \exp(\xi P_0)\cdot h_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right).$$

What remains is to show the expression of $h_{1,s}(\xi)$. The expression of M_s in Equation (3.11) integrated between time 0 and time $s < \tau_R$ gives

$$\int_0^s M_s\,du = \int_0^s\sum_{i=1}^{M_0}\mathbb{1}_{\{E^i_{\sigma_1}>u\}}\,du + \int_0^s \mathcal{N}(\Delta_u)\,du.$$

As $h_{1,s}$ is the moment-generating function of $\int_0^s M_s\,du$, it follows that for any $\xi < 0$:

$$\begin{aligned}h_{1,s}(\xi) &= \mathbb{E}_{M_0,P_0}\left[\exp\left(\xi\int_0^s\sum_{i=1}^{M_0}\mathbb{1}_{\{E^i_{\sigma_1}>u\}}du\right)\right]\mathbb{E}\left[\exp\left(\xi\int_0^s\mathcal{N}(\Delta_u)\,du\right)\right]\\ &= \mathbb{E}\left[\exp\left(\xi E^1_{\sigma_1}\wedge s\right)\right]^{M_0}\mathbb{E}\left[\exp\left(\xi\int_0^s\mathcal{N}(\Delta_u)\,du\right)\right].\end{aligned}$$

We recall that $\mathcal{N}$ a MPPP of intensity $\nu = \lambda_1 dx \otimes \sigma_1 e^{-\sigma_1 y} dy$ and that Δ_s is defined as $\Delta_s = \{(x,y),\, 0 < x < s,\, y > s - x\}$.

Let's begin with the first term of $h_{1,s}(\xi)$:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\xi E^1_{\sigma_1} \wedge s\right)\right]^{M_0} &= \left(\int_0^s e^{\xi u}\sigma_1 e^{-\sigma_1 u}\, du + \int_s^\infty e^{\xi s}\sigma_1 e^{-\sigma_1 u}\, du\right)^{M_0} \\ &= \left(\frac{\sigma_1 - \xi e^{-(\sigma_1-\xi)s}}{\sigma_1 - \xi}\right)^{M_0}.\end{aligned}$$

Let's continue with the second term of $h_{1,s}(\xi)$. The integration of $\mathcal{N}(\Delta_u)$ on $[0,s[$ gives:

$$\begin{aligned}\int_0^s \mathcal{N}(\Delta_u)\, du &= \int_0^s \iint_{\mathbb{R}^2} \mathbb{1}_{x<u} \cdot \mathbb{1}_{y<u-x}\, \mathcal{N}(dx,dy)\, du \\ &= \iint_{\mathbb{R}^2} \int_0^s \mathbb{1}_{x<u<x+y}\, du\, \mathcal{N}(dx,dy) \\ &= \iint_{\mathbb{R}^2} ((x+y)\wedge s - x \wedge s)\, \mathcal{N}(dx,dy).\end{aligned}$$

We then consider the Laplace functional for this MPPP:

$$\mathbb{E}\left[\exp\left(\mathcal{N}(g)\right)\right] = \exp\left[\lambda_1 \int_0^\infty \int_0^\infty \left(e^{g(x,y)} - 1\right) dx\, \sigma_1(dy)\right]$$

with $\sigma_1(dy)$ the density distribution of an exponential random variable of parameter σ_1. In our case $g(x,y) := \xi((x+y)\wedge s - x \wedge s)$:

$$\begin{aligned}\mathbb{E}\left[\exp\left(\xi \int_0^s \mathcal{N}(\Delta_u)\, du\right)\right] &= \exp\left[\lambda_1 \int_0^\infty \int_0^\infty \left(\exp\left[\xi\left((x+y)\wedge s - x\wedge s\right)\right] - 1\right) dx\, \sigma_1(dy)\right] \\ &= \exp\left[\lambda_1 \left(\int_0^s \int_0^\infty \left(\exp\left[\xi\left(y \wedge (s-x)\right)\right]\right)\, \sigma_1(dy)\, dx - s\right)\right] \\ &= \exp\left[\lambda_1 \left(\int_0^s \int_0^{s-x} \exp\left[\xi y\right]\, \sigma_1(dy)\, dx + \int_0^s \int_{s-x}^\infty \exp\left[\xi(s-x)\right]\, \sigma_1(dy)\, dx - s\right)\right] \\ &= \exp\left[\lambda_1 \frac{\xi}{\sigma_1 - \xi}\left(s - \frac{1 - e^{-(\sigma_1-\xi)s}}{\sigma_1 - \xi}\right)\right].\end{aligned}$$

As a consequence the moment-generating function of $\int_0^s M_s\, du$ is given by:

$$h_{1,s}(\xi) = \exp\left[M_0 \log\left[\frac{\sigma_1 - \xi e^{-(\sigma_1-\xi)s}}{\sigma_1 - \xi}\right] + \lambda_1 \frac{\xi}{\sigma_1 - \xi}\left(s - \frac{1 - e^{-(\sigma_1-\xi)s}}{\sigma_1 - \xi}\right)\right]$$

□

As the moment generating function of P_s has been characterised, it is possible to deduce, by derivation, the first two moments of P_s knowing (M_0, P_0) for any time s before the gene replication.

Corollary 3.1. *At equilibrium, for a time $s \in [0, \tau_R[$, knowing the state of the cell at birth (M_0, P_0), the first two moments of P_s are:*

$$\begin{aligned}
\mathbb{E}_{M_0,P_0}[P_s] &= P_0 + \lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right), \\
\mathbb{E}_{M_0,P_0}\left[P_s^2\right] &= \left(\mathbb{E}_{M_0,P_0}[P_s]\right)^2 + M_0\frac{\lambda_2}{\sigma_1}\left(1 - e^{-\sigma_1 s} + \frac{\lambda_2}{\sigma_1}\left[1 - e^{-\sigma_1 s}\left(e^{-\sigma_1 s} + 2s\sigma_1\right)\right]\right) \\
&\quad + \frac{\lambda_1\lambda_2}{\sigma_1^2}\left[s\sigma_1 - 1 + e^{-\sigma_1 s} + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1 s\left(1 + e^{-\sigma_1 s}\right) - 2\left(1 - e^{-\sigma_1 s}\right)\right)\right]
\end{aligned}$$

Proof. From Proposition 3.5, it follows that the first two moments of P_s can be obtained by derivation of the moment-generating function:

$$\begin{aligned}
\mathbb{E}_{M_0,P_0}[P_s] &= \lim_{\xi\to 0}\frac{d}{d\xi}\left[\exp(\xi P_0)\, h_{1,s}\left(\lambda_2\left(e^{\xi} - 1\right)\right)\right] \\
&= \lim_{\xi\to 0}\left[\exp(\xi P_0)\cdot\left(P_0 h_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right) + \lambda_2 e^{\xi} h'_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right)\right)\right] \\
&= P_0 + \lambda_2 h'_{1,s}(0)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{M_0,P_0}\left[P_s^2\right] &= \lim_{\xi\to 0}\frac{d^2}{d\xi^2}\left[\exp(\xi P_0)\, h_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right)\right] \\
&= \lim_{\xi\to 0}\Big[\exp(\xi P_0)\times\left(P_0^2 h_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right) + 2P_0\lambda_2 e^{\xi} h'_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right)\right. \\
&\qquad \left.+\lambda_2 e^{\xi} h'_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right) + \left(\lambda_2 e^{\xi}\right)^2 h'_{1,s}\left(\lambda_2\left(e^{\xi}-1\right)\right)\right)\Big] \\
&= P_0^2 + (2P_0+1)\lambda_2 h'_{1,s}(0) + (\lambda_2)^2 h''_{1,s}(0) \\
&= \left(\mathbb{E}_{M_0,P_0}[P_s]\right)^2 + \lambda_2 h'_{1,s}(0) + (\lambda_2)^2\left(h''_{1,s}(0) - h'_{1,s}(0)^2\right)
\end{aligned}$$

Consider the expression of $h_{1,s}$ of Proposition 3.5, and let's search its derivatives:

$$\begin{aligned}
h_{1,s}(\xi) &= \exp\left(M_0\log\left[\frac{\sigma_1 - \xi e^{-(\sigma_1-\xi)s}}{\sigma_1 - \xi}\right] + \lambda_1\frac{\xi}{\sigma_1-\xi}\left(s - \frac{1-e^{-(\sigma_1-\xi)s}}{\sigma_1-\xi}\right)\right), \\
h'_{1,s}(\xi) &= h_{1,s}(\xi)\left\{M_0\left[\frac{1}{(\sigma_1-\xi)} - \frac{(1+\xi s)\,e^{-(\sigma_1-\xi)s}}{\left(\sigma_1 - \xi e^{-(\sigma_1-\xi)s}\right)}\right]\right. \\
&\qquad \left.+\lambda_1\left(\frac{\sigma_1}{(\sigma_1-\xi)^2}s - \frac{1-e^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^2} + \xi\cdot\frac{se^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^2} - 2\xi\cdot\frac{1-e^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^3}\right)\right\}, \\
h''_{1,s}(\xi) &= h_{1,s}(\xi)\left\{h'_{1,s}(\xi)^2 + M_0\left[\frac{1}{(\sigma_1-\xi)} - \frac{(1+\xi s)\,e^{-(\sigma_1-\xi)s}}{\left(\sigma_1 - \xi e^{-(\sigma_1-\xi)s}\right)}\right]\right. \\
&\qquad \left.+\lambda_1\left(\frac{\sigma_1}{(\sigma_1-\xi)^2}s - \frac{1-e^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^2} + \xi\cdot\frac{se^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^2} - 2\xi\cdot\frac{1-e^{-(\sigma_1-\xi)s}}{(\sigma_1-\xi)^3}\right)\right\}.
\end{aligned}$$

By taking the limit for $\xi \to 0$, it follows:

$$\begin{aligned} h'_{1,s}(0) &= M_0 \frac{1-e^{-\sigma_1 s}}{\sigma_1} + \frac{\lambda_1}{\sigma_1}\left(s - \frac{1-e^{-\sigma_1 s}}{\sigma_1}\right), \\ h''_{1,s}(0) &= h'_{1,s}(0)^2 + \frac{M_0}{\sigma_1^2}\left[1 - e^{-\sigma_1 s}\left(e^{-\sigma_1 s} + 2s\sigma_1\right)\right] + 2\frac{\lambda_1}{\sigma_1^3}\left(\sigma_1 s\left(1+e^{-\sigma_1 s}\right) - 2\left(1-e^{-\sigma_1 s}\right)\right) \end{aligned}$$

hence the result. □

The previous corollary gives expressions for $\mathbb{E}_{M_0,P_0}[P_s]$ and $\mathbb{E}_{M_0,P_0}\left[P_s^2\right]$. In the next proposition, we integrate these expressions over all birth states (M_0, P_0) to find formulas for $\mathbb{E}[P_s]$ and $\mathbb{V}\mathrm{ar}[P_s]$ for any time $s < \tau_R$ before replication. These expression depends on joint moments of M_0 and P_0.

Proposition 3.6. *Let's consider the functions:*

$$\begin{aligned} f_1(s) &:= \left(\frac{\lambda_1}{\sigma_1}s + \left(x_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right), \\ g_1(s) &:= \left(\lambda_2 \frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)^2 x_0 \\ &\quad + x_0\frac{\lambda_2}{\sigma_1}\left(1 - e^{-\sigma_1 s} + \frac{\lambda_2}{\sigma_1}\left[1 - e^{-\sigma_1 s}\left(e^{-\sigma_1 s} + 2s\sigma_1\right)\right]\right) \\ &\quad + \frac{\lambda_1\lambda_2}{\sigma_1^2}\left[s\sigma_1 - 1 + e^{-\sigma_1 s} + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1 s\left(1+e^{-\sigma_1 s}\right) - 2\left(1-e^{-\sigma_1 s}\right)\right)\right] \end{aligned}$$

with x_0 defined in Proposition 3.4. At any time $s \in [0, \tau_R[$ before replication, depending on joint moments of P_0 and M_0, the mean and the variance of P_s are given by:

$$\begin{aligned} \mathbb{E}[P_s] &= \mathbb{E}[P_0] + \lambda_2 f_1(s), \\ \mathbb{V}\mathrm{ar}[P_s] &= \mathbb{V}\mathrm{ar}[P_0] + 2\lambda_2 \frac{1-e^{-\sigma_1 s}}{\sigma_1}\mathbb{C}\mathrm{ov}[P_0, M_0] + g_1(s). \end{aligned}$$

Proof. By considering the mean of the random variable $\mathbb{E}[P_s|(M_0,P_0)]$ in Corollary 3.1, it directly comes the result for $\mathbb{E}[P_s]$. For the variance, let's consider the expression of $\mathbb{E}\left[P_s^2|(M_0,P_0)\right]$

$$\begin{aligned} \mathbb{E}\left[P_s^2\right] &= \mathbb{E}\left[\left(\mathbb{E}_{M_0,P_0}[P_s]\right)^2\right] + \mathbb{E}[M_0]\frac{\lambda_2}{\sigma_1}\left(1 - e^{-\sigma_1 s} + \frac{\lambda_2}{\sigma_1}\left[1 - e^{-\sigma_1 s}\left(e^{-\sigma_1 s} + 2s\sigma_1\right)\right]\right) \\ &\quad + \frac{\lambda_1\lambda_2}{\sigma_1^2}\left[s\sigma_1 - 1 + e^{-\sigma_1 s} + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1 s\left(1+e^{-\sigma_1 s}\right) - 2\left(1-e^{-\sigma_1 s}\right)\right)\right] \\ \mathbb{V}\mathrm{ar}[P_s] &= \mathbb{E}\left[\mathbb{E}_{M_0,P_0}[P_s]^2\right] - \mathbb{E}[P_s]^2 + \mathbb{E}[M_0]\frac{\lambda_2}{\sigma_1}\left(1 - e^{-\sigma_1 s} + \frac{\lambda_2}{\sigma_1}\left[1 - e^{-\sigma_1 s}\left(e^{-\sigma_1 s} + 2s\sigma_1\right)\right]\right) \\ &\quad + \frac{\lambda_1\lambda_2}{\sigma_1^2}\left[s\sigma_1 - 1 + e^{-\sigma_1 s} + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1 s\left(1+e^{-\sigma_1 s}\right) - 2\left(1-e^{-\sigma_1 s}\right)\right)\right]. \end{aligned}$$

Now let's consider the expression of $\mathbb{E}\left[\mathbb{E}_{M_0,P_0}\left[P_s\right]^2\right] - \mathbb{E}\left[P_s\right]^2$:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{M_0,P_0}\left[P_s\right]^2\right] - \mathbb{E}\left[P_s\right]^2 &= \mathbb{E}\left[P_0^2\right] + \mathbb{E}\left[\left(\lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right)^2\right] \\
&\quad + 2\mathbb{E}\left[P_0 \times \lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right] \\
&\quad - \mathbb{E}\left[P_0\right]^2 + \mathbb{E}\left[\lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right]^2 \\
&\quad - 2\mathbb{E}\left[P_0\right]\mathbb{E}\left[\lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right] \\
&= \mathbb{V}\mathrm{ar}\left[P_0\right] + \mathbb{V}\mathrm{ar}\left[\lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right] \\
&\quad + 2\mathbb{C}\mathrm{ov}\left[P_0, \lambda_2\left(\frac{\lambda_1}{\sigma_1}s + \left(M_0 - \frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1 s}}{\sigma_1}\right)\right].
\end{aligned}$$

Finally, one has just to remark that due to Proposition 3.4 $\mathbb{E}\left[M_0\right] = \mathbb{V}\mathrm{ar}\left[M_0\right] = x_0$. □

Protein Number After Replication

Let's continue to the case after replication, for a time s such as $\tau_R \leq s < \tau_D$. We adopt a similar approach as for the previous case: let's consider that the state just after replication (M_{τ_R}, P_{τ_R}) is known, and we want to determine the first two moments of P_s for any time s after the replication.

Proposition 3.7. *At equilibrium, for a time $s \in [\tau_R, \tau_D[$, knowing the state of the cell at replication (M_{τ_R}, P_{τ_R}), the first two moments of P_s are:*

$$\begin{aligned}
\mathbb{E}_{M_{\tau_R},P_{\tau_R}}\left[P_s\right] &= P_{\tau_R} + \lambda_2\left(2\frac{\lambda_1}{\sigma_1}\left(s-\tau_R\right) + \left(M_{\tau_R} - 2\frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1(s-\tau_R)}}{\sigma_1}\right), \\
\mathbb{E}_{M_{\tau_R},P_{\tau_R}}\left[P_s^2\right] &= \left(\mathbb{E}_{M_{\tau_R},P_{\tau_R}}\left[P_s\right]\right)^2 \\
&\quad + M_{\tau_R}\frac{\lambda_2}{\sigma_1}\left(1 - e^{-\sigma_1(s-\tau_R)} + \frac{\lambda_2}{\sigma_1}\left[1 - e^{-\sigma_1(s-\tau_R)}\left(e^{-\sigma_1(s-\tau_R)} + 2\left(s-\tau_R\right)\sigma_1\right)\right]\right) \\
&\quad + 2\frac{\lambda_1\lambda_2}{\sigma_1^2}\left[\left(s-\tau_R\right)\sigma_1 - 1 + e^{-\sigma_1(s-\tau_R)}\right. \\
&\qquad \left. + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1\left(s-\tau_R\right)\cdot\left(1 + e^{-\sigma_1(s-\tau_R)}\right) - 2\left(1 - e^{-\sigma_1(s-\tau_R)}\right)\right)\right].
\end{aligned}$$

Proof. After the replication, the rate of mRNA production is doubled, but otherwise, the dynamic is identical as it was before the replication. One can hence easily adapt the proofs of Proposition 3.5 and Corollary 3.1, by replacing the initial state by the state at replication (M_{τ_R}, P_{τ_R}), by considering that the mRNA production rate is $2\lambda_1$, and that the time spent since the initial state is $s - \tau_R$. □

We can then integrate the previous expressions on all possible states at replication (M_{τ_R}, P_{τ_R}). It follows that:

Proposition 3.8. *Let's consider the functions:*

$$\begin{aligned}
f_2(s) &:= \left(2\frac{\lambda_1}{\sigma_1}(s-\tau_R) + \left(x_{\tau_R} - 2\frac{\lambda_1}{\sigma_1}\right)\frac{1-e^{-\sigma_1(s-\tau_R)}}{\sigma_1}\right),\\
g_2(s) &:= \left(\lambda_2\frac{1-e^{-\sigma_1(s-\tau_R)}}{\sigma_1}\right)^2 x_{\tau_R}\\
&\quad + x_{\tau_R}\frac{\lambda_2}{\sigma_1}\left(1-e^{-\sigma_1(s-\tau_R)} + \frac{\lambda_2}{\sigma_1}\left[1-e^{-\sigma_1(s-\tau_R)}\left(e^{-\sigma_1(s-\tau_R)} + 2(s-\tau_R)\sigma_1\right)\right]\right)\\
&\quad + 2\frac{\lambda_1\lambda_2}{\sigma_1^2}\left[(s-\tau_R)\sigma_1 - 1 + e^{-\sigma_1(s-\tau_R)}\right.\\
&\qquad \left. + 2\frac{\lambda_2}{\sigma_1}\left(\sigma_1(s-\tau_R)\left(1+e^{-\sigma_1(s-\tau_R)}\right) - 2\left(1-e^{-\sigma_1(s-\tau_R)}\right)\right)\right].
\end{aligned}$$

with x_{τ_R} as defined in Theorem 3.3. At any time $s \in [\tau_R, \tau_D[$ after replication, depending on joint moments of P_{τ_R} and M_{τ_R}, the mean and the variance of P_s are given by:

$$\begin{aligned}
\mathbb{E}[P_s] &= \mathbb{E}[P_{\tau_R}] + \lambda_2 f_2(s),\\
\mathbb{V}\mathrm{ar}[P_s] &= \mathbb{V}\mathrm{ar}[P_{\tau_R}] + 2\lambda_2\frac{1-e^{-\sigma_1(s-\tau_R)}}{\sigma_1}\mathbb{C}\mathrm{ov}[P_{\tau_R}, M_{\tau_R}] + g_2(s).
\end{aligned}$$

Proof. It is similar to the proof of Proposition 3.6. □

Protein Number in the Whole Cell Cycle

Now we can gather the two previous cases, we are able to propose an expression for $\mathbb{E}[P_s]$ and $\mathbb{V}\mathrm{ar}[P_s]$ for any time s of the cell cycle.

Theorem 3.4. *At any time s of the cell cycle, the mean and the variance of the protein number P_s are*

$$\begin{aligned}
\mathbb{E}[P_s] &= \mathbb{E}[P_0] + \lambda_2\left(f_1(\tau_R \wedge s) + \mathbb{1}_{s\geq\tau_R} f_2(s)\right)\\
\mathbb{V}\mathrm{ar}[P_s] &= \mathbb{V}\mathrm{ar}[P_0] + 2\lambda_2\frac{1-e^{-\sigma_1 s\wedge\tau_R}}{\sigma_1}\mathbb{C}\mathrm{ov}[P_0, M_0] + g_1(s\wedge\tau_R)\\
&\quad + \mathbb{1}_{s\geq\tau_R}\left(2\lambda_2\frac{1-e^{-\sigma_1(s-\tau_R)}}{\sigma_1}\mathbb{C}\mathrm{ov}[P_{\tau_R}, M_{\tau_R}] + g_2(s)\right)
\end{aligned}$$

with f_1 and g_1 defined in Proposition 3.6 and f_2 and g_2 defined in Proposition 3.8.

Proof. For $s < \tau_R$, the expressions correspond to those of Proposition 3.6.

For $\tau_R \leq s < \tau_D$, one can remark that a direct consequence of this proposition gives expressions for $\mathbb{E}[P_{\tau_R}]$ and $\mathbb{V}\mathrm{ar}[P_{\tau_R}]$ only depending on $\mathbb{E}[P_0]$ and $\mathbb{V}\mathrm{ar}[P_0]$. Indeed, since almost surely it comes that $P_{\tau_R-} \overset{a.s.}{=} P_{\tau_R}$, one can consider the limit for $s \to \tau_R$ in the case of $s \in [0, \tau_R[$ in Proposition 3.6:

$$\mathbb{E}[P_{\tau_R}] = \mathbb{E}[P_0] + \lambda_2 f_1(\tau_R), \tag{3.13}$$

$$\mathbb{V}\mathrm{ar}[P_{\tau_R}] = \mathbb{V}\mathrm{ar}[P_0] + 2\lambda_2\frac{1-e^{-\sigma_1\tau_R}}{\sigma_1}\mathbb{C}\mathrm{ov}[P_0, M_0] + g_1(\tau_R). \tag{3.14}$$

Consequently, it is possible to write expressions of $\mathbb{E}[P_s]$ and $\mathbb{V}\mathrm{ar}[P_s]$ as depending only on $\mathbb{E}[P_0]$, $\mathbb{V}\mathrm{ar}[P_0]$, $\mathbb{C}\mathrm{ov}[P_0, M_0]$ and $\mathbb{C}\mathrm{ov}[P_{\tau_R}, M_{\tau_R}]$. □

The formulas of $\mathbb{E}[P_s]$ and $\mathbb{V}\mathrm{ar}[P_s]$ of the theorem still depend on $\mathbb{E}[P_0]$, $\mathbb{V}\mathrm{ar}[P_0]$, $\mathbb{C}\mathrm{ov}[P_0, M_0]$ and $\mathbb{C}\mathrm{ov}[P_{\tau_R}, M_{\tau_R}]$. Explicit expressions for these quantities are still unknown. But as we are at equilibrium, it comes that $(P_0, M_0) \overset{\mathcal{D}}{=} (P_{\tau_D}, M_{\tau_D})$ which allows to find expressions for all these quantities that explicitly only depend on the model parameters: λ_1, σ_1, λ_2, τ_R and τ_D. These expressions are given in the appendix Section 3.B: the expressions of $\mathbb{E}[P_0]$, $\mathbb{V}\mathrm{ar}[P_0]$ are given in Proposition 3.9, as the expressions of $\mathbb{C}\mathrm{ov}[P_0, M_0]$ and $\mathbb{C}\mathrm{ov}[P_{\tau_R}, M_{\tau_R}]$ are determined in Proposition 3.10.

3.4.4 Parameter Computation

In this subsection, we explain how to fix the parameters λ_1, σ_1, λ_2 and τ_R to make them correspond to the exponential measures. As in Subsection 3.3.5, we characterise the mean and the variance of either mRNA or protein concentration, not at a given time s, but over the whole cell cycle. To do so, we define the global average and global variance of mRNAs as

$$\widehat{\mathbb{E}}[M/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \mathbb{E}\left[\frac{M_s}{V(s)}\right] ds, \tag{3.15}$$

$$\widehat{\mathbb{V}}ar[M/V] := \frac{1}{\tau_D}\int_0^{\tau_D} \left[\mathbb{E}\left[\left(\frac{M_s}{V(s)}\right)^2\right] - \widehat{\mathbb{E}}[M/V]^2\right] ds; \tag{3.16}$$

and consider analogue definitions for the proteins: $\widehat{\mathbb{E}}[P/V]$ and $\widehat{\mathbb{V}}ar[P/V]$.

As in the previous models, we set the doubling time τ_D to 150 min and the volume at birth $V_0 = 1.3\,\mu\mathrm{m}^3$. For each gene, we have to determine four different parameters $\lambda_1, \sigma_1, \lambda_2$ and τ_R. We have considered the genes of Taniguchi et al. (2010) for which the empirical mean of messengers μ_m and proteins μ_p concentrations, as well as the mRNA half-life time τ_m have been measured. We still deduce the mRNA degradation rate σ_1 with the mRNA half-life time τ_m (such that $\sigma_{1,i} = \log 2/\tau_{M,i}$).

Param	Median	Mean	Maximum	Minimum
λ_1^{-1}	$5.52 \cdot 10^1$	$1.71 \cdot 10^2$	$7.06 \cdot 10^3$	0.69
λ_2^{-1}	0.62	7.89	$1.70 \cdot 10^3$	$8.96 \cdot 10^{-3\star}$
σ_1^{-1}	5.05	6.68	52.1	0.91
τ_R	88	89	110	70

Table 3.1: Quantitative summary of the parameters in min (*: this value of the gene *yjiY* is biologically unrealistic ; maybe due to an error on the measure of its mRNA in Taniguchi et al. (2010))

To determine the time τ_R of replication of each gene, we have looked at the gene position. We first estimate the time at which the DNA begins its replication in the cell cycle; as the speed of replication is relatively constant, we determine the time τ_R only with the distance from the gene to the origin of DNA replication (for more details, see the appendix Section 4.A).

We still have to determine the transcription rate λ_1 and the translation rate λ_2. One can interpret the empirical average mRNA and protein concentration of the experiment (respectively μ_m and μ_p) as the global average of mRNA and protein concentrations of the model (respectively $\widehat{\mathbb{E}}[M/V]$ and $\widehat{\mathbb{E}}[P/V]$). The global averages are known through the integration, over the cell cycle, of the mean formulas of Theorem 3.3 and Theorem 3.4:

$$\widehat{\mathbb{E}}[M/V] = \frac{\lambda_1}{\sigma_1}\frac{1}{\tau_D}\int_0^{\tau_D} \frac{1}{V_0 2^{s/\tau_D}}\left(1 - \frac{e^{-(s+\tau_D-\tau_R)\sigma_1}}{2 - e^{-\tau_D\sigma_1}} + \mathbb{1}_{s\geq\tau_R}\left(1 - e^{-(s-\tau_R)\sigma_1}\right)\right) ds,$$

$$\widehat{\mathbb{E}}[P/V] = \lambda_2\frac{1}{\tau_D}\int_0^{\tau_D} \frac{1}{V_0 2^{s/\tau_D}}\left(f_1(\tau_R) + f_2(\tau_D) + f_1(\tau_R \wedge s) + \mathbb{1}_{s\geq\tau_R} f_2(s)\right) ds.$$

As a consequence, parameters λ_1 and λ_2 can be expressed as follows

$$\begin{aligned}\lambda_1 &= \sigma_1\tau_D\mu_m\left(\int_0^{\tau_D}\frac{1}{V_0 2^{s/\tau_D}}\left(1-\frac{e^{-(s+\tau_D-\tau_R)\sigma_1}}{2-e^{-\tau_D\sigma_1}}+\mathbb{1}_{s\geq\tau_R}\left(1-e^{-(s-\tau_R)\sigma_1}\right)\right)ds\right)^{-1},\\ \lambda_2 &= \tau_D\mu_p\left(\int_0^{\tau_D}\frac{1}{V_0 2^{s/\tau_D}}\left(f_1(\tau_R)+f_2(\tau_D)+f_1(\tau_R\wedge s)+\mathbb{1}_{s\geq\tau_R}f_2(s)\right)ds\right)^{-1}.\end{aligned}$$

For each gene, all the parameters can be hence determined.

3.4.5 Biological Interpretation of the Results

In this subsection we determine the variability added by the gene replication and we compare the results with the variability measured in Taniguchi et al. (2010). Unlike the previous model of Section 3.3, we can directly compute the protein variance thanks to the explicit expressions of Theorem 3.3 and Theorem 3.4. The profile during the cell cycle and the global noise of each protein concentration are then analytically computed.

3.4.5.1 Environmental State Decomposition in Profile of Protein Concentration

The analytical expressions of the mean and the variance allow to study the evolution of every protein production during the whole cell cycle. In Figure 3.3a, we take the example of the protein Adk: we show its average concentration (thick line) and its standard deviation (blue area) during the cell cycle. It appears that the mean concentration at a given time s of the cell cycle $\mathbb{E}[P_s/V(s)]$ is not constant during the cell cycle, as it was the case for the first two models. The curve of $\mathbb{E}[P_s/V(s)]$ fluctuates around 2% of the global average protein production $\widehat{\mathbb{E}}[P/V]$. Experimental measures of average protein expression during the cell cycle show similar results: the article Walker et al. (2016) for instance measures the expression of genes at different positions on the chromosome and shows a similar profile during the cell cycle and depicts a fluctuation also around 2% of the global average (see figure 1.d and figure S6.b of the article).

This profile shows that there is, in this model, two origins for the global variability $\widehat{\mathbb{V}}ar[P/V]$: one which is induced by the production mechanism itself, and the other that is due to cell cycle effect. The first one is represented in the figure by the standard deviation at any time of the cell cycle $\sqrt{\mathbb{V}\mathrm{ar}[P_s/V(s)]}$, the other is represented by the distance of $\mathbb{E}[P_s/V(s)]$ at any time s of the cell cycle around the global mean $\widehat{\mathbb{E}}[P/V]$. We can have a quantitative description of these two contributions to the variability through the notion of *environmental state decomposition* that is used in the literature (see Hilfinger and Paulsson (2011)). For any type of protein, one can decompose the global variance $\widehat{\mathbb{V}}ar[P/V]$ (as it is defined in Equation (3.16)) as:

$$\widehat{\mathbb{V}}ar[P/V] = \widehat{\mathbb{V}}ar_1[P/V] + \widehat{\mathbb{V}}ar_2[P/V]$$

with $\widehat{\mathbb{V}}ar_1[P/V]$ and $\widehat{\mathbb{V}}ar_2[P/V]$ representing the quantities

$$\widehat{\mathbb{V}}ar_1[P/V] := \frac{1}{\tau_D}\int_0^{\tau_D}\mathbb{V}\mathrm{ar}\left[\frac{P_s}{V(s)}\right]ds, \tag{3.17}$$

$$\widehat{\mathbb{V}}ar_2[P/V] := \frac{1}{\tau_D}\int_0^{\tau_D}\left(\mathbb{E}\left[\frac{P_s}{V(s)}\right]\right)^2 ds - \widehat{\mathbb{E}}[P/V]^2. \tag{3.18}$$

The two terms represent the two natures of the global variance:

$\widehat{\mathbb{V}}ar_1[P/V]$ represents the deviation of $P_s/V(s)$ around its "local mean" $\mathbb{E}[P_s/V(s)]$ for any time s of the cell cycle. This variability is the direct result of the stochastic events that occurs in the protein production mechanism: stochastic changes in the mRNA number as well as events of translation.

$\widehat{\mathbb{V}}ar_2[P/V]$ represents the impact of the cell cycle on the global variability. It takes into account the distance between $\mathbb{E}[P_s/V(s)]$ and the global average production $\widehat{\mathbb{E}}[P/V]$. This distance is deterministic and is due to periodic external events of the cell cycle (in our case, the gene replication) that change "local mean" $\mathbb{E}[P_s/V(s)]$. This term is hence interpreted as the contribution of the external influence of the cell cycle to the global variability.

Remark 3.5. *Such decomposition was not considered for the previous models. The reason is that if such decomposition is used back then, the term $\widehat{\mathbb{V}}ar_2[P/V]$ would be null (because in these models, the means $\widehat{\mathbb{E}}[P/V(s)]$ remain constant across the cell cycle). We will discuss again the use of the environmental state decomposition in these case in subsubsection 3.4.5.3.*

In the case of Figure 3.3a, the term $\widehat{\mathbb{V}}ar_1[P/V]$ of the *environmental state decomposition* is higher than $\widehat{\mathbb{V}}ar_2[P/V]$, meaning that most of the noise is explained by the protein production itself. It is the case for all genes of the set, the ratio

$$\frac{\widehat{\mathbb{V}}ar_2[P/V]}{\widehat{\mathbb{V}}ar[P/V]}$$

is very small (99% of the cell have such ratio below 2%). It confirms the previous results: the gene cycle has almost no effect on the variability, its contribution to the global variability is negligible compared to the effect of the stochastic nature of the protein synthesis mechanism and the binomial sampling at division.

3.4.5.2 Proteins with Higher Cell Cycle Effect

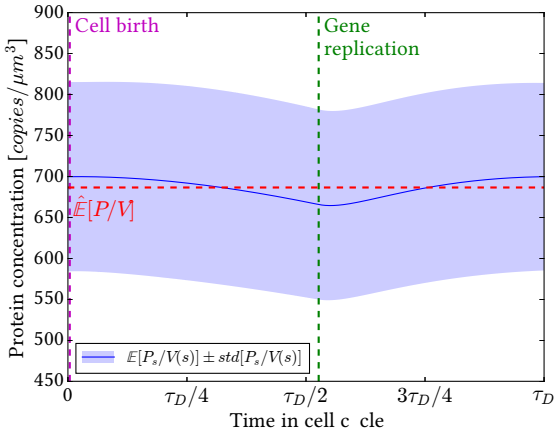
The previous result shows that for the genes considered, there is no significant contribution of the cell cycle to the protein variability with the parameters obtained through Taniguchi et al. (2010). Yet, some proteins have been proposed to have cycle-dependent concentrations and could trigger periodic events such as DNA replication initiation, or division. We investigate what range of parameters of our model would give such proteins. We show below that with our model of gene replication and division, such protein can hardly be obtained with realistic biological parameters.

In order to have such cycle-dependent proteins, one needs at least to have a reliable periodic signal: the concentration should follow a predictable path across the cell cycle, with minimal fluctuations around this path. In our case, it means that protein concentration across cell cycle $P_s/V(s)$ should be close to its mean protein production $\mathbb{E}[P_s/V(s)]$. To have so, the term $\widehat{\mathbb{V}}ar_1[P/V]$ of the environmental state decomposition should as low as possible. As we have analytical solutions for protein concentration mean and variance, we can investigate which range of parameters indeed decrease $\widehat{\mathbb{V}}ar_1[P/V]$.

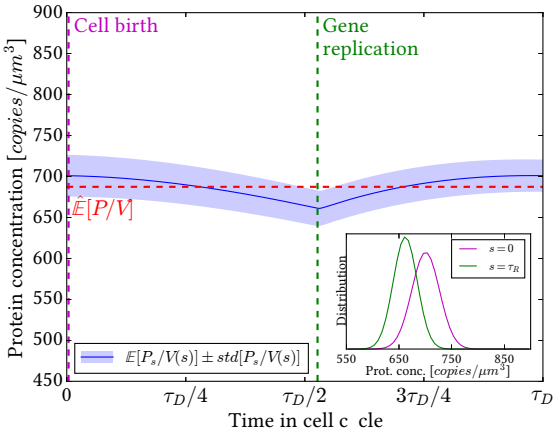
Based on the protein Adk, while keeping the global average concentration $\widehat{\mathbb{E}}[P/V]$ constant, we have analysed the following effects on the ratio $\widehat{\mathbb{V}}ar_2[P/V]/\widehat{\mathbb{V}}ar[P/V]$ (see Figure 3.3c):

- Gene position: we have changed the time of gene replication, by changing the gene position from close to the origin of replication up to the termination. We have adapted the gene activity λ_1 in order to keep the same average protein production $\widehat{\mathbb{E}}[P/V]$. Changes on this parameter make no changes in the ratio $\widehat{\mathbb{V}}ar_1[P/V]/\widehat{\mathbb{V}}ar[P/V]$: the variability is still largely due to the protein production mechanism and not the cell cycle.

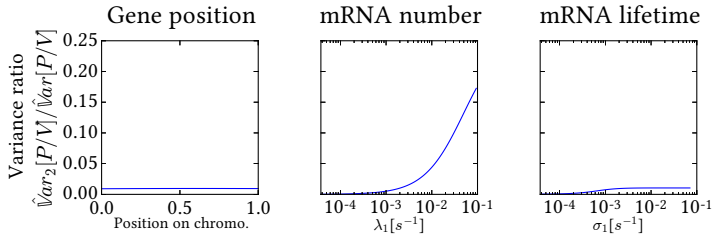
- mRNA number: we have increased the mRNA number by increasing the gene activity λ_1, while decreasing the mRNA activity λ_2 in order to keep the average $\widehat{\mathbb{E}}[P/V]$ constant. It appears that a high mRNA number indeed decreases the ratio: 50 times more mRNAs can gain give a a profile where 20% of the variance is due to the term $\widehat{\mathbb{V}}ar_2[P/V]$.



(a) Adk concentration profile



(b) Protein concentration profile with higher $\widehat{\mathbb{V}}ar_2\left[P/V\right]$



(c) Ratio $\widehat{\mathbb{V}}ar_2\left[P/V\right]/\widehat{\mathbb{V}}ar\left[P/V\right]$ for different parameters

Figure 3.3: Protein profile. **(a)**: Protein profile concentration of Adk. The mean concentration $\mathbb{E}\left[P_s/V(s)\right]$ is not constant across cell cycle and fluctuates across the global average $\widehat{\mathbb{E}}\left[P/V\right]$ (in red). The large standard deviation of $P_s/V(s)$ (coloured area) indicates a large term $\widehat{\mathbb{V}}ar_1\left[P_i/V\right]$ in the environmental state decomposition. **(b)**: profile of a modified version of AdK. In this version, there is a higher number of mRNAs (approximately ten times more) that last less time. The effect is a higher term $\widehat{\mathbb{V}}ar_2\left[P_i/V\right]$ in the environmental state decomposition (main figure), but the it is not enough to clearly separate between the distributions at birth (at time $s = 0$) and at the replication of the gene (at time $s = \tau_R$) (inset). **(c)**: Show the ratio $\widehat{\mathbb{V}}ar_2\left[P/V\right]/\widehat{\mathbb{V}}ar\left[P/V\right]$ while varying successively the gene position, the mRNA number and the mRNA lifetime while keeping $\widehat{\mathbb{E}}\left[P/V\right]$ constant.

- mRNAs lifetime: we have increased the mRNA degradation rate σ_1 while increasing mRNA activity λ_2 in order to keep the average $\widehat{\mathbb{E}}[P/V]$ constant. The ratio $\widehat{\mathbb{V}}ar_2[P/V]/\widehat{\mathbb{V}}ar[P/V]$ change of few percents but the effect has much less impact on the outcome compared to the mRNA number.

It appears that only a higher mRNA number, and to a lesser extent, a lower mRNA lifetime can increase the ratio $\widehat{\mathbb{V}}ar_2[P/V]/\widehat{\mathbb{V}}ar[P/V]$. The protein production of such protein is shown in Figure 3.3b: this protein is based on Adk but with approximately ten times more mRNAs that last ten times shorter (we also diminished the mRNA activity rate λ_2 in order to keep the same average protein production $\widehat{\mathbb{E}}[P/V]$). It represents around one transcription every 4 seconds (which is among the speediest transcription rates). Even if the profile is more gathered around the mean concentration $\mathbb{E}[P_s/V(s)]$ curve, it is still not providing a reliable enough signal of protein concentration. Indeed, as we consider the protein concentration at time 0 and at time τ_R (times where the distribution are the most distant from each other), the two distributions are still greatly overlapping (see inset of Figure 3.3b).

This part shows that, with biologically relevant parameters, it is not difficult to have a cycle-dependent protein with reliable enough signal to be able to trigger periodic events. As our model only represents gene replication and division, it is possible that other mechanisms, such as complex formation, feedback or proteolysis might give a more precise signal. In all cases, these observations support the previous results: gene replication seems to play a limited role in the protein variability.

3.4.5.3 Environmental State Decomposition and Intrinsic/Extrinsic Decomposition

The introduction of the environmental state decomposition in this section, largely used in literature, brings us to the following comment as it often used as a way to distinguish the extrinsic to the intrinsic noise. We show here that this decomposition does not separate exactly what is usually considered as extrinsic noise from the intrinsic noise.

The environmental state decomposition is used in literature (Swain et al., 2002, Elowitz et al., 2002, Hilfinger and Paulsson, 2011) as a way to decompose the two natures of the protein variability: the intrinsic noise due to the stochastic nature of birth and death of mRNAs and proteins and the extrinsic due to randomising external effect from the biological environment. In our case, the environmental state decomposition states that the average global variance of a protein $\widehat{\mathbb{V}}ar[P/V]$ is the sum of the two terms $\widehat{\mathbb{V}}ar_1[P/V]$ and $\widehat{\mathbb{V}}ar_2[P/V]$ respectively defined by Equation (3.17) and Equation (3.18). With the usual interpretation of literature, $\widehat{\mathbb{V}}ar_1[P/V]$ would be interpreted as the "intrinsic variance" and $\widehat{\mathbb{V}}ar_2[P/V]$ as the "extrinsic variance".

But it is noticeable that the second term only captures a part of what is generally accepted as the extrinsic noise. The binomial sampling (studied in Subsection 3.3.6) for instance, is not directly due to the protein production mechanism and can hence naturally be considered as having an external effect on the protein noise. And yet, the additional variance of this mechanism is not added in the second term $\widehat{\mathbb{V}}ar_2[P/V]$: indeed the binomial division has no effect on $\mathbb{E}[P_s/V(s)]$, it only affects the variance $\mathbb{V}\mathrm{ar}[P_s/V(s)]$. As a consequence, by definition of $\widehat{\mathbb{V}}ar_1[P/V]$ and $\widehat{\mathbb{V}}ar_2[P/V]$, the binomial division only increases the first term in the environmental state decomposition.

In the model of the present section, the gene replication is only external effect to be separate in the environmental state decomposition as it is the only mechanism that makes the mean $\mathbb{E}[P_s/V(s)]$ change across the cell cycle.

3.4.5.4 Effect of the Population Distribution

As noticed, in Theorem 3.2, the definitions of $\widehat{\mathbb{E}}[P/V]$ and $\widehat{\mathbb{V}}ar[P/V]$ implicitly represent the mean and the variance of protein concentration in a population of cells whose ages are uniformly distributed. In real

experimental populations of cells, like in Taniguchi et al. (2010), the number of cells in the population is exponentially growing: every dividing cell gives birth to two daughter cells. The distribution of ages is therefore not uniform.

To take into account this effect, one needs to correct the definitions of $\widehat{\mathbb{E}}[P/V]$ and $\widehat{\mathbb{V}}ar[P/V]$ by weighting them according to a typical age distribution of exponentially growing populations. Let's consider ν the distribution in age of such population (i.e., the probability that the age of the cell is between s and $s+ds$ is given by $\nu(ds)$), then we can propose new definitions for the global mean and variance:

$$\begin{aligned}\widehat{\mathbb{E}}[P/V] &:= \int_0^{\tau_D} \mathbb{E}\left[\frac{P_s}{V(s)}\right]\nu(ds),\\ \widehat{\mathbb{V}}ar[P/V] &:= \int_0^{\tau_D}\left[\mathbb{E}\left[\left(\frac{P_s}{V(s)}\right)^2\right]-\widehat{\mathbb{E}}[P/V]^2\right]\nu(ds).\end{aligned}$$

The distributions of age and length in exponentially growing populations have been studied in the literature (see Collins and Richmond (1962), Sharpe et al. (1998), Robert et al. (2014) for instance), and we have deduced the typical age distribution ν from it.

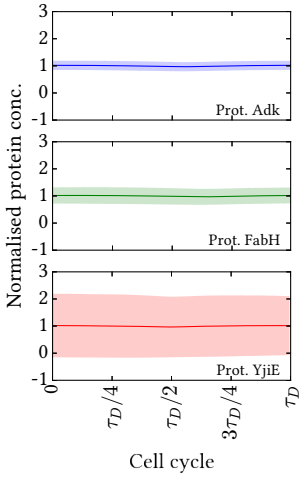
By remaking the analyses using these definitions, we compute, for every gene, the variance in the case an uniform population divided by the one obtained in an exponentially growing population. This ratio is centred around 1 with a standard deviation of $8\cdot 10^{-3}$. The distribution considered has therefore not significant impact on the variance of the model. This is due to the fact that the mean concentration of every gene remains approximately constant during the cell cycle, there is therefore no difference of protein concentration dosage at the beginning or at the end of the cell cycle.

As this effect seems negligible, for further models, we will still consider the mean and the variance as uniformly averaged over the cell cycle when we will have to estimate the parameters.

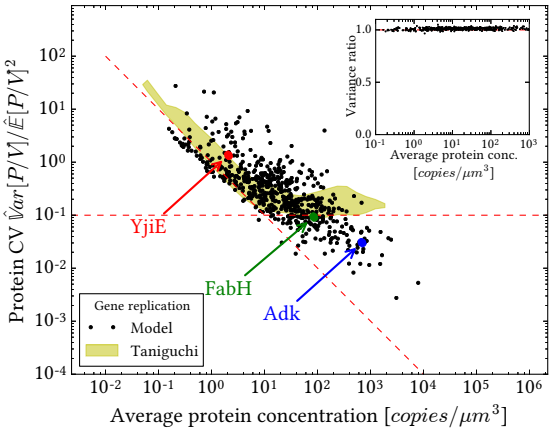
3.4.5.5 Comparison with experimental results

The profile during the cell cycle and the global noise of each protein concentration are analytically computed and can be compared with the previous models and Taniguchi et al. (2010) measures. As previously said, the gene replication has little contribution to the protein noise; therefore it is not surprising that little changes appears in Figure 3.4a compared with the previous models. Figure 3.4a shows the profile of the three representative genes : *yjiE*, *adk*, *fabH* (the non-normalised profile of *adk* was shown in Figure 3.3a). These three genes are respectively close, distant and opposite to the origin of replication (besides having very different average productions). The mean concentration seems still having less relative variability for the highest produced protein. Even if we have shown that the mean concentration of each protein (thick lines) changes across the cell cycle, it is only in barely perceptible proportions.

Figure 3.4b shows, for each protein, the global coefficient of variation of protein concentration (defined as $\widehat{\mathbb{V}}ar[P/V]/\widehat{\mathbb{E}}[P/V]^2$) as a function of the average concentration. The global tendency is still approximately inversely scaling the average protein concentration, and there is still no lower bound limit as it is the case in Taniguchi et al. (2010) experiment. It is confirmed with the inset of Figure 3.4b where the ratio between the variances for the protein concentration of this model is compared with the model with binomial sampling of Figure 3.5a. This ratio is always around 1, indicating that the replication does not contribute significantly to the global variance.



(a) Normalised protein concentration profile.



(b) CV of protein concentrations

Figure 3.4: Result of simulations of the model with cell replication, compared with Taniguchi et al. (2010) experimentation. Figure 3.4a: normalised protein concentration profile over the cell cycle for three representative proteins. The thick line represents the mean concentration over the cell cycle, and the coloured area represents the standard deviation. The profile shows no significant differences with the previous models (see Figure 3.4a for the more details on Adk). Figure 3.4b: coefficient of variation (CV) of the protein concentration as a function of the average protein concentration. There are, once again, no significant differences as before (see Figure 3.4b). In particular, there is no lower "extrinsic" noise plateau for highly expressed proteins, as it is the case for Taniguchi et al. (2010) (yellow area). Inset: ratio between the CV of protein concentration of the model with binomial sampling (Subsection 3.3.6) divided by the one of this section. Gene replication adds no particular noise compared with the model with only binomial sampling.

3.5 Conclusions on Models with Cell Cycle

In this chapter, we have produced a series of models that have taken into account the cell cycle, with a growing volume and division. Unlike classical models, this feature permits a quantitative comparison between theoretical models and experiments in terms of variability.

It clearly appears in this chapter analysis that the main contributor to the noise of protein expression so far is the protein production mechanism itself, what literature refers as the "intrinsic noise". The only significant external effect for the noise of some protein expressions the binomial segregation at division. The effect of gene replication was studied in the last section. The analytical expressions for the mean and the variance for this last model allow direct comparisons with experimental measures without simulation. It clearly appears that the variability added by the gene replication represents only a very little proportion of the global variability. Proteins with a significant proportion of the variability due to gene replication are difficult to obtain in a biological context. We have also determined in the environmental state decomposition the binomial division effect on the variance is not separate from the term usually attributed to intrinsic variability.

Some aspects of the protein production process are not considered neglected in these models. In particular, like classical models they "gene-centred": the production of a type of protein has no influence on the others. In reality, it is not the case. To produce mRNAs, genes sequester an RNA-polymerases during the elongation; and

so does the mRNAs with ribosomes to produce proteins. Both RNA-polymerases and ribosomes are common resources shared among all genes; and fluctuations of these quantities may have repercussions on protein variability. In the next chapter will propose a model production of all types of proteins that takes into account this aspect.

3.A Appendix: Gillespie Algorithms for Non-Homogenous Process

Gillespie (1977) describes an algorithm to simulate stochastic trajectories such as the quantities of different chemical species interacting together. The main idea is to consider the state of a system (for instance the number of each chemical compounds) and to compute the first reaction to occur, as well as the time when it happens. Once both pieces of information computed, one change the current state of the system accordingly with the reaction, and update the time.

One important hypothesis is that all reactions occur at exponential times (even if the rates of these exponential times may depend on the current state of the system). In this work, we have encountered a case where it was not the case. In the model with cell cycle of Subsection 3.3.3, at any time s, the state is described by (M_s, P_s) (respectively, the number of mRNAs and proteins), and the rate of mRNA production is $\Lambda(s) = \lambda_1 V(s)$ with λ_1 a parameter and $V(s)$ the non-constant volume of the cell. The parameter$\Lambda(s)$ does not depend on the state (M_s, P_s) but is time dependent through $V(s)$; for this reason, it is not an exponential time.

In this case, the time laps T until the next mRNA production is characterised by

$$\forall x > 0, \quad \mathbb{P}\left[T > x\right] = \exp\left(-\int_0^x \Lambda(s)\, ds\right)$$

which is not a exponential distribution as Λ is non-constant. To compute T, let's consider that $\Lambda(s)$ is strictly positive for any $s \in \mathbb{R}_+$, as a consequence $F(x) := \int_0^x \Lambda(s)\, ds$ is strictly increasing. Let's sample the exponential random variable E of parameter 1. We have hence

$$\forall y > 0 \quad \mathbb{P}\left[E > y\right] = \exp\left(-y\right).$$

If we consider the case of $y = F(x)$, since F is strictly increasing, it comes

$$\mathbb{P}\left[E > y\right] = \exp\left(-F(x)\right) \qquad \text{and} \qquad \mathbb{P}\left[E > y\right] = \mathbb{P}\left[F^{-1}(E) > x\right].$$

As a consequence the random variable $F^{-1}(E)$ has the same distribution as T.

Based on that we can propose a new version of the algorithm of Gillespie that can take into account non-exponential times such as T:

Algorithm 1. *The equivalent of Gillespie algorithm that considers non-homogenous events is*

1. *Initialisation: Initialise time of molecules in the system and the time.*
2. *Next exponential event: determine the next event that occurs at exponential time as in Gillespie algorithm.*
3. *Next non-homogeneous event: determine the next event that occurs at non-homogenous rates with the method previously described.*
4. *Update: choose between events of Step 2 or Step 3 that happen first. Update the time and the molecule count accordingly.*
5. *Iterate: Consider again the Step 2 unless it has reached the end of the simulation.*

3.B Appendix: Means, Variances and Covariances of (M_0, P_0) and (M_{τ_R}, P_{τ_R})

The Subsection 3.4.3 was considering protein number the model with cell cycle and gene replication. The subsection ends up to the Theorem 3.4 that gives expressions for the mean and the variance of P_s for any time s of the cell cycle depending on $\mathbb{E}[P_0]$, $\mathbb{V}\text{ar}[P_0]$, $\mathbb{C}\text{ov}[P_0, M_0]$ and $\mathbb{C}\text{ov}[P_{\tau_R}, M_{\tau_R}]$.

In this appendix, we propose to give explicit expression for these quantities if we are at equilibirum. Between times τ_D- and τ_D, the proteins undergo a binomial segregation, and since the system is at equilibrium, the distribution of the number of proteins after division P_{τ_D} is the same as the distribution of proteins at birth P_0. As a consequence:

$$\sum_{i=1}^{P_{\tau_D-}} B_{i,1/2} \overset{\mathcal{D}}{=} P_0$$

with $\left(B_{i,1/2}\right)$ being independent Bernoulli random variables of parameter $1/2$ and being all independent of P_{τ_D-} . It comes the following Lemma:

Lemma 3.1. *The mean and the variance of P_0 depend on the mean and the variance of P_{τ_D-} such as:*

$$\begin{aligned}\mathbb{E}[P_{\tau_D-}] &= 2\mathbb{E}[P_0] \\ \mathbb{V}\text{ar}[P_{\tau_D-}] &= 4\mathbb{V}\text{ar}[P_0] - 2\mathbb{E}[P_0]\,.\end{aligned}$$

Proof. With the moment-generating function of P_0, it comes that

$$\mathbb{E}[\exp[\xi P_0]] = \mathbb{E}\left[\prod_{1=1}^{P_{\tau_D-}} \mathbb{E}\left[\exp\left[B_{i,1/2}\right]\right]\right] = \mathbb{E}\left[\left(\frac{1+e^\xi}{2}\right)^{P_{\tau_D-}}\right] = \mathbb{E}\left[\exp\left[\log\left(\frac{1+e^\xi}{2}\right) P_{\tau_D-}\right]\right]$$

As a consequence, by calling $\eta(\xi) := \mathbb{E}[\exp[\xi P_{\tau_D-}]]$ the moment generating function of P_{τ_D-}, it follows:

$$\begin{aligned}\frac{d}{d\xi}\mathbb{E}[\exp[\xi P_0]] &= \frac{e^\xi}{1+e^\xi}\cdot\eta'\left(\log\left(\frac{1+e^\xi}{2}\right)\right) \\ \frac{d^2}{d\xi^2}\mathbb{E}[\exp[\xi P_0]] &= \frac{e^\xi}{(1+e^\xi)^2}\cdot\eta'\left(\log\left(\frac{1+e^\xi}{2}\right)\right) + \left(\frac{e^\xi}{1+e^\xi}\right)^2\cdot\eta''\left(\log\left(\frac{1+e^\xi}{2}\right)\right).\end{aligned}$$

As ξ goes to 0, one finds:

$$\begin{aligned}\mathbb{E}[P_0] &= \frac{\mathbb{E}[P_{\tau_D-}]}{2} \\ \mathbb{E}\left[P_0^2\right] &= \frac{1}{4}\cdot\mathbb{E}[P_{\tau_D-}] + \frac{1}{4}\cdot\mathbb{E}\left[P_{\tau_D-}^2\right].\end{aligned}$$

Hence comes the result. □

Using this Lemma allows to determine the mean and the variance of P_0:

Proposition 3.9. *At equilibrium, the mean and the variance of the protein number at birth are:*

$$\begin{aligned}\mathbb{E}[P_0] &= \lambda_2\left(f_1(\tau_R) + f_2(\tau_D)\right), \\ \mathbb{V}\text{ar}[P_0] &= \frac{1}{3}\Big\{2\mathbb{E}[P_0] + 2\frac{\lambda_2}{\sigma_1}\left[\left(1-e^{-\sigma_1\tau_R}\right)\mathbb{C}\text{ov}[P_0, M_0] + \left(1-e^{-\sigma_1(\tau_D-\tau_R)}\right)\mathbb{C}\text{ov}[P_{\tau_R}, M_{\tau_R}]\right] \\ &\qquad + g_1(\tau_R) + g_2(\tau_D)\Big\}\end{aligned}$$

with f_1, f_2, g_1 and g_2 defined in Proposition 3.6 and Proposition 3.8.

Proof. By considering the expressions of Proposition 3.8 for $s = \tau_D-$, it comes:

$$\begin{aligned}\mathbb{E}\left[P_{\tau_D-}\right] &= \mathbb{E}\left[P_{\tau_R}\right] + \lambda_2 f_2\left(\tau_D\right), \\ \mathbb{V}\mathrm{ar}\left[P_{\tau_D-}\right] &= \mathbb{V}\mathrm{ar}\left[P_{\tau_R}\right] + 2\lambda_2 \frac{1-e^{-\sigma_1(\tau_D-\tau_R)}}{\sigma_1}\mathbb{C}\mathrm{ov}\left[P_{\tau_R}, M_{\tau_R}\right] + g_2\left(\tau_D\right).\end{aligned}$$

By Equation (3.13) and Equation (3.14), we have that

$$\begin{aligned}\mathbb{E}\left[P_{\tau_D-}\right] &= \mathbb{E}\left[P_0\right] + \lambda_2\left(f_1\left(\tau_R\right) + f_2\left(\tau_D\right)\right), \\ \mathbb{V}\mathrm{ar}\left[P_{\tau_D-}\right] &= \mathbb{V}\mathrm{ar}\left[P_0\right] + 2\frac{\lambda_2}{\sigma_1}\left[\left(1-e^{-\sigma_1\tau_R}\right)\mathbb{C}\mathrm{ov}\left[P_0, M_0\right] + \left(1-e^{-\sigma_1(\tau_D-\tau_R)}\right)\mathbb{C}\mathrm{ov}\left[P_{\tau_R}, M_{\tau_R}\right]\right] \\ &\quad + g_1\left(\tau_R\right) + g_2\left(\tau_D\right).\end{aligned}$$

Lemma 3.1 describes the effect of the binomial sampling between τ_D- and τ_D on the mean and the variance of P. Since, we are at equilibrium of the cell cycles, it comes that

$$\begin{aligned}\mathbb{E}\left[P_0\right] &= \lambda_2\left(f_1\left(\tau_R\right) + f_2\left(\tau_D\right)\right) \\ 3\mathbb{V}\mathrm{ar}\left[P_0\right] &= 2\mathbb{E}\left[P_0\right] + 2\frac{\lambda_2}{\sigma_1}\left[\left(1-e^{-\sigma_1\tau_R}\right)\mathbb{C}\mathrm{ov}\left[P_0, M_0\right] + \left(1-e^{-\sigma_1(\tau_D-\tau_R)}\right)\mathbb{C}\mathrm{ov}\left[P_{\tau_R}, M_{\tau_R}\right]\right] \\ &\quad + g_1\left(\tau_R\right) + g_2\left(\tau_D\right).\end{aligned}$$

□

The only two remaining quantities to determine are $\mathbb{C}\mathrm{ov}\left[P_0, M_0\right]$ and $\mathbb{C}\mathrm{ov}\left[P_{\tau_R}, M_{\tau_R}\right]$.

Proposition 3.10. *Let's define the functions*

$$\begin{aligned}k_1\left(s\right) := &\frac{\lambda_1\lambda_2}{\sigma_1^2}\mathbb{E}\left[M_0\right]e^{-s\sigma_1}\left(s\sigma_1 - \left(1-e^{-\sigma_1 s}\right)\right) \\ &+\frac{\lambda_1}{\sigma_1}\mathbb{E}\left[P_0\right]\left(1-e^{-s\sigma_1}\right) \\ &+\frac{\lambda_1\lambda_2}{\sigma_1^2}\mathbb{E}\left[M_0\right]\left(1-e^{-s\sigma_1}\right)^2 \\ &+\frac{\lambda_2}{\sigma_1}e^{-s\sigma_1}\left(\left(\mathbb{E}\left[M_0^2\right] - \mathbb{E}\left[M_0\right]\right)\left(1-e^{-\sigma_1 s}\right) + \sigma_1 s\mathbb{E}\left[M_0\right]\right) \\ &+\frac{\lambda_1\lambda_2}{\sigma_1^2}\left[\frac{\lambda_1}{\sigma_1}\left(1-e^{-s\sigma_1}\right)\left(s\sigma_1 - \left(1-e^{-\sigma_1 s}\right)\right) + \left(1-e^{-\sigma_1 s}\left(s\sigma_1+1\right)\right)\right].\end{aligned}$$

and

$$\begin{aligned} k_2(s) \quad := \quad & \frac{2\lambda_1\lambda_2}{\sigma_1^2}\mathbb{E}[M_{\tau_R}]\, e^{-(s-\tau_R)\sigma_1}\left((s-\tau_R)\sigma_1 - \left(1-e^{-\sigma_1(s-\tau_R)}\right)\right) \\ & + \frac{2\lambda_1}{\sigma_1}\mathbb{E}[P_{\tau_R}]\left(1-e^{-(s-\tau_R)\sigma_1}\right) \\ & + \frac{2\lambda_1\lambda_2}{\sigma_1^2}\mathbb{E}[M_{\tau_R}]\left(1-e^{-(s-\tau_R)\sigma_1}\right)^2 \\ & + \frac{\lambda_2}{\sigma_1}e^{-(s-\tau_R)\sigma_1}\left(\left(\mathbb{E}\left[M_{\tau_R}^2\right]-\mathbb{E}[M_{\tau_R}]\right)\left(1-e^{-\sigma_1(s-\tau_R)}\right)+\sigma_1(s-\tau_R)\,\mathbb{E}[M_{\tau_R}]\right) \\ & + \frac{2\lambda_1\lambda_2}{\sigma_1^2}\left[\frac{2\lambda_1}{\sigma_1}\left(1-e^{-(s-\tau_R)\sigma_1}\right)\left((s-\tau_R)\sigma_1-\left(1-e^{-\sigma_1(s-\tau_R)}\right)\right)\right. \\ & \qquad \left. + \left(1-e^{-\sigma_1(s-\tau_R)}\left((s-\tau_R)\sigma_1+1\right)\right)\right]. \end{aligned}$$

Then comes the covariances:

$$\mathbb{C}\mathrm{ov}[M_0, P_0] = \frac{1}{(4-e^{-\tau_D\sigma_1})}\left\{k_1(\tau_R)\,e^{-(\tau_D-\tau_R)\sigma_1}+k_2(\tau_D)\right\} - \mathbb{E}[M_0]\,\mathbb{E}[P_0]$$

and

$$\mathbb{C}\mathrm{ov}[M_{\tau_R}, P_{\tau_R}] = \left(\mathbb{C}\mathrm{ov}[M_0, P_0] + \mathbb{E}[M_0]\,\mathbb{E}[P_0]\right)e^{-\tau_R\sigma_1} + k_1(\tau_R) - \mathbb{E}[M_{\tau_R}]\,\mathbb{E}[P_{\tau_R}].$$

Proof. Let's first determine $\mathbb{E}[P_s M_s]$ for any time $0 < s < \tau_R$ before replication. For this proof, we consider another description of the protein production that the one proposed in Equation (3.12). We consider three groups of proteins :

- Proteins present at birth.
- Proteins created during the cell cycle by mRNAs present at birth. Each of the M_0 mRNAs present at birth is able to produce proteins at rate λ_2 until its degradation that occurs in an exponential time of parameter σ_1.
- Proteins created during the cell cycle by mRNAs also created during the cell cycle. Each of the $\mathcal{N}([0,s[\times\mathbb{R}_+)$ mRNAs created since birth is able to create proteins at rate λ_2 during its existence that lasts an exponential time of parameter σ_1.

The protein number hence decomposed can be written as

$$P_s \quad = \quad P_0 + \sum_{i=0}^{M_0}\mathcal{N}_{\lambda_2}^{0,i}\left([0,\theta^{0,i}\wedge s[\right) + \sum_{i=1}^{\mathcal{N}([0,s[\times\mathbb{R}_+)}\mathcal{N}_{\lambda_2}^{1,i}\left([0,\theta^{1,i}\wedge\left(s-t^{1,i}\right)[\right).$$

with for any $i\in\mathbb{N}$ and $l\in\{0,1\}$, $t^{l,i}$ being the mRNA birth time (they are uniformly distributed in $[0,\tau_R[$), $\theta^{l,i}$ being the lifetime of mRNAs ($\theta^{l,i}\sim\mathcal{E}(\sigma_1)$), and $\mathcal{N}_{\lambda_2}^{l,i}$ denote Point Poisson Process of parameter λ_2.

Furthermore, we recall the process M_s as described in Equation (3.11) for $s<\tau_R$:

$$M_s = \sum_{i=1}^{M_0}\mathbb{1}_{\{E_{\sigma_1}^i > s\}} + \mathcal{N}(\Delta_s).$$

When making the mean of the product $M_s P_s$, there is three crossed terms that are the product of independent variables:

$$\begin{aligned} A &= \sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i} > s\}} \times \sum_{i=1}^{\mathcal{N}([0,s[\times\mathbb{R}_+)} \mathcal{N}_{\lambda_2}^{1,i}\left([0, \theta^{1,i} \wedge \left(s - t^{1,i}\right)[\right), \\ B &= \mathcal{N}(\Delta_s) \times P_0, \\ C &= \mathcal{N}(\Delta_s) \times \sum_{i=0}^{M_0} \mathcal{N}_{\lambda_2}^{0,i}\left([0, \theta^{0,i} \wedge s[\right); \end{aligned}$$

and three crossed terms that are the product of non-independent random variables:

$$\begin{aligned} D &= \sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i} > s\}} \times P_0, \\ E &= \sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i} > s\}} \times \sum_{i=0}^{M_0} \mathcal{N}_{\lambda_2}^{0,i}\left([0, \theta^{0,i} \wedge s[\right), \\ F &= \mathcal{N}(\Delta_s) \times \sum_{i=1}^{\mathcal{N}([0,s[\times\mathbb{R}_+)} \mathcal{N}_{\lambda_2}^{1,i}\left([0, \theta^{1,i} \wedge \left(s - t^{1,i}\right)[\right). \end{aligned}$$

The crossed terms with independent variable are calculated with the mean of each term which does not present any difficulties. It gives

$$\begin{aligned} \mathbb{E}[A] &= \lambda_2 \frac{\lambda_1}{\sigma_1} \mathbb{E}[M_0]\, e^{-s\sigma_1}\left(s - \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right), \\ \mathbb{E}[B] &= \frac{\lambda_1}{\sigma_1} \mathbb{E}[P_0]\left(1 - e^{-s\sigma_1}\right), \\ \mathbb{E}[C] &= \frac{\lambda_1 \lambda_2}{\sigma_1^2} \mathbb{E}[M_0]\left(1 - e^{-s\sigma_1}\right)^2. \end{aligned}$$

Let's now consider the three remaining terms. The term D gives

$$\begin{aligned} \mathbb{E}[D] &= \mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{E}\left[\mathbb{1}_{\{\theta^{0,i} > s\}} | (M_0, P_0)\right] \times P_0\right] \\ &= \mathbb{E}[M_0 P_0]\, e^{-s\sigma_1}. \end{aligned}$$

The term E gives:

$$\begin{aligned}
\mathbb{E}[E] &= \mathbb{E}\left[\sum_{i=1}^{M_0}\sum_{j\neq i}^{M_0} \mathbb{1}_{\{\theta^{0,i}>s\}}\mathcal{N}_{\lambda_2}^{0,j}\left([0,\theta^{0,j}\wedge s[\right)\right] \\
&\quad +\mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i}>s\}}\mathcal{N}_{\lambda_2}^{0,i}\left([0,\theta^{0,i}\wedge s[\right)\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}^{M_0} \mathbb{E}\left[\mathbb{1}_{\{\theta^{0,i}>s\}}|M_0\right]\mathbb{E}\left[\mathcal{N}_{\lambda_2}^{0,j}\left([0,\theta^{0,j}\wedge s[\right)|M_0\right]\right] \\
&\quad +\mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i}>s\}}\mathbb{E}\left[\mathcal{N}_{\lambda_2}^{0,i}\left([0,\theta^{0,i}\wedge s[\right)|M_0,\theta^{0,i}\right]\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}^{M_0} e^{-s\sigma_1}\frac{\lambda_2}{\sigma_1}\left(1-e^{-\sigma_1 s}\right)\right] + \mathbb{E}\left[\sum_{i=1}^{M_0} \mathbb{1}_{\{\theta^{0,i}>s\}}\mathbb{E}\left[\mathcal{N}_{\lambda_2}^{0,i}\left([0,s[\right)|M_0,\theta^{0,i}\right]\right] \\
&= \frac{\lambda_2}{\sigma_1}\left(\mathbb{E}\left[M_0^2\right]-\mathbb{E}\left[M_0\right]\right)e^{-s\sigma_1}\left(1-e^{-\sigma_1 s}\right)+\lambda_2 s\mathbb{E}\left[M_0\right]e^{-s\sigma_1} \\
&= \frac{\lambda_2}{\sigma_1}e^{-s\sigma_1}\left[\left(\mathbb{E}\left[M_0^2\right]-\mathbb{E}\left[M_0\right]\right)\left(1-e^{-\sigma_1 s}\right)+\sigma_1 s\mathbb{E}\left[M_0\right]\right].
\end{aligned}$$

The case of the last term F is more complicated; we separate the sum on $\mathcal{N}([0,s[\times\mathbb{R}_+)$ into two sums: one on $\mathcal{N}(\Delta_s)$ and the other on $\mathcal{N}(\widetilde{\Delta_s})$ with $\widetilde{\Delta_s} := \{(x,y),\, 0<x<s,\, y<s-x\}$ (in order to have $\Delta_s \cup \widetilde{\Delta_s} = [0,s[\times\mathbb{R}_+$ and $\mathcal{N}(\Delta_s)$ independent of $\mathcal{N}(\widetilde{\Delta_s})$). Hence it follows that:

$$\begin{aligned}
\mathbb{E}[F] &= \mathbb{E}\left[\mathcal{N}(\Delta_s)\times\sum_{i=1}^{\mathcal{N}(\Delta_s)}\mathcal{N}_{\lambda_2}^{1,i}\left([0,\theta^{1,i}\wedge\left(s-t^{1,i}\right)[\right)\right] \\
&\quad +\mathbb{E}\left[\mathcal{N}(\Delta_s)\right]\times\mathbb{E}\left[\sum_{i=1}^{\mathcal{N}(\widetilde{\Delta_s})}\mathcal{N}_{\lambda_2}^{1,i}\left([0,\theta^{1,i}\wedge\left(s-t^{1,i}\right)[\right)\right] \\
&= \mathbb{E}\left[\mathcal{N}(\Delta_s)\times\sum_{i=1}^{\mathcal{N}(\Delta_s)}\mathbb{E}\left[\mathcal{N}_{\lambda_2}^{1,i}\left([0,\theta^{1,i}\wedge\left(s-t^{1,i}\right)[\right)|\theta^{1,i},t^{1,i},\mathcal{N}(\Delta_s)\right]\right] \\
&\quad +\lambda_2\frac{\lambda_1}{\sigma_1}\left(1-e^{-s\sigma_1}\right)\times\mathbb{E}\left[\sum_{i=1}^{\mathcal{N}(\widetilde{\Delta_s})}\theta^{1,i}\wedge\left(s-t^{1,i}\right)\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[F] &= \lambda_2 \mathbb{E}\left[\mathcal{N}(\Delta_s) \times \sum_{i=1}^{\mathcal{N}(\Delta_s)} \mathbb{E}\left[\theta^{1,i} \wedge \left(s - t^{1,i}\right) \mid \left(\theta^{1,i}, t^{1,i}\right) \in \Delta_s\right]\right] \\
&\quad + \lambda_2 \mathbb{E}[\mathcal{N}(\Delta_s)] \times \iint_{\widetilde{\Delta_s}} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&= \lambda_2 \mathbb{E}\left[\mathcal{N}(\Delta_s)^2\right] \frac{1}{\nu(\Delta_s)} \mathbb{E}\left[\theta^{1,i} \wedge \left(s - t^{1,i}\right) \mid \left(\theta^{1,i}, t^{1,i}\right) \in \Delta_s\right] \\
&\quad + \lambda_2 \mathbb{E}[\mathcal{N}(\Delta_s)] \times \iint_{\widetilde{\Delta_s}} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&= \lambda_2 \mathbb{E}\left[\mathcal{N}(\Delta_s)^2\right] \frac{1}{\nu(\Delta_s)} \iint_{\Delta_s} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&\quad + \lambda_2 \mathbb{E}[\mathcal{N}(\Delta_s)] \times \iint_{\widetilde{\Delta_s}} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta.
\end{aligned}$$

We know that $\mathcal{N}(\Delta_s)$ is a Poisson random variable of parameter $\nu(\Delta_s) = \lambda_1/\sigma_1 \times (1 - e^{-s\sigma_1})$, hence $\mathbb{E}[\mathcal{N}(\Delta_s)] = \nu(\Delta_s)$ and $\mathbb{E}\left[\mathcal{N}(\Delta_s)^2\right] = \nu(\Delta_s)\left(\nu(\Delta_s) + 1\right)$. As a consequence:

$$\begin{aligned}
\mathbb{E}[F] &= \lambda_2 \left(1 + \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right)\right) \iint_{\Delta_s} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&\quad + \lambda_2 \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right) \times \iint_{\widetilde{\Delta_s}} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta. \\
&= \lambda_2 \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right) \iint_{[0,s[\times\mathbb{R}_+} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&\quad + \lambda_2 \iint_{\Delta_s} \theta \wedge (s-t)\, \lambda_1 dt \otimes \sigma_1 e^{-\sigma_1 \theta} d\theta \\
&= \lambda_2 \frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right) \frac{\lambda_1}{\sigma_1}\left(s - \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right) + \lambda_2 \int_0^s \int_{s-t}^{\infty} (s-t)\ \sigma_1 e^{-\sigma_1 \theta} d\theta\, \lambda_1 dt \\
&= \lambda_2 \left(\frac{\lambda_1}{\sigma_1}\right)^2 \left(1 - e^{-s\sigma_1}\right)\left(s - \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right) + \lambda_2 \int_0^s t \int_t^{\infty} \sigma_1 e^{-\sigma_1 \theta} d\theta\, \lambda_1 dt \\
&= \lambda_2 \left(\frac{\lambda_1}{\sigma_1}\right)^2 \left(1 - e^{-s\sigma_1}\right)\left(s - \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right) + \lambda_2 \lambda_1 \int_0^s t e^{-\sigma_1 t} dt \\
&= \lambda_2 \left(\frac{\lambda_1}{\sigma_1}\right)^2 \left(1 - e^{-s\sigma_1}\right)\left(s - \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right) + \lambda_2 \frac{\lambda_1}{\sigma_1}\left(-e^{-\sigma_1 s} s + \frac{1 - e^{-\sigma_1 s}}{\sigma_1}\right) \\
&= \lambda_2 \frac{\lambda_1}{\sigma_1^2}\left[\frac{\lambda_1}{\sigma_1}\left(1 - e^{-s\sigma_1}\right)\left(s\sigma_1 - \left(1 - e^{-\sigma_1 s}\right)\right) + \left(1 - e^{-\sigma_1 s}\left(s\sigma_1 + 1\right)\right)\right]
\end{aligned}$$

so, it comes that for any s before replication that:

$$\mathbb{E}[M_s P_s] = k_1(s) + \mathbb{E}[M_0 P_0]\, e^{-s\sigma_1}. \tag{3.19}$$

Similarly for $\tau_R \leq s < \tau_D$ after replication, one can show that:

$$\mathbb{E}[M_s P_s] = k_2(s) + \mathbb{E}[M_{\tau_R} P_{\tau_R}]\, e^{-(s-\tau_R)\sigma_1}. \tag{3.20}$$

With these two relations, one can determine the expression of $\mathbb{E}\left[M_{\tau_D-}P_{\tau_D-}\right]$:

$$\begin{aligned}\mathbb{E}\left[M_{\tau_D-}P_{\tau_D-}\right] &= k_2(\tau_D) + \mathbb{E}\left[M_{\tau_R}P_{\tau_R}\right] e^{-(\tau_D-\tau_R)\sigma_1} \\ &= \mathbb{E}\left[M_0P_0\right] e^{-\tau_D\sigma_1} + k_0^1\left(\tau_R\right) e^{-(\tau_D-\tau_R)\sigma_1} + k_{\tau_R}^2\left(\tau_D\right).\end{aligned}$$

Since M_{τ_D-} and P_{τ_D-} undergo a binomial sampling between τ_D- and τ_D, and since at equilibrium $(M_0,P_0) \overset{\mathcal{D}}{=} (M_{\tau_D},P_{\tau_D})$, by considering $(B_{k,i})_{k\in\{1,2\},i\in\mathbb{N}}$ i.i.d. Bernoulli random variables of parameter $1/2$, it comes

$$\begin{aligned}\mathbb{E}\left[M_0P_0\right] &= \mathbb{E}\left[\sum_{i=0}^{M_{\tau_D-}} B_{1,i} \sum_{i=0}^{P_{\tau_D-}} B_{2,i}\right] \\ &= \mathbb{E}\left[\sum_{i=0}^{M_{\tau_D-}} \sum_{i=0}^{P_{\tau_D-}} \mathbb{E}\left[B_{1,i}B_{2,i}\middle|\left(M_{\tau_D-},P_{\tau_D-}\right)\right]\right] \\ &= \frac{1}{4}\mathbb{E}\left[M_{\tau_D-}P_{\tau_D-}\right].\end{aligned}$$

As by definition of the covariance

$$\mathbb{C}\mathrm{ov}\left[P_0,M_0\right] := \mathbb{E}\left[M_0P_0\right] - \mathbb{E}\left[M_0\right]\mathbb{E}\left[P_0\right],$$

it comes the result for $\mathbb{C}\mathrm{ov}\left[P_0,M_0\right]$. For $\mathbb{C}\mathrm{ov}\left[P_{\tau_R},M_{\tau_R}\right]$, one can simply use the expression for Equation (3.19) with $s=\tau_R$ since the quantity $\mathbb{E}\left[M_0P_0\right]$ is now known.

□

3.C Appendix: Simple Model to Predict the Effect of Binomial Sampling

In Figure 3.5b, we have considered the ratio of protein noise: the noise in the model where the division is exact, divided by the noise with the case where the division is binomial. We have figured in cyan dash line the prediction of a simplified model of such a ratio. In this appendix, we described the model used.

Let's consider a quantity P that goes through a division. The division can be performed by two means, either through exact division, or through binomial sampling (see Subsection 3.3.6). The result of these divisions will respectively denoted by P_e and P_b. During division, the volume is divided by two, changing from $2V_0$ to V_0. In order to be plotted in Figure 3.5b, we need to consider the coefficient of variation of protein concentration after division

$$\eta := \frac{\mathbb{V}\mathrm{ar}\left[P_e/V_0\right]/\mathbb{E}\left[P_e/V_0\right]^2}{\mathbb{V}\mathrm{ar}\left[P_b/V_0\right]/\mathbb{E}\left[P_b/V_0\right]^2} \quad \text{as a function of} \quad x := \frac{\mathbb{V}\mathrm{ar}\left[P/(2V_0)\right]}{\mathbb{E}\left[P/(2V_0)\right]}.$$

Proposition 3.11. *The coefficient of variation ratio η as a function of x is given by*

$$\eta = \frac{2V_0x}{2V_0x+1}.$$

Proof. We have the quantity before division P. Let's observe the effect of exact division on its concentration. Since by definition, we have that $P_e = P/2$, it comes that

$$\begin{aligned}\mathbb{E}\left[P_e/V_0\right] &= \mathbb{E}\left[P/(2V_0)\right], \\ \mathbb{V}\mathrm{ar}\left[P_e/V_0\right] &= \mathbb{V}\mathrm{ar}\left[P/(2V_0)\right].\end{aligned}$$

For the effect of binomial division, one can refer to Lemma 3.1, that describes the effect of the binomial division on the mean and the variance of any number. With it, it comes that

$$\begin{aligned}\mathbb{E}\left[P_b\right] &= \frac{\mathbb{E}\left[P\right]}{2},\\ \mathbb{V}\mathrm{ar}\left[P_b\right] &= \frac{\mathbb{V}\mathrm{ar}\left[P\right] + 2\mathbb{E}\left[P_b\right]}{4}.\end{aligned}$$

We then divide by the volume in order to observe the concentrations:

$$\mathbb{E}\left[P_b/V_0\right] = \frac{\mathbb{E}\left[P\right]}{2V_0} = \mathbb{E}\left[P/\left(2V_0\right)\right]$$

and

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[P_b/V_0\right] &= \frac{\mathbb{V}\mathrm{ar}\left[P_b\right]}{V_0^2} = \frac{\mathbb{V}\mathrm{ar}\left[P\right] + 2\mathbb{E}\left[P_b\right]}{4V_0^2}\\ &= \mathbb{V}\mathrm{ar}\left[P/\left(2V_0\right)\right] + \frac{\mathbb{E}\left[P/(2V_0)\right]}{2V_0}.\end{aligned}$$

As a consequence, it comes that

$$\eta = \frac{\mathbb{V}\mathrm{ar}\left[P_e/V_0\right]}{\mathbb{V}\mathrm{ar}\left[P_b/V_0\right]} = \frac{\mathbb{V}\mathrm{ar}\left[P/(2V_0)\right]}{\mathbb{V}\mathrm{ar}\left[P/\left(2V_0\right)\right] + \mathbb{E}\left[P/(2V_0)\right]/(2V_0)} = \frac{2V_0x}{2V_0x+1}$$

□

Chapter 4

Multi-protein Model

The previous chapter has considered models with cell cycle, that allow the comparison of the predicted variance with real experimental measures of Taniguchi et al. (2010). But it seems that the noise obtained in these models do not reproduce the protein variability, especially for highly expressed genes. It has been proposed in Taniguchi et al. (2010), and even earlier in the literature (Elowitz et al., 2002, Swain et al., 2002), that fluctuations of commonly shared resources in the protein production, such as RNA-polymerases and ribosomes (macromolecules required respectively for every transcription and translation) can add significant variability in gene expression. We describe in this chapter a model that extends the previous models with the introduction of this key feature: the limited amount of RNA-polymerases and ribosomes for the production of every protein. As the models of the previous chapter were "gene-centred", each class of proteins was considered independently from each other; the common sharing or RNA-polymerases and ribosomes advocates here for the consideration of a multi-protein model where all the genes are considered altogether.

Models that consider multi-protein production are rare in literature. Two examples are Mather et al. (2013) that describes a production of two types of proteins, and Fromion et al. (2015) that includes the translation of a large number of classes of proteins, both consider with a limited number of ribosomes available for translation. Both articles carried out mathematical analysis, but they both focus on the translation part, considering the number of mRNAs as constant, or (in the case of Mather et al. (2013)) as independently fluctuating. In reality, mRNAs production is neither constant, neither independent, as it depends on RNA-polymerase dynamic. Moreover, the models of the articles implicitly take place in a fixed volume; there is no notion growth of the cell volume, nor replication and division, which is also the case for classical models. As we have said in Subsection 3.2.2, it is difficult to quantitatively compare their results with experimental measurements in this case.

We hence propose a model that is in the direct continuation of the previous models: with cell cycle, binomial sampling and gene replication; but we also introduce the notion of limited amount of RNA-polymerases and ribosomes. We will investigate the potential impact they can have on gene expression variability. At first, in Section 4.1, the model is described in detail. As it appears that its complete mathematical analysis is complex, we will propose to examine a simplified description in Section 4.2 that helps to fix the parameters on experimental measures. In Section 4.3, we will examine the impact on the protein variance of free ribosomes and free RNA-polymerases. Globally, it will appear little additional noise in protein production compared

with the models of the previous section and that this small contribution will appear mainly due to the low number of free ribosomes and high number of RNA-polymerases. In Section 4.4 we will present the results of simulations with different different modelling choice; we will show that they have globally no apparent impact on the protein heterogeneity.

4.1 Description of the Main Model

The aim of this model is to integrate the production of all proteins of the cell and their interactions. It is therefore a model at the scale of the whole bacteria: all genes, mRNAs and proteins are considered, as well as all ribosomes and RNA-polymerases. We are interested in the intertwined effects between local units of production of a particular type of proteins, and the global behaviour of common quantities such as the number of free RNA-polymerases and ribosomes. In this section, we present the model: firstly in Subsection 4.1.1 are presented the main biological aspects included in the model, then in Subsection 4.1.2 are described in detail all the mechanisms and notation used in the model of the chapter.

4.1.1 Main Features of the Main Model

The introduction of RNA-polymerases and ribosomes has several consequences on the model: several features have to be added or changed in order to have a consistent representation of the cell. We present below these different aspects.

First of all, RNA-polymerases and ribosomes are explicitly present in the model. As these macromolecules are shared among all types of proteins, we cannot consider each gene as independent from each other as it was the case in the previous chapter. One therefore has to take into account all the different types of genes, mRNAs and proteins of the cell altogether; and the production of each type of proteins depends on the availability of RNA-polymerases and ribosomes. In the model, we will suppose that there is no notion of operons (an operon is a single mRNA strand on which several genes are coded): each gene is therefore considered as having its own specific promoter.

In the model, RNA-polymerases can be allocated to a gene or not: if it is an allocated (or sequestered), then it is specifically bound on a gene in a transcription process; if it is non-allocated (or free), then it is either moving freely in the cytoplasm or is sliding on the DNA non-specifically (the sliding on the DNA has been proposed as taking part in the kinetics of promoter binding (Kabata et al., 1993)). In the first part of the chapter, we will gather all these non-allocated RNA-polymerases into one single group of free RNA-polymerases (in Subsection 4.4.3 we will study the case where there are two separate cases for the cytoplasmic and the non-specifically bound polymerases).

As explained in Section 1.1, in order to produce an mRNA, the RNA-polymerase has to bind on the gene promoter, initiate the transcription; then elongation occurs in which the mRNA chain is polymerised; finally, the termination releases both the RNA-polymerase and the mRNA in the medium. In the model of this chapter, we separate this process in two parts: the binding and initiation on one side and the elongation and termination on the other side.

As the binding and the initiation are gathered in one single event in the model, one has to represent its rate of occurrence. The probability to bind on a specific promoter depends on the *concentration* of free RNA-polymerases: the same number of free RNA-polymerases has a lower tendency to bind on a promoter as the volume of the cell is higher. It also depends on promoter specific aspects such as its sequence affinity for RNA-polymerases or the chromosome architecture. Also, as the DNA is replicated, the gene has twice more promoters, hence increasing the occurrence this encounter event. As a consequence, the binding and initiation of an RNA-polymerase on a specific gene is gathered in the model as a single event whose rate is considered as depending on three quantities:

- the gene copy number of this gene
- the free RNA-polymerase concentration
- a parameter that takes into account the specificity of the promoter such as the promoter affinity with RNA-polymerase, the chromosome architecture or the propensity of initiation

Once the binding-initiation event previously described is finished, the mRNA production can begin. In our model, the process of elongation and the termination are considered together as a single event. During this time, the RNA-polymerase is considered as sequestered on the DNA. In real cells, the speed of elongation is relatively constant in stable environmental conditions; as a consequence, in our model, we consider that it depends only to the gene length. In the model, there is no notion of operon and each transcript contains only one gene, we choose to consider the gene length as representative of the length of the transcript. Therefore the rate of the event is considered in the model as only depending on the length of the gene.

Ribosome mechanic has a lot in common with RNA-polymerase in the model. They are also grouped into two categories: the non-allocated ribosomes (or free ribosomes) that evolve in the medium and the allocated ribosomes (or sequestered ribosomes) which are bound on mRNA involved in a translation process. The translation is considered as occurring in two steps. At first, the ribosome binds on the RBS and initiate the elongation; the rate at which this event occurs depends on the free ribosome concentration, on the mRNA copy number and on some mRNA specific parameter that takes into account aspects like RBS-ribosome affinity for instance. Secondly, once the elongation initiated, the process of elongation and termination is considered as only depending on the length of the gene.

In the models of the previous chapter, the volume was defined as an external and deterministic object: the idea was to study the behaviour of one gene immersed into a "background environment" where the cell grow and divide. One key assumption in this case is that the gene of interest has no influence on the overall bacteria dynamic. In the current model, it is not any more the case: all the genes are considered simultaneously and their production represents the production of all proteins. It is not possible to consider that the production of a single type of protein has no effect on the global performance of the cell in this model. The volume growth depends now on the global production of proteins, and not as an independent and deterministic feature.

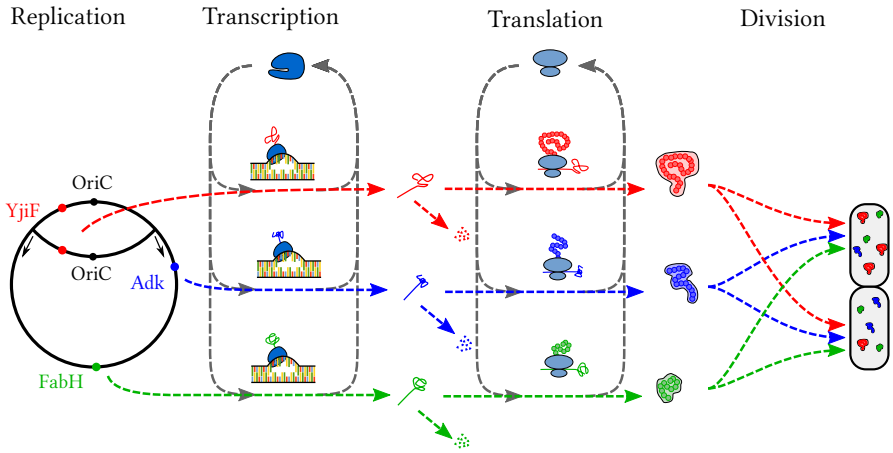
To represent the volume growth we rely on the "density constraint": it appears that the cell tends to maintain constant its density of cell components in order to have an efficient intracellular diffusion (Marr, 1991). So the total mass of compounds in the cell (proteins, metabolites, DNA, etc.) can be considered as proportional to the cell volume. Most of the dry mass of the cell is due the total amount of proteins (Neidhardt and Umbarger, 1996), we therefore consider that the volume is proportional to the total mass of proteins (each protein contributes in proportion to its own mass to the total protein mass). [1]

4.1.2 Model Presentation

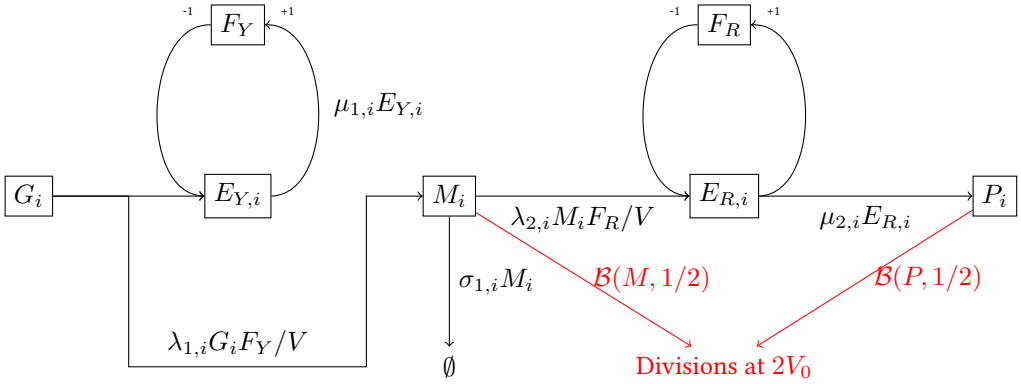
Let's present more exhaustively the mechanisms and the notations of the model. The model has some global variables such as the volume, the number of free RNA-polymerases and ribosomes; and some variables that are gene specific, like the number of mRNAs, of proteins, etc. One can refer to Figure 4.1 for an overview of the model.

Units of Production In this model, we consider K types of proteins; each protein is produced in a single production unit, with a particular type of mRNAs and a specific gene associated with. In each unit of production

[1] Ribosome components also represent an important part of the cell dry mass (Neidhardt and Umbarger, 1996). But as it will be seen in the following section, the concentration of ribosomes is considered as constant in the model; therefore, the total mass of ribosomes and proteins taken altogether is still proportional to the volume. It is still consistent with the "density constraint" hypothesis.



(a) Biological mechanism for three typical genes.



(b) Production of production unit of the i-th protein with the common pools of free RNA-polymerases and ribosomes.

Figure 4.1: Multi-protein model. **(a)** The model of this chapter considers interdependent genes through the common sharing of RNA-polymerases and ribosomes. **(b)** For i-th type of protein, the number of genes, mRNAs and proteins are respectively G_i, M_i and P_i; elongation event depends on free RNA-polymerase concentration F_Y/V and free ribosome concentrations F_R/V (see main text for more details).

$i \in \{1, ..., K\}$ we denote by $G_i(s)$, $M_i(s)$ and $P_i(s)$ respectively the number of gene copies, of messengers and of proteins at time s.

Volume Increase As previously said, the volume $V(s)$ is no longer deterministic at it was the case in the models of the previous chapter and it is considered as proportional to the current total mass of proteins in the cell. We denote by β_P represents ratio mass-volume and by w_i the mass of a type i protein. In that case, we have by definition

$$V(s) = \sum_{i=1}^{K} w_i P_i(s)/\beta_P. \tag{4.1}$$

That means that each protein of type i created increases the total volume of the cell with respect to the factor w_i/β_P.

Global Variables The total number of RNA-polymerases and ribosomes (whether allocated or not) are respectively denoted by $N_Y(s)$ and $N_R(s)$. In a first step, we consider that the both these quantities are in constant concentration, that is to say

$$N_Y(s) = \lfloor \beta_Y V(s) \rfloor \quad \text{and} \quad N_R(s) = \lfloor \beta_R V(s) \rfloor,$$

with β_Y and β_R constant parameters and where $\lfloor . \rfloor$ is the notation for the floor function. It means that as the cell grows, new RNA-polymerases and ribosomes are added to the system in the corresponding proportion. In Subsection 4.4.2 we will consider the case more complex where both RNA-polymerases and ribosomes are directly produced through a gene expression process.

At any time s, we denote by $F_Y(s)$ the random variable that represents the number of free polymerases, that is to say RNA-polymerases that are not specifically sequestered on a messenger. In the same way, let's denote by $F_R(s)$ the random variable that represents the number of free ribosomes at time s.

Reaction Rates Let's then define all the reactions that are specific to each gene. As previously said, the rate at which an RNA-polymerase binds on a specific promoter and initiate elongation depends on the concentration of free RNA-polymerases $F_Y(s)/V(s)$ and the copy number of the gene $G_i(s)$. The rate is therefore $\lambda_{1,i} G_i(s) F_Y(s)/V(s)$ where $\lambda_{1,i}$ accounts for the specificity of the promoter (its affinity for the RNA-polymerase etc.). As elongation begins, the RNA-polymerase is considered as sequestered (decreasing the number of free polymerases F_Y by one unit) on the DNA until the termination. The total number of RNA-polymerases currently elongating a messenger of type i is denoted by the random variable $E_{Y,i}(s)$. As a consequence, the total amount of RNA-polymerases N_Y is given by

$$N_Y = F_Y + \sum_{i=1}^{K} E_{Y,i}.$$

The elongation time is given by an exponential random variable of rate $\mu_{1,i}$. Once the elongation terminates, the RNA-polymerase is released in the cytoplasm (increasing the number of free RNA-polymerase F_Y by one unit). A messenger is considered created as soon as its elongation begins: the reason for it is that in bacteria (unlike eukaryotes), since transcription and translation happen in the same medium, a translation can begin on an mRNA on which the transcription is not finished. Each messenger of type i has a lifetime given by an exponential random variable of rate $\sigma_{1,i}$.

Similarly to transcription, the rate at which a ribosome encounters an mRNA of type i and initiate translation depends on the number of mRNAs $M_i(s)$ and on the ribosome concentration $F_R(s)/V(s)$. The rate

for translation initiation is therefore $\lambda_{2,i} M_i(s) F_R(s)/V(s)$ where $\lambda_{2,i}$ will account for mRNA specific aspects (RBS affinity for ribosome, etc.). The total number of ribosomes sequestered on messengers of type i is $E_{Y,i}(s)$ and each elongation time follows an exponential distribution of rate $\mu_{2,i}$. Here we consider that the protein is created after termination (since the protein is usually fully functional once its translation is completed); the number of proteins $P_i(s)$ is then increased by one unit. As in the models of the previous chapter, we do not consider protein proteolysis since it usually occurs at much longer timescale than cell cycle.

Remark 4.1. *It can be remarked that in the models of the previous chapter, the mRNA initiation rate per gene was considered as constant. It implicitly meant that we have considered that the concentration of free RNA-polymerases remains constant across the cell cycle. Similarly with the translation part, free ribosomes were also considered in constant concentrations. Now, the consequence common pool of non-allocated RNA-polymerases and ribosomes make these rates explicitly depends on these varying concentrations.*

DNA Replication and Division of the Cell At a time s, each gene $i \in \{1, ..., K\}$ is characterised by the gene copy number $G_i(s)$. As in the gene-centred model of Section 3.4, there is only one replication per gene in the cell cycle: as a consequence, the for each $i \in \{1, ..., K\}$, $G_i(s)$ is constantly equal to 1 until the gene is replicated; from this instant until the division, it is set to 2. There is two modelling choice for when the DNA replication is initiated: it can occur at a fixed time after the last division or it can or when the cell reaches a certain volume V_I. The first simulations are made by considering the volume-dependent initiation event, but as we will see in Subsection 4.4.5, simulations with the two possibilities show no noticeable difference. The volume V_I is fixed to $1.8\,\mu m^3$ (see Wallden et al. (2015) and the appendix Section 4.A about this choice). We consider the speed of DNA replication as constant; as a consequence, once known the replication time τ_I, the delay until the replication of i-th gene is fixed, and is given by the gene position.

For the division, we considered at first that, like in the previous models, the division occurs when the reaches exactly the volume $2V_0$ (with $V_0 = 1.3\,\mu m^3$ as it was the case for the previous models). We will consider in Subsection 4.4.4 the case where the division is not as precise. As in the previous models, the effect of septation is a binomial sampling of messengers and proteins: each of them as an equal chance to be in the next cell or not. The volume of the new cell is proportional to the total mass of the remaining proteins. Moreover, at division, all gene copy number are anew set to one; ribosomes and RNA-polymerases will anew set accordingly to the new volume.

The model of this section is more complex than its counterparts of the previous chapter. It is due in part to the feedback loop that proteins have on their own production: the more proteins, the more the volume increases, thereby increasing the total amount of ribosomes and hence the translation rates. This complexifies the complete analytical description of mRNA and protein mean productions, which has a detrimental impact on the search of parameters based on Taniguchi et al. (2010) experiments. The next section proposes a model that mimics the average behaviour of our stochastic model: the goal is to be able to fit parameters to real measures and use them for stochastic simulations.

4.2 Simple Deterministic Model for Protein Production

Expressions for the mean of mRNA and protein concentrations were used in the models of the previous chapter as a way to fix the model parameters based on experimental measures. But as the multi-protein model of this chapter is difficult to describe analytically, the exact mean of protein and mRNA concentrations are unknown. As a consequence, it is not possible to directly adjust the parameters to make the protein productions of the model correspond to those of Taniguchi et al. (2010).

To address this problem, we propose a representation to reflect the average behaviour of the stochastic model; for that, we use the classical framework of systems of Ordinary Differential Equations (ODEs)[2]. This framework is usually used in literature to describes the average behaviour associated to gene expression (see Borkowski et al. (2016), Goelzer et al. (2011) for instance): the evolution of mRNA and protein concentrations are described through the chemical kinetics representation. The goal is to have a relatively correct representation of the average production, so that the parameters of the stochastic model can be correctly fitted.

We will present this simplified model in Subsection 4.2.1 and describe its dynamics in Subsection 4.2.2. We will use these results to deduce a set of parameters that corresponds to Taniguchi et al. (2010) experiment as it will be explained in Subsection 4.2.3. Finally, in Subsection 4.2.4, we will validate the global correspondence between the average behaviour of the stochastic model and model presented in this section.

4.2.1 Presentation of the Deterministic Production Model

The model chosen to reflect the average behaviour of the stochastic model previously described is a system of ordinary differential equations (ODEs) that describes the kinetics of each compound concentration of the system.

We still consider K genes, each of them associate with a particular type of mRNA and protein. For a gene of type i, the concentrations of gene copies is given by $g_i(s)$; mRNAs and protein concentrations are denoted by $m_i(s)$ and $p_i(s)$. Similarly, $f_Y(s)$ and $f_R(s)$ respectively represent the concentrations of free RNA-polymerases and free ribosomes; while $e_{Y,i}(s)$ and $e_{R,i}(s)$ denote the concentrations of RNA-polymerases and ribosomes currently sequestered in the i-th protein production unit. It is important tot note that, contrary to the stochastic model of the previous section, all these quantities correspond to concentration and not numbers of entities (their stochastic counterparts would be the concentrations $G_i(s)/V(s)$, $M_i(s)/V(s)$, $P_i(s)/V(s)$, etc.).

The reactions between different compounds are given by the *law of mass action*, that is to say that the rate of chemical reaction is proportional to the reactant abundance and their activities. We are interested for instance in the evolution of m_i that denotes the concentration of mRNAs of type i . The creation of a type i mRNA is the result of a reaction between a free RNA-polymerase (whose concentration is $f_Y(s)$) and the gene i (whose concentration is $g_i(s)$); $\lambda_{1,i}$ is interpreted as the affinity constant of the reaction. The type i mRNA degradation is the result of a reaction that occurs at rate $\sigma_{1,i}$.

As in the usual description of the cell (see Goelzer et al. (2011) for instance), one also must consider the dilution: without any molecule creation, the concentration of the compound still decreases as the cell grows due to dilution. If we consider that the cell is growing exponentially, doubling of volume in a time τ_D, then the rate of dilution is $\log 2/\tau_D$. The exponential growth corresponds to the volume dynamics of real bacteria (Wang et al., 2010), and we will see in Subsection 4.2.4 that it is a good approximation of the growth of cells in stochastic simulations.

All these aspects considered altogether, the kinetics of the concentration of mRNAs of type i is given by the ODE:

$$\frac{dm_i}{ds}(s) = \lambda_{1,i} g_i(s) \cdot f_Y(s) - \sigma_{1,i} m_i(s) - \frac{\log 2}{\tau_D} \cdot m_i(s). \tag{4.2}$$

The first term represents the mRNA creation; the second, the mRNA degradation; and the last, the dilution.

Similarly, for the other reactions, it comes for $i \in \{1, ..., K\}$, at any time s:

[2]The model presented here would rather be a fluid limit model. Since the stochastic model is non-linear it is theoretically not corresponding to the average production of the stochastic model. Nonetheless, we will see in Subsection 4.2.4 that it is still a good prediction of the behaviour of the average protein production.

$$\frac{dp_i}{ds}(s) = \mu_{2,i} e_{R,i}(s) - \frac{\log 2}{\tau_D} \cdot p_i(s), \tag{4.3}$$

$$\frac{de_{Y,i}}{ds}(s) = \lambda_{1,i} g_i(s) \cdot f_Y(s) - \mu_{1,i} e_{Y,i}(s) - \frac{\log 2}{\tau_D} \cdot e_{Y,i}(s), \tag{4.4}$$

$$\frac{de_{R,i}}{ds}(s) = \lambda_{2,i} m_i(s) \cdot f_R(s) - \mu_{2,i} e_{R,i}(s) - \frac{\log 2}{\tau_D} \cdot e_{R,i}(s). \tag{4.5}$$

Similarly to the stochastic model of the previous section, we consider that the RNA-polymerases (whether allocated or not) are still considered in constant concentration β_Y. It means that

$$\beta_Y = f_Y(s) + \sum_{i=1}^{K} e_{Y,i}(s), \tag{4.6}$$

since $\sum_i e_{Y,i}$ and f_Y represent the concentrations of respectively the allocated and non-allocated RNA-polymerases. It is similar to the ribosomes as we have:

$$\beta_R = f_R(s) + \sum_{i=1}^{K} e_{R,i}(s). \tag{4.7}$$

The classical strategy in literature to study such system (an analogue model is done in Borkowski et al. (2016)) is to consider the system in steady state growth: the gene concentration g_i is considered as constantly equal to its average value during the cell cycle, and then one can calculate the concentrations of m_i, p_i, $e_{Y,i}$ and $e_{R,i}$ at steady-state by writing the equation Equations (4.2) to (4.7) with the derivative term as null. We have tried such methods to determine the concentrations. But, even if the protein concentrations then predict are around the stochastic simulation; they are not precise enough: there is a clear shift between the stochastic protein concentration and the one predicted by this method. In fact, it would be a good approximation if the gene concentration is constant during the cell cycle, as it was the case in the first model of the last chapter (Section 3.3).

So we have decided to describe more precisely the cell cycle with a non-constant gene concentration. We place ourselves in one cell cycle: at a time s such as $0 \leq s < \tau_D$. We also consider known all times $\tau_{R,i}$ of each gene replication; it means in particular that for every time s, the i-th gene copy number is known: $g_i(s) = \left(1 + \mathbb{1}_{s \geq \tau_{R,i}}\right) / (V_0 2^{s/\tau_D})$ (the factor $V_0 2^{s/\tau_D}$ represents the volume). By analogy with the equilibrium condition presented in Subsection 3.3.2, we expect that a large number of cell cycles have already occurred, so that the concentration of any entities is the same at the beginning and at the end of the cell cycle. It means that, for each unit of production, the concentrations m_i, p_i, $e_{Y,i}$ and $e_{R,i}$ are such as

$$\forall i \in \{1, ..., K\} \quad \begin{cases} p_i(0) = p_i(\tau_D), & m_i(0) = m_i(\tau_D), \\ e_{Y,i}(0) = e_{Y,i}(\tau_D), & e_{R,i}(0) = e_{R,i}(\tau_D). \end{cases} \tag{4.8}$$

With these considerations, we have a system of ODEs that aims to emulate the average behaviour of stochastic model of Section 4.1 during the cell cycle. In the next section, under some simplifications, we propose to gives expressions for $m_i(s)$, $p_i(s)$, $e_{Y,i}(s)$, $e_{R,i}(s)$, $f_Y(s)$ and $f_R(s)$ as a function of all the parameters ($\lambda_{1,i}$, $\sigma_{1,i}$, etc.) and $g_i(s)$.

4.2.2 Dynamics of the Average Production Model

In order to estimate the parameters, one needs to have expressions for m_i, $e_{Y,i}$, p_i, $e_{R,i}$, f_R and f_Y of the previous ODEs for any time s of the cell cycle. But the interdependence between $e_{Y,i}$ and f_Y on one hand and $e_{R,i}$ and f_R on the hand raise difficulties when integrating these equations. Explicit solution for the dynamics m_i, $e_{Y,i}$, p_i, $e_{R,i}$, f_R and f_Y are therefore not easy to obtain directly.

In order to have expressions for these quantities anyway, we choose to make some biologically reasonable simplifications that permit to give explicit expressions for f_Y and f_R. In the next subsections, the stochastic simulations will show a good correspondence between their average concentration of free RNA-polymerase and ribosomes and the ones predicted here; it will therefore justify *a posteriori* the simplifications that we make in this section.

Let's consider at first the RNA-polymerases. We denote by $\langle \mu_1 \rangle := \sum_i \mu_{1,i}/K$ the mean elongation rates of transcription and the function h such as

$$h(s) := \sum_{i=1}^{K} e_{Y,i}(s) \frac{\langle \mu_1 \rangle}{\mu_{1,i}}.$$

The dynamic of h is given by summing the equations Equation (4.4) for i from 1 to K, and by using Equation (4.6):

$$\frac{d}{ds} h(s) = f_Y(s) \cdot \langle \mu_1 \rangle \left(1 + \sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s) \right) - \beta_Y \langle \mu_1 \rangle - \frac{\log 2}{\tau_D} \cdot h(s). \tag{4.9}$$

The h is simply a weighted sum of the allocated RNA-polymerases $e_{Y,i}$. We decided to consider that such weighting has little influence, and that h does not greatly differ from the uniform sum $\sum_i e_{Y,i}$, that is to say:

$$h(s) = \sum_{i=1}^{K} e_{Y,i}(s) \frac{\langle \mu_1 \rangle}{\mu_{1,i}} \quad \simeq \quad \sum_{i=1}^{K} e_{Y,i}(s) = \beta_Y - f_Y(s).$$

It would be in particular true if all elongation rates $\mu_{1,i}$ are identical for every genes (i.e. if $\mu_{1,i} \equiv \langle \mu_1 \rangle$ for all i).

With this simplification, from Equation (4.9), on obtain a differential equation on f_Y:

$$\frac{d}{ds} f_Y(s) = \langle \mu_1 \rangle \beta_Y \left(\frac{\log 2}{\langle \mu_1 \rangle \tau_D} + 1 \right) - \langle \mu_1 \rangle \left(1 + \frac{\log 2}{\langle \mu_1 \rangle \tau_D} + \sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s) \right) f_Y(s). \tag{4.10}$$

One can remark that the concentrations of free RNA-polymerases is on a quick timescale. Indeed, Table 1.A.2 in Chapter 1 gives the total numbers of transcriptions and translations in the whole cell: there are a dozen of transcriptions, and hundreds of translations per second. As a consequence, one can expect that that f_Y quickly reach their equilibrium compared to the cell cycle. This consideration will be justify *a posteriori* with the correspondence with the stochastic simulations.

With these considerations, we consider that the derivative term of Equation (4.10) is null and it comes:

$$f_Y(s) = \beta_Y \frac{1 + \dfrac{\log 2}{\langle \mu_1 \rangle \tau_D}}{\displaystyle\sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s) + 1 + \frac{\log 2}{\langle \mu_1 \rangle \tau_D}}.$$

Moreover, as the elongation rates $\mu_{1,i}$ will be determined in the next section, it will appears that $\log 2/(\langle\mu_1\rangle \times \tau_D) \sim 10^{-3} \ll 1$. Therefore, it is possible to neglected the contribution of this term. Taking this aspect into consideration, and with the same logic for the free ribosomes, it follows:

$$f_Y(s) = \beta_Y \frac{1}{1 + \sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s)} \quad \text{and} \quad f_R(s) = \beta_R \frac{1}{1 + \sum_{i=1}^{K} \frac{\lambda_{2,i}}{\mu_{2,i}} m_i(s)}.$$

With the global quantities f_Y and f_R known, we are able to give expression for gene-specific variables. For each $i \in \{1, ..., K\}$, the number of mRNAs of type i, one can integrate Equation (4.2) and find that:

$$\frac{dm_i}{ds}(s) = \lambda_{1,i} g_i(s) \cdot f_Y(s) - \sigma_{1,i} m_i(s) - \frac{\log 2}{\tau_D} \cdot m_i(s).$$

With the boundary conditions (Equation (4.8)), it is easy to deduce that:

$$m_i(s) = \lambda_{1,i} \frac{e^{-\sigma_{1,i} s}}{2^{s/\tau_D}} \left[\int_0^s 2^{u/\tau_D} e^{\sigma_{1,i} u} g_i(u) f_Y(u)\, du + \frac{\int_0^{\tau_D} 2^{u/\tau_D} e^{\sigma_{1,i} u} g_i(u) f_Y(u)\, du}{2e^{\sigma_{1,i}\tau_D} - 1} \right]. \tag{4.11}$$

Since the quantities g_i , f_Y are known, we have an explicit solution for m_i.

Similarly for $e_{Y,i}(s)$ and $e_{R,i}(s)$, it comes

$$e_{Y,i}(s) = \lambda_{1,i} \frac{e^{-\mu_{1,i} s}}{2^{s/\tau_D}} \left[\int_0^s 2^{u/\tau_D} e^{\mu_{1,i} u} g_i(u) f_Y(u)\, du + \frac{\int_0^{\tau_D} 2^{u/\tau_D} e^{\mu_{1,i} u} g_i(u) f_Y(u)\, du}{2e^{\mu_{1,i}\tau_D} - 1} \right],$$

$$e_{R,i}(s) = \lambda_{2,i} \frac{e^{-\mu_{2,i} s}}{2^{s/\tau_D}} \left[\int_0^s 2^{u/\tau_D} e^{\mu_{2,i} u} m_i(u) f_R(u)\, du + \frac{\int_0^{\tau_D} 2^{u/\tau_D} e^{\mu_{2,i} u} m_i(u) f_R((u)\, du}{2e^{\mu_{2,i}\tau_D} - 1} \right].$$

Let's now consider the type i protein concentration. By integrating the equation Equation (4.2), and by considering the Equation (4.8), it comes:

$$p_i(s) = \frac{\mu_{2,i}}{2^{s/\tau_D}} \int_0^{\tau_D} (1 + \mathbb{1}_{u<s})\, 2^{u/\tau_D} e_{R,i}(u)\, du. \tag{4.12}$$

As in the models of the previous chapter, we are interested in the average concentrations over the cell cycle. Since, in the system of ODEs, we define the average concentrations over the cell cycle of free RNA-polymerases and ribosomes respectively as

$$\overline{f_Y} = \frac{1}{\tau_D} \int_0^{\tau_D} \beta_Y \frac{1}{\sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s) + 1}\, ds \quad \text{and} \quad \overline{f_R} = \frac{1}{\tau_D} \int_0^{\tau_D} \beta_R \frac{1}{\sum_{i=1}^{K} \frac{\lambda_{2,i}}{\mu_{2,i}} m_i(s) + 1}\, ds. \tag{4.13}$$

We defined similarly the concentrations $\overline{m_i}$ and $\overline{p_i}$ averaged over the cell cycle. By integrating Equation (4.12) and Equation (4.12), it follows:

$$\overline{m_i} = \frac{\lambda_{1,i}}{\sigma_{1,i}\tau_D + \log 2} \int_0^{\tau_D} g_i(u) f_Y(u)\, du \quad \text{and} \quad \overline{p_i} = \frac{\lambda_{2,i}\mu_{2,i}\tau_D}{\log 2\,(\mu_{2,i}\tau_D + \log 2)} \int_0^{\tau_D} m_i(u) f_R(u)\, du. \tag{4.14}$$

Now we have expressions of the average concentrations of $\overline{m_i}$, $\overline{p_i}$, $\overline{f_R}$ and $\overline{f_Y}$ for any time s in the cell cycle that will be used in the next subsection to determine the parameters.

4.2.3 Parameters Estimation

Parameters that will be used in the stochastic simulations are determined in this section. In total, we have to determine all reaction rates for every protein type ($\lambda_{1,i}$, $\mu_{1,i}$, $\sigma_{1,i}$, $\lambda_{2,i}$ and $\mu_{2,i}$ for $i \in \{1, ..., K\}$) as well as concentration parameters of RNA-polymerases, and ribosomes (respectively β_Y and β_R), the proportion between the volume and the proteic mass β_P, the mass of each proteins w_i and the copy number g_i of every gene.

To do so, the idea is to use the measures of Taniguchi et al. (2010): the average concentration of mRNAs and proteins for each gene, as well as the mRNA halftime. As explained in the previous chapter (Section 3.1), 1081 genes were considered in the experiment, among which 841 have their mRNA production measured. The genome of *E. coli* is about approximately 4000 expressed genes so the measures permit to represent only a part of the total protein production mechanism. In a first step, we only take into account the 841 genes with protein and mRNA production measured and consider that it would represent the whole genome; in Subsection 4.4.1 we will study the case of a simulation with a completed set of genes.

Taniguchi et al. (2010) gives no measures about the quantities of non-allocated RNA-polymerases or ribosomes. So, to be able to completely determine a set of parameters, we fix the average concentration of free RNA-polymerases and ribosomes. It means we can have multiple sets of parameters depending on this choice. During the simulations, we will examine several simulations with different values for average free RNA-polymerase and ribosome concentrations to see their impact on the dynamic of the model Section 4.3

As in the models of the previous chapter, the rate $\sigma_{1,i}$ of mRNA degradation of type i is still deduced from its half-life measured in Taniguchi et al. (2010) $\tau_{m,i}$ through the expression $\sigma_{1,i} = \log 2/\tau_{m,i}$ (for more information, refer to subsubsection 3.3.4.2); the doubling time $\tau_D = 150\,\text{min}$ is given by the article. The gene copy number $g_i(s)$ at time s is deduced from the position of the gene position of the i-th gene (see Section 4.A for more details).

The rates $\mu_{1,i}$, $\mu_{2,i}$ of mRNAs and protein elongation rates can be deduced from the gene length of the i-th gene. In the description of model, we have considered that the length of the mRNA is characterised by its length; so a rate the parameter $\mu_{1,i}$ is given by the mRNA elongation speed (39 Nucl/s in Bremer and Dennis (1996) for slow growing cells) divided by the length of the i-th gene. Similarly, $\mu_{1,i}$ is given by the protein elongation speed (12 aa/s in Bremer and Dennis (1996) for slow growing cells) divided by the number of amino-acid coded by the i-th gene divided . The mass of each protein w_i is also deduced from the length of the gene as it determines the number of amino-acids of the protein.

What remains to determine are the concentration parameters of RNA-polymerases, and ribosomes (β_Y and β_R), the proportion between the volume and the mass of proteins β_P, as well as the activities of the gene and the mRNA (respectively $\lambda_{1,i}$ and $\lambda_{2,i}$) in each unit of production $i \in \{1, ..., K\}$. To do so, we interpret the mRNA and protein concentration of each type measured in Taniguchi et al. (2010) as the average concentration of each mRNA and proteins over the cell cycle of this model (respectively $\overline{m_i}$and $\overline{p_i}$). Moreover, as previously said, the average concentrations of free RNA-polymerases $\overline{f_Y}$ and free ribosomes $\overline{f_R}$ are fixed.

We want now to compute β_Y, β_R, $\lambda_{1,i}$ and $\lambda_{2,i}$ based on known values for $\overline{f_Y}$, $\overline{f_R}$, $\overline{m_i}$ and $\overline{p_i}$. Let's first determine the parameter β_P. In the description of the stochastic model, Equation (4.1) states that at every moment, the volume is considered to be proportional to the total mass of proteins. Interpreting $\overline{p_i}$ as the average concentration of the protein of type i leads by integration of Equation (4.1) to

$$\beta_P = \sum_{i=1}^{K} w_i \overline{p_i}.$$

Let's continue with the parameters relevant to the transcription: $\lambda_{1,i}$ and β_Y. With Equation (4.13) and

Equation (4.14), if we consider the vector $[\beta_Y, \lambda_{1,1}, ..., \lambda_{1,K}]$, it can be considered as a solution of the system

$$\begin{cases} \beta_Y & = \overline{f_Y} \left(\dfrac{1}{\tau_D} \displaystyle\int_0^{\tau_D} \left(\sum_{i=1}^{K} \frac{\lambda_{1,i}}{\mu_{1,i}} g_i(s) + 1 \right)^{-1} ds \right)^{-1} \\ \lambda_{1,i} & = \overline{m_i} \cdot (\sigma_{1,i}\tau_D + \log 2) \cdot \left(\displaystyle\int_0^{\tau_D} g_i(u) f_Y(u)\, du \right)^{-1} \quad \forall i \in \{1, ..., K\}. \end{cases} \tag{4.15}$$

Since $\overline{f_Y}$, $\overline{m_i}$ and $g_i(s)$ have already been settled, we can use a fixed point optimisation procedure to determine β_Y and all $\lambda_{1,i}$. Then, as these parameters are determined, we now have an explicit expression for $f_Y(s)$ for any time s of the cell cycle.

Let's finish with parameters relevant to translation, that is to say $\lambda_{2,i}$ and β_R. Here again, we use a fixed point optimisation procedure to deliver the result. With Equation (4.13) and the expression of $\overline{p_i}$ in Equation (4.14), if we consider now the vector $[\beta_R, \lambda_{2,1}, ..., \lambda_{2,K}]$, it is solution of the system

$$\begin{cases} \beta_R & = \overline{f_R} \times \left(\dfrac{1}{\tau_D} \displaystyle\int_0^{\tau_D} \left(\sum_{i=1}^{K} \frac{\lambda_{2,i}}{\mu_{2,i}} m_i(s) + 1 \right)^{-1} ds \right)^{-1} \\ \lambda_{2,i} & = \overline{p_i} \times \left(\dfrac{\mu_{2,i}\tau_D}{\log 2\,(\mu_{2,i}\tau_D + \log 2)} \displaystyle\int_0^{\tau_D} m_i(u) f_R(u)\, du \right)^{-1} \quad \forall i \in \{1, ..., K\}. \end{cases} \tag{4.16}$$

By fixing the average amount of free RNA-polymerases and ribosomes, it is possible, through this procedure to settle sets of parameters that are fitted to the experimental measures.

4.2.4 Validation of the Average Production Model

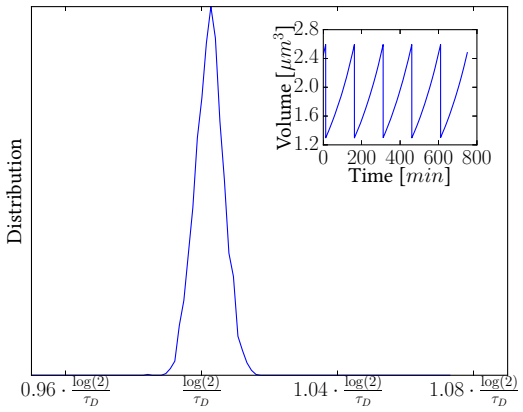
The description of the average production through the system of ODE make the computation of parameters of the stochastic model possible. Yet the good correspondence between the model of average production of this section and the average behaviour of the stochastic model of Section 4.1 has to be supported. For instance, one has to check that the stochastic simulations are producing the good quantities of mRNAs and proteins (which is the main purpose of this section).

Here, we present the results of a particular simulation with parameters determined using the previous protocol of parameters. The quantitative description of the set of parameters is given in Table 4.1. Its average behaviour will be compared with expressions of the expression derived from the system of ODEs. The simulation presented here takes only the 841 genes with protein and mRNA production described in Taniguchi et al. (2010), and we have fixed the number of free RNA-polymerases and ribosomes in order to compute the parameters; but the results of this subsection remain true with the other sets of parameters later presented.

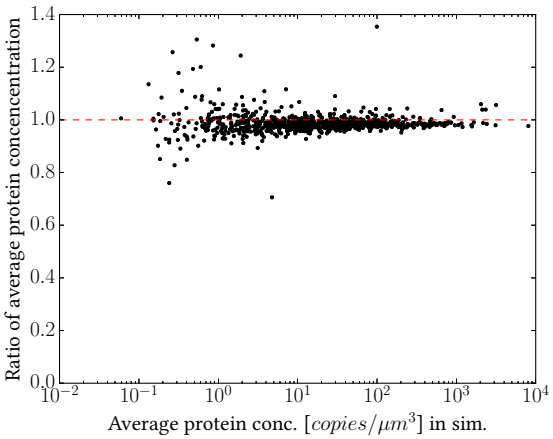
The system of ODEs supposes that the volume growth is exponential with rate $\log 2/\tau_D$. In Figure 4.1a, the volume of the cell indeed seems to grow exponentially in the simulation; the growth rate is centred around which corresponds to the expected a doubling time of τ_D.

For each type of gene, Figure 4.1b shows the ratio between the protein production observed in the simulation divided by the protein production expected. It appears that the correspondence is correct, especially for the highly expressed proteins. It is less precise for the protein less expressed (which may be due in part to the longer time needed for their empirical mean to converge). Globally, the correspondence between the productions seems good enough.

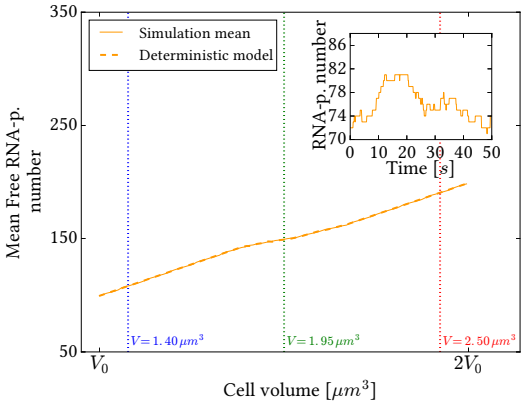
The stochastic simulation displays relative quick timescale for the evolution of free RNA-polymerases (of the order of the second) and even quicker for the free ribosomes (insets of Figure 4.1c and Figure 4.1d).



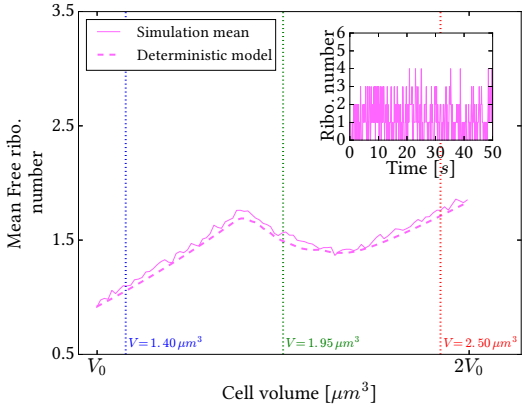
(a) Growth rate distribution and volume



(b) Protein production in simulations and in experiments.



(c) Free RNA-polymerase profile



(d) Free ribosome profile

Figure 4.1: Average correspondence of the stochastic model to the system of ODEs. **(a)**: a simulation sample that shows that cells grow exponentially from around V_0 up to around $2V_0$ (inset); the growth rate distribution is centred around the expected growth rate $\log 2/\tau_D$ (main figure). **(b)**: Ratio between the average protein concentration in simulation and in experiments. **(c)** and **(d)**: the mean of respectively of free RNA-polymerases and ribosomes at each moment of the cell cycle in the simulations (solid lines) and is predicted by the system of ODEs (dashed lines). Insets: example of dynamics of respectively free RNA-polymerases and ribosomes for one simulation.

Quantity	Param.	Median	Mean	Maximum	Minimum
Waiting time for a transcription per gene*	$(\lambda_{1,i}\overline{f_Y})^{-1}$	65.1	$3.90 \cdot 10^2$	$3.22 \cdot 10^4$	0.69
Waiting time for a translation per mRNA*	$(\lambda_{2,i}\overline{f_R})^{-1}$	0.57	9.75	$1.75 \cdot 10^3$	$8.79 \cdot 10^{-3}$**
mRNA lifetime	$\sigma_{1,i}^{-1}$	5.15	6.63	52.1	0.91
mRNA elongation	$\mu_{1,i}^{-1}$	0.41	0.49	1.97	$7.05 \cdot 10^{-2}$
Protein elongation	$\mu_{2,i}^{-1}$	0.44	0.53	2.14	$7.64 \cdot 10^{-2}$

Table 4.1: Quantitative summary of the parameters in min. (*: show little changes with other choice of $\overline{f_Y}$ and $\overline{f_R}$; **: this value of the gene *yjiY* is biologically unrealistic, maybe due to an error on the measure of one type of mRNA in Taniguchi et al. (2010); removing this aberrant value does not change the simulations).

Computed from the stochastic simulations, the main Figure 4.1c and Figure 4.1d present the mean number of free RNA-polymerases and ribosomes as a function of the cell volume. The mean of each free entity is not constant during the cell cycle. The dashed lines represent the expected value of free entities given by the model of average production of this section (Equation (4.13)). It is indeed a good approximation for the behaviour of free RNA-polymerases and ribosomes.

All these results support the idea that the expressions deactivated from the system of ODEs are a good way to describe the average behaviour of the stochastic model. In the next sections, interest in the results of the simulation in terms of distribution.

4.3 Impact of Free RNA-polymerases and Ribosomes

As said in Subsection 4.2.3, the parameter computation supposes the fixation of the average concentration of free RNA-polymerases and ribosomes. In this section we propose several simulations where the average concentration of these free entities are changed and see the influence it has on their distribution and on the protein variability.

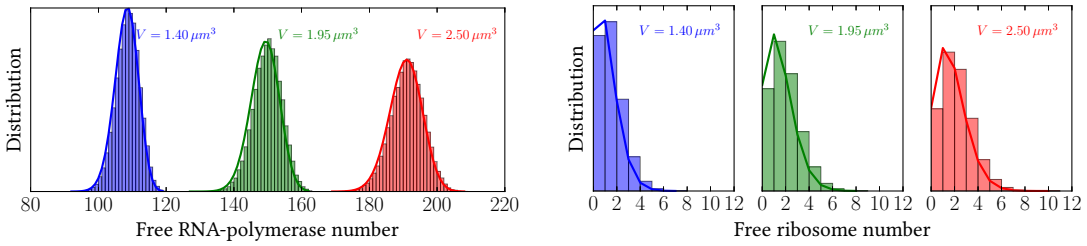
4.3.1 Few Free Ribosomes and Many Free RNA-polymerases

We begin with a simulation with a low concentration of non-allocated ribosomes as it seems a reasonable biological assumption. Indeed since ribosomes are composed of multiple subunits which comes with high costs for the cell, they are present in limited amount. Consequently, they are subject to a large competition between transcripts (see Warner et al. (2001) in the case of the yeast); therefore it is reasonable to take a low concentration of free ribosomes. At the same time, the parameters are settled in such a way that most of the RNA-polymerases are non-allocated: it appears in real cells at every instant, most of the RNA-polymerases are not specifically bound on the DNA (Klumpp and Hwa, 2008). In that, this simulation aims to represent a case tat is close to what happens in real cells.

Free RNA-polymerase and Ribosome Distributions

In these simulation, we look at the distributions of free RNA-polymerases and ribosomes. In Figure 4.1a and Figure 4.1b we show these distributions at three different phase in the cell cycle: we have selected cells of a given volume (either $1.40\,\mu m^3$,$1.95\,\mu m^3$ or $2.50\,\mu m^3$, which correspond to the beginning, middle and end

of the cell cycle). These distributions change as the volume increase (so that the average follows the curves shown in Figure 4.1c and Figure 4.1d).



(a) Free RNA-polymerase histograms

(b) Free ribosome histograms

Figure 4.1: Distributions of free RNA-polymerases and ribosomes for cells of volume $1.40\,\mu m^3$, $1.95\,\mu m^3$ and $2.50\,\mu m^3$. Free RNA-polymerase (fig. **(a)**) and free ribosomes (fig. **(b)**) number distribution for cells each of the volumes. In thick lines the binomial distribution predicted for a simplified model (see main text and Section 4.C).

In order to interpret the observed distributions of free RNA-polymerases and ribosomes at a certain extend, we can propose a simplified model of RNA-polymerase and ribosome allocation (it is greatly inspired by the model described in Fromion et al. (2015)). It is a simplification of the stochastic model of the chapter, mainly in that translation and the translation are considered separately, and that there is no notion of cell growth. The idea would be to approach the "local" equilibrium of RNA-polymerases and ribosomes before any significant change in the volume.

This simplified description predicts that for a given volume V, the distribution of free RNA-polymerases and ribosomes would be both a binomial distribution (see their parameters in Section 4.C). These predicted binomial distributions are plotted in Figure 4.1 in thick lines. In the RNA-polymerase case, the binomial distribution globally fit the histograms. The ribosome distribution is denatured: the parameters of the binomial distribution (N, ϕ) are such that $\phi \ll N$. It is due to the low concentration of free ribosomes chosen for the parameters computation. But even this denatured case shows a good correspondence between the binomial distribution and the simulation histograms.

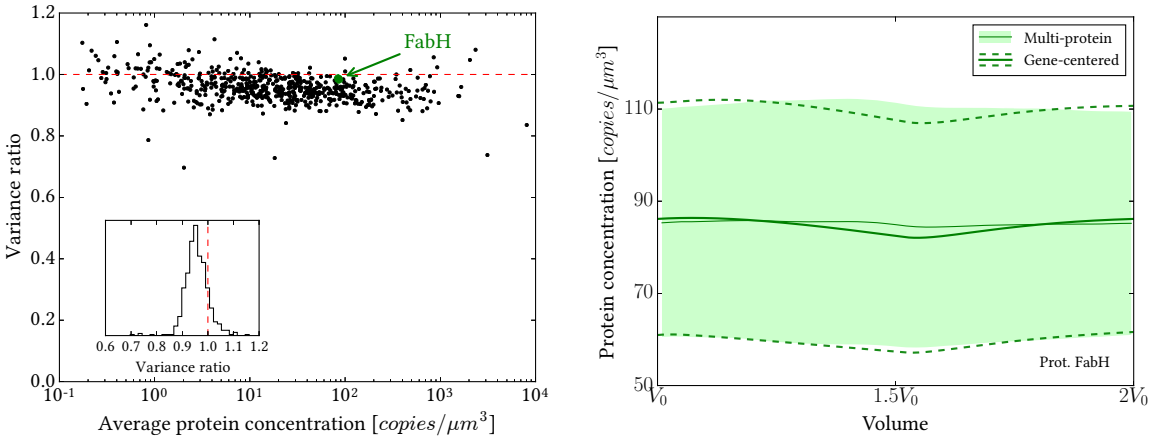
Noise of Proteins

By performing the simulations, the global noise of each protein concentration is measured. In order to estimate the variance added by the interactions between the different protein production units, we compare this noise with the one obtained in the gene-centred model of Section 3.4 (with cell-cycle, binomial division and gene replication). This subsection compares these two models.

Figure 4.2a shows, for each gene, the variance of protein concentration in the gene-centred model, divided by the one in the multi-protein model. It appears that 90% of the genes have a variance ratio above 0.9 (the mean of the ratio is 0.96 in the set of genes). It means that the interactions between protein productions only represents at most 10% of variability.

This good concordance between the two models in terms of protein average expression (Figure 4.1a) and the protein variance (Figure 4.2b) is confirmed by the general aspect of the protein profiles. Taking the example of protein FabH, Figure 4.2b shows a comparison of its profiles between both models: the figure shows, for each volume of the cell cycle, the mean and from either side the standard deviation of protein concentration. The evolution of mean protein production (the thick lines in the figures) differs: the effect of gene replication

is less marked in the case of the multi-protein model. But globally the gene-centred model of the previous chapter seems globally fit the simulations of the multi-protein model both in terms of mean production and variability.



(a) Variance ratio of protein concentration.

(b) FabH production profile in both models.

Figure 4.2: Comparison with the model with gene replication of section §Section 3.4. **(a)** Ratio between the protein variance of the gene-centred model of Section 3.4 and the variance of the multi-protein model of this chapter. The simulations show a limited tendency for significant additive variability in the multi-protein model. Inset: the histogram of the variances. **(b)** The profile of the protein FabH as a function of the cell volume. In blue is the profile obtained in the multi-protein model through simulation; in cyan the profile obtained in the model of section §Section 3.4. Even if the standard deviation seems similar, the profile seems less sensible to the cell-cycle than in the case of the previous model.

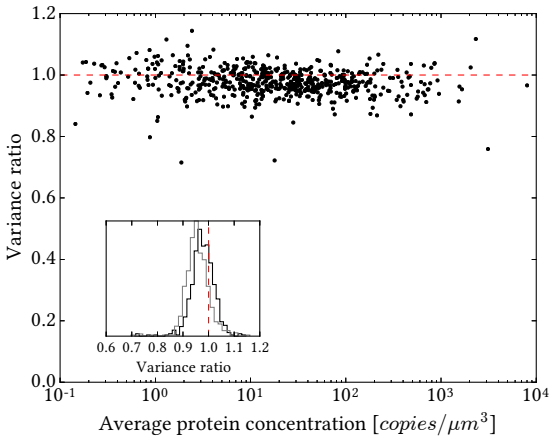
This result seems to support the idea that globally, the gene-centred model is a correct first approximation of the dynamic of protein concentration during the cell cycle. This reassembles to a mean-field property where the interdependent productions of protein can be approximated by independent processes (Fromion et al. (2015) proved such result in the case of their own model).

This simulation with low abundance of free ribosomes and a large concentration of free RNA-polymerase seems, as previously explained, relatively biologically pertinent. In what follows, all the simulations take the set of parameters of this subsection and change one particular simulation aspect for each of them.

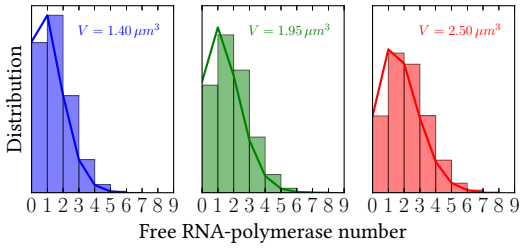
4.3.2 Influence of Free RNA-polymerase Concentration

In this subsection, we interested in the effect of the abundance of free RNA-polymerases on the protein variability. We have produced a series of parameters where the average concentration of free RNA-polymerases was fixed successively to 1, 10, 100 and 1000 copies/μm^3. In each case, we have deduced a set of parameters, where the affinity constants $\lambda_{1,i}$, $\lambda_{2,i}$ are still calculated in such a way that average mRNA and protein concentrations still correspond to the experimental measures.

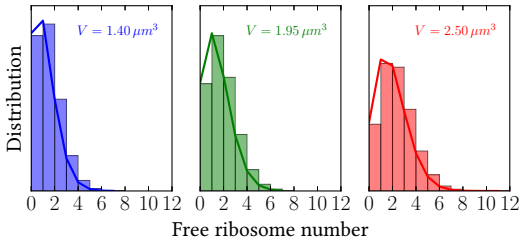
In Figure 4.3 are shown the results for a very low free RNA-polymerase concentration. As the number of free RNA-polymerases is low, its distribution reassembles the distribution of ribosomes (see Figure 4.3b) and is still well predicted by the simplified model (presented in Section 4.C). The variance of protein seems to decrease as the free RNA-polymerase concentration is lower. The gap between between the multi-protein



(a) Variance ratio of protein concentration.



(b) Free RNA-polymerase histograms



(c) Free ribosome histograms

Figure 4.3: Simulations with a low concentration of free RNA-polymerases. **(a)**: Ratio between the protein variance of the gene-centred and the multi-protein models. The ratio has little difference with higher concentration for RNA-polymerases. Inset: in black the histogram of the variances; in grey, the corresponding histogram in the case of Figure 4.2a. Free RNA-polymerase (fig. **(a)**) and free ribosome (fig. **(b)**) number distribution for cells each of the volumes. In thick lines the binomial distribution predicted for the simplified model (see Section 4.C).

model and the gene-centred model is reduced: now, on average the variance ratio is 0.98 (90% of the genes have a variance ratio above 0.92).

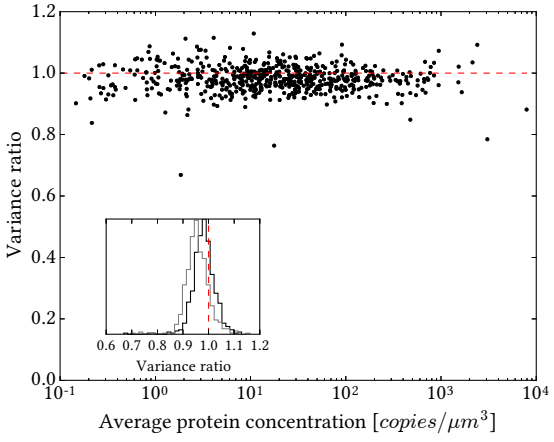
One the contrary, a very high number of free RNA-polymerases show similar results as in Subsection 4.3.1.

4.3.3 Influence of Free Ribosome Concentration

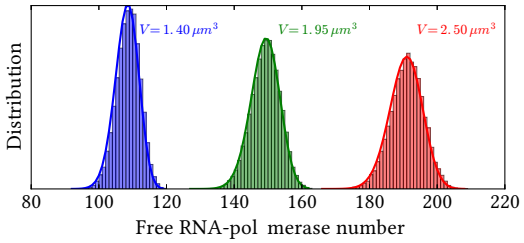
Analyse similar to the previous section has been performed for the case of free ribosomes: we computed a set of parameters based on average concentrations of non-allocated ribosomes of 1, 10, 100 and 1000 copies/µm^3. It can be first remarked that for very high free concentrations, the binomial fit of the simplified model (described in Section 4.C) is not relevant to describe the free ribosome distribution.

In this case again, changes to the average concentration of free ribosomes show a little but noticeable difference. As the average concentration of free ribosomes increases, the variance of each protein decreases. As shown in Figure 4.4a, for a concentration of 1000 copies/µm^3, the variance of the multi-protein represent on average 0.98 of the one predicted by the gene-centred model (90% of the genes have a variance ratio above 0.93).

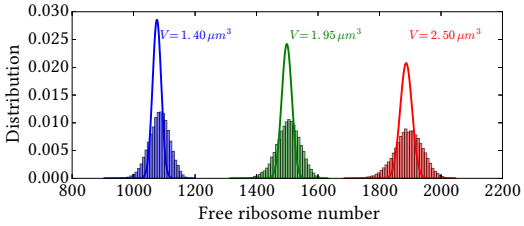
Fluctuations in the number of free ribosomes seem the be the main source of the additional variability observed in the multi-protein model; and this effect seems less important as the number of free ribosomes is high. But in real bacteria, the number of free ribosomes usually seems quite low due to the high cost of ribosome production; then, the a low number of free ribosomes (like in the simulation of Subsection 4.3.1)



(a) Variance ratio of protein concentration.



(b) Free RNA-polymerase histograms



(c) Free ribosome histograms

Figure 4.4: Simulations with a high concentration of free ribosomes. **(a)**: Ratio between the protein variance of the gene-centred and the multi-protein models. Inset: in black the histogram of the variances; in grey, the corresponding histogram in the case of Figure 4.2a. Free RNA-polymerase (fig. **(a)**) and free ribosomes (fig. **(b)**) number distribution for cells each of the volumes. In thick lines the binomial distribution predicted for a simplified model (see Section 4.C).

seems more plausible than this simulation.

To confirm the specific influence of fluctuations of ribosomes on the protein variability, we have performed a simulation with a modified version of the model. The multi-protein model has been changed in such a way that the concentration of non-allocated ribosomes is fixed as constant during the whole simulation (meanwhile the free RNA-polymerases are still fluctuating). Results about protein variability are similar of what is shown in Figure 4.4a: the variance of each protein concentration is equivalent to what was described by the gene-centred model.

The conclusion of this section is that the interaction between the different productions of proteins add little additional noise to the model: in the best case, the gene have an increase of 10% of variability compared to the case where all the production are considered independently. This additional variability seems to be less important as the concentration of free RNA-polymerases is low and the concentration of free ribosomes is high.

4.4 Other Possible Influence on Protein Variability

In this section, based on the set of parameters of the simulation Subsection 4.3.1 (with few free ribosomes and more RNA-polymerases), we make variations on some modelling choices for some cellular mechanisms: a large set of genes, RNA-polymerases and ribosomes as a result of gene expression, the introduction of RNA-polymerase non-specific binding on the DNA, considering uncertainty in the division and DNA replication

processes, etc. We will show that the protein variability is quite robust to any of these changes. We will show that most of these changes do not seem to bring a significant addition variability source for the protein noise: as for the results presented in Subsection 4.3.1, the protein variance is still increased by at most 10% compared to the gene-centred model.

4.4.1 Additional Genes

The genome of *E. coli* has approximately 4000 expressed genes. But the measures of Taniguchi et al. (2010) take into account only a part of it. Only 1018 protein types were considered in the article, and among them, only 841 types have the mRNA production estimated. In order to better represent the complete genome of the bacteria, we have created a set of parameters with an extended pool of additional randomly created genes so that the total number of genes would be 4000.

For each new gene, we have sampled an average protein and mRNA concentration, an mRNA lifetime and a gene position. By studying the data of Taniguchi et al. (2010), we have investigated all possible statistical correlations between these quantities; it appears that only the mRNA and protein concentration are positively correlated. We therefore have sampled the mRNAs lifetime and the gene position and length independently from the two other quantities.

As in the dataset, the genes appears evenly distributed on the chromosome; we have sampled the gene position uniformly. The empirical mRNA lifetime distribution fitted a log-normal distribution; we have chosen the mRNA lifetime accordingly.

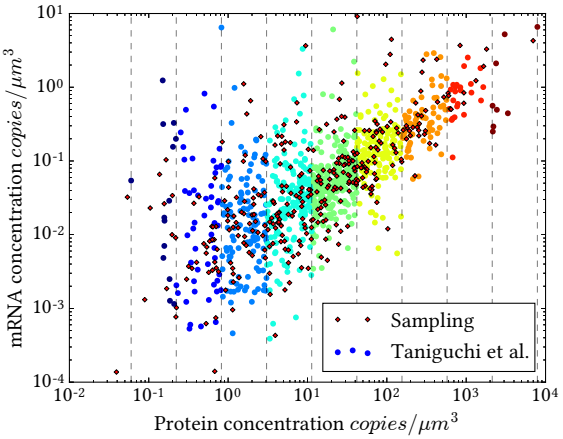


Figure 4.1: Sampling example of mRNA and protein concentration.

For the mRNA and the protein, we have taken into account their correlation. The dataset was binned according to the protein production (the different colours in Figure 4.1). At first, the protein production is sampled according to its empirical distribution. Depending on which bin the obtain protein production falls in, the corresponding mRNA production is sampled according to the mRNA empirical distribution in the bin. By these procedures, the created genome seems representative to the original dataset (see (Figure 4.1)).

Simulations with the completed genome show no significant difference in terms of protein variability. In particular, the variance ratio between protein concentration of the gene-centred model and the multi-protein model is not different as in Subsection 4.3.1.

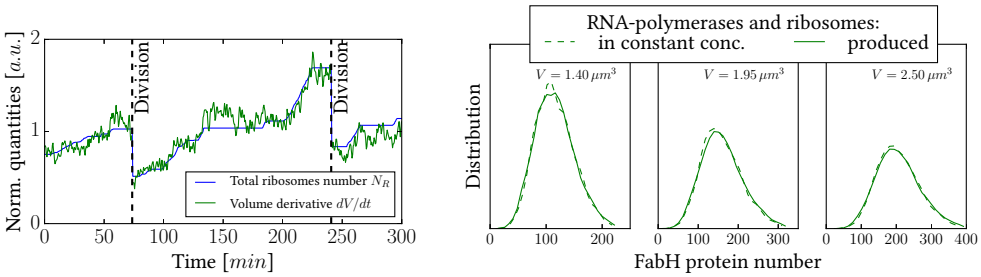
4.4.2 Production of RNA-polymerase and Ribosomes

In the stochastic model of the chapter, all ribosomes and all RNA-polymerases are supposed to have constant concentrations (respectively β_R and β_Y). In reality, both RNA-polymerases and ribosomes are composed of different subunits, each subunit is either a protein or, in the case of ribosomes, a functional RNA. The variability of the production of these subunits can have an overall impact on the global production.

We have performed a preliminary simulation that takes into account this aspect: the goal is not to have a precise description of mechanisms of RNA-polymerase and ribosome production, but rather to have an insight in the magnitude of additional variability it can induce. In this version of the model, the expression of one

gene represents RNA-polymerase production and the expression of another gene represents the ribosome production. It refers to a case where the RNA-polymerases and ribosomes would be composed of only one proteic subunit.

We therefore created two genes, whose protein production was fixed to correspond to the wanted concentration of RNA-polymerases and ribosomes. The mRNA production and lifetime, the gene position and length have been chosen by the same procedure as described in to the previous subsection.



(a) Ribosome number and growth rate

(b) FabH distribution for three volumes in the cell cycle

Figure 4.2: Model with production of RNA-polymerases and ribosomes. **(a)**: correlation between the number of ribosomes and the growth rate in the modified version of the stochastic model. **(b)**: distributions of the number of FabH protein for cells of volume $1.40\,\mu m^3$,$1.95\,\mu m^3$ or $2.50\,\mu m^3$. For each volume, the distribution in the model with production of RNA-polymerases and ribosomes is similar to the model with RNA-polymerase and ribosomes in constant concentration.

This simulation brings an additional variability in the growth rate: the cell growth is more fluctuant. These fluctuations are directly correlated with the number of ribosomes in the cell (Figure 4.2a). But surprisingly, these additional variability has no significant impact in the protein variability. The Figure 4.2b shows the distribution of the protein FabH for cells of different volumes. This case does not differ from the case where the total amount of RNA-polymerases and ribosomes were in constant concentration.

We can propose a possible interpretation of these results. The fluctuations in the total number of ribosomes seems influence primarily the speed of growth (as shown in Figure 4.2a): when the ribosomes are produced, it accelerates the global production of every types of proteins thus increasing the volume. As a consequence, both the production of each type of protein and the volume are co-regulated. Fluctuations in the total number of ribosomes affect the volume growth and the production of the i-th protein in the same way such as in a cell of a given volume, the i-th protein distribution is relatively unchanged.

4.4.3 Non-specifically Bound Polymerases

In the stochastic model of this chapter, as described in Subsection 4.1.2, RNA-polymerases are either on the DNA involved in a transcription process, or is among the F_Y free RNA-polymerases that freely evolve in the cytoplasm. But it has been shown that a lot of RNA-polymerases can bind non-specifically on the DNA, without initiating transcriptions. For instance, Klumpp and Hwa (2008) estimated that around 90% of the RNA-polymerases are non-specifically bound to the DNA.

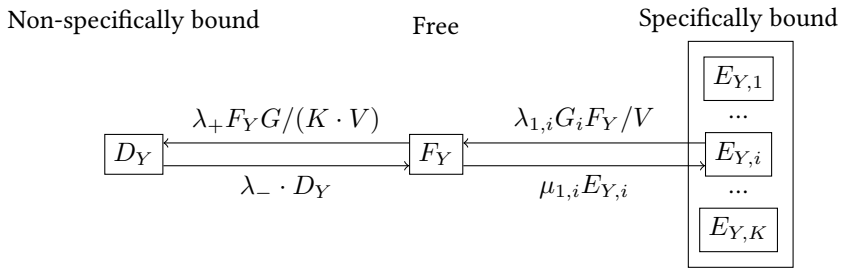


Figure 4.3: RNA-polymeses dynamics in the case of non-specific binding. An RNA-polymerases can be either among the specifically bound on the DNA (in $E_{Y,i}$ for $i \in \{1, ..., K\}$) or free (among F_Y) or non specifically bound on the DNA (among D_Y).

We have created an alternative version of the stochastic model where is introduced a third possible class for RNA-polymerases: RNA-polymerases can also bound non-specifically on the DNA. The binding rate is modelled as follows: at any time s, a free RNA-polymerase bind on the DNA at a rate that depends on the number of free RNA-polymerases $F_Y(s)$ and on the DNA concentration $G(s)/(K \cdot V(s))$; the global rate is hence $\lambda_+ F_Y(s) G(s)/(K \cdot V(s))$ where λ_+ is a parameter that represents the natural affinity of RNA-polymerases for the DNA. Once an RNA-polymerase is bound, it is released in a time represented by an exponential random variable of rate λ_- (see (Figure 4.3)).

We performed a simulation where the parameters λ_+ and λ_- are chosen such that around 90% of the RNA-polymerases are sequestered on the DNA at any time (as it was the case in Klumpp and Hwa (2008)). The protein noise does not seem to be impacted in this case either.

4.4.4 Uncertainty in the Replication Initiation and Division

In the stochastic model as it was initially described, replication initiation and division occur when the cell reaches the respective volumes of V_I and $2V_0$. In reality, these cell decisions are not exact. We propose here a modification of the stochastic model that takes into account this aspect.

The replication and division decision is still a topic of research (for example, see (Tyson and Diekmann, 1986, Wang et al., 2010, Soifer et al., 2014, Osella et al., 2014)). For the division, one hypothesis (referred as "sizer model") is that the division decision depends on the current size of the cell (the size can refer to the mass or the volume, but as explained in Subsection 4.1.1, the density constrain (Marr, 1991) ensures the close proportionality between these two quantities). With this hypothesis, at every instant, the instantaneous probability to divide depends only on the current cell size. It appears that, at least in a first approximation, the cell size distributions observed experimentally can be explained by this "size model" (Robert et al., 2014, Osella et al., 2014). It is therefore this framework that we have considered to represent the cell division decision.

At every moment s of the simulation, with a cell of volume $V(s)$, we introduce an instantaneous division rate $b_D(V(s))$ with b_D is a positive function (it means that the probability to divide between times s and $s+ds$ is given by $b_D(V(s))\, ds$). The division decision is hence only volume dependent. The function b_D is chosen so that the division occurs around the volume $2V_0$ with $V_0 = 1.3\,\mu m^3$ and division precision can be fixed (for more information about the function b_D, see Section 4.B).

Similarly for the replication initiation decision, the stochastic model initially described consider a fixed volume V_I at which the DNA-replication is initiated. We introduce variability in this cell decision, in the same way as we do for the division: at every moment s, we consider a replication initiation rate $b_I(V(s))$ such as the function b_I is chosen in order to have a replication initiation that occurs around volume V_I.

We perform several simulation where we consider different function b_D and b_I in order to have different precision in the division and replication initiation decisions. All these aspects did not seem to have a determinant influence on the protein variability.

4.4.5 Deterministic Time for Replication

When the stochastic model has been initially presented (in Subsection 4.1.2), we have proposed two ways to model the time of DNA replication initiation τ_I . It can either occurs at a deterministic time after the last division, or it can happen when the cell reaches the specific volume V_I. We have checked that this modelling choice has no significant influence on the global dynamic of the system, in particular in the protein noise.

4.5 Conclusions on the Different Sources of Variability

In the two last chapters, we have investigated a series of models of increasing complexity that incorporate different cellular mechanisms that interfere with gene expression. The goal has been to propose a large description of many different possible contributions to the protein variability. Below, we sum up all the results we have obtained in these last two chapters.

- Transcription and translation (Section 3.3): to begin with, we have proposed a gene-centred model where only the mechanisms of transcription and translation are considered. The protein variability predicted is only due internal to gene expression mechanism itself, it is usually referred as "intrinsic noise". It has given us the basic model on which different external aspects have been added. Globally, the noise predicted by this model globally fit the first "intrinsic regime" predicted by Taniguchi et al. (2010) (characterised by a coefficient of variation, defined as the variance divided by the mean squared, inversely proportional to the mean). But the noise of highly expressed proteins seems still underestimated for this simple model.

- Division (Subsection 3.3.6): then the effect of division has been introduced. The binomial sampling of each protein appears to have potentially substantial additional variability for some proteins. Proteins with a low Fano factor (defined as the variance divided by the mean) of protein concentration have a significant increase in their variability: for the set of proteins studied, this effect can double the noise coming from the transcription and translation processes.

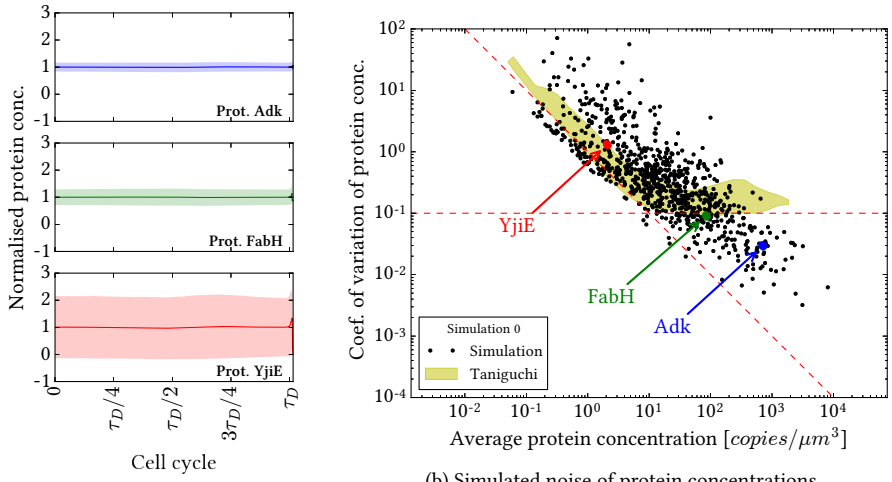
- Gene replication (Section 3.4): the third model has considered the replication of each gene at a certain point in the cell cycle. The consequence is, for each type of protein, to have the mean protein concentration that changes across the cell cycle. But the additional variability due to this effect is very small in regards to the heterogeneity induce by the protein production mechanism and the division.

- Fluctuations of ribosomes and RNA-polymerases (Subsection 4.3.1): in this chapter, we have observed the influence of the global sharing of limited amount of RNA-polymerases and ribosomes in the production of proteins. An additional variability is observed but seems limited: the protein concentration increases its variance of at most 10% compared to the gene-centred model with gene replication. But globally, the gene-centred model approximates correctly the behaviour of the protein production. It is analogue to a mean-field property, such as the one that was demonstrated in the simpler model of Fromion et al. (2015).

- Other sources (Section 4.4): other possible sources of potential variability such as RNA-polymerases and ribosomes produced through gene expression, the non-specific binding and random division and

replication decision have been studied. It appears that none of these effects influence significantly the protein production variability.

Through this work, we have examined the most usual external sources of variability proposed in the literature. Of all the possible origin of extrinsic heterogeneity tested, the binomial sampling seems the prevalent one. As for the free ribosomes and RNA-polymerase fluctuation, yet often proposed as being the principal source of external noise (Kærn et al., 2005, Swain et al., 2002, Taniguchi et al., 2010), their impact seems quite limited. The results of this chapter seems indeed to show that the only noticeable additional variability is due to the low concentration of free ribosomes and high concentration of RNA-polymerases, and that, in any cases, their contributions is inferior to the protein mechanism itself or on the binomial division.



(a) Normalised protein concentration profile.

(b) Simulated noise of protein concentrations

Figure 4.1: Results of simulations of the multi-protein model, compared to Taniguchi et al. (2010) experimentation (dataset of (Subsection 4.3.1), but similar for all model variations). **(a)** Normalised protein concentration profile over the cell cycle for three representative proteins. **(b)** Coefficient of variation (defined as the variance divided by the mean squared) of the protein concentration as a function of the average protein concentration. There is no particular difference with the model with cell cycle and gene replication (see Figure 3.4). It does not replicate Taniguchi et al. (2010) experiments, indicated by the yellow area (corresponding to the point cloud of Figure 3.1c), especially for highly expressed proteins.

At the end, it is not surprising that the protein profile and the global protein coefficient of variation for the multi-protein model (Figure 4.1) has little difference with their counterparts of the previous models. In particular, the additional variability induced by ribosome fluctuations cannot explain the second regime observed in Taniguchi et al. (2010) experiment: the protein coefficient of variation (the variance divided by the mean squared) still globally inversely scales the average production and there is no lower bound limit.

To understand this decay, we can propose two possible explanations: either an experimental bias when measuring the data, or another process not consider in our model. The exhaustive measures of Taniguchi et al. (2010) have not been fully replicated and covers a large range of fluorescence intensities. As this effect

mainly affects proteins with the highest fluorescence, it is possible that some saturation induces a bias in the estimation of variance of highly produced proteins.

Another possibility is to suggest that some important cellular aspect, not modelled here, can have an important impact on protein variance. Even if we have represented the mechanisms that are usually referred in literature as potential important sources of noise, the model proposed here is far from taking into account all the aspects of gene expression. Many other possible sources of heterogeneity can be proposed. For instance, changes in the availability of amino-acids in the medium can induce fluctuations in the translation speed. Also, in the model of this chapter, we have considered the binding and initiation as a single event for both the transcription and translation. A more precise representation would be to describe them as two different processes (Siwiak and Zielenkiewicz (2013) gives for instance a median transcriptional initiation time of 15 s which is of the same order of magnitude as the elongation time). Another aspect not modelled here is the gene regulation procedure: the activation and deactivation of the gene can induce an additional variability in the protein production; it is even more true since the transcription factor is itself a protein and, as a consequence, is itself subject to variability. The assumption that every event occurs at exponential times can also be discuss: for instance, the elongation time would be better represented as having an Erlang distribution (see Chapter 2 of Leoncini (2013)).

4.A Appendix: Gene Replication Times

In simulations, the time at which a gene is replicated is estimated as follows: we first determine the time of DNA replication initiation (the time τ_I in the cell cycle); the, as we consider that the DNA-polymerase replicates DNA at constant speed, we can deduce the time of replication of each gene only by knowing its position.

The article Wallden et al. (2015) investigate the replication initiation; it appears that the initiation occurs at a relative fixed volume per replication origin, and thus independently from the time since the previous division. This relative volume seems, furthermore, constant for different conditions. For slow growing bacteria (with only one DNA replication per cell-cycle), such as those in Taniguchi et al. (2010), the volume at which DNA replication initiation occurs is $V_I = 1.8\,\mu\text{m}^3$. The time of replication initiation τ_I can hence either be considered as a deterministic or stochastic quantity:

- if the volume growth is considered as deterministic and exponential, such as $V(s) = V_0 2^{s/\tau_D}$, then the time of initiation can be considered as being equal to

$$\tau_I = \frac{\tau_D}{\log 2} \log \frac{V_I}{V_0}.$$

- we can also consider that the replication initiation is only volume dependent. One can trigger the initiation when the volume of the cell is around V_I (as it is suggested in Wallden et al. (2015) and which corresponds to the classical model suggested by Donachie (1968)). In that case, the moment of initiation τ_I is stochastic: the probability to divide between s and $s + ds$ is given by $b_I(V(s))\,ds$ (with $V(s)$ the volume at time s of the cell cycle). The function b_I is chosen such as the average volume of initiation is indeed V_I. This mechanism is comparable to the one use to determine the moment of division, in particular, b_I is similar to b_D (see Section 4.B).

In the previous chapter, in the model Section 3.4, as the volume growth is considered as deterministic, we use the first method. In the current chapter, both of these possibilities have been tested without any significant differences (see Section 4.4).

Once known the initiation of DNA replication τ_I, the remaining delay to gene replication of the gene i is considered as deterministic as we consider the speed of DNA replication as relatively constant. We consider that the whole chromosome is replicated in around 40 min (Grant et al., 2011). As a consequence, the distance of the gene from the origin of replication is sufficient to determine the time it takes for the DNA-polymerase to replicate it. The position of each gene was determined with Ecogene database (Zhou and Rudd, 2013).

4.B Appendix: Stochastic Division

The previous simulations considered that division as occurring exactly when the cell reaches the volume $2V_0$. In real cells, the division is not exact. The cell division decision has been studied in the literature (Tyson and Diekmann, 1986, Sharpe et al., 1998, Wang et al., 2010, Osella et al., 2014, Robert et al., 2014, Soifer et al., 2014).

In this literature, it clearly appears the need for a division decision that takes into account, at least partially, the cell volume in order to maintain robust size distributions. If the division mechanism only relies on time in the cell cycle, small variations in the growth produce generations of cells with unstable size distribution.

We propose in the model of the chapter a division that only depends on the volume: at a time s, if the cell is of volume $V(s)$ the rate of division is given by $b_D(V(s))$. We want to choose b_D such as the distribution of cell size at division would be a log-normal distribution centred around $2V_0$. The parameters of this log-normal distribution are referred as μ and σ. The parameter σ that indicates the spreading of the log-normal distribution is considered as settled (we have studied the impact on protein variability of this parameter by making several simulation with different value for it, see Section 4.4). As a consequence, the parameter μ can be determined in order to have a distribution centred around $2V_0$.

Let's now propose a function of division rate b_D in order to obtain such log-normal distribution for the cell division that gives such log-normal distribution for the volumes at division. To do so, let's place in a particular cell and we consider that it grows exponentially at rate α with $v(0)$ its volume at birth: in that case we would have $v(s) = v(0)e^{\alpha s}$ (in the case of this chapter, we take $\alpha := \log 2/\tau_D$). Let's denote by T_D and V_D respectively the time and the volume at division. The goal is to determine the distribution of V_D depending on the division function b_D .

Proposition 4.1. *For an exponentially growing cell, the distribution of V_D , the volume at division, has its probability density function d_D that depends on the division function b_D such as*

$$d_D(v) = \frac{b_D(v)}{\alpha v}$$

at any volume v.

Proof. A rate of division $b_D(v(s))$ is inhomogeneous, the distribution of the time of division T_D is therefore given by

$$\forall t > 0 \quad \mathbb{P}\left[T_D > t\right] = \exp\left(-\int_0^t b_D(v(s))\, ds\right). \tag{4.17}$$

We recall that $V_D = v(0)\exp\left(\alpha T_D\right)$. As d_D denotes the probability density function of V_D, it comes that for any $v > v(0)$

$$d_D(v) = \frac{d\left(\mathbb{P}\left[V_D < v\right]\right)}{dv} = \frac{d\left(\mathbb{P}\left[T_D < \dfrac{\log\left(v/v(0)\right)}{\alpha}\right]\right)}{dv} = \frac{d\left(1 - \mathbb{P}\left[T_D > \dfrac{\log\left(v/v(0)\right)}{\alpha}\right]\right)}{dv}.$$

With the (Equation (4.17)), it follows that

$$d_D(v) = \frac{b_D(v)}{\alpha v}.$$

□

As the function d_D is known to be a log-normal distribution of parameters μ and σ, we now have an expression for b_D.

4.C Appendix: Simple Models for Transcription and Translation

The article Fromion et al. (2015) proposes a multi-protein model for translation, with a shared limiting number of ribosomes, where each type of mRNA is supposed to be in constant quantities and where the maximum number of ribosomes on one single mRNA is limited. The system also evolves in a fixed volume as it is the case for classical models.

In order to have a prediction for the number of respectively free RNA-polymerases and free ribosomes; we consider two analogue models that are slightly simplified versions of the model of Fromion et al. (2015). The two analogue models respectively represent the transcription and translation part; they are completely independent.

The goal is, for each of the model, to provide to reproduce the equivalent of the first results of Fromion et al. (2015) and we will show that the expected distribution of free RNA-polymerases (or free ribosomes) is binomial in these simplified cases.

Model for Transcription

As explained in Subsection 4.3.1, one can interpret the model of Fromion et al. (2015) as taking place in a fixed volume V (it would correspond to a small portion of the cell cycle in the stochastic model of this chapter, a portion where the volume of the cell does not change much). We also consider that the gene copy of each unit of production remains constant; as a consequence, the gene copy number of the i-th gene G_i is constant and known. As in the stochastic model of the chapter, and contrary to the model of Fromion et al. (2015), we consider that there is no limiting number of elongating RNA-polymerases on one gene.

In a pool of K genes, let's denote by N_Y the constant total number of polymerases. We consider the random variables $E_{Y,i}$ for $i \in \{1, ..., K\}$ be the number of RNA-polymerases attached to the i-th gene. As a consequence, the random variable

$$F_Y := N_Y - \sum_{i=1}^{K} E_{Y,i} \tag{4.18}$$

is the number of free RNA-polymerases in the system.

The process $X(t) = (E_{Y,i}(t), i \in \{1, ...K\})$ takes place in the state place S the subset of $\mathbb{N}^K$ such as

$$S := \left\{ x \in \mathbb{N}^K, \sum_{i=1}^{K} x_i \leq N_Y \right\}.$$

It means that there is at most N_Y RNA-polymerases that can be attached to genes at the same time. We can describe the Markov process transition by the following Q-matrix: by setting the vector $e_i = (\delta_{i'=i})_{i' \in \{1,...,K\}}$

(δ is used here as the Kronecker delta), for any $x, y \in S$,

$$\begin{cases} q(x, x+e_i) = \lambda_{1,i} G_i \lambda_{1,i} f(x)/V & \text{for any } i \in \{1, ...K\}, \\ q(x, x-e_i) = \mu_{1,i} x_i & \text{for any } i \in \{1, ...K\}, \text{ if } x_i > 0, \\ q(x, y) = 0 & \text{if } \|x - y\| > 1. \end{cases}$$

where

$$f_Y(x) := N_Y - \sum_{i=1}^{K} x_i$$

the number of free RNA-polymerases. Equation (4.18) leads in particular to $f(x - e_i) = f(x) + 1$ for all i.

We search here an invariant reversible probability measure π for the Markov process, that is to say a measure that verifies for any $i \in \{1, \dots, K\}$:

$$\pi(x)\mu_{1,i}x_i \quad = \quad \pi(x - e_i) \cdot \lambda_{1,i} G_i \left(f_Y(x) + 1\right)/V. \tag{4.19}$$

Proposition 4.2. *The invariant measure π of the number of RNA-polymerases in each gene has the following form*

$$\pi(x) = \frac{1}{Z} \cdot \frac{1}{f_Y(x)!} \prod_{i=1}^{K} \frac{(G_i \lambda_{1,i}/(V\mu_{1,i}))^{x_i}}{x_i!}$$

for any $x \in S$ and with $Z > 0$ the normalisation constant.

Proof. Let's consider such a distribution π and verify that it verifies Equation (4.19). For a gene $i \in \{1, ..., K\}$, we take $x \in S$ such that $x - e_i \in S$.

$$\begin{aligned} \pi(x)\mu_{1,i}x_i &= \frac{1}{Z} \cdot \frac{1}{f_Y(x)!} \prod_{i'=1}^{K} \frac{(G_{i'}\lambda_{1,i'}/(V\mu_{1,i'}))^{x_{i'}}}{x_{i'}!} \mu_{1,i} x_i \\ &= \frac{1}{Z} \cdot \frac{1}{f_Y(x)!} \prod_{i'=1}^{K} \frac{(G_{i'}\lambda_{1,i'}/(V\mu_{1,i'}))^{x_{i'}}}{x_{i'}!} \cdot \frac{x_i}{G_i\lambda_{1,i}/(V\mu_{1,i})} \cdot \frac{G_i\lambda_{1,i}}{V} \\ &= \frac{1}{Z} \cdot \frac{1}{(f_Y(x)+1)!} \prod_{i' \neq i}^{K} \left(\frac{(G_{i'}\lambda_{1,i'}/(V\mu_{1,i'}))^{x_{i'}}}{x_{i'}!} \right) \cdot \frac{(G_i\lambda_{1,i}/(V\mu_{1,i}))^{(x_i-1)}}{(x_i-1)!} \cdot \frac{G_i\lambda_{1,i}}{V} (f_Y(x)+1) \\ &= \pi(x - e_i) \cdot G_i\lambda_{1,i}(f_Y(x)+1)/V. \end{aligned}$$

So the measure π indeed verifies Equation (4.19). □

We can now derive from the previous proposition the equilibrium distribution of F_Y, the number of free RNA-polymerases of the process.

Proposition 4.3. *The number of free polymerases F_Y follows*

$$\mathbb{P}\left[F_Y = n\right] = \binom{N_Y}{n} \frac{\Lambda_Y^{N_Y - n}}{(1 + \Lambda_Y)^{N_Y}},$$

with Λ defined such as

$$\Lambda_Y := \sum_{i=1}^{K} G_i \lambda_{1,i}/(V\mu_{1,i}).$$

It means that F_Y follows a binomial distribution $\mathcal{B}(\phi, N)$ for which $\phi = (1+\Lambda_Y)^{-1}$ and $N = N_Y$.

Remark 4.2. *For a given volume V, the average number of free polymerases is*

$$\mathbb{E}[F_Y] = \frac{N_Y}{1+\Lambda_Y} = \frac{N_Y}{1+\sum_{i=1}^{K} G_i\lambda_{1,i}/(V\mu_{1,i})}.$$

It is identical to the number of free RNA-polymerases obtained in a cell of volume V in the deterministic model (see Equation (4.13)).

Proof of Proposition 4.3. Now we search the distribution of the random variable F_Y, the number of free RNA-polymerases. Indeed, with Equation (4.19), it follows that for a given $n \in \{1, ..., N_Y\}$:

$$\begin{aligned}
\mathbb{P}[F_Y = n] &= \sum_{x\in S} \pi(x) \mathbb{1}_{\sum_i x_i = N_Y - n} \\
&= \sum_{x\in S} \frac{1}{Z}\cdot\frac{1}{n!}\prod_{i=1}^{K}\frac{(G_i\lambda_{1,i}/(V\mu_{1,i}))^{x_i}}{x_i!}\mathbb{1}_{\sum_i x_i = N_Y - n} \\
&= \frac{1}{Z}\cdot\frac{1}{n!}\sum_{x\in S}\prod_{i=1}^{K}\frac{(G_i\lambda_{1,i}/(V\mu_{1,i}))^{x_i}}{x_i!}\mathbb{1}_{\sum_i x_i = N_Y - n} \\
&= \frac{1}{Z_{F_Y}}\cdot\frac{1}{n!}\mathbb{P}\left[\sum_{i=1}^{K}\sum_{k=1}^{G_i} C_{i,k} = N_Y - n\right]
\end{aligned}$$

with $\forall i \in \{1, \dots, K\}$ and $\forall k \in \{1, \dots, G_i\}$, $C_{i,k} \sim \mathcal{P}(\lambda_{1,i}/(V\mu_{1,i}))$.

Since the random variables $C_{p,k}$ are following a Poisson distribution, so does their sum with the parameter $\Lambda := \sum_{i=1}^{K} G_i\lambda_{1,i}/(V\mu_{1,i})$. We can, moreover, find an expression for Z_{F_Y}:

$$\begin{aligned}
1 &= \sum_{n=0}^{N_Y}\mathbb{P}[F_Y = n] \\
&= \frac{1}{Z_{F_Y}}\cdot\sum_{n=0}^{N_Y}\frac{1}{n!}\mathbb{P}[C_{1,1} = N_Y - n] \\
&= \frac{1}{Z_{F_Y}}\cdot\sum_{n=0}^{N_Y}\frac{1}{n!}e^{-\Lambda}\frac{\Lambda^{N_Y-n}}{(N_Y-n)!} \\
&= \frac{1}{Z_{F_Y}}\cdot\frac{1}{N_Y!}e^{-\Lambda}\cdot\sum_{n=0}^{N_Y}\binom{N_Y}{n}\Lambda^{N_Y-n}
\end{aligned}$$

so $Z_{F_Y} = \frac{e^{-\Lambda}}{N_Y!}\cdot(1+\Lambda)^{N_Y}$. □

Model for Translation

The model for translation considered here is completely analogue to the transcription case. We still consider that the volume V is fixed and that for each gene, the number M_i of mRNA of type i is known and constant (because of these, the process describe here is independent from transcription). As in the stochastic model of the chapter, and contrary to the model of Fromion et al. (2015), we consider that there is no limiting number of elongating ribosomes on one mRNA.

Similarly to the transcription, we can define N_R (the total number of ribosomes), $E_{R,i}$ (the number of ribosomes elongating an mRNA of type i) and F_R (the number of free ribosomes) such as

$$F_R := N_R - \sum_{i=1}^{K} E_{R,i}.$$

The rate at which a ribosome is sequestered on a type i mRNA is therefore $M_i\lambda_{2,i}/V$, and the rate at which an elongation terminates on a type i mRNA is $\mu_{2,i}E_{R,i}$.

As this model is analogue to the transcription case, we can also prove that

Proposition 4.4. *The number of free ribosomes F_R follows*

$$\mathbb{P}\left[F_R = n\right] = \binom{N_R}{n} \frac{\Lambda_R^{N_R-n}}{(1+\Lambda_R)^{N_R}},$$

with Λ_R defined such as

$$\Lambda_R := \sum_{i=1}^{K} M_i\lambda_{2,i}/(V\mu_{2,i}).$$

It means that F_R follows a binomial distribution $\mathcal{B}\left(\phi, N\right)$ for which $\phi = (1+\Lambda_R)^{-1}$ and $N = N_R$.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Acar, M., Mettetal, J. T., and van Oudenaarden, A. (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4):471–475.

Arias, A. M. and Hayward, P. (2006). Filtering transcriptional noise during development: concepts and mechanisms. *Nature Reviews Genetics*, 7(1):34–44.

Austin, D. W., Allen, M. S., McCollum, J. M., Dar, R. D., Wilgus, J. R., Sayler, G. S., Samatova, N. F., Cox, C. D., and Simpson, M. L. (2006). Gene network shaping of inherent noise spectra. *Nature*, 439(7076):608–611.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., and Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science (New York, N.Y.)*, 305(5690):1622–1625.

Balázsi, G., van Oudenaarden, A., and Collins, J. J. (2011). Cellular Decision-Making and Biological Noise: From Microbes to Mammals. *Cell*, 144(6):910–925.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643.

Becskei, A. and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593.

Berg, O. G. (1978). A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of theoretical biology*, 71(4):587–603.

Billingsley, P. (1999). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, New York, 2nd edition edition.

Blake, W. J., Kærn, M., Cantor, C. R., and Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462.

Bokes, P., King, J. R., Wood, A. T. A., and Loose, M. (2011). Multiscale stochastic modelling of gene expression. *Journal of Mathematical Biology*, 65(3):493–520.

Borkowski, O., Goelzer, A., Schaffer, M., Calabre, M., Mäder, U., Aymerich, S., Jules, M., and Fromion, V. (2016). Translation elicits a growth rate-dependent, genome-wide, differential protein production in Bacillus subtilis. *Molecular Systems Biology*, 12(5):870.

Bremer, H. and Dennis, P. P. (1996). Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, 2 edition.

Cai, L., Friedman, N., and Xie, X. S. (2006). Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362.

Camas, F. M., Blázquez, J., and Poyatos, J. F. (2006). Autogenous and nonautogenous control of response in a genetic network. *Proceedings of the National Academy of Sciences*, 103(34):12718–12723.

Collins, J. F. and Richmond, M. H. (1962). Rate of Growth of Bacillus cereus Between Divisions. *Journal of General Microbiology*, 28(1):15–33.

Dawson, D. (1993). Measure-valued markov processes. In *École d'été de probabilités de Saint-Flour XXI-1991*, pages 1–260. Springer.

Deutscher, M. P. (2006). Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research*, 34(2):659–666.

Donachie, W. D. (1968). Relationship between Cell Size and Time of Initiation of DNA Replication. *Nature*, 219(5158):1077–1079.

Dublanche, Y., Michalodimitrakis, K., Kümmerer, N., Foglierini, M., and Serrano, L. (2006). Noise in transcription negative feedback loops: simulation and experimental analysis. *Molecular Systems Biology*, 2(1):41.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186.

Fournier, T., Gabriel, J. P., Mazza, C., Pasquier, J., Galbete, J. L., and Mermod, N. (2007). Steady-state expression of self-regulated genes. *Bioinformatics*, 23(23):3185–3192.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J., and Eisen, M. B. (2004). Noise Minimization in Eukaryotic Gene Expression. *PLOS Biol*, 2(6):e137.

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.

Fromion, V., Leoncini, E., and Robert, P. (2013). Stochastic gene expression in cells: a point process approach. *SIAM Journal on Applied Mathematics*, 73(1):195–211.

Fromion, V., Leoncini, E., and Robert, P. (2015). A Stochastic Model of the Production of Multiple Proteins in Cells. *SIAM Journal on Applied Mathematics*, 75(6):2562–2580.

George, W. K. J., Beuther, P. D., and Lumley, J. L. (1978). Processing of Random Signals. In M.Sc, B. W. H., editor, *Proceedings of the Dynamic Flow Conference 1978 on Dynamic Measurements in Unsteady Flows*, pages 757–800. Springer Netherlands.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Goelzer, A., Fromion, V., and Scorletti, G. (2011). Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6):1210–1218.

Goldberger, R. F. (1979). Strategies of Genetic Regulation in Prokaryotes. In Goldberger, R. F., editor, *Biological Regulation and Development*, number 1 in Biological Regulation and Development, pages 1–18. Springer US. DOI: 10.1007/978-1-4684-3417-0_1.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036.

Grant, M. A., Saggioro, C., Ferrari, U., Bassetti, B., Sclavi, B., and Lagomarsino, M. C. (2011). DnaA and the timing of chromosome replication in Es-cherichia coli as a function of growth rate. *BMC Systems Biology*, 5(1):201.

Halford, S. E. (2009). An end to 40 years of mistakes in DNA–protein association kinetics? *Biochemical Society Transactions*, 37(2):343.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598.

Hilfinger, A. and Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–12172.

Hornos, J. E. M., Schultz, D., Innocentini, G. C. P., Wang, J., Walczak, A. M., Onuchic, J. N., and Wolynes, P. G. (2005). Self-regulating gene: An exact solution. *Physical Review E*, 72(5):051907.

Innocentini, G. C. P. and Hornos, J. E. M. (2007). Modeling stochastic gene expression under repression. *Journal of Mathematical Biology*, 55(3):413–431.

Jacod, J. and Shiryaev, A. N. (1987). *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Jansen, M. (2014). Stochastic equations for a self-regulating gene. *arXiv:1403.3265 [math, q-bio]*.

Kabata, H., Kurosawa, O., Arai, I., Washizu, M., Margarson, S. A., Glass, R. E., and Shimamoto, N. (1993). Visualization of single molecules of RNA polymerase sliding along DNA. *Science*, 262(5139):1561–1563.

Kaczanowska, M. and Rydén-Aulin, M. (2007). Ribosome Biogenesis and the Translation Process in Escherichia coli. *Microbiology and Molecular Biology Reviews*, 71(3):477–494.

Kærn, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews. Genetics*, 6(6):451–464.

Kalisky, T., Dekel, E., and Alon, U. (2007). Cost–benefit theory and optimal design of gene regulation functions. *Physical Biology*, 4(4):229.

Keilson, J. (1974). Markov chain models–rarity and exponentiality. Technical report, DTIC Document.

Kingman, J. F. C. (1993). *Poisson Processes*. Number 3 in Oxford Studies in Probability. Oxford University Press, Oxford.

Klumpp, S. and Hwa, T. (2008). Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–20250.

Koch, A. L. and Levy, H. R. (1955). Protein Turnover in Growing Cultures of Escherichia Coli. *Journal of Biological Chemistry*, 217(2):947–958.

Kurtz, T. G. (1992). Averaging for martingale problems and stochastic approximation. In *Applied Stochastic Analysis*, pages 186–209. Springer.

Leoncini, E. (2013). *Towards a global and systemic understanding of protein production in prokaryotes*. PhD thesis, Ecole Polytechnique X.

Li, G.-W. and Elf, J. (2009). Single molecule approaches to transcription factor kinetics in living cells. *FEBS Letters*, 583(24):3979–3983.

Madan Babu, M. and Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, 31(4):1234–1244.

Maloy, S. and Stewart, V. (1993). Autogenous regulation of gene expression. *Journal of Bacteriology*, 175(2):307–316.

Marr, A. G. (1991). Growth rate of Escherichia coli. *Microbiological Reviews*, 55(2):316–333.

Mather, W. H., Hasty, J., Tsimring, L. S., and Williams, R. J. (2013). Translational Cross Talk in Gene Networks. *Biophysical Journal*, 104(11):2564–2572.

McAdams, H. H. and Arkin, A. (1997). Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819.

Miller, C. G. (1996). Protein Degradation and Proteolytic Modification. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, 2 edition edition. chapter 62.

Miller, S. and Childers, D. (2012). *Probability and Random Processes: With Applications to Signal Processing and Communications*. Academic Press.

Neidhardt, F. C. and Umbarger, H. E. (1996). Chemical Composition of Escherichia coli. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, 2 edition edition. chapter 3.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846.

Novick, A. and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7):553–566.

Osella, M., Nugent, E., and Lagomarsino, M. C. (2014). Concerted control of Escherichia coli cell division. *Proceedings of the National Academy of Sciences*, 111(9):3431–3435.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73.

Papanicolaou, G. C., Stroock, D., and Varadhan, S. S. (1977). Martingale approach to some limit theorems. In *Duke Turbulence Conference (Duke Univ., Durham, NC, 1976), Paper*, volume 6.

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

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175.

Pedraza, J. M. and Oudenaarden, A. v. (2005). Noise Propagation in Gene Networks. *Science*, 307(5717):1965–1969.

Raj, A. and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226.

Raser, J. M. and O'Shea, E. K. (2004). Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 304(5678):1811–1814.

Rigney, D. R. and Schieve, W. C. (1977). Stochastic model of linear, continuous protein synthesis in bacterial populations. *Journal of Theoretical Biology*, 69(4):761–766.

Robert, L., Hoffmann, M., Krell, N., Aymerich, S., Robert, J., and Doumic, M. (2014). Division in Escherichia coli is triggered by a size-sensing rather than a timing mechanism. *BMC Biology*, 12(1):17.

Robert, P. (2010). *Stochastic networks and queues*. Springer, Berlin; New York.

Rogers, L., Williams, D., and Wiley, J. (1987). Diffusions, Markov processes and martingales, vol 2: Ito calculus.

Rosenfeld, N., Elowitz, M. B., and Alon, U. (2002). Negative Autoregulation Speeds the Response Times of Transcription Networks. *Journal of Molecular Biology*, 323(5):785–793.

Rudin, W. (1986). Real and complex analysis (3rd).

Russell, J. B. and Cook, G. M. (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiological Reviews*, 59(1):48–62.

Savageau, M. A. (1974). Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, 252(5484):546–549.

Schleif, R. (2000). Regulation of the L-arabinose operon of Escherichia coli. *Trends in genetics: TIG*, 16(12):559–565.

Schrödinger, E. (1944). *What is Life?: With Mind and Matter and Autobiographical Sketches*. Cambridge University Press, Cambridge ; New York, reprint edition.

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261.

Sharpe, M. E., Hauser, P. M., Sharpe, R. G., and Errington, J. (1998). Bacillus subtilis cell cycle as studied by fluorescence microscopy: constancy of cell length at initiation of DNA replication and evidence for active nucleoid partitioning. *Journal of Bacteriology*, 180(3):547–555.

Singh, A., Razooky, B. S., Dar, R. D., and Weinberger, L. S. (2012). Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. *Molecular Systems Biology*, 8:607.

Siwiak, M. and Zielenkiewicz, P. (2013). Transimulation - Protein Biosynthesis Web Service. *PLOS ONE*, 8(9):e73943.

Soifer, I., Robert, L., Barkai, N., and Amir, A. (2014). Single-cell analysis of growth in budding yeast and bacteria reveals a common size regulation strategy. *arXiv:1410.4771 [cond-mat, q-bio]*. arXiv: 1410.4771.

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.

Süel, G. M., Garcia-Ojalvo, J., Liberman, L. M., and Elowitz, M. B. (2006). An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(7083):545–550.

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.

Thattai, M. and van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530.

Tyson, J. J. and Diekmann, O. (1986). Sloppy size control of the cell division cycle. *Journal of Theoretical Biology*, 118(4):405–426.

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.

Walker, N., Nghe, P., and Tans, S. J. (2016). Generation and filtering of gene expression noise by the bacterial cell cycle. *BMC Biology*, 14:11.

Wallden, M., Fange, D., Ullman, G., Marklund, E. G., and Elf, J. (2015). Fluctuations in growth rates determine the generation time and size distributions of E. coli cells. *arXiv:1504.03145 [q-bio]*. arXiv: 1504.03145.

Wang, P., Robert, L., Pelletier, J., Dang, W. L., Taddei, F., Wright, A., and Jun, S. (2010). Robust growth of Escherichia coli. *Current biology: CB*, 20(12):1099–1103.

Warner, J. R., Vilardell, J., and Sohn, J. H. (2001). Economics of Ribosome Biosynthesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 66:567–574.

Weiss, J. N. (1997). The hill equation revisited: uses and misuses. *The FASEB Journal*, 11(11):835–841.

Yildirim, N. and Mackey, M. C. (2003). Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophysical Journal*, 84(5):2841–2851.

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, X. S. (2006). Probing Gene Expression in Live Cells, One Protein Molecule at a Time. *Science*, 311(5767):1600–1603.

Yvinec, R., Zhuge, C., Lei, J., and Mackey, M. C. (2013). Adiabatic reduction of a model of stochastic gene expression with jump Markov process. *Journal of Mathematical Biology*, 68(5):1051–1070.

Zhou, J. and Rudd, K. E. (2013). EcoGene 3.0. *Nucleic Acids Research*, 41(Database issue):D613–624.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modèles stochastiques pour la production des protéines : l'impact de l'autorégulation, du cycle cellulaire et des interactions entre les productions de protéines sur l'expression génétique

Keywords : Expression génétique, modèle stochastique, production des protéines, autorégulation

Résumé : Le mécanisme de production des protéines, qui monopolise la majorité des ressources d'une bactérie, est hautement stochastique : chaque réaction biochimique qui y participe est due à des collisions aléatoires entre molécules, potentiellement présentes en petites quantités. La bonne compréhension de l'expression génétique nécessite donc de recourir à des modèles stochastiques qui sont à même de caractériser les différentes origines de la variabilité dans la production ainsi que les dispositifs biologiques permettant éventuellement de la contrôler.

Dans ce contexte, nous avons analysé la variabilité d'une protéine produite avec un mécanisme d'autorégulation négatif : c'est-à-dire dans le cas où la protéine est un répresseur pour son propre gène. Le but est de clarifier l'effet de l'autorégulation sur la variance du nombre de protéines exprimées. Pour une même production moyenne de protéine, nous avons cherché à comparer la variance à l'équilibre d'une protéine produite avec le mécanisme d'autorégulation et celle produite en « boucle ouverte ». En étudiant un modèle limite, avec une mise à l'échelle (scaling), nous avons pu faire une telle comparaison de manière analytique. Il apparaît que l'autorégulation réduit effectivement la variance, mais cela reste néanmoins limité : un résultat asymptotique montre que la variance ne pourra pas être réduite de plus de 50%. L'effet sur la variance à l'équilibre étant modéré, nous avons cherché un autre effet possible de l'autorégulation : nous avons observé que la vitesse de convergence à l'équilibre est plus rapide dans le cadre d'un modèle avec autorégulation.

Les modèles classiques de production des protéines considèrent un volume constant, sans phénomènes de division ou de réplication du gène, avec des ARN-polymérases et les ribosomes en concentrations constantes. Pourtant, les variation au cours du cycle de chacune de ces quantités a été proposée dans la littérature comme participant à la variabilité des protéines. Nous proposons une série de modèles de complexité croissante qui vise à aboutir à une représentation réaliste de l'expression génétique. Dans un modèle avec un volume suivant le cycle cellulaire, nous intégrons successivement le mécanisme de production des protéines (transcription et traduction), la répartition aléatoire des composés à la division et la réplication du gène. Le dernier modèle intègre enfin l'ensemble des gènes de la cellule et considère leurs interactions dans la production des différentes protéines à travers un partage commun des ARN-polymérases et des ribosomes, présents en quantités limitées. Pour les modèles où cela était possible, la moyenne et la variance de la concentration de chacune des protéines ont été déterminées analytiquement en ayant eu recours au formalisme des Processus Ponctuels de Poisson Marqués. Pour les cas plus complexes, nous avons estimé la variance au moyen de simulations stochastiques. Il apparaît que, dans l'ensemble des mécanismes étudiés, la source principale de la variabilité provient du mécanisme de production des protéines lui-même (bruit dit « intrinsèque »). Ensuite, parmi les autres aspects « extrinsèques », seule la répartition aléatoire des composés semble avoir potentiellement un effet significatif sur la variance ; les autres ne montrent qu'un effet limité sur la concentration des protéines. Ces résultats ont été confrontés à certaines mesures expérimentales, et montrent un décalage encore inexpliqué entre la prédiction théorique et les données biologiques, ce qui appelle à de nouvelles hypothèses quant aux possibles sources de variabilité.

En conclusion, les processus étudiés ont permis une meilleure compréhension des phénomènes biologiques en explorant certaines hypothèses difficilement testables expérimentalement. Des modèles étudiés, nous avons pu dégager théoriquement certaines tendances, montrant que la modélisation stochastique est un outil important pour la bonne compréhension des mécanismes d'expression génétique.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Title: Stochastic models for protein production: the impact of autoregulation, cell cycle and protein production interactions on gene expression

Keywords: Gene expression, stochastic model, protein production, autoregulation

Abstract: The mechanism of protein production, to which is dedicated the majority of resources of the bacteria, is highly stochastic: every biochemical reaction that is involved in this process is due to random collisions between molecules, potentially present in low quantities. The good understanding of gene expression requires therefore to resort to stochastic models that are able to characterise the different origins of protein production variability as well as the biological devices that potentially control it.
In this context, we have analysed the variability of a protein produced with a negative autoregulation mechanism: i.e. in the case where the protein is a repressor of its own gene. The goal is to clarify the effect of this feedback on the variance of the number of produced proteins. With the same average protein production, we sought to compare the equilibrium variance of a protein produced with the autoregulation mechanism and the one produced in "open loop". By studying the model under a scaling regime, we have been able to perform such comparison analytically. It appears that the autoregulation indeed reduces the variance; but it is nonetheless limited: an asymptotic result shows that the variance won't be reduced by more than 50%. The effect on the variance being moderate, we have searched for another possible effect for autoregulation: it havs been observed that the convergence to equilibrium is quicker in the case of a model with autoregulation.
Classical models of protein production usually consider a constant volume, without any division or gene replication and with constant concentrations of RNA-polymerases and ribosomes. Yet, it has been suggested in the literature that the variations of these quantities during the cell cycle may participate to protein variability. We propose a series of models of increasing complexity that aims to reach a realistic representation of gene expression. In a model with a changing volume that follows the cell cycle, we integrate successively the protein production mechanism (transcription and translation), the random segregation of compounds at division, and the gene replication. The last model integrates then all the genes of the cell and takes into account their interactions in the productions of different proteins through a common sharing of RNA-polymerases and ribosomes, available in limited quantities. For the models for which it was possible, the mean and the variance of the concentration of each proteins have been analytically determined using the Marked Poisson Point Processes. In the more complex cases, we have estimated the variance using computational simulations. It appears that, among all the studied mechanisms, the main source of variability comes from the protein production mechanism itself (referred as "intrinsic noise"). Then, among the other "extrinsic" aspects, only the random segregation of compounds at division seems to have potentially a significant impact on the variance; the other aspects show only a limited effect on protein concentration. These results have been confronted to some experimental measures, and show a still unexplained decay between the theoretical predictions and the biological data; it instigates the formulations of new hypotheses for other possible sources of variability.
To conclude, the processes studied have allowed a better understanding of biological phenomena by exploring some hypotheses that are difficult to test experimentally. In the studied models, we have been able to indicate theoretically some trends; hence showing that the stochastic modelling is an important tool for a good understanding of gene expression mechanisms.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France