Towards a global and systemic understanding of protein production in prokaryotes

Emanuele Leoncini

École Doctorale de l'École Polytechnique

Thèse présentée en vue de l'obtention du grade de
Docteur de l'École Polytechnique
Spécialité Mathématiques Appliquées

Towards a global and systemic understanding of protein production in prokaryotes

Emanuele Leoncini

Vincent Fromion	INRA	supervisor
Philippe Robert	INRIA	supervisor
Loïc Chaumont	Université d'Angers	reviewer
Patricia Reynaud-Bouret	CNRS	reviewer
Jean Bérard	Université Lyon 1	jury member
Olivier Martin	Université Paris Sud	jury member
Lydia Robert	UPMC	jury member

Remerciements

Je tiens tout d'abord à remercier Vincent Fromion et Philippe Robert pour avoir accepté de m'accueillir dans leurs équipes et d'encadrer ma thèse. Je les remercie pour le sujet de recherche passionnant qu'ils m'ont proposé et pour m'avoir guidé dans le monde fascinant des mathématiques appliquées à la biologie. Ils ont su me laisser une grande liberté dans mon travail et m'ont guidé dans le développement de mon indépendance et de mon esprit critique.

Je remercie Philippe pour son exigence scientifique, son esprit critique et ses conseils, qui m'ont motivé quand j'ai été confronté à des problèmes difficiles.

Vincent, d'autre part, a su me donner le plaisir de l'application et m'a aidé à m'approprier le problème biologique. Sa vision claire du système biologique m'a toujours inspiré et poussé à ne jamais perdre de vue la question d'origine et à transférer les résultats obtenus théoriquement dans une perspective biologique.

Tous mes remerciements vont à Patricia Reynaud-Bouret et Loïc Chaumont qui ont accepté d'être rapporteurs de ma thèse. Je suis également très honoré que Jean Bérard, Olivier Martin et Lydia Robert aient accepté de faire partie de mon jury.

Je remercie Hugues Berry et Vincent Calvez pour m'avoir invité et accueilli lorsque du colloque "Cell Biology" à Lyon. Je remercie également Olivier Gandrillon pour les échanges et les discussions scientifiques que j'ai eus avec lui, ainsi que Fabien Crauste pour sa disponibilité et sympathie. Je tiens à remercier Amandine Veber pour les échanges scientifiques et sa disponibilité et Jonathan Touboul pour les précieux conseils.

Je remercie tous les membres de l'équipe RAP-Inria, Nicolas Broutin, Virginie Collette, Renaud Dessalles, Sarah Eugène, Mathieu Feuillet, Christine Fricker, Jelena Pesic, Jim Roberts, Nada Sbihi, Henning Sulzbach et Philippe, pour la belle ambiance qui m'a accompagné pendant ces trois ans. Je suis heureux de les avoir initiés au plaisir du bon café (ristretto). Un grand merci à Mathieu pour avoir supporté un tel italien bavard, pour nos longues discussions et pour les greatest hits allemands des années 80s. Merci à Virginie pour son aide, sa patience lorsque je devais me préparer pour une mission et sa bonne humeur communicative. Merci à Christine pour les discussions culturelles et sa gentillesse. Merci à Jim pour ses nombreux conseils et pour son English aplomb lorsqu'il m'écoutait présenter mes travaux. Merci à Jelena, Sarah et Henning pour avoir ravivé l'atmosphère et pour les moments partagés; je regrette de ne les avoir rencontrés que pendant ma troisième année de thèse.

Je remercie tous les membres permanents, les doctorants, les post-doc et les stagiaires de l'équipe MIG-INRA pour leur accueil et et les débats autour des sucreries faites-maison au coin café. Je me suis senti comme à la maison parmi tous ces amateurs de café. Parmi tous, je tiens à remercier Anne Goelzer et Laurent Tournier pour leur amabilité et leur soutien. Je tiens aussi à remercier Juliette Degrouard pour m'avoir aidé dans les différentes démarches administratives au sein de l'INRA, Eric Montaudon pour son aide sur le matériel informatique et Véronique Martin pour nos échanges "routieres". Merci aussi à Olivier Borkowski pour sa sympathie débordante et à Pierre Flores, tous les trois profitant de l'encadrement vivant de Vincent.

Un merci aux amis et collègues du Bâtiment 16 où j'ai commencé mon expérience à Inria par un stage sous la direction de Dirk Drasdo, qui je remercie également. Merci aussi à Gregory Batt et aux autres membres de l'équipe Contraintes pour nos échanges scientifiques. Merci à l'équipe Gallium pour m'avoir accueilli comme "membre honoraire", grâce aussi à Jonathan Protzenko et Alexandre Pilkiewicz, et pour m'avoir initié aux mystères de l'informatique.

Un remerciement particulier à Jonathan avec qui j'ai vécu côte à côte tout au long de ces trois ans : merci pour le sport, les cafés, le soutien informatique, les restos chinois, les fipettes et les soirées en compagnie de Gina. Grazie di cuore!

Merci à tous les coureurs et nageurs pour avoir partagé de bons moments pour se déstresser des joies de la thèse : Jonathan, Mathieu, Thierry, Pauline, Saverio, Luna, Damiano, Elisa,

Philippe... Un remerciement particulier à la Uno de Mathieu, bagnole officielle de l'équipe de natation Inria.

Merci aux amis avec lesquels j'ai eu occasion de partager de beaux moments au cours de ces dernières années, souvent autour d'une table bien garnie de gourmandises (surtout italiennes). Un grand merci à Giacomo, maître incontesté de style et "d'astio", et à Alicia. Merci à Filippo et Clara pour les inoubliables dîners sur l'avenue Jean Jaures : je te serai toujours reconnaissant de m'avoir appris les secrets de l'Amatriciana. Merci à Luna et Nick, Paris n'est plus la même ville sans vous ! Merci à Daniele e Raphaëlle : "love is like that, like a fifty fifty". Merci à Daniel, ami d'aventures entre Chevaleret et Jussieu. Merci Anne-Lise pour sa touche parisienne et les saveurs du Sud-Ouest. Grazie a Giustino e Marco per le serate passate fra pacifici orsacchiotti e mazurche d'altri tempi, grazie anche a Cristina e Irene per portare un po' di sana toscanità tra gli indigeni. Grazie a Carolina e Ciccio, dal nostro incontro fortuito a quel crocicchio nel 5^o arondissement ai vicoli di Locorotondo.

Ringrazio i miei genitori per avermi incoraggiato negli studi e aver sostenuto le mie scelte, anche quando mi hanno portato lontano da casa. Grazie per avermi insegnato ad essere esigente e responsabile. Grazie per aver sempre creduto in me.

Un pensiero speciale a Yasmine, grazie per avermi spinto a partire e per avermi raggiunto, per essermi stata vicina e avermi sostenuto, in particolare in questo difficile terzo anno di dottorato. Solo io e te.

Résumé

Les réactions biochimiques sous-jacentes au fonctionnement des cellules sont des processus intrinsèquement stochastiques. En conséquence, le fonctionnement de la cellule, considérée comme un système, est aléatoire en raison des fluctuations de ses composantes fondamentales. Parmi ces dernières se trouvent les protéines, qui jouent un rôle majeur dans les cellules. Le caractère stochastique des protéines est tel qu'il est même responsable des différences observées dans le phénotype et ce même dans le cas de cellules clonées exposées à des conditions environnementales identiques. La grande difficulté rencontrée dans le développement de techniques quantitatives fiables pour la mesure des fluctuations de l'expression génétique au niveau cellulaire a favorisé le développement et l'utilisation de modèles stochastiques essayant de capturer les principales caractéristiques du système. Il est donc crucial que les modèles adoptent des hypothèses réalistes, afin de pouvoir les utiliser comme un véritable outil d'investigation.

Dans ce travail de thèse nous avons mis en place un nouveau cadre mathématique basé sur les Processus Ponctuels de Poisson Marqués (MPPP) pour décrire les principales étapes de la production d'une protéine spécifique, grâce à une analogie entre le système de production de protéines et les réseaux de files d'attente. Cette approche s'est avérée être très adaptée à la tâche, car elle permet de considérer des hypothèses générales pour certaines étapes, tout en gardant le caractère analytique des modèles présents dans la littérature, qui se réduisent à un cas particulier de cette approche générale pour des hypothèses spécifiques. Avec ce cadre, nous avons réussi à surmonter l'hypothèse fondamentale et restrictive des modèles classiques, qui exige une durée exponentielle de toutes les étapes. La description non-markovienne de l'expression génétique obtenue grâce à ce nouveau cadre a permis d'aborder le problème d'une manière plus satisfaisante et, en particulier, de proposer un modèle plus réaliste qui comprend des hypothèses réalistes de l'étape d'élongation de la protéine et de la dilution des protéines en raison de la croissance du volume. Eu égard aux résultats obtenus, cette nouvelle approche a montré que les modèles classiques ont su capturer qualitativement les caractéristiques du bruit dans l'expression d'un gène, mais leur pouvoir de prédiction est limité par le fait que les formules quantitatives obtenues sont incorrectes. L'utilisation des MPPP a permis de d'évaluer l'impact de différents choix de modélisation, tout en gardant la capacité d'obtenir des formules analytiques pour les premiers moments des distributions des différents processus en fonction des paramètres biophysiques.

La modélisation du processus de production d'une seule protéine, bien que puissante, ne prend pas en compte la description des interactions qui peuvent se produire en raison de la production simultanée des différents types de protéines. Pour cette raison, nous avons proposé une première modélisation de la production de plusieurs protéines en considérant les interactions comme le résultat de la compétition pour des ressources communes. Plus précisément, les ribosomes sont responsables de la traduction de tous les types de messagers et leur nombre est limité et strictement contrôlé dans la cellule. En pratique, le coût élevé, en termes de ressources, associé à la production des ribosomes oblige la cellule à optimiser leur usage et il s'avère qu'ils sont presque toujours en train de traduire des messagers et restent très peu inactifs après l'achèvement de leur tâche. En conséquence, il y a une rude compétition entre les messagers pour avoir accès à des ribosomes libres et la production globale en est affectée. Le système de production est étudié par une approche de champ moyen à la fois dans les régimes de sous-charge et de surcharge d'utilisation des ribosomes. Le modèle multi-protéines est une approche novatrice dans le

domaine qui ouvre une nouvelle direction dans l'étude des fluctuations des protéines au niveau cellulaire.

En conclusion, la thèse a porté sur l'étude de la nature stochastique de l'expression génétique, en développant différents modèles afin de progresser vers une description plus réaliste des phénomènes. Toutes ces études ont été menées en essayant de mettre la biologie au premier plan, car nous croyons que ces modèles représentent un outil fondamental dans l'étude et la compréhension des processus biologiques complexes.

Abstract

Biochemical reactions underlying the functioning of cells are inherently stochastic processes. As a consequence, the whole system is noisy and undergoes fluctuations in its fundamental components. Proteins are major players for the living. The behavior of cells as well as their stochastic character manifests itself via striking differences in phenotype. The differences are apparent even in the case of identical, cloned cells which underwent the same environmental conditions. The extreme difficulty in obtaining reliable experimental quantitative results concerning fluctuations in gene expression has fostered the development and intense use of stochastic models. These models aim at capturing the main characteristics of the system. Given the context, it is crucial that models embrace realistic assumptions.

We introduced a new mathematical framework based on Marked Poisson Point Processes (MPPP) to describe the main steps of the production of a specific protein. We leveraged the similarities between the protein production system and queueing networks. This approach has proven to be perfectly suited, since it allowed us to consider general assumptions, while still permitting the derivation of analytical formulas. This is one of the key features of the models found in literature. Furthermore, the classic models are incorporated in this approach and well-known results can be obtained for specific assumptions. This has been possible, since we were able to overcome the restrictive assumption, crucial in classic framework, of exponentially distributed duration of all steps. The non-Markovian description of gene expression obtained through this new framework has permitted to tackle the problem in a more satisfying way and, in particular, propose a more realistic model of gene expression, which includes realistic assumptions of protein elongation step and protein dilution due to volume growth. Such a realistic model shows that if on one hand the classic models have captured the qualitative behavior of the underlying biological processes, on the other hand their quantitative results might have been inaccurate, resulting in a limited predictive power. The MPPP framework has proved to serve as a testing platform, allowing to quantify precisely the impact of different modeling choices, while keeping intact the ability to obtain analytical formulas of statistics depending on the biophysical parameters.

The single-protein modelisation, although powerful, fails to describe the possible interactions deriving from the simultaneous production of different types of proteins. For this reason, we moved the first steps towards a modelisation of the production of many proteins, considering interactions as the result of the competition for common resources. In particular, ribosomes translate any type of messengers and turn to be present in limited and strictly controlled number within a cell. In fact, the high cost in term of resources associated with the production of ribosomes forces the system to have almost always active ribosomes translating messengers into proteins. As a consequence, messengers compete against each other for the rare resource of free ribosomes and the global production is affected. The system of production is studied via a mean-field approach both in the underloaded and overloaded regimes of use of the ribosomes. The multi-protein model brings a completely new approach in the domain and marks a new direction in the investigation of protein fluctuations at the cellular level.

In conclusion, the thesis has focused on the study of the stochastic nature of gene expression, by developing different models in order to progress towards a more realistic description of the phenomena. All these studies have been conducted trying to put biology in the foreground, since we believe these models represent a fundamental step in the investigation and understanding of complex biological processes and are a complementary tool to biological experiments.

Contents

Chapter 1

Introduction

> Aussi la biologie est-elle, pour l'homme, la plus signifiante de toutes les sciences; celle qui a déjà contribué, plus que toute autre sans doute, à la formation de la pensée moderne, profondément bouleversée et définitivement marquée dans tous les domaines [...]
>
> *Le hasard et la nécessité*
> Jacques Monod, 1970

Nicolaus Copernicus proposed, in his *De revolutionibus orbium coelestium*, an heliocentric model for the celestial objects. This theory led to a deep wound in the narcissistic simulacrum that men built around the Ptolemaic model by putting away the man from the center of the Universe.

The chance is then found to be responsible of the mutations and so of the evolution itself. If the Theory of Evolution was well established and phenomenologically accepted since the end of the XIX century, it required a physical theory of heredity in order to have a deeper acceptance and solid foundations. This is the aim of the *molecular theory of the genetic code*, which tries to connect concepts related to the chemical structure of the genetic material, the information it contains and the molecular mechanisms for the morphogenetic and physiologic expression of this information. The *molecular biology* has further pushed the man out of the center of the Universe, redefining life in terms of molecular interactions.

In 1953, Watson and Crick proposed their helicoidal model of DNA structure, which, together with subsequent works, has given insights on how genetic information is stocked in cells and how it is possible to pass this information from a generation to the next. If DNA contains the morphogenetic and functional information of an individual, the fulfillment of the cell project is accomplished through the *gene expression*. The products of gene expression mainly consist of proteins and functional RNAs in the case of non-protein coding genes such as *ribosomal RNA* (rRNA) or *transfer RNA* (tRNA). The protein production is the central topic of this PhD work and we will focus on its description via a mathematical modeling.

Since the early works of the 50's, major advancements have taken place in the understanding of gene expression leading to an extremely detailed biochemical model of the underlying processes. However, if the description of the system is impressively highly detailed, there are still open questions on the fundamental mechanisms, urging a systemic analytic description in order to

identify the laws underlying gene expression. The encounter of different components is the fundamental step common to all elementary biochemical reactions. This, together with the fact that cytoplasm is a stiff disordered medium where different compounds move by diffusion, makes the whole process inherently stochastic. Although the first indirect experimental evidences can be traced back to the end of the 50's [50], stochasticity has been clearly proven experimentally only in recent years [40]. Noise has become a central topic in molecular biology, bearing new challenges in order to reconcile the random biochemical reactions to the precise development of organisms. It is clear that cells and in particular gene expression are robust with respect to fluctuations both in the environment and in its fundamental components. If noise is firstly an obstacle to the accomplishment of the genetic program, in some specific situations fluctuations can be exploited, as in the case of the decision making process. However, despite observable positive or negative consequences of noise, those mechanisms represent a small portion of the whole picture and a characterization of the status of noise is still an open question.

Q: Before this knowledge how can we characterize the underlying stochastic process via a mathematical analysis?

The objective of mathematical biology is to extract information on the system and on its specific mechanisms, via a mathematical description and characterization. Nevertheless, there is a number of difficulties connected with gene expression due to the colossal number of elementary steps and the mixing of deterministic and stochastic processes. Moreover, a complex interaction pathway links different processes both at local and global scale making the analysis more and more difficult. Hence the need to combine elementary steps into effective steps and focus on a part of the system, once the level of detail has been decided and pertinent questions have been formulated, with the necessity to be able to perform the analysis in some mathematical framework.

If on one side experimental evidence and measures on the fluctuations in protein production are recent, nevertheless the problem has been tackled since the end of 70's and a quite large literature has followed in subsequent years. These models of stochastic gene expression have addressed the problem considering effective main steps in the production of a specific protein under reasonable assumptions. These assumptions and this modeling approach have been proposed when the process was not yet well studied and understood. A Markovian description of the system has been then adopted, allowing a simplification of the analysis and leading to derive analytic formulas of the mean and variance of the different processes modeled, but such description comes with a strong assumption, i.e. each step has an exponentially distributed duration. If the obtained theoretical results are consistent with experiments with corresponding effective biophysical parameters, however these models are "semi-quantitative". In fact,if on one side it is possible to find appropriate parameters to fit experimental data, however these models have not been used in a predictive way, i.e. choose parameters beforehand and compare the outcomes with the experiments.

If the Markovian description was adopted to simplify the mathematical analysis of the model, the strong exponential assumption that comes with it has not been discussed in the light of the findings of the last decades. However, in order to investigate the pertinence of the Markovian description and of the strong assumption which comes with, we are forced to leave such classic framework for a more general one. The aim of the thesis work is on one side to revisit the reference model of stochastic gene expression [54] and discuss the assumptions with respect to the knowledge acquired in the last years and, on the other side, to consider the coupling of different proteins. The first objective lead us to introduce a new mathematical framework based on Marked Poisson Point Processes (MPPP), in order to overcome the theoretical limitations of the classic approach, providing a general context to test assumptions and mix processes of different

nature. The second objective is instead connected to the observation that all types of proteins share common cellular machinery, such as *ribosomes*, present in limited numbers. The limited availability of these resources lead to a rude competition between different protein production chains. In particular, we analyze rigorously the impact of the competition of messengers for ribosomes on the resulting fluctuations in protein copies and on the number of free ribosomes.

Presentation of subsequent chapters

Chapter 1: Introduction. In this chapter we give few elements of the complex biology that underlies gene expression. In this perspective, this introductory biology-oriented chapter should allow the non-biologist reader to better follow the next chapters. However, this chapter is not intended to be exhaustive from the biological point of view, since the presentation of the biology is functional to the following chapters. Important experimental results concerning stochasticity in gene expression are then presented as well as a short description of few of the fundamental models in the area. Few key concepts of stochastic gene expression are then recalled.

Chapter 2: MPPP description of gene expression. In this chapter we introduce the MPPP mathematical framework and derive a non-Markovian description of the gene expression of a single protein based on a three-stage model. Analytic formulas of mean and variance of the various modeled processes at equilibrium are derived under general assumptions. Tests in this first approach of possible choices of general steps point towards a counter-intuitive result in terms of protein fluctuations: if the exponential duration of one step is replaced with a deterministic one, the corresponding fluctuations in protein number result increased.

Chapter 3: Realistic model of gene expression. The framework introduced in Chapter 2 is used to build a realistic model of gene expression, the *four-stage* model, including the supplementary step of the protein elongation step and the description of protein dilution due to volume growth. It is then discussed the choice of realistic biological assumptions and the model is therefore specified and analyzed. In particular, the counter-intuitive result of Chapter 2 is recovered when analyzing the impact of protein elongation on the overall fluctuations. Moreover, the formulas derived with an accurate description of protein dilution prove that the formulas in literature are oversimplified and rely possibly on too strong assumptions.

Chapter 4: Multi-protein model. In this chapter the problem of the interactions deriving from the simultaneous production of many proteins types. Unlike the previous chapters, where the amount of free ribosomes is constant, we suppose ribosomes are present in limited number and consider their switching between actively elongating proteins and freely diffusing in the cytoplasm. Messengers compete against each other for the (rare) resource of free ribosomes and the protein production is globally affected. The system of production is studied, by considering a large number of protein types with specific concentrations, via a mean-field approach both in the underloaded and overloaded regimes of use of the ribosomes. In particular, under the realistic biological assumption of overloaded ribosomes, we find that at equilibrium the number of free ribosomes follows a Poisson distribution and the rate of production of each protein type is obtained.

Appendix. Appendix A recall few fundamental definitions and theorems used in the manuscript. Appendix B is devoted to recalls of biology. In particular, Section B.1 gives a more-in-depth description of few fundamental biological mechanisms, while Section B.2 serves as a biological glossary, describing few specific cellular components.

1.1 Gene expression: main mechanisms

The rest of the chapter should introduce the non-biologist reader to the topic of gene expression and its modeling, through a tour of the main biological mechanisms involved and the state-of-the-art of experiments and models. However, this is not intended to be exhaustive and the biological mechanisms described are oversimplified, since they are introduced in the perspective of the mathematical modeling of the next chapters.

In 1953 James D. Watson and Francis Crick published in the journal *Nature* an article [72] in which they expose their model of the structure of the DNA, which featured the anti-parallel double helix held together by hydrogen bonds between pairing nucleotides. In those turbulent years, theoretical biologists proposed models using the partial information at their disposal, years before the first experimental results of molecular genetics were available. All the models and mechanisms proposed were based on the information that could be extracted by indirect observations and this was also the case for Watson and Crick, who used X-ray diffraction images to propose their model.

The Watson-Crick model provided also key insights to explain how genetic information is transferred from a generation to the next and how this information may be spread within the cell. The authors, in the *Nature* article [72], write

> It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

1.1.1 Central Dogma: fifty years of molecular biology

The *central dogma of molecular biology*, that was first stated by Crick in 1958 [11], was then re-stated by the author in 1970 as follows:

> The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

The two main concepts that were produced in the late 50s were those of *sequential information* and of *defined alphabets*. At the time was already known that proteins have a specific three-dimensional configuration, which affects the activity of the protein itself. The researchers decoupled the problem supposing that the amino acid chain was able to fold itself up, reducing the problem to a one dimensional one and allowing to focus just on the assembly of the polypeptide chain. It was well-established at that time that the alphabet of the proteins is composed by twenty amino acids, but it was unknown the mechanisms that lead to their encoding.

At that time it was well known that DNA, RNA and proteins play a leading role in gene expression and the central dogma is a possible solution to the problem consisting in the formulation of general rules for the information transfer from a polymer with a defined alphabet to another one.

Crick represents the flow of detailed sequence information from one chain to the other using arrows, in a schema as in figure 1.1a, where all possible transfers are plotted. The transfers could be divided in three group following the general opinion in the late fifties: those for which that seemed to exist because of direct or indirect evidence, those which have no experimental evidence nor a strong theoretical need and those which were unlikely to exist. Crick carries out a progressive simplification of this scheme excluding first the processes in the last class and validating those more likely to happen.

(a) All possible flows of detailed sequence information in gene expression.

(b) Classification of the transfers by Crick in 1970 [12].

Figure 1.1: All possible flows are showed in figure (a). Figure (b) shows the picture resulting in Crick's paper [12]. The Solid arrows are "general transfers" (first class), dotted arrows are "special transfers" (second class) and the absent arrows are the undetected transfers.

Using the classification made by Crick in 1970 [12], we can draw the schema shown in figure 1.1b. Here the solid arrows represent the "general transfers" (first class), while the dotted arrows are the "special transfers" (second class). The absent arrows are the undetected transfers.

The central dogma has to be read as a negative statement saying that there are no information transfers from protein, stressing out which are the most likely transfers (solid lines) and which are the probable ones (dotted lines). Nevertheless the central dogma does not say anything about the machinery involved and the control mechanisms. It was an attempt to give theoretical insights on the main principles which lead to the expression of a gene, using the partial information available at the time and represents the very foundations of molecular biology.

Experiments have confirmed the correctness of the main principles stated by Crick and new technologies have considerably increased the knowledge on the subject and have given a detailed description of the underlying biochemical reactions. This descriptive approach seems to have no end: finer mechanisms pop up when more accurate techniques are available and take their place in the already complex scenario of gene expression.

Despite extensive researches in the field and the many knowledge acquired, little is known about fundamental mechanisms and strategies underlying protein production, because of the extreme complexity of the whole process and the stochastic nature of the elementary biochemical reactions. For all these reasons, mathematical models represent a tool of investigation, in order to isolate mechanisms and check hypothesis based on the acquired knowledge.

1.1.2 Gene expression: main biological mechanisms

The present section is devoted to a general short introduction of the main steps of gene expression and of the main biological mechanisms which intervene in such complex process. This is not intended to be exhaustive, but to introduce the basic terminology which will be used in the following chapters. Specific biological mechanisms will be introduced through the manuscript when needed.

Despite the Central Dogma gives the fundamental principles of information transfer in gene expression in any cell type, the description of the process via mathematical modeling should take into account the specificity of the cell types. In particular, models need to distinguish between *prokaryotic* and *eukaryotic* cells, at least for specific mechanisms and for their different

geometric organization. This PhD work focuses on *prokaryotes* and the subsequent modelisation is therefore affected. However, we will make clear when a modeling choice is strictly connected with prokaryotes; all other choices must be understood as common to both cell types.

Gene activation

Gene activation is the process which allows a gene to be expressed at a specific time. The way this activation may occur varies a lot from gene to gene and from organism to organism. The main mechanisms causing gene activation are the dissociation of a repressor and the association of an activator.

A *repressor* is a DNA-binding protein that regulates the expression of a specific gene by binding the operator, which is a segment of DNA that a regulator binds to. The binding of the repressor blocks the attachment of RNA polymerase to the *promoter* and prevents the transcription of the genes. If an *inducer* molecule is present, it can interact with the repressor and inhibit its action by detaching it or preventing its binding to the operator.

An *activator* is a DNA-binding protein that regulates one or more genes by increasing the transcription rate. RNA polymerase binds to the *promoter region* of the gene, forming a complex which sometimes proceed to gene transcription. An activator recruits the RNA polymerase to its promoter region.

If the two previous mechanisms are shared between prokaryotic and eukaryotic cells, *chromatin remodeling* is specific to eukaryotes. Chromatin is the complex of DNA and histone proteins with which it associates. Histones are highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called *nucleosomes*. Chromatin on one side serves as a way to condense DNA within the cellular nucleus and, on the other side, as a control of gene expression. Raser and O'Shea [62] hypothesize that chromatin remodeling is the key regulation mechanism for certain eukaryotic promoters.

Gene activation is a complex process resulting from different mechanisms and it is gene and organism specific. Despite genes may show different states, in first approximation it can be described as a two-states process, i.e. the gene may show only two possible states, active or inactive.

The number of copies of a gene within bacteria is a fundamental factor and should be considered in a model describing the expression of a specific gene. When bacteria are growing they duplicate their DNA, that leads to a number of at least two copies per gene, since the genetic information has to be split between daughter cells.

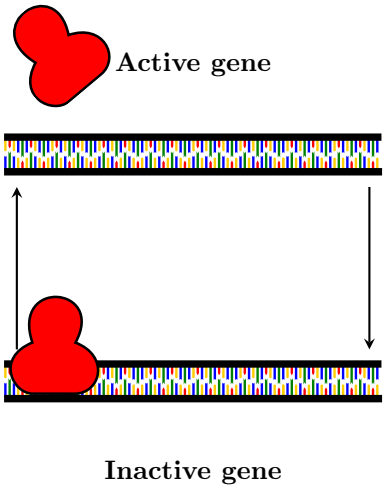


Figure 1.2: Gene activation. When the *repressor* (red cartoon) binds to the gene, it inhibits the mRNA transcription while the gene is activated when the *repressor* unbinds.

Remark. Bacteria are often obliged to have more than two copies of DNA, since the duration of replication ($\sim$ 40 minutes) is sometimes longer than cell cycle time, which is $\sim$ 20 minutes in *Escherichia coli* in fast growth conditions. For this reason, in "normal" growth conditions we observe DNA regions with one, two or four copies of genes, while in "regeneration" regime, where the cell division cycle takes about 20 minutes, we

have up to eight copies of genes localized closer to the origin of replication.

Transcription

The transcription process can be described through the following fundamental steps:

1. **initiation:** the polymerase binds to one of the specificity factors σ to form a "holoenzyme" in order to attach to a specific promoter in the DNA. The more similar is a sequence to a "consensus sequence" the stronger is the binding to the DNA. After the first bond has been synthesized, the RNA polymerase must clear the promoter (this phase is called *promoter clearance*). During this time it may occur that a truncated transcript, called *abortive initiation*, is released;

2. **elongation:** after the promoter clearance, the polymerase assembles in a controlled fashion the mRNA chain;

3. **termination:** the *ρ-independent transcription termination* or the *ρ-dependent transcription termination*. The first involves terminator sequences within the RNA that signals the RNA polymerase to stop. The latter uses the ρ terminator factor to stop RNA synthesis.

For further details we refer to Appendix B.1.2.

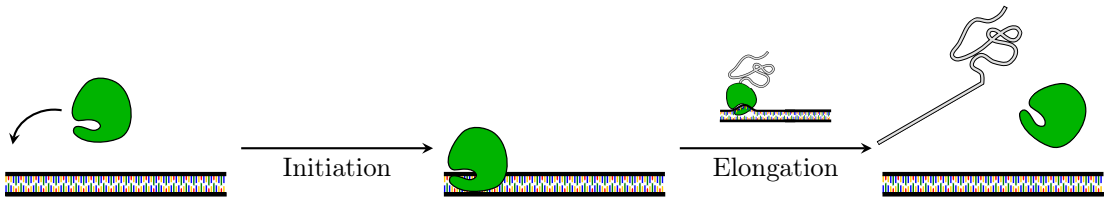


Figure 1.3: **Transcription.** The polymerase binds on the active gene and transcription initiation takes place. Once the initiation step is completed, the polymerase starts copying one DNA strand and elongates the mRNA, which is eventually released into the cytoplasm.

The transcription regulation controls the frequency and the number of produced messengers. The gene transcription is subject to many control mechanisms and we just recall the most common. The *specificity factors* alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them, i.e. sigma factors in prokaryotic transcription. Other regulations are made at gene level and have been enumerated in the previous section. In post-transcriptional phase the regulatory machine controls the number of mRNAs that are translated into proteins. The stability and distribution of the different transcripts is regulated (*post-transcriptional regulation*) by means of RNA binding protein (RBP) that controls the various steps and rates of the transcripts.

Prokaryotic and eukaryotic transcription shows peculiar characteristics. In fact, since there is no precise spatial organization in prokaryotes, translation step can start when the polymerase is still building the messenger. This is not possible in eukaryotes since transcription occurs in the nucleus and, therefore, the messenger needs first to be exported out of the nucleus in order that the translation can take place.

1.1.3 Translation

Schematically prokaryotic translation consists of the following steps (see Figure 1.4 for schematic representation):

1. **initiation:** which involves the assemblage of components such as ribosomal subunits (50S and 30S), mRNA, the first aminoacyl tRNA, GTP (energy) and initiation factors (IF1, IF2, IF3). The tRNA (transfer RNA) serves as the physical link between the nucleotide sequence of mRNA and the amino acid sequence of proteins. In particular, the aminoacyl tRNA (or charged tRNA) carries an amino acid to the ribosome as directed by the three-nucleotide sequence (codon) read by the ribosome. The ribosome has three sites: A, P and E sites. The A site is the entry-point for aminoacyl tRNA, except for the first that binds directly on the P site. In the P site the peptidyl tRNA is formed, i.e. a tRNA bound to the peptide being synthesized, and in the E site the uncharged tRNA detaches from the ribosome;

2. **elongation:** it is a controlled process in which the polypeptide chain is elongated with the addition of amino acids to the carboxyl end of the growing end. Elongation involves several elongation factors, a conformal change, bond formations, etc. The aminoacyl tRNA attaches in the A site, then moves to the P site where the polypeptide is attached to the growing chain and the uncharged tRNA is moved to the E site where exits from the complex;

3. **termination:** occurs when one of three terminating codons moves to the A site. These codons are not recognized by any tRNA but by the so called release factors. These factors trigger hydrolysis of the ester bond and release the newly produced protein in the cytoplasm. The ribosome recycling step is responsible of ribosome disassembly in such a way to be ready to start translation of other messengers.

Translation is carried out by more than one ribosome simultaneously. Because of relative large size of ribosomes, they can only attach sites on mRNA at least 35 nucleotides apart. The so called *polysome* is the complex of one mRNA and a number of ribosomes attached to it.

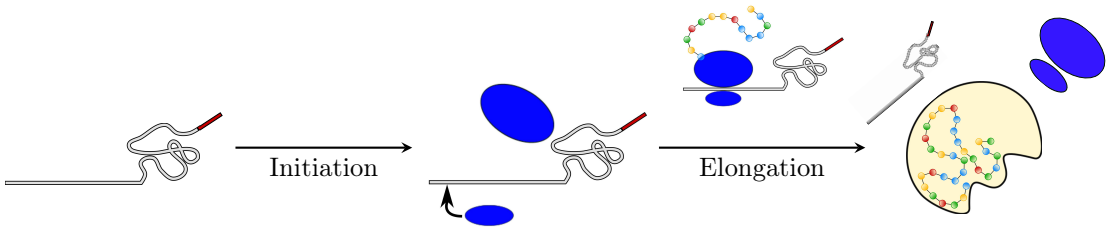


Figure 1.4: **Translation.** The ribosome binds on the messenger and translation initiation takes place, involving a series of controls in order to assure the right progress of translation. Once the initiation step has completed, the ribosome starts the elongation of the protein chain and eventually releases it into the cytoplasm.

The translation of mRNA can also be controlled by a number of mechanisms, mostly at the level of initiation. Recruitment of the small ribosomal subunit can be modulated by mRNA secondary structure, anti-sense RNA binding or protein binding. In both prokaryotes and eukaryotes there is a large number of RNA binding proteins, which are often directed to their target sequence by the secondary structure of the transcript. This structure may change depending on

certain conditions, such as temperature or the presence of a ligand. Moreover, some transcripts act as ribozymes and self-regulate their expression.

mRNA degradation

The process of messenger degradation is an essential function for recycling nucleotides and for regulating the level of gene expression and is performed by *RNase*. The decay process occurs on short time scales, i.e. the typical half-life of a messenger is of about two minutes at 37°C in most cases. This rapid decay process serves to permit to continuously adjust the number of specific messengers to the population needs depending on the specific environmental conditions.

The decay process consists of two main steps [14]:

1. **initiation:** primarily due to endonucleolytic attack mediated by the *RNase E* enzyme in *E. Coli* [41].

2. **break-down:** following the initial endonucleolytic cleavage, which is thought to inactivate the message for translation. Additional cleavages take place and result in breakdown of the mRNA into fragments.

Experiments have shown that prokaryotic mRNAs are more unstable and have shorter lives than in eukaryotes. This is probably connected to the absence of a physical separation between the sites of RNA synthesis and RNA function; decay is possibly the major form of post-transcriptional control in these organisms. The stability of the mRNA is connected also to the competition of *RNases* and ribosomes to bind to messengers [58], i.e. genes with a weak affinity of ribosomes and mRNAs show higher levels of mRNA degradation, since ribosome binding protects the messenger from decay [3, 9, 74].

Protein decay: *proteolysis* and volume dilution

Two main mechanisms of profoundly different nature are responsible of the decay of proteins: *proteolysis* and volume dilution.

The first mechanism is analogous of the degradation mechanism of messengers, but it is now mainly used to destroy possibly dangerous proteins, such as misfolded proteins, since protein's structure determines not only its specific cellular function, but also its intracellular stability. The degradation machinery differs between eukaryotes and prokaryotes, as shown in the review article of Goldberg [21]. Prokaryotes, in particular, have developed an elaborate proteolytic machinery to quickly destroy misfolded proteins. If *protease* is the enzyme that conducts proteolysis, nevertheless, the machinery is much more complex, since if *proteases* were free to act in the cytosol, "they would quickly convert the cell into a bag of amino-acids" [21]. In any case, the proteolysis appears as a control mechanism to prevent the release/survival of malfunctioning proteins or to remove damaged proteins. Actually, proteins are continuously subjected to stress, such as temperature, that eventually causes the protein denaturation. The denatured protein needs to be removed since its functioning has been compromised. This aging phenomenon of proteins occurs on long timescales: the average protein lifetime is usually bigger than the protein cell cycle.

The second mechanism, protein dilution, is of completely different nature. Both prokaryotic and eukaryotic cells double their internal components in order to give rise to two daughter cells. The volume growth associated to the doubling affects the concentration of each cellular component because of volume dilution. Intuitively, if we stop the production of proteins at some point, their concentration will drop down as a consequence of the increase of cell volume. This mechanism is therefore very different from the biochemical interactions which lead to *proteolysis*. Dilution is strictly connected with the cell growth rate and it turns out to be continuous and

deterministic, since growth rate is fixed, as long as environmental conditions are kept unchanged. In normal conditions and for stable proteins, dilution is the leading degradation mechanism.

1.2 Stochasticity in gene expression: experiments

Randomness and determinism are constantly present in the development, growth and life of cells: random biochemical reactions have to be reconciled to the precise development of organisms. The biological implications of the stochastic fluctuations in gene expression has boosted researches in the field, that have multiplied both theoretical and experimental works.

If the stochastic fluctuations were often taken apart by considering statistics on large numbers and reducing the analysis to deterministic models, experimental scientists have become more and more aware of the inherent stochastic nature of the gene expression. Researchers have found that variability among cells in a genetically identical population is strongly connected with fluctuations in the expression of single genes. Stochasticity in the protein production is often just considered as a danger for the normal development of organisms. Nevertheless, some living organisms may exploit the stochastic fluctuations in the expression of genes to introduce phenotypic diversity in genetically identical cells. This variability can be advantageous in specific cases, like face to drastic variations in environment or stress conditions, but it can also be very dangerous when it turns out to be an obstacle to the realization of the cell program.

We focus on experimental results concerning stochasticity in the gene expression, from experimental evidences of the stochastic nature of the phenomenon to negative or positive consequences of fluctuations. Few models are considered here and we refer to Sections 1.3 and 2.A for a detailed description.

In the late 50s Novick and Weiner [50] showed that the production *beta-galactosidase* (β-*galactosidase* or β-*gal*) was variable and random in individual cells, but those studies were hindered by the lack of reliable measures and were not considered conclusive to prove stochasticity in gene expression. One of the first studies which use an expression reporter in single cells was the work of Ko et al.[40] in early 90s. In this work researchers have examined the effect of different doses of glucocorticoid on the expression of the transgene encoding β-*gal* and have found a large cell-to-cell variability by directly measuring the amount of protein in different cells. Moreover, as in Novick's work, increasing the dose does not increase uniformly the expression in every cell, but it increases the frequency of cells displaying high level of expression. The dose dependence has been interpreted by authors as a change in the probability that an individual cell would express the gene at high level, concluding that the gene expression is a stochastic process.

In 2002, Ozbudak et al.[51] studied the fluctuations of gene expressing green fluorescent protein (GFP) driven by an inducible promoter in *Bacillus Subtilis*. The authors tune the rate of transcription by varying the level of induction of the promoter. Translation rate was modulated by introducing mutations in the *ribosome binding site* (RBS). It results that the transcription and translation rates affect the protein fluctuations and the results were interpreted using the theoretical model proposed by Thattai and van Oudernaarden [71]. This model predicts that the protein relative variance depends on the transcription rate, but it remains unchanged because of variations in the rate of translation.

Elowitz et al.[16] introduced the *dual-reporter technique* to measure the stochastic fluctuations of proteins in *Escherichia Coli*. This technique allows to express two different fluorescent proteins, the CFP and YFP, from identical promoters. Since the two proteins share the same regulatory control, the differences between their expression can be attributed to the "intrinsic" stochasticity of the gene expression process, because of the random microscopic events which govern each reaction. On the other side, the "extrinsic" noise, which derives from cellular heterogeneity,

such as regulatory proteins, ribosomes and polymerases, or stochastic events in upstream signal transduction, will affect both proteins. We refer to Section 1.3 for a deeper analysis on these concepts. Authors showed that extrinsic noise represents a non-negligible portion of the overall fluctuations and stressed the necessity to take into account both sources of noise when controlling or minimizing the fluctuations of a system.

Jonathan M. Raser and Erin K. O'Shea [62] used the same technique to study gene expression in yeast. The authors analysed three different promoters in the budding yeast *Saccharomyces cervisiae*: the *PHO5*, *PHO84* and *GAL1* promoters. The total noise on the three promoters was found to be dominated by the contribution of the external factors, such as cell shape and size, cell cycle stage or gene-specific signaling. The authors reduced these possible factors of heterogeneity by using experimental techniques, like *flow cytometry*, used to isolate sub-populations with homogeneous sizes. The extrinsic noise resulted to be diminished, but non dramatically. Moreover the extrinsic noise was found to be not promoter-specific, since it resulted correlated when the two fluorescent proteins were associated with promoters that are distinctly regulated. This leads to hypothesize that the extrinsic noise will cause proteins to be maintained in constant relative concentrations. In order to analyze the noise in eukaryotes, the authors use a three-stage model similar to the model presented by Paulsson [54], see Section 2.A.2 for details. The authors claim the applicability of the three-stage model to both prokaryotes and eukaryotes, the main difference being the specific mechanisms of gene regulations. Relative differences in the parameters can lead to different scenarios which can be biologically interpreted. It can be easily showed that two promoters can produce the same average number of mRNAs with different fluctuation characteristics in this number: a promoter that undergoes frequent activation processes followed by inefficient transcription will show smaller variability with respect to a promoter which has rare activation processes followed by stable active state. The authors find three characteristic regimes of gene regulation, defined in terms of the rates of the three-stage model and which result in different noise profiles. In this paper, extrinsic noise seems to be predominant with respect to intrinsic and seems to be of global nature, i.e. it affects the expression of any gene.

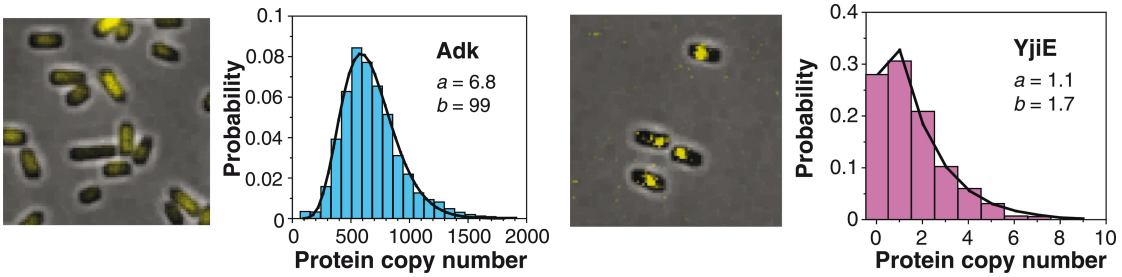
A global analysis of the production of proteins in *Saccharomyces cerevisiae* was conducted independently by Bar et al.[2] and by Newman et al.[48] in 2006. Newman and collaborators [48] studied fluctuations in more than 2500 proteins, using the pairing of high-throughput flow cytometry and a library of GFP-tagged yeast strains to monitor protein levels at single cell resolution. This new strategy for large-scale protein abundance measurements allowed the scientists to deduce that abundance is the major factor governing protein variation, which most likely originates from the stochastic production and destruction of mRNAs. Bar and collaborators [2] studied 43 different proteins under 11 experimental conditions, founding that the variance is roughly proportional to the mean, as predicted by models of stochastic gene expression. Highly expressed genes seem differentiate from this trend since their variance appears uncorrelated with respect to abundance, as showed also in [48]. The researches point to low-copy mRNA fluctuations and gene regulation as the main responsible for protein fluctuations, which is consistent with the scaling property observed. Moreover, using a dual-reporter diploid strain in similar fashion than Elowitz [16], they show that intrinsic noise contributes substantially to the overall protein noise in the case of proteins with intermediate expression level, while it is much smaller than extrinsic fluctuations for highly produced proteins. Both works [48] and [2] stress how proteasome genes are characterized by low noise levels, while stress proteins are very noisy, which indicates a precise structure in protein-specific variation and suggests that noise levels have been selected to reflect costs and potential benefits.

The works of Yu et al.[75] and Cai et al.[8] have instead focused on the development of techniques allowing real time observations with single cell sensitivity, in order to analyze gene expressed at low levels. *β-galactosidase* is the protein studied in both works, since this is the

standard reporter for gene expression both in prokaryotes and eukaryotes. A single molecule *β-gal* can produce a large number of fluorescent product molecules by hydrolysing a synthetic fluorogenic substrate, which makes *β-gal* an high-sensitivity cellular reporter. However, the drawback of its use is the fast diffusion of the fluorescent products which are quickly dispersed. Cai and collaborators [8] propose to trap the cells into a microfluidic device: cells are trapped into closed microfluidic chambers, such that the fluorescent products expelled from the cells can accumulate in the small volume of each chamber. Yu and collaborators [75] suggest another technique: they designed a fusion protein consisting of YFP (yellow fluorescent protein) and a membrane protein (*tsr*), slowing down the dispersion of fluorescent material and allowing to take measures. Both works were performed on *Escherichia coli* cells with a target polypeptide expressed under repressed conditions. Thanks to the use of single-molecule fluorescence microscopy on mRNA [61, 22, 42] and on proteins [75, 8], Taniguchi et al.[70] have performed a quantitative system-wide analysis of mRNA and protein expression in individual cells in *Eschierichia coli*, see Figure 1.5. The authors, after normalization to account for cell size and gene copy number variation due to cell cycle, have measured protein abundances ranging between 10^{-1} and 10^4 copies per cell. They found that while the noise scales with protein abundance for low expressed proteins < 10, as in [2, 48], this is not the case for proteins produced in higher quantities, where noise reaches a plateau suggesting that each protein has at least 30% of variation. They made striking real-time measurements of mRNAs, using FISH technique, and proteins at same time on 137 strains for high expressed proteins, analyzing both mRNA production and mRNA-protein correlation.

Noise and, in particular, intrinsic noise is an obstacle to the genetic program since the stochasticity of biochemical reactions leads to uncertainty in the resulting amount of proteins which could be deleterious to the achievement of the cell program. On the other hand, in specific cases these fluctuations positively exploited by cells, as a source of heterogeneity or as fundamental tool of decision making.

Starting with positive effects, fluctuations in gene expression are pointed as a major mechanism to obtain different phenotypes in an identical population. This differentiation can lead to the spring of sub-populations which are committed to different responses to environmental changes. Cell variability can be boosted in the presence of networks that can produce mutually exclusive profiles such as ON and OFF expression of a gene: small variations in the gene expression can not cause the switch from one state to the other, but rare and large fluctuations can lead to a transition. This is the case, for example, of the *lysis - lysogeny* decision in lambda phage-infected *E. Coli* [43, 29] or of the *lac* operon in *E. Coli* [52, 45] or the *galactose* utilization network in yeast [33]. In particular, for the *lysis - lysogeny* decision the stochastic effects in the expression of some regulatory factors could explain the “decision" of cells to take the lysic or lysogenic pathway. The reason to choose for a stochastic based decisional network can be connected with the performances of the resulting strategy. For example, in the presence of food cells can adopt two different strategies: they can sense food in the environment and then activate the metabolic machinery or they can stochastically decide to activate the metabolic networks in some sub-populations in anticipation of possible food arrival. The first strategy is more effective but it can be slow, while the second sacrifices few resources for a quick response. Researchers have shown how the stochastic switching strategy could be a good alternative to the sensing machinery in the cases in which stochastic fluctuations were more or less synchronized with the environment fluctuations [1]. Cellular stress, such as lack of food or exposure to antibiotics, is another case where stochastic decision could explain the observed behavior in bacterial populations, as shown in the case of *competence* in *Bacillus subtilis* [42, 68]. In particular, it is shown how the reduction of fluctuations results in lower percentage of competent cells, reducing the chances of the survival of the population under stress conditions. Although the utilization



(a) Cytoplasmic protein *Adk* (b) DNA-binding protein *YjiE*

Figure 1.5: Stochasticity in gene expression. Quantitative imaging of a YFP-fusion library. Figures (a) and (b) are representative fluorescence images of two library strains, with respective single-cell protein level histograms fitted to gamma distributions with parameters a and b. (a) The cytoplasmic protein *Adk* uniformly distributed intracellularly. (b) The DNA-binding protein *YjiE* with clear intracellular localization. Adapted from Taniguchi et al. [2010] [70].

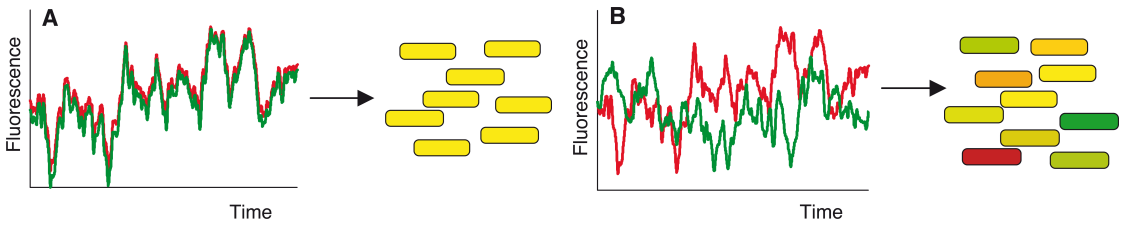
of noise for specific mechanisms, noise in gene expression has to be thought as deleterious for organisms and for gene expression in particular, which reveals a robustness with respect to fluctuations. The genome-wide works of Newman et al.[48] and Bar et al.[2] point how the variability is gene specific, in particular stress-genes, which are non essential for cell functioning, show high fluctuations, while proteasome genes are much less variables. This allocation of noise indicates that different production strategies have been selected and are possibly the result of the tradeoff between low level of noise in the protein production and the cost in term of resources of producing a large number of proteins at any time.

1.3 Intrinsic and extrinsic noise

Biological systems are constituted by individuals interacting in changing environments. In particular, fluctuations in gene expression are due to the probabilistic nature of the underlying biochemical reactions ("intrinsic noise") as well as to the effect of environment on this production ("extrinsic noise"). The measured fluctuations are therefore the result of the combined effect of these two sources of randomness, which lead to a system hardly treatable. Decomposing noise into separate terms, even if it does not provide information on the latent mechanisms, it allows to evaluate models without the obligation to specify simultaneously extrinsic or intrinsic mechanisms.

The concepts of intrinsic and extrinsic noise were introduced by Michael B. Elowitz et al.[16] and Swain et al.[69] in 2002. The stochasticity inherent in biochemical processes underlying gene expression, such as transcription and translation, is referred to as *intrinsic noise*, while the fluctuation in local environment or in the states of any other cellular factor that affects gene expression results in *extrinsic noise*. We make these definitions more clear by describing the simple and ingenuous experimental approach designed by Elowitz et al.[16] to perform this separation. The researchers used two equivalent independent gene reporters placed in the same cell and observed the two copies simultaneously. Correlations between the outcomes of the two reporters reflect the influence of the common environment, *extrinsic* see Figure 1.6a, while differences in their expressions are the consequences of the random microscopic events governing each reaction, *intrinsic* see Figure 1.6b.

If X denotes the number of proteins of interest in a given cell, we can always write the cell-to-



(a) Extrinsic noise. (b) Intrinsic noise.

Figure 1.6: **Intrinsic and extrinsic noise.** Two independent identical promoters marked with CFP (green) and YFP (red) controlled by the same regulatory sequences. (a) In the absence of intrinsic noise, the proteins of the two promoter experiment fluctuate in a synchronous fashion, because of changes in environment or on global factors impacting gene expression, i.e. number of free ribosomes or gene copy number. This correlated fluctuations lead to a population with same amounts of proteins in each cell, even if this amount could change from cell to cell because of extrinsic factors. (b) When considering the random nature of biochemical reactions the levels of the two proteins vary in uncorrelated fashion and produce eventually heterogeneity in the cell population. Figures from Elowitz et al.[16].

cell variability σ^2_X by conditioning the data on the state $\mathbf{Z}$ of the extrinsic variables, i.e. number of polymerases, ribosomes, ... Therefore, the cell-to-cell variability can be decomposed as

$$\sigma^2_X = \underbrace{\langle \sigma^2_{X|\mathbf{Z}} \rangle}_{\text{unexplained by } \mathbf{Z}} + \underbrace{\sigma^2_{\langle X|\mathbf{Z} \rangle}}_{\text{explained by } \mathbf{Z}} , \tag{1.3.1}$$

where we used the notation of [27] and where we used the law of total variance. In particular, $\langle \sigma^2_{X|\mathbf{Z}} \rangle$ is the variance of the random variable X in the subpopulation characterized with extrinsic variables $\mathbf{Z}$ and the angular brackets denote the averages over all such subpopulations. The term $\sigma^2_{\langle X|\mathbf{Z} \rangle}$ is the variance of the conditional expectation of X given $\mathbf{Z}$. The decomposition (1.3.1) is equivalent to the decomposition in the original theoretical paper of Swain et al.[69]. However, conditioning on the state of the environment captures the correct contributions only under the case of slow environmental fluctuations, but it is not well suited in the case of dynamic environment.

The main issue of decomposition (1.3.1) is that it looks at the environment at a precise point in time, but the whole history matters and, to keep track of it, Hilfinger et al.[27] propose a new decomposition

$$\sigma^2_X = \underbrace{\langle \sigma^2_{X_t|\mathbf{Z}[0,t]} \rangle}_{\sigma^2_{\text{int}}} + \underbrace{\sigma^2_{\langle X_t|\mathbf{Z}[0,t] \rangle}}_{\sigma^2_{\text{ext}}}, \tag{1.3.2}$$

where the first term in the right hand side is the variance of X_t in a subpopulation sharing an environmental history $\mathbf{Z}[0,t]$ averaged over all possible histories and the second term is the variance of the conditional expectation of X_t given a history $\mathbf{Z}[0,t]$, where $t = 0$ corresponds to the infinite past. In ergodic systems, the term σ^2_{ext} can be interpreted as the time variation of the average, while σ^2_{int} results the fluctuations around the average. By applying the decomposition (1.3.2) to the two-promoter reporter, see Figure 1.6, we obtain that $\text{Cov}(X,Y) \equiv \sigma^2_{\text{ext}}$, where X and Y are the numbers of the two identical and independent reporters in the same cell. This brings back the original ideas of Elowitz et al.in [16], where they interpreted the extrinsic contribution as the correlation between the two reporters, while the intrinsic noise is seen as uncorrelated fluctuations of the reporters under investigation.

The terms *intrinsic* and *extrinsic* remind immediately to "inside" and "outside" of a specific system. Therefore, their meaning depends on the definition of the system and the environment, as pointed out by Paulsson [54] and Hilfinger [27], i.e. if we consider the proteins as a system, then the contributions of gene activation/deactivation and of mRNAs on the protein fluctuations can be classified as extrinsic, while all these fluctuations are intrinsic as long as we consider them in the system under analysis. From a biological viewpoint, this noise classification could be done according to the importance of a cellular compound for the gene expression. Nevertheless in such a classification, ribosomes should be more likely considered intrinsic, since they are essential to gene expression, while gene activation/inactivation may result from the regulation machinery making it dependent on the state of the cell. The intrinsic noise could be connected to the specificity of a protein with respect to the others. More in detail, the specificity of the fluctuations of a protein comes from its birth and death processes and by its average number, by its messengers fluctuations and, lastly, by the specificity of the activation/deactivation of the relative gene. This last process can be less specific for some proteins, since it happens that certain genes have a common regulator which can affect all genes in an *operon* (see Section B.2.6 for details). Operon regulation may lead to complex correlations in the expression of different proteins. Moreover there are fundamental factors of gene expression which are commonly shared as ribosomes, polymerases, tRNAs, amino acids, ... For these reasons, when analyzing different sources of noise, it is necessary to clearly define the system under analysis.

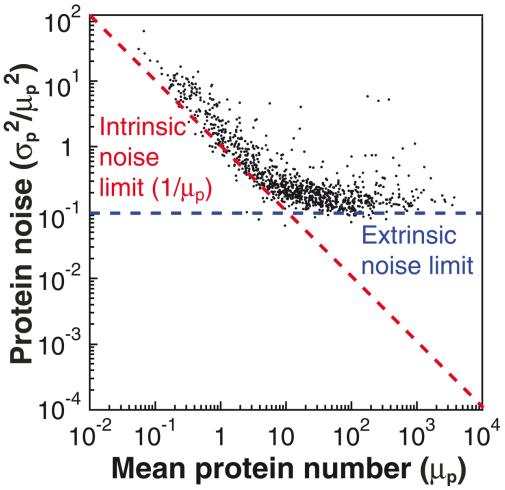


Figure 1.7: Intrinsic and extrinsic contributions to noise with respect to protein concentration. Noise for proteins produced in small quantities scales with the inverse of protein abundance, as predicted by theoretical models of intrinsic noise, see [54, 57, 71] and results in Chapters 2 and 3. However, it reaches a plateau for proteins produced in high quantities, as shown in the plot. Figures from Taniguchi et al.[70].

In 2005 Rosenfeld et al.[66] focused on the dynamic evolution of fluctuations and were able to estimate the time scales of both intrinsic and extrinsic noises. Extrinsic fluctuations exercise their effect on the cell-cycle time scales, i.e. about 40 minutes, while intrinsic noise intervenes on shorter time scales, less than 10 minutes. The slow dynamics and the amplitude of extrinsic fluctuations cause the genetic networks to have a memory of about one cell cycle.

Large scale experiments have investigated the gene expression at large scale [2, 48, 70]. Despite

differences of the organisms under analysis or quantitative differences in the results, all studies show that at low or intermediate expression levels the intrinsic fluctuations are relevant, while extrinsic fluctuations are prevalent for highly expressed genes. In particular, when intrinsic noise is relevant the noise scales with protein abundance, while it reaches a plateau when extrinsic noise is predominant, see Figure 1.7 for details. The extrinsic noise accounts for a large number of possible fluctuations and can be divided into *global*, i.e. events that affect expression of all genes, and *gene specific*, where for example fluctuations of a particular transcription factor affect the expression of one specific gene. The large scale experiments point to global effects of noise affecting all measured proteins, for this reason the noise is associated to fluctuations in cellular components such as *polymerases*, *metabolites* and *ribosomes*.

In conclusion, the concepts of intrinsic and extrinsic noise are important when dealing with experiments, since they permit to separate different sources of noise and allow to analyze independently the two components using specific modeling. However, these concepts do not give any help in the characterization of the underlying mechanisms and further modeling is needed to understand in deep the major steps affecting gene expression both at the level of a single protein and of the whole cell.

1.4 Stochasticity in gene expression: models

Stochastic models of gene expression have been proposed long before the first experimental evidence of the stochasticity of protein production. The works of Rigney and Schieve [1977] [64] and Berg [1978] [5] in the late seventies are the first papers proposing a stochastic modeling of gene expression.

At the time the experimental techniques did not allow to measure quantities in single cell, but the approach was to measure the number of proteins by averaging over a subpopulation. This approach clearly do not allow to take into account the eventual fluctuations between the individuals of a bacterial population. On the other hand, few experiments aimed to show fluctuations in protein production by indirect means, as done by Novick in [1957] [50], and there was a growing belief that the gene expression is a complex stochastic process. This theory was corroborated by physical reasons such as the fact that the chaotic motion of the molecules should reflect into fluctuations in protein levels. Moreover, a direct visual evidence of fluctuations was the observation of randomly distributed transcription initiations using electron miographs of microbial genetic activity.

Given the previous context Rigney and Schieve proposed two models of gene expression. In the simpler one, see [64], the authors consider the stochastic transcription as the main process which affects the production of a specific protein. More in detail, they consider a two state promoter:“occupied” status, when a polymerase is bounded on the gene, and “free” status otherwise. They then compute characteristic quantities such as the average time of transcription inter-initiation and, more importantly, they compute the average number of mRNAs initiated per promoter and its variance at equilibrium. Moreover, they were able to recover few experimental results, such as the fact that the average number of molecules per cell increases linearly proportional to time, and to give a characterisation of the fluctuations around the mean value. The description of the processes is Markovian: no matter of the history of the process, the state at present is sufficient to determine the future path of our process. Despite the simplicity of such model, Rigney and Schieve introduce the main tools of the classic approach. In fact, they describe the time evolution of the probability density function via the Kolmogorov backward equations, also referred to as master equation, and derive all the statistics of interest by using classical tools of renewal theory, see [10] for details. These results can also be derived by using the

generating function approach whose main ideas will be recalled in Sections 2.1.3 and 2.A . The more complete model of Rigney [63] is a more detailed model of protein production in bacteria. By using the same approach as in [64], the author proposes a complete description including protein production and degradation and protein partitioning at cell division, thus proposing a population description of protein production under constant conditions.

In the same period, Berg studied the production of proteins in a microbial population, see [5]. The aim of the author was to analyze more in detail the fluctuations of the number of proteins at steady state, using a finer description of the process. The lack of single-cell experiments made stochastic description the sole mean of investigation of the behavior of individual cells and, because of its quantitative character, the characterisation of the fluctuations can bear information on some mechanisms in protein production. Berg considered a microbial population at steady state with specific doubling time, which is supposed to be deterministic. More in detail the author described the dynamics of mRNAs, which, in turn, determine the number of proteins produced in the population. Since the specific protein is assumed to be stable, the *proteolysis* is not considered and the only mechanism which prevents an infinite accumulation is the binomial partitioning of protein at cell division. The probability distributions of messengers and proteins are then formally derived via the master equation and an explicit characterization of the average and of variance of protein copy number is given by means of generating functions. All the results depend on the cell age in the cell cycle and are valid for synchronously dividing cells. Moreover, as for the work of Rigney and Schieve, the assumption of exponentially distributed durations lead to a Markovian description of the whole process. One of the main results of the paper is the confutation of the Poisson assumption for the produced proteins. In fact, up to that date, it was supposed that the protein number follows a Poisson process. Berg showed how the fluctuations of the protein number, when considering the transcription step, can be much wider of the Poisson distribution, emphasizing the importance of an appropriate description of gene expression.

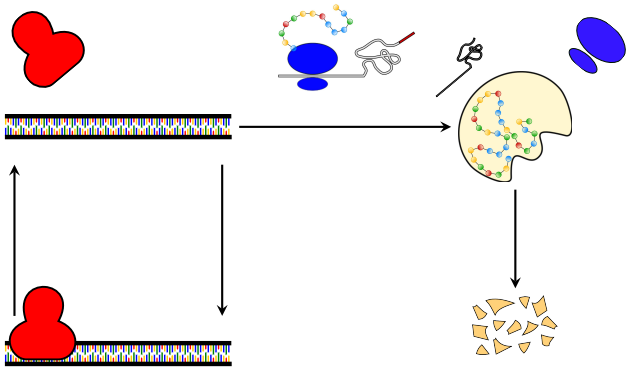


Figure 1.8: **Peccoud & Ycart model.** The model describes the gene induction, represented as a transition between active and inactive states, and the production of proteins. The level of gene expression is proportional to the fraction of the regulatory factor in the activated state. Protein degradation is considered as well.

Since the introduction of reliable single-cell experiments in the nineties such as the pioneering work of Ko et al.[39], we have assisted to a renewed interest into stochastic models of gene expression. In 1995, Peccoud and Ycart [57], inspired by the work of Ko and collaborators, proposed a simple Markovian description of gene expression. With respect to Rigney and Berg works, the

authors focus on a simple gene regulation mechanism: the gene induction. An inducible gene is normally silent, but can be expressed under the control of external signals, the reaction being usually mediated by a regulatory factor. The activation of such regulatory factor is reversible and is described as the switching between active and inactive states, as shown in figure 1.8. The messenger dynamics are not considered, but the level of proteins is assumed to be proportional to the time the gene spends in active state. The analysis is restricted to a single cell and it is not considered the population dynamics, however the degradation of proteins allows the protein number not to explode to infinity as long as we look at equilibrium dynamics. The statistics are derived using similar tools as Rigney and Berg and, in particular, they give close formula for the transient behavior of the mean number of protein copies and their variance. The asymptotic statistics are then derived and a simple method of parameter estimation, using the information that can be derived by experiments.

Paulsson in 2005 [54] gave a review on stochastic gene expression and proposed a model which summarizes the main features of the models found in literature. As the works presented above, the model aims to characterise the average and fluctuations of the number of copies of a specific protein when equilibrium is reached. This model, with respect to the model of Peccoud and Ycart presented above, considers also the dynamics of messengers. The described processes of gene expression are the activation and deactivation of the gene by means of a repressor, the dynamics of the corresponding mRNA and of protein. More in detail the gene can show two different states, active and inactive, where the transition from one state to the other is supposed to be exponentially distributed. If the gene is in active state, then it produces a new mRNA following a Poisson process, which describes as well the mRNA degradation by means of *RNase*. Proteins are translated by mRNAs and are degraded, both processes showing exponentially distributed duration.

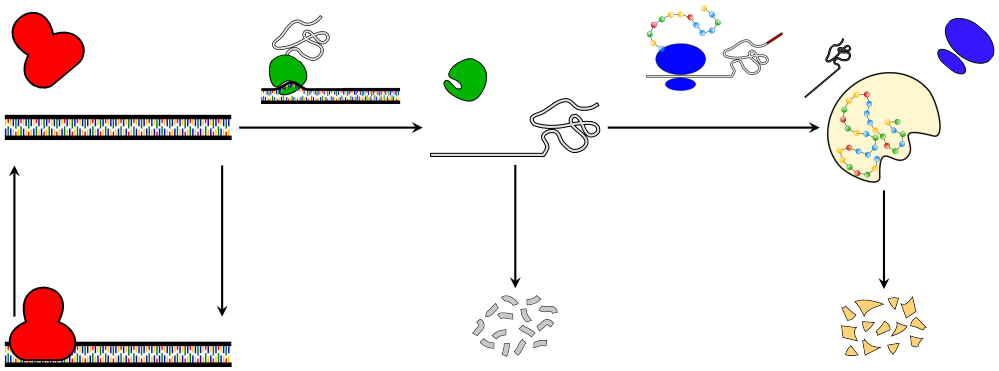


Figure 1.9: **Paulsson's model.** In this model gene activation/deactivation is described and is represented here via the binding/unbinding of a repressor, in red. When the gene is in active state it can undergo transcription and it produces a new mRNA at a certain rate. The mRNA can then be translated by ribosomes, represented in blue in the figure, and produces proteins at a certain rate. Both mRNAs and proteins can be degraded. Note that the duration of all the described steps is exponentially distributed, which allows for a Markovian description of gene expression and which is the fundamental common characteristic of classic modeling.

1.4.1 Limits of classic models

In Section 1.1.2 we have seen how the encounter of two cellular molecules is the fundamental mechanism of the biochemical reactions underlying gene expression. The objective of mathematical biology is to extract information on such system and on its specific mechanisms, via a mathematical description and characterization. However, gene expression comprises a colossal number of elementary steps, hence the need to combine those steps into effective steps and focus on a part of the system. The intricate interaction pathway linking numerous processes both at small and global scale contributes to the complexity of the system. The elementary processes underlying gene expression are therefore grouped into critical steps, as we have seen in the Section 2.A, where the main steps of transcription and translation are modeled globally as a first-order chemical reaction, i.e. they are supposed to be exponentially distributed.

The models in literature, up to our knowledge, describe the duration of each step to be exponentially distributed. Nevertheless, the suitability of these assumptions and of the description of the gene expression process has been poorly investigated. We recall briefly the main ideas and mathematical framework that stand behind classic models and the crucial assumption, that we will refer to as "*exponential assumption*".

The classic three stage model and mathematical toolbox

In the beginning of this section we have described few of the fundamental models used to describe gene expression in literature, among which the three-stage model [54] is the reference model for experimentalists. All these models share a common underlying description, that can be already found in the first systematic and accurate studies of stochastic models for gene expression, see Rigney [63, 64] and Berg [5]. In recent years the three-stage model has been used as the fundamental structure in most well-known works of Shahrezaei and Swain [67], Paulsson [54] and Peccoud and Ycart [57].

In these studies the promoter of the gene, corresponding to the specific protein of interest, can be in one of two possible states, active or inactive, and the switching occurs up to a exponentially distributed random time. Moreover, transcription, translation and the degradation of proteins and messengers are supposed to be exponentially distributed (or geometrically distributed in case of a discrete time setting).

The fundamental assumption of exponentially distributed duration of the various steps of the three-stage model allows a Markovian modelling. Without loss of generality, we consider the one gene case and denote with $Y(t) \in \{0,1\}$ the state of the gene at time t, where $Y(t) = 1$ indicates that the gene is active at time t, while $Y(t) = 0$ if it is inactive. If we denote by $N_2(t)$ the number of mRNAs and by $N_3(t)$ the number of proteins, then it turns out that $(X(t)) = (Y(t), N_2(t), N_3(t))$ is a Markov process with values in $\{0,1\} \times \mathbb{N}^2$ and this representation covers most of the models described in the beginning of this section, see Section 2.A for further details.

As a consequence of the Markovian description, we obtain a system of linear differential equations of order 1, the Fokker-Planck equations, of $p(t,(y,n_2,n_3))$, i.e. the probability that $X(t)$ is in state (y,n_2,n_3) at time t. The solution of the system has a unique stable point $(\pi(y,n_2,n_3) : (y,n_2,n_3) \in \{0,1\} \times \mathbb{N}^2)$, the invariant distribution of the Markov process. Although it is not possible to obtain an explicit expression of the invariant distribution, it is still possible to get information about the moments of the invariant distribution. In particular, since the coefficients of the obtained Fokker-Planck equation are affine with respect to the variables n, the moments of the invariant distribution satisfy a recurrence equation, giving an explicit expression for the first two moments and, in particular, for the variance. This is the main theoretical result that has been used in many papers in literature, see [54, 64, 5, 69, 71, 57], and is possible exclusively under

the assumption that *all* the durations of the main steps, such as mRNA and protein production, are exponentially distributed.

Exponential assumption

We refer to *exponential assumption* when the time to produce a particular cellular component and its lifetime are assumed to be exponentially distributed.

The exponential assumption is natural in the following simple situation: a large number of trials are necessary to achieve some goal (like transcription or translation initiation) and each trial requires some duration D and succeeds with probability α. This scheme describes correctly the duration of time to establish a binding and, therefore, it may describe properly the time required for a successful binding of RNA polymerase to the gene and of ribosome to mRNA.

On the other side, this assumption may not be true if we consider the elongation time of an mRNA or protein chain. In particular, the protein elongation results in an iterative procedure in which each codon of the messenger chain is coupled with a particular tRNA, which adds a new amino-acid to the growing polypeptide chain by means of ribosome, see Figure 1.10. The insertion of a new amino-acid requires firstly the encounter of a charged tRNA and the resulting distribution of the elongation step, which is the sum of exponentially distributed random variables, is no longer exponential.

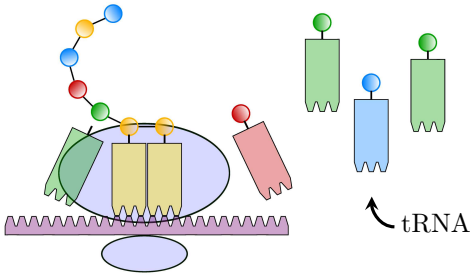


Figure 1.10: Protein elongation.

Another phenomenon is not described in the classic models: protein dilution. The volume growth associated to cell doubling affects the concentration of all cellular component, and in particular of proteins, because of dilution. During exponential growth phase, dilution is proven to be the main cause of protein disappearing, while *proteolysis* play a minor role. Few papers in literature [43, 71] affirm to consider dilution as the protein degradation mechanism, but its description is wrong since it is modeled as a discrete stochastic process, while there is a wide agreement on the fact that it is deterministic and continuous.

Unfortunately, there is no hope to include these two phenomena in the classic mathematical framework, since this approach is strongly limited by the exponential assumption, which limits the possibility to extend the existing models and to consider possibly more general mechanisms, using the knowledge acquired on gene expression through experiments.

Chapter 2

MPPP description of gene expression

The study of the fluctuations of the number of a specific protein in a cell is a crucial problem for biologists in order to better understand the production of proteins and has already been tackled in the past. In particular, researcher have obtained close-form analytic expressions of the mean and variance of the number of copies of a protein for a simplified stochastic model of gene expression (see Section 2.A for further details). A Markovian approach (via Fokker-Planck equations) is classically used to derive analytic formulas of mean and variance of the number of proteins at equilibrium, under the assumption that the duration of all modeled steps is exponentially distributed. This assumption is, however, not completely satisfactory from the modeling point of view since the duration of some steps is more likely to be Gaussian, if not quasi deterministic. In such a setting, Markovian methods can no longer be used and a new approach allowing for more general assumptions is required. This pathway is essential for obtaining a finer characterization of the fluctuations of the number of proteins, which is of primary interest to understand the general economy of the cell and to analyze the solutions imposed by the evolutionary force.

The present chapter is devoted to the introduction of a new description of gene expression which uses the Marked Poisson Point Processes (MPPP) as the main mathematical tool and allows to get rid of the exponential assumption. In the following sections, starting from the description of gene expression of the classic *three-stage model* (see Section 2.A.2), we will introduce the technical tools of the MPPP description and derive the main theoretical results.

Plan of the Chapter. In Section 2.1 we describe the main biological processes, the effective steps of the mathematical model and the limits of the classic models. The mathematical description of gene expression via MPPP is introduced in Section 2.2 and the main general results are derived in Section 2.3. In Section 2.4 we derive results for specific choices of distributions.

The Appendix 2.A is devoted to the presentation of few techniques used in classical models. In particular, the original approach of Rigney [64, 63] and the consolidated approach used in more recent papers [54, 57] are presented.

2.1 Biology and mathematical assumptions

We present now the biological mechanisms which have motivated our description and show the limitations of the classic approach, when it comes to include more realistic assumptions. In the

biological section we will briefly recall the main steps of gene expression, for further details see Section 1.1.2 or Appendix B.

2.1.1 Biological context

The *gene expression* is the biological process by which the genetic information contained into the DNA of a cell is synthesized into a functional product, the proteins. The production of proteins is the most important cellular activity, both for the functional role of its products and the high cost in term of resources.

The information flow from DNA genes to proteins is a fundamental process, common to all kinds of cells, and consists of two main elementary steps: *transcription* and *translation*. During the transcription process, the *RNA polymerase* binds to an active gene, which corresponds to a specific protein, and makes a complementary copy of a specific portion of a DNA strand, a *messenger RNA* (mRNA). Each mRNA, which is a long chain of nucleotides, is a chemical "blueprint" for a particular protein. The synthesis of the protein chain from the mRNA is achieved by a large and complex molecule, the *ribosome* during the *translation* step. More in detail, the ribosome binds to the messenger and assembles the polypeptide chain using the mRNA as a template: to each mRNA *codon*, a triplet of nucleotides, corresponds a specific *amino acid*, the fundamental brick of proteins. Once the polypeptide chain has been completed, the amino acids fold spontaneously or with the help of *chaperons*, so that the protein may assume its functional three-dimensional structure.

The gene expression is a highly stochastic process and results from the realization of a very large number of elementary stochastic processes of different nature. The thermal excitation affects many processes, since it implies for example the free diffusion in the cytoplasm in which particles behave basically as if they were plunged into a viscous fluid. In a first approximation of the production process, three fundamental mechanisms are combined in the protein production. The first is the pairing of two cellular components freely diffusing through the cytoplasm and is a direct consequence of the diffusion. The second mechanism is the "spontaneous" rupture of such a binding, as the result of thermal excitation, and the subsequent release of the two components. The last involved stochastic mechanism corresponds to the processing capability of both polymerases and ribosomes and results in an active process, since it requires energy in order to take place. The active steps associated to the polymerase and ribosome activity are highly sophisticated and include, for example, dedicated proof reading mechanisms. In order to proceed to transcription initiation, see Section 1.1.2, is required a successful pairing of the polymerase to a specific DNA motif. Once initiation step has succeeded, the elongation step can take place and this results as a series of specific stochastic processes, in which the polymerase recruits one of the four nucleotides corresponding to the DNA template. In a similar fashion, the *translation* step consists in the pairing/unpairing of ribosome to a specific sequence of the mRNA and the subsequent elongation step. In particular, the protein elongation results in an iterative energy-consuming procedure in which each codon of the messenger chain is associated to a specific *tRNA*, which adds the corresponding amino acid to the growing polypeptide chain by means of ribosome.

In summary, most of the elementary processes can be schematically seen as the encounter of two components in a viscous fluid. However, the classic approach to gene expression modeling is to group those elementary processes into critical steps as initiation, elongation and degradation, which are common to both transcription and translation.

2.1.2 Mathematical model of gene expression: three-stage model

The main steps of gene expression, as described shortly in the previous section, can be translated into a mathematical model, whose main steps are now described. We will not dwell here in detail, on the contrary, the aim of this section is to introduce the basic notation that will be used throughout the chapter.

The total number of proteins of the specific type is a random variable P. The cell can be seen as a complex system producing a given protein with an average concentration $\mathbb{E}[P]$, where $\mathbb{E}[X]$ denotes the expected value of a random variable X. The protein concentration is directly connected with the main parameters describing the different processes and this connection is made explicit through a quite simple formula, as will be shown later on. We recall that the main objective of the mathematical descriptions of gene expression is to derive explicit formulas of the variance of the number of proteins in terms of the main model parameters.

Notation 2.1.1. *If not specified differently, a process is said to occur at some rate λ, means that the duration of such process is exponentially distributed with parameter λ.*

Gene activation. The initiation of transcription is strongly regulated by various molecular mechanisms like the association/dissociation of a *repressor* (see Figure 3.1) and the association of an *activator*. The gene is said "inactive" if a repressor prevents the polymerase binding and is said "active" otherwise. Usually the whole process is described as a telegraph process for which a transition from inactive state 0 to active state 1 occurs at rate λ_1^+ and, similarly from state 1 to state 0 at rate λ_1^-. Here the fundamental assumption is that the distribution of these steps is exponential. In a prokaryotic cell, there may be several copies of a specific gene and this fact has been included in few models in the past years, see Paulsson [54]. Nevertheless, since we are interested in the variance of the number of proteins, we will assume in the following sections that there is only one copy of the gene. The analogous result for the case with multiple copies is straightforward to obtain since, by independence, the variance of protein number is proportional to the number of copies of the gene.

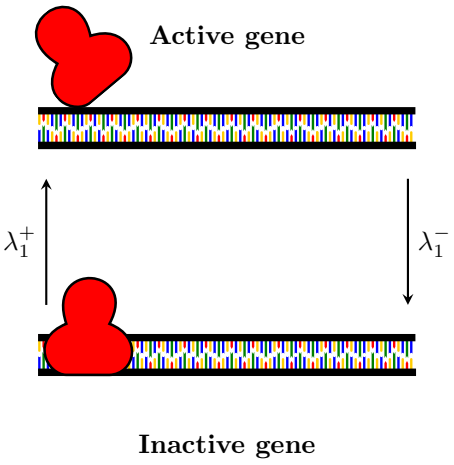


Figure 2.1: Gene activation. The gene activation/deactivation occur at rate λ_1^+ and λ_1^- respectively.

Transcription. A RNA polymerase binds on an active gene in an exponential time with rate λ_2. This effective rate measures the frequency of transcription initiation and takes into account several physical parameters, including, for example, the *affinity* between the specific gene and the polymerase. The distribution F_2 on $\mathbb{R}_+$ of the lifetime σ_2 of a mRNA is assumed to be general.

Translation. Similarly, the binding of a ribosome on an mRNA occurs in an exponentially distributed time with rate λ_3, which measures the frequency of translation initiation and includes also the affinity between messenger and ribosome. The distribution F_3 of the lifetime σ_3 of the protein is also general. The decay of the protein concentration occurs for two main reasons:

by *proteolysis*, *i.e.* the protein degradation into amino acids, or by cellular dilution, due to the cellular volume increase of the bacterium during the exponential growth phase.

Remark. The analysis on gene expression of Chapters 2 and 3 is focused on the complex process of production of a specific protein. For this reason, the interaction of the specific protein production chain with the production process of other proteins is not considered.

2.1.3 Limits of classic models: the exponential assumption

In the previous section we have recalled the main steps of gene expression and we have seen how the elementary processes in the protein production can be seen schematically as the encounter of two cellular components in a viscous fluid.

The classic approach is to group these elementary processes into critical steps, see Section 2.A, and will be recalled in Section 2.2. In particular, the transcription (resp. translation) step is modeled globally as a first-order chemical reaction, i.e. it is supposed to be exponentially distributed. The same assumption is usually applied to the messenger and protein degradation.

Nevertheless, the suitability of these assumptions and of the chosen description of the gene expression process has been poorly investigated. In this chapter we focus mainly on the model assumptions and develop an alternative description of gene expression that allows to retrieve analytic close formulas of mean and variance of proteins in a more general context.

Before introducing this new mathematical description of gene expression, we recall briefly the main ideas and mathematical framework that stands behind classic models and discuss the crucial "*exponential assumption*" that comes with.

The classic three-stage model and mathematical toolbox

The *three stage model* described in Section 2.1.2 is the fundamental approach to describe gene expression in literature, as testified by the theoretical research on this model and its large use by experimentalists. The key steps of this description can be already found in the first systematic and accurate studies of stochastic models for gene expression, as Rigney [64, 63] and Berg [5]. In recent years the *three-stage model* has been used as the fundamental structure in most well-known works of Shahrezaei and Swain [67], Paulsson [54] and Peccoud and Ycart [57]. See Figure 2.2.

The promoter of the gene, corresponding to the specific protein of interest, can be in one of two possible states: active or inactive. In these studies transcription, translation and the degradation of proteins and messengers are modeled as first-order chemical reactions, *i.e.* they are supposed to be exponentially distributed (or geometrically distributed in case of a discrete time setting). With the above notations, this amounts to saying that both σ_2 and σ_3 are exponentially distributed.

The assumption of exponentially distributed duration of the various phases of the *three-stage model* leads naturally to a Markovian modeling. The overall dynamic of gene activation can be described, see Paulsson [54], by the random variable $Y(t) \in \{0, 1\}$, where $Y(t) = 1$ indicates that the gene is active at time t, while $Y(t) = 0$ if it is inactive. Recall that we consider, without loss of generality, only the one gene case. If we denote by $N_2(t)$ the number of mRNAs and by $N_3(t)$ the number of proteins, then it turns out that $(X(t)) = (Y(t), N_2(t), N_3(t))$ is a Markov process with values in $\{0, 1\} \times \mathbb{N}^2$. This representation is common to most of the models of the literature. Some of them have, in fact, a lower dimensional state space because of assumptions on the number of mRNAs for example. We denote with $p(t, (y, n_2, n_3))$ the probability that $X(t)$ is in state (y, n_2, n_3) at time t, i.e.

$$p(t, (y, n_2, n_3)) = \mathbb{P}\left[X(t) = (y, n_2, n_3)\right].$$

As a consequence, the general theory of Markov processes gives a system of linear differential equations of order 1, the Fokker-Planck equations, for the functions $p(t,(y,n_2,n_3))$. The system of equations has the general form

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}p(t,(y,n_2,n_3)) &= \lambda_1(y)p(t,(1-y,n_2,n_3)) + \lambda_2 p(t,(y,n_2-1,n_3))\mathbb{1}_{\{y=1\}} \\ &\quad + \mu_2(n_2+1)p(t,(y,n_2+1,n_3)) + \lambda_3 n_2 p(t,(y,n_2,n_3-1)) \\ &\quad + \mu_3(n_3+1)p(t,(y,n_2,n_3+1)), \quad (2.1.1)\end{aligned}$$

where $\lambda_1(0)=\lambda_1^-$ and $\lambda_1(1)=\lambda_1^+$, $\mu_i = 1/\mathbb{E}(\sigma_i)$ for $i=1,\,2$. The solution of the system has a unique stable point $(\pi(y,n_2,n_3),(y,n_2,n_3)\in\{0,1\}\times\mathbb{N}^2)$, the invariant distribution of the Markov process, whose explicit expression is not known to the best of our knowledge. Nevertheless, since the coefficients of the right-hand-side of this equation are affine with respect to the number of messengers n_2 and the number of proteins n_3, the moments of the invariant distribution satisfy a recurrence equation. This equation is not trivial, but gives an explicit expression for the first two moments and, in particular, for the variance, which is the key quantity to investigate fluctuations. This is the main theoretical result that has been used in many papers in literature, see [54, 64, 5, 69, 71].

It should be kept in mind that this approach is possible only under the assumption that *all* the durations of the main steps (like the production time of an mRNA or of a protein) are exponentially distributed. This assumption is now discussed.

Exponential assumption

We refer to *exponential assumption* when the time to produce a particular cellular component and its lifetime are assumed to be exponentially distributed.

We discuss now the appropriateness of the use of the exponential assumption, with respect to the biophysical process described. The exponential assumption is natural in the following simple situation: a large number of trials are necessary to achieve some goal (like transcription or translation initiation) and each trial requires some duration D and succeeds with probability α. If G_α is the total number of attempts to succeed, i.e. $\mathbb{P}(G_\alpha \geq n) = (1-\alpha)^n$, then

$$\lim_{\alpha\to 0}\mathbb{P}(\alpha G_\alpha \geq x) = e^{-x} \qquad \text{for } x\geq 0.$$

In other words, if α is small then $\alpha G_\alpha \approx E_1$, where E_1 is an exponential random variable with mean 1. Consequently, the total duration of time necessary to realize the objective is, due to the averaging of the law of large numbers (G_α is large),

$$\sum_{i=1}^{G_\alpha} D_i \approx G_\alpha \mathbb{E}\left[D\right] \approx \frac{\mathbb{E}\left[D\right]}{\alpha} E_1$$

and is therefore exponentially distributed with mean $\mathbb{E}\left[D\right]/\alpha$.

As is seen, this scheme may describe correctly the duration of time to establish a binding of a polymerase or of a ribosome. More in detail, it may describe properly the time required for a successful binding of RNA polymerase to the gene and of ribosome to mRNA.

On the other side, this assumption may not be true if we consider the elongation time of an mRNA or protein chain. In particular, during the polypeptide chain elongation, each tRNA, transporting a specific amino acid (see B.2.13 for details), should bind to the ribosome. If the distribution of the duration of this step is indeed exponential, nevertheless the fact that

elongation steps requires an average number of $300 - 400$ steps, one for each amino acid, then the resulting distribution of the duration of the whole process is no longer exponential. In first approximation, because of the large number of elongation steps, a deterministic elongation time with a small Gaussian perturbation should be considered.

The description of elongation step and the impact of different choices for the distribution of the duration of this step will be analyzed in detail in Chapter 3. In the present chapter we assume that both messenger and protein are instantaneously produced, i.e. the duration of elongation is zero. However, one of the main results of this chapter is to show, via convenient mathematical tools, that the choice of the distributions has an important impact on the quantification of protein fluctuations and, therefore, on the qualitative properties of the gene expression.

2.2 MPPP Description of Gene Expression

In this section, the various stochastic processes are introduced. The main results and notations concerning the marked Poisson point processes (MPPP) is introduced in Section A.0.3. In the entire manuscript, all the Poisson point process and the associated random variables are supposed to be independent.

Gene activation

It is assumed that there is one active gene, which is activated at rate λ_1^+ and inactivated at rate λ_1^-. Recall that the assumption that $n_{\max}$, the maximum number of active genes, is 1 does not restrict the generality of our results since the quantities analyzed in this paper (expected values and variances) are proportional to $n_{\max}$. Let (E_n) and (F_n) be i.i.d. exponential random variables with respective rates λ_1^- and λ_1^+. The process of activation of the gene at equilibrium can be represented as a stationary process $(Y(t), t \in \mathbb{R})$ with values in $\{0, 1\}$. Note that $(Y(t))$ is defined on the whole real line, i.e. that the activation/inactivation process has started at $t = -\infty$. As it will be seen, this is a convenient representation to describe properly the equilibrium of the protein production process. The increasing sequence of the instants of activation of the gene is denoted by $(t_n, n \in \mathbb{Z})$ with the convention that $t_0 \leq 0 < t_1$. In particular

$$\{t_n, n \in \mathbb{Z}\} = \{s \in \mathbb{R} : Y(s-) = 0 \text{ and } Y(s) = 1\}$$

and $t_{n+1} - t_n = E_n + F_n$. Because of our assumption (t_n) is a stationary renewal point process.

Since we are interested to study the system at equilibrium and since $(Y(t))$ is defined on $\mathbb{R}$, we can suppose that the process $Y(t)$ started at $-\infty$ and is at equilibrium at the instant 0. So, instead of start the processes at time 0 and suppose they reach equilibrium after an infinite time, we may suppose that the production machinery has started at $-\infty$ and it is then at equilibrium at time 0. This convention will be used for all the processes investigated.

Production of mRNAs

When the gene is active, it produces mRNAs at rate $\lambda_2 > 0$ and $F_2(\mathrm{d}y)$ denotes the distribution on $\mathbb{R}_+$ of the lifetime of an mRNA. Let $\mathcal{N}_{\lambda_2}=((s_n, \sigma_{2,n}), n \in \mathbb{Z})$ be a MPPP on $\mathbb{R} \times \mathbb{R}_+$ with intensity measure $\lambda_2 \, \mathrm{d}x \otimes F_2(\mathrm{d}y)$.

If the gene is always active throughout the time interval $[s, t]$, then the formula

$$\mathcal{N}_{\lambda_2}([s, t] \times \mathbb{R}) = \sum_{n \in \mathbb{Z}} \mathbb{1}_{\{s \leq s_n \leq t\}} = \int_{\mathbb{R} \times \mathbb{R}_+} \mathbb{1}_{\{s \leq u \leq t\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v)$$

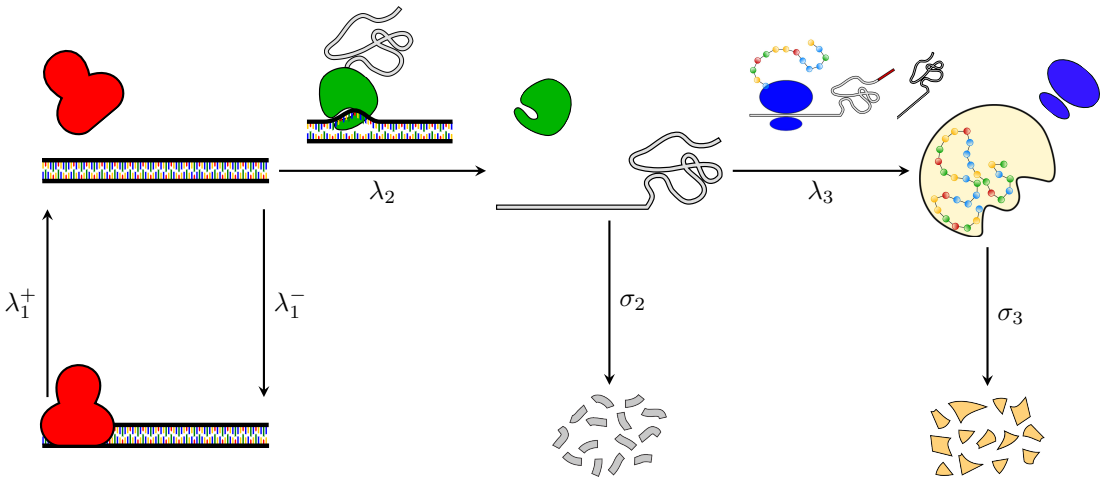


Figure 2.2: Three-Stage model. The gene activation/deactivation occur at rate λ_1^+ and λ_1^- respectively. Transcription and translation occur at rates λ_2 and λ_3 respectively. The degradation times of mRNAs and proteins have probability distributions $F_2(dt)$ and $F_3(dt)$ respectively.

represents the total number of mRNAs created between time s and time t, and

$$\sum_{n\in\mathbb{Z}} \mathbb{1}_{\{s\le s_n\le t\le s_n+\sigma_{2,n}\}} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le u\le t\le u+v\}}\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v)$$

is the number of mRNAs still alive at time t. More generally, if we include the gene dynamics into the formula, we find that the number of messengers created in the time interval $[s,t]$ and still alive at time t is

$$\sum_{n\in\mathbb{Z}} \mathbb{1}_{\{s\le s_n\le t\le s_n+\sigma_{2,n},Y(s_n)=1\}} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le u\le t\le u+v,Y(u)=1\}}\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v).$$

Production of Proteins

A given mRNA produces proteins at rate λ_3 and $F_3(\mathrm{d}y)$ is the distribution of the duration of the lifetime of a protein.

For $u\in\mathbb{R}$, denote by $\mathcal{N}_{\lambda_3}^u$ a MPPP with intensity $\lambda_3\,\mathrm{d}x\otimes F_3(\mathrm{d}y)$, this process describes in the following the creation of proteins associated to a mRNA created at time u. We assume that $\mathcal{N}_{\lambda_3}^{u_1}$ and $\mathcal{N}_{\lambda_3}^{u_2}$ are independent for $u_1\neq u_2$. In particular, if mRNA lifetime is v then

$$\mathcal{N}_{\lambda_3}^u([u,u+v]\times\mathbb{R}_+) = \int_{[u,u+v]\times\mathbb{R}_+} \mathcal{N}_{\lambda_3}^u(\mathrm{d}x,\mathrm{d}y)$$

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

Remark. If the gene is always active, i.e. $Y(t)\equiv 1$, the whole process of production of mRNAs and proteins under this specific assumption can thus be described by the sequence

$$\mathcal{A} = \left(s_n,\sigma_{2,n},\mathcal{N}_{\lambda_3}^{s_n}\right).$$

Recall that $\mathcal{N}^0_{\lambda_3} : (\Omega, \mathcal{F}, \mathcal{P}) \to \mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+)$, where $\mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+)$ is the set of point processes on $\mathbb{R} \times \mathbb{R}_+$. If we denote with $\mathbb{Q}$ the distribution of $\mathcal{N}^0_{\lambda_3}$ on $\mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+)$, the process $\mathcal{A}$ can be seen as a marked Poisson point process on $\mathbb{R} \times \mathbb{R}_+ \times \mathcal{M}_p(\mathbb{R} \times \mathbb{R}_+)$ with intensity measure $\lambda_3 \, \mathrm{d}x \otimes F_3(\mathrm{d}y) \otimes \mathbb{Q}$. This observation will not be used in the following to keep the setting as simple as possible but the proof of Proposition 2.3.5 below could be shortened by using it together with Proposition A.0.5.

The notations with some definitions for the stochastic models used in this paper are now summarized.

Notation 2.2.1.

- *Gene activation.*
  The activation rate is [resp. inactivation rate] is λ_1^+ [resp. λ_1^-] and
  $$\delta_+ = \frac{\lambda_1^+}{\Lambda} \text{ and } \Lambda = \lambda_1^+ + \lambda_1^-.$$
- *mRNA production.*
  The rate of production of mRNAs by an active gene is λ_2, $F_2(\mathrm{d}x)$ is the distribution of an mRNA lifetime, σ_2 denotes a random variable with distribution F_2 and
  $$\rho_2 \overset{\text{def.}}{=} \lambda_2 \mathbb{E}\left[\sigma_2\right] = \lambda_2 \int_{\mathbb{R}_+} x F_2(\mathrm{d}x).$$
- *Protein production.*
  The rate of production of proteins by an mRNA is λ_3, the lifetime distribution of a protein is $F_3(\mathrm{d}x)$, σ_3 denotes a random variable with distribution F_3 and
  $$\rho_3 \overset{\text{def.}}{=} \lambda_3 \mathbb{E}\left[\sigma_3\right] = \lambda_3 \int_{\mathbb{R}_+} x F_3(\mathrm{d}x).$$

2.3 General results of MPPP description of gene expression

In the previous section, using basic definitions about Poisson point processes, we have given a first description of the main processes modeled in the *three-stage model*. In this section, using the MPPP description, we derive general results concerning gene expression. In particular, the number of the different processes at equilibrium and general formulas for the mean and the variance of each process. We will not discuss here the results for specific assumptions of the general distributions and we refer to the following section.

2.3.1 Gene state

The behavior of the process $(Y(t))$ describing the state of the gene is well known in literature. In particular, $Y(t) \in \{0, 1\}$ and is Bernoulli distributed at equilibrium. In order to obtain the stationary distribution π_Y of the process Y it is sufficient to write the detailed balance equation, since $(Y(t))$ is a reversible Markov process. Therefore,

$$\begin{cases} \lambda_1^+ \pi_Y(0) = \lambda_1^- \pi_Y(1) \\ \pi_Y(0) + \pi_Y(1) = 1, \end{cases}$$

where $\pi_Y(0) = \mathbb{P}[Y = 0]$ and $\pi_Y(1) = \mathbb{P}[Y = 1]$. Solving the previous system we obtain, at equilibrium,

$$\mathbb{P}[Y(0) = 1] = \delta_+ = \frac{\lambda_1^+}{\lambda_1^+ + \lambda_1^-} = 1 - \mathbb{P}(Y(0) = 0).$$

From now on, it will be assumed that $(Y(t))$ is defined on $\mathbb{R}$ and is at equilibrium.

In order to obtain analytic expression of the number of messengers and proteins, we need also to compute the covariance of the process $(Y(t))$. For $t \geq 0$,

$$\mathbb{P}(Y(t) = 1|Y(0) = 1) = \delta_+ + (1 - \delta_+)e^{-\Lambda t}, \tag{2.3.1}$$

with $\Lambda = \lambda_1^+ + \lambda_1^-$, in fact, since the process $(Y(t))$ is stationary, then

$$\begin{aligned}\mathbb{E}[Y(u)Y(s)] &= \mathbb{E}[Y(|u-s|)Y(0)|Y(0) = 1]\,\mathbb{P}[Y(0) = 1] \\ &= \mathbb{P}[Y(|u-s|) = 1|Y(0) = 1]\,\mathbb{P}[Y(0) = 1]\end{aligned}$$

and we obtain the previous formula by solving the Kolmogorov forward equations, see Note 2.3.1. See Norris [49] and Peccoud and Ycart [57] for detailed computations.

Note 2.3.1. *The generator matrix Q_Y of the two-state continuous time Markov chain $Y(t) \in \{0, 1\}$ is given by*

$$Q_Y = \begin{pmatrix} -\lambda_1^+ & \lambda_1^+ \\ \lambda_1^- & -\lambda_1^- \end{pmatrix}.$$

Denote with $p_{ij}(t) = \mathbb{P}(Y(t) = j|Y(0) = i)$, with $i, j \in \{0, 1\}$. Recall that we want to compute $p_{11}(t) = \mathbb{P}(Y(t) = 1|Y(0) = 1)$. The Kolmogorov forward equation $P'(t) = P(t)Q_Y$, where $P(t) = (p_{ij}(t))_{i,j\in\{0,1\}}$, together with the identity $p_{01} = 1 - p_{11}$ gives

$$\begin{cases} p'_{11}(t) = \lambda_1^+ - (\lambda_1^+ + \lambda_1^-)p_{11}(t), & \forall t > 0 \\ p_{11}(0) = 1, \end{cases}$$

whose solution is formula (2.3.1).

2.3.2 Messengers

Number of mRNAs

A result on the number of mRNAs at equilibrium and its distribution is derived in this section. The techniques used to prove it will also be used to investigate the distribution of the number of proteins in the next section. In order to present the MPPP approach, we will develop computations for mRNAs. They are simpler from the point of view of notations and they include the main ideas.

Proposition 2.3.2. *The number M of mRNA's at equilibrium can be represented as*

$$M = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\leq 0\leq u+v, Y(u)=1\}}\, \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v), \tag{2.3.2}$$

where $\mathcal{N}_{\lambda_2}$ is a Poisson marked point process with intensity $\lambda_2\,\mathrm{d}x \otimes F_2(\mathrm{d}y)$.

Proof. Let M_t denote the number of messengers born after time $t = 0$ and still alive at time t. Then M_t is given by

$$M_t = \sum_n \mathbb{1}_{\{0 \le s_n \le t \le s_n + \sigma_{2,n}, Y(s_n)=1\}} = \int_{\mathbb{R}_+} \int_0^{+\infty} \mathbb{1}_{\{u \le t \le u+v, Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v), \tag{2.3.3}$$

where $\mathcal{N}_{\lambda_2} = (s_n, \sigma_{2,n})$, see Section 2.2. Recall that s_n is the (potential) n^{th} binding time of a polymerase on the gene: an mRNA is created only if the gene is active, i.e. $Y(s_n) = 1$. The term $\sigma_{2,n}$ represents the lifetime of the newly produced mRNA. The right-hand-side of the previous equation accounts for the number of mRNAs produced in the interval $[0, t]$ and still alive at time t ($u + v \ge t$).

Since the process $(Y(t))$ is stationary as well as the Poisson marked point process, they are both invariant by translation. By translating by $-t$, one gets that M_t has the same distribution as

$$M_t \overset{\text{dist.}}{=} \int_{-t}^{0} \int_{\mathbb{R}_+} \mathbb{1}_{\{0 \le u+v, Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v),$$

by letting t go to infinity, we obtain the desired result. □

It is crucial that the distribution of M_t can be explicitly expressed as a functional of the marked Poisson process $\mathcal{N}_{\lambda_2}$. The same property is true for its limit. In this context, with the help of the coupling argument, there is no need of a Markovian setting to prove that M_t converges in distribution as t goes to infinity. As will be seen, the distribution of the limit M can be obtained by using some properties of Poisson point processes. For all these reasons, there is no need to assume σ_2 and σ_3 are exponentially distributed.

In the proof of the above result, we have in fact proved a more general following result. The point process $\mathcal{M}$ representing the instants of creation of mRNAs and the associated lifetime at equilibrium can be represented as

$$\mathcal{M} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \delta_{(u,v)} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v), \tag{2.3.4}$$

where δ_z is the Dirac mass at z [1].

The number of mRNAs alive at equilibrium can thus be represented as

$$M = \int \mathbb{1}_{\{u \le 0 \le u+v\}} \mathcal{M}(\mathrm{d}u, \mathrm{d}v) = \int \mathbb{1}_{\{u \le 0 \le u+v, Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v)$$

which is precisely the expression of Proposition 2.3.2. When the activation rate of the gene goes to infinity, the point process $\mathcal{M}$ is simply a marked Poisson point process and M has a Poisson distribution with parameter $\rho_2 = \lambda_2 \mathbb{E}(\sigma_2)$.

We now use this representation to get an explicit expression of the variance of the number of mRNAs at equilibrium.

[1] If $f : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ is integrable with respect to the measure $\mathcal{M}(\mathrm{d}u, \mathrm{d}v)$, then

$$\mathcal{M}(f) \overset{\text{def}}{=} \int_{\mathbb{R}\times\mathbb{R}_+} f(u,v)\, \mathcal{M}(\mathrm{d}u, \mathrm{d}v) = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} f(u,v)\, \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v).$$

The point process $\mathcal{M}$ is identified by the sequence $((r_n, \sigma_{2,n}), n \in \mathbb{Z})$, where r_n are the times of messenger birth conditionally to the gene activation and $\sigma_{2,n}$ the corresponding lifetimes. For this reason the point process $\mathcal{M}$ can be represented also as $\mathcal{M} = \sum_n \delta_{(r_n, \sigma_{2,n})}$ hence the expression (2.3.4).

Proposition 2.3.3. *If the distribution of the lifetime of a mRNA is $F_2(\mathrm{d}x)$, the average of the number M of mRNAs at equilibrium is given by*

$$\mathbb{E}(M) = \delta_+\rho_2 = \delta_+\lambda_2\mathbb{E}(\sigma_2).$$

The variance of M is

$$\mathrm{var}(M) = \mathbb{E}(M) + 2\rho_2^2\delta_+(1-\delta_+)\int_0^{+\infty}\int_0^{+\infty} e^{-\Lambda v}\overline{F}_2(u)\overline{F}_2(u+v)\,\mathrm{d}u\,\mathrm{d}v \tag{2.3.5}$$

where $F_2(x) = F_2([0,x])$ and $\overline{F}_2(x) = (1-F_2(x))/\mathbb{E}(\sigma_2)$, $\Lambda = \lambda_1^+ + \lambda_1^-$ and $\delta_+ = \lambda_1^+/\Lambda$.

The formula for $\mathbb{E}(M)$ is in fact quite intuitive: $\delta_+\lambda_2$ is the production rate of mRNAs and $\mathbb{E}(\sigma_2)$ is their mean lifetime.

Proof. Conditionally on the process $(Y(t))$, M follows a Poisson distribution, hence for $z \in [0,1]$,

$$\begin{aligned}\mathbb{E}\left(z^M \mid (Y(t))\right) &= \exp\left(-\lambda_2(1-z)\int_{-\infty}^0\int_{\mathbb{R}_+} \mathbb{1}_{\{u\le 0<u+v,Y(u)=1\}}\,\mathrm{d}u\,F_2(\mathrm{d}v)\right)\\ &= \exp\left(-\lambda_2(1-z)\int_{-\infty}^0 \mathbb{1}_{\{Y(u)=1\}}\mathbb{P}(\sigma_2\ge -u)\,\mathrm{d}u\right)\end{aligned} \tag{2.3.6}$$

by taking $f(u,v) = -\log(z)\mathbb{1}_{\{Y(u)=1,u\le 0,u+v>0\}}$ in Relation (A.0.1). If we differentiate formula (2.3.6) with respect to z and take $z=1$, we obtain

$$\mathbb{E}\left(M \mid (Y(t))\right) = \lambda_2\int_{-\infty}^0 \mathbb{1}_{\{Y(u)=1\}}\mathbb{P}(\sigma_2\ge -u)\,\mathrm{d}u.$$

Since $(Y(t))$ is at equilibrium, $\mathbb{P}(Y(u)=1) = \delta_+$, hence integrating the last relation gives

$$\mathbb{E}(M) = \delta_+\lambda_2\int_{-\infty}^0 \mathbb{P}(\sigma_2\ge -u)\,\mathrm{d}u = \delta_+\lambda_2\mathbb{E}(\sigma_2).$$

If we differentiate twice Formula (2.3.6) and substitute $z=1$, we obtain

$$\begin{aligned}\mathbb{E}(M(M-1)\mid (Y(t))) &= \lambda_2^2\left(\int_0^{+\infty}\mathbb{1}_{\{Y(-u)=1\}}\mathbb{P}(\sigma_2\ge u)\,\mathrm{d}u\right)^2\\ &= \lambda_2^2\int_{\mathbb{R}_+^2}\mathbb{1}_{\{Y(-u)=1,Y(-v)=1\}}\mathbb{P}(\sigma_2\ge u)\mathbb{P}(\overline{\sigma_2}\ge v)\,\mathrm{d}u\,\mathrm{d}v,\end{aligned}$$

which, integrated with respect to $(Y(t))$, gives

$$\mathbb{E}\left(M^2\right) - \mathbb{E}(M) = \lambda_2^2\int_{\mathbb{R}_+^2}\mathbb{P}(Y(-u)=1,Y(-v)=1)\mathbb{P}(\sigma_2\ge u,\overline{\sigma_2}\ge v)\,\mathrm{d}u\,\mathrm{d}v,$$

where the random variable $\overline{\sigma_2}$ is independent of σ_2 and has the same distribution. Using relation (2.3.1), for $u\le v$ and $\Lambda = \lambda_1^+ + \lambda_1^-$, we get

$$\begin{aligned}\mathbb{P}(Y(-u)=1,Y(-v)=1) &= \mathbb{P}(Y(-v)=1)\mathbb{P}(Y(-u)=1\mid Y(-v)=1)\\ &= \delta_+\left(\delta_+ + (1-\delta_+)e^{-\Lambda(v-u)}\right).\end{aligned}$$

Therefore $\mathbb{E}(M^2) - \mathbb{E}(M)$ is the sum of

$$\lambda_2^2 \delta_+^2 \int_{\mathbb{R}_+^2} \mathbb{P}(\sigma_2 \geq u, \overline{\sigma_2} \geq v) \, \mathrm{d}u \, \mathrm{d}v = (\lambda_2 \delta_+ \mathbb{E}(\sigma_2))^2 = (\mathbb{E}(M))^2$$

and, up to the multiplicative factor $2\lambda_2^2 \delta_+ (1 - \delta_+)$, of

$$\int_{\mathbb{R}_+^2} \mathbb{P}(\sigma_2 \geq u, \overline{\sigma_2} \geq v) e^{-\Lambda(v-u)} \mathbb{1}_{\{u \leq v\}} \, \mathrm{d}u \, \mathrm{d}v.$$

The proposition is proved. □

Relative variance

The relative variance of the number M of mRNAs at equilibrium, see (2.3.2), is defined as

$$\frac{\mathrm{var}(M)}{\mathbb{E}[M]^2} = \frac{1}{\mathbb{E}[M]} + 2 \frac{1 - \delta_+}{\delta_+} I_{F_2}, \tag{2.3.7}$$

by relation (2.3.5), with

$$I_{F_2} = \int_0^{+\infty} e^{-\Lambda v} \overline{F}_2(u) \overline{F}_2(u+v) \, \mathrm{d}u \, \mathrm{d}v,$$

where $\Lambda = \lambda_1^+ + \lambda_1^-$. When the mean $\mathbb{E}(M)$ is fixed, I_{F_2} is the only quantity which depends on the distribution of the lifetime of an mRNA.

To conclude this section, we now apply the previous general formulas to specific choices of the probability distribution. In particular, we will get an analytic formula of the previous for exponential and deterministic distributions. These assumptions are not completely realistic from a biological point of view; nevertheless they are used to stress the impact of probability distribution on the messenger variance. If the distribution of the lifetime of an mRNA is the exponential distribution E_{μ_2} with parameter μ_2, one gets

$$I_{E_{\mu_2}} = \frac{1}{2\mu_2(\Lambda + \mu_2)}.$$

If the lifetime of an mRNA is the deterministic distribution D_{μ_2} with a unit mass at $1/\mu_2$, the above formula yields

$$I_{D_{\mu_2}} = \frac{1}{\Lambda^2} \left(e^{-\Lambda/\mu_2} - 1 + \frac{\Lambda}{\mu_2} \right).$$

Straightforward calculations with these formulas show that $I_{E_{\mu_2}} \leq I_{D_{\mu_2}}$. The ratio $I_{D_{\mu_2}} / I_{E_{\mu_2}}$ varies in fact between 1 and 2, see Figure 2.3. The variance for the exponential distribution is smaller than the one for the deterministic distribution with the same mean. This result is not quite intuitive if one takes into account that the variance of the exponential distribution is quite large.

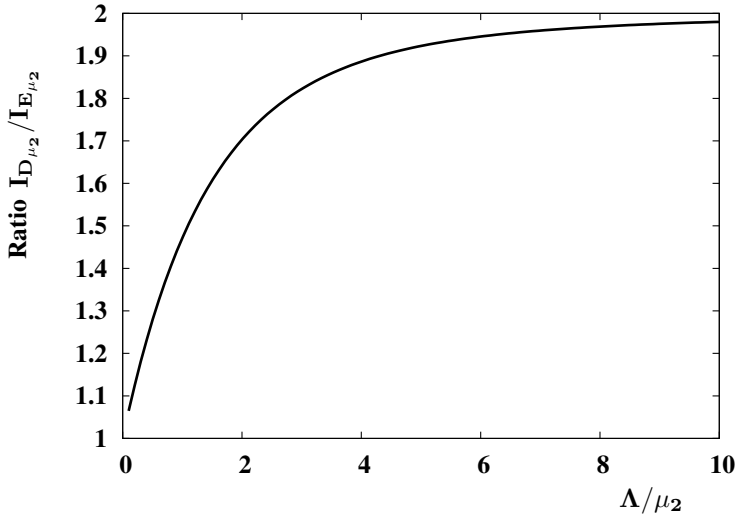


Figure 2.3: Variance of M: Comparison of $I_{D_{\mu_2}}$ and $I_{E_{\mu_2}}$ defined by (2.3.7). Deterministic versus Exponential.

2.3.3 Proteins

Recall that if an mRNA is created at time u and has a lifetime v, then on the time interval $[u, u+v]$ proteins are created according to the marked Poisson point process $\mathcal{N}^u_{\lambda_3}$ with intensity $\lambda_3 \, \mathrm{d}x \otimes F_3(\mathrm{d}y)$. Denote with $(s_n, n \in \mathbb{Z})$ the sequence of mRNA births, the point processes $\mathcal{N}^{s_n}_{\lambda_3}$ are supposed to be independent. The instants of creation of proteins together with their lifetimes can thus be represented by the following point process

$$\mathcal{P} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathcal{M}(\mathrm{d}u, \mathrm{d}v) \int_{[u,u+v]\times\mathbb{R}_+} \delta_{(x,y)} \, \mathcal{N}^u_{\lambda_3}(\mathrm{d}x, \mathrm{d}y), \tag{2.3.8}$$

where $\mathcal{M}$ is the point process defined by formula (2.3.2).

Proposition 2.3.4. *The number P of proteins at equilibrium can be represented by the random variable*

$$P = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \, \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v) \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x\le 0\le x+y, u\le x\le u+v\}} \mathcal{N}^u_{\lambda_3}(\mathrm{d}x, \mathrm{d}y). \tag{2.3.9}$$

Proof. The derivation is quite straightforward. If an mRNA alive between time u and $u+v$ generates a protein at time x with lifetime y, this protein will be present at time 0 if $x \le 0 \le x+y$. The argument that this is indeed the representation of the number of proteins at equilibrium follows the same lines of the proof of Proposition 2.3.2. □

Theorem 2.3.5. *If the distribution of the lifetime of a mRNA [resp. protein] is $F_2(\mathrm{d}x)$ [resp. $F_3(\mathrm{d}y)$], then the expected value of the random variable P, which is the number of proteins at equilibrium, is given by*

$$\mathbb{E}(P) = \delta_+ \rho_2 \rho_3 = \delta_+ \lambda_2 \lambda_3 \mathbb{E}(\sigma_2) \mathbb{E}(\sigma_3)$$

and its variance $\mathrm{var}(P)$ *can be expressed as*

$$\mathrm{var}(P) = \mathbb{E}(P) + \lambda_2\rho_3^2\delta_+ \int_0^{+\infty} \int_{\mathbb{R}_+} \left[\int_{-s}^{(-s+t)\wedge 0} \overline{F}_3(u)\,\mathrm{d}u\right]^2 \mathrm{d}s F_2(\mathrm{d}t) \tag{2.3.10}$$
$$+ \rho_2^2\rho_3^2\delta_+(1-\delta_+) \int_{\mathbb{R}_+^4} e^{-\Lambda|(u_1-u_2)+(v_1-v_2)|} \prod_{i=1}^{2} \overline{F}_2(u_i)\overline{F}_3(v_i)\,\mathrm{d}u_i\,\mathrm{d}v_i,$$

where, for $j = 2,\ 3,\ F_j(x) = F_j([0,x])$ *and* $\overline{F}_j(x) = (1-F_j(x))/\mathbb{E}(\sigma_j)$, $\Lambda = \lambda_1^+ + \lambda_1^-$ *and* $\delta_+ = \lambda_1^+/\Lambda$.

The expression of $\mathbb{E}(P)$ can also be understood intuitively. Recall that $\mathbb{E}(M) = \delta_+\lambda_2\mathbb{E}(\sigma_2)$ is the mean number of mRNAs, $\lambda_3\mathbb{E}(M)$ is therefore the production rate of proteins and $\mathbb{E}(\sigma_3)$ is their mean lifetime.

Proof. We start with the simple case of the mean. For fixed $u,\ v \in \mathbb{R}_+$, formula (A.0.3) gives

$$\mathbb{E}\left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}} \mathcal{N}_{\lambda_3}^u(\mathrm{d}x,\mathrm{d}y)\right] = \lambda_3 \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y)$$

Integrating this expression with respect to $\mathbb{1}_{\{Y(u)=1\}}\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v)$ and taking its expectation, we get

$$\begin{aligned}
&\mathbb{E}[P \mid (Y(t))] = \\
&= \lambda_3\ \mathbb{E}\left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \mathbb{E}\left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}} \mathcal{N}_{\lambda_3}^u(\mathrm{d}x,\mathrm{d}y)\,\middle|\,(\mathcal{N}_{\lambda_2}),(Y(t))\right]\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v)\,\middle|\,(Y(t))\right] \\
&= \lambda_3\ \mathbb{E}\left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y)\right]\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v)\,\middle|\,(Y(t))\right] \\
&= \lambda_3 \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \left[\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y)\right]\lambda_2\,\mathrm{d}u\,F_2(\mathrm{d}v) \\
&= \lambda_2\lambda_3 \int_{\mathbb{R}_-\times\mathbb{R}_-} \mathbb{1}_{\{Y(u+x)=1\}}\mathbb{1}_{\{x\le 0\}}\mathbb{1}_{\{u\le 0\}}\mathbb{P}(\sigma_2\ge -u)\mathbb{P}(\sigma_3\ge -x)\,\mathrm{d}x\,\mathrm{d}u,
\end{aligned}$$

where we have used again formula (A.0.3). A further integration gives finally the expectation

$$\begin{aligned}
\mathbb{E}(P) &= \lambda_2\lambda_3 \int_{\mathbb{R}_-\times\mathbb{R}_-} \mathbb{P}(Y(u+x)=1)\mathbb{P}(\sigma_2\ge -u)\mathbb{P}(\sigma_3\ge -x)\,\mathrm{d}x\,\mathrm{d}u \\
&= \lambda_2\lambda_3\delta_+ \int_{\mathbb{R}_-\times\mathbb{R}_-} \mathbb{P}(\sigma_2\ge -u)\mathbb{P}(\sigma_3\ge -x)\,\mathrm{d}x\,\mathrm{d}u \\
&= \delta_+\lambda_2\mathbb{E}(\sigma_2)\lambda_3\mathbb{E}(\sigma_3).
\end{aligned}$$

Recall that $\mathcal{N}_{\lambda_2}$ can also be represented as $\mathcal{N}_{\lambda_2} = (s_n, \sigma_{2,n})$ and

$$P = \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(s_n)=1\}}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ s_n\le x\le s_n+\sigma_{2,n}}\right\}} \mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y).$$

Denote by $\widehat{\mathbb{E}}$ the conditional expectation $\mathbb{E}[\,\cdot\mid (Y(t)),(s_n,\sigma_{2,n})]$. The conditional generating function $\widehat{\mathbb{E}}\left[z^P\right]$ can be written as

$$\widehat{\mathbb{E}}\left(\prod_{n\in\mathbb{Z}}\exp\left(-\log(z)\int_{\mathbb{R}\times\mathbb{R}_+}\mathbb{1}_{\{Y(s_n)=1\}}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ s_n\le x\le s_n+\sigma_{2,n}}\right\}}\mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y)\right)\right)$$
$$=\prod_{n\in\mathbb{Z}}\widehat{\mathbb{E}}\left[\exp\left(-\log(z)\int_{\mathbb{R}\times\mathbb{R}_+}\mathbb{1}_{\{Y(s_n)=1\}}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ s_n\le x\le s_n+\sigma_{2,n}}\right\}}\mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y)\right)\right],$$

since the point processes $\mathcal{N}_{\lambda_3}^{s_n}$, $n\in\mathbb{Z}$, are independent.

The nth term of this product is, applying Proposition A.0.5 to the marked Poisson point processes $\mathcal{N}_{\lambda_3}^{s_n}$,

$$\exp\left(-\lambda_3(1-z)\mathbb{1}_{\{Y(s_n)=1\}}\int_{\mathbb{R}\times\mathbb{R}_+}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ s_n\le x\le s_n+\sigma_{2,n}}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y)\right).$$

By integrating $\widehat{\mathbb{E}}\left(z^P\right)$ with respect to $\mathcal{N}_{\lambda_2}$, the generating function can thus be written as

$$\mathbb{E}\left[z^P|(Y(t))\right]=\mathbb{E}\left[\exp\left(-\int_{\mathbb{R}\times\mathbb{R}_+}g(u,v)\mathcal{N}_{\lambda_2}(\mathrm{d}u,\mathrm{d}v)\right)\right],$$

where

$$g(u,v)=\lambda_3(1-z)\mathbb{1}_{\{Y(u)=1\}}\int_{\mathbb{R}\times\mathbb{R}_+}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le x\le u+v}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y).$$

Applying again Proposition A.0.5 to the marked Poisson point process $\mathcal{N}_{\lambda_2}$, we get

$$\mathbb{E}\left(z^P|(Y(t))\right)=$$
$$=\exp\left(-\lambda_2\int_{\mathbb{R}}\mathrm{d}u\int_{\mathbb{R}_+}F_2(\mathrm{d}v)\left(1-\exp\left(-\lambda_3(1-z)\int_{\mathbb{R}\times\mathbb{R}_+}\mathbb{1}_{\left\{\substack{x\le 0\le x+y,\\ u\le 0\le u+v,\\ Y(u+x)=1}\right\}}\,\mathrm{d}x\,F_3(\mathrm{d}y)\right)\right)\right).$$

In order to obtain an expression for $\mathbb{E}\left(P(P-1)|(Y(t))\right)$, we have to differentiate twice the previous formula with respect to z and evaluate it at $z=1$. The resulting formula should then be integrated with respect to $(Y(t))$ and we can get formula (2.3.10), by using similar arguments as in the proof of Proposition 2.3.3 (with more technical calculations). □

2.4 Results of MPPP description of Three-Stage model

To show the effectiveness of the analytic formula (2.3.10) of the protein variance, we consider the cases of exponential and deterministic distributions. More realistic distributions are considered in Figure 2.6 and 2.5. This specific analysis will give insights of the impact of the distribution choice on the protein variance.

Explicit formulas of protein variance

In each case the average lifetime of an mRNA [resp. protein] is $1/\mu_2$ [resp. $1/\mu_3$]. Recall that $\delta_+=\lambda_1^+/\Lambda$ and $\Lambda=\lambda_1^+-\lambda_1^-$. As for the case of mRNAs above, even if from a biological point

of view these assumptions are not completely realistic, this analysis shows the impact of the distribution on the variance, and therefore of the necessity of having closed form expressions for a large set of distributions.

Exponential Distribution. If the distribution of the lifetime of an mRNA [resp. protein] is exponential with parameter μ_2 [resp. μ_3], then formula (2.3.10) gives the classical result on the variance, see Paulsson [54],

$$\mathrm{var}^{(\mathrm{E})}(P) = \mathbb{E}(P)\left(1 + \frac{\lambda_3}{\mu_2+\mu_3} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_3)}{(\mu_2+\mu_3)(\Lambda+\mu_2)(\Lambda+\mu_3)}\right). \tag{2.4.1}$$

Deterministic Case. If the lifetime of an mRNA is exponentially distributed with parameter μ_2 and the protein lifetime is deterministic, equal to $1/\mu_3$, then formula (2.3.10) gives the identity

$$\begin{aligned}\mathrm{var}^{(\mathrm{D})}(P) = \mathbb{E}(P)\Bigg[1 + 2\frac{\lambda_3}{\mu_2}\left(1 - \frac{\mu_3}{\mu_2}(1-e^{-\mu_2/\mu_3})\right) \\ + \frac{2\lambda_2\lambda_3(1-\delta_+)\mu_2}{\Lambda^2-\mu_2^2}\left(\frac{\mu_3}{\Lambda^2}\left[1-e^{-\Lambda/\mu_3}\right]\right. \\ \left.\left. - \frac{\mu_3}{\mu_2^3}\Lambda\left[1-e^{-\mu_2/\mu_3}\right] + \Lambda\left[\frac{1}{\mu_2^2}-\frac{1}{\Lambda^2}\right]\right)\Bigg].\end{aligned} \tag{2.4.2}$$

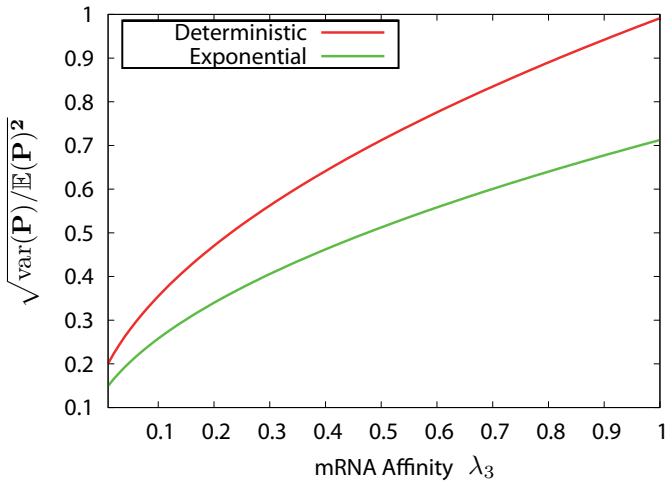


Figure 2.4: Square root of relative variance of the number of proteins with a fixed mean. These curves are obtained using the analytic formulas (2.4.1) and (2.4.2).

Numerical analysis: a counter-intuitive result

Relation (2.3.10) gives an explicit, but intricate expression for the variance, we present some numerical experiments based on this formula. Figures 2.6, 2.5 and 2.4 consider the case when the average number of proteins at equilibrium is fixed and equal to 300, that $\lambda_2 = 0.02\mathrm{sec.}^{-1}$,

$\lambda_1^- = 0.01\text{sec.}^{-1}$ and that the average of the lifetime of an mRNA [resp. protein] is 172sec. [resp. 1000sec.]. We have considered several possible choices for the distribution F_3, it is assumed that all the other distributions are exponential. When the distribution F_3 is Gaussian, S denotes its variance.

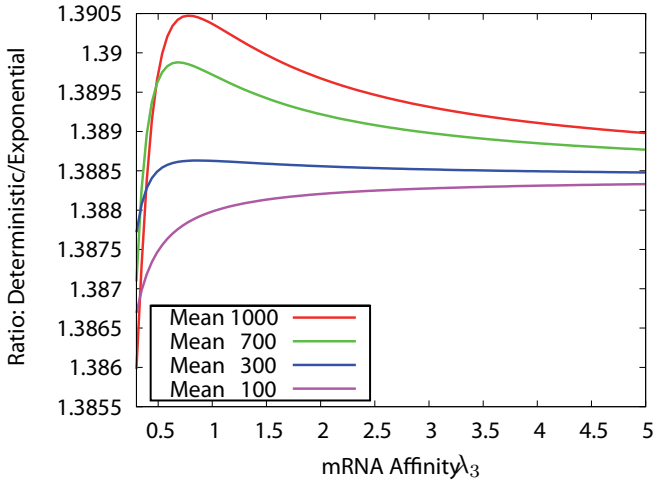


Figure 2.5: Ratio of the square root of relative variances of the number of proteins with deterministic and exponential lifetimes respectively. Each curve corresponds to different level of protein expression and is obtained using the analytic formulas (2.4.1) and (2.4.2). Note that the mean is fixed along each curve.

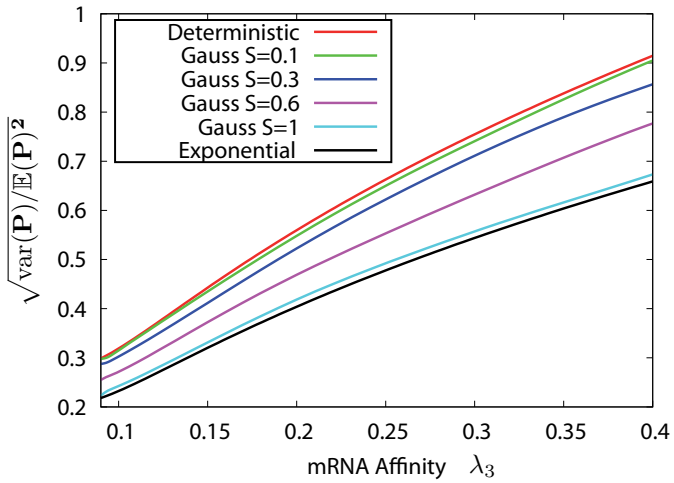


Figure 2.6: Square root of relative variance of the number of proteins with a fixed mean. The curves corresponding to Gaussian distribution are indexed with respect to the parameter $S = \sigma_{\mathcal{N}}/\mathbb{E}(\mathcal{N})$, where $\mathbb{E}(\mathcal{N})$ is the average of the Gaussian lifetime and $\sigma_{\mathcal{N}}$ its standard deviation. The curves are obtained via Monte Carlo simulations and the statistics are obtained numerically.

The counter-intuitive result studied at the end of Section 2.3.2 is now observed in the case of proteins. The exponential distribution has large variance, in fact if X is exponentially distributed with parameter λ, then $\mathrm{var}(X) = \mathbb{E}^2(X)$, while a deterministic lifetime has obviously no fluctuations. Surprisingly, by replacing a noisy exponentially distributed protein lifetime with a deterministic one, the fluctuations of the protein number result increased. This unexpected result is valid for any choice of parameters and is shown in figure 2.4.

If we consider a Gaussian distribution for the protein lifetime, the curves are comprised in between the exponential and deterministic cases. Moreover, the profiles of variance corresponding to a normal distribution show a precise behavior: the profile corresponding to a Gaussian random variable with small variance is close to the deterministic curve and we obtain a monotone decreasing sequence of curves as long as we increase the variance. These results are shown in both figures 2.6 and 2.5.

2.A Appendix: survey of few classic models of gene expression

A systematic and accurate use of stochastic models to investigate gene expression can be traced since the late 70s with the models proposed by David R. Rigney in the late 70s ([64], [63]) and by Otto G. Berg [5]. These models are still key references for the models of stochastic gene expression, since, despite the many simplifications and assumptions, the authors develop a complete mathematical analysis of their model, proposing analytic formulas for the first moments of protein numbers in a single bacterium and within a population. Both works predicted mRNA and proteins fluctuations taking into account deterministic cell growth, chromosome replication at fixed time and partitioning of proteins between daughter cells.

These pioneering works and the techniques they used, were brushed up in the 90s by the works of Peccoud and Ycart [57] and McAdams and Arkin [43]. In the last paper the authors model gene expression using a stochastic formulation of chemical kinetics derived by Gillespie [19], predicting noticeable variability in protein numbers between individual cells. This was the beginning of an intense research and studies on stochastic models of gene expression (Paulsson *et alii* [55, 53, 54], Swain *et alii* [69, 67], ...).

In this section we present some of the most relevant models that can be found in the literature that deal with stochastic modeling of gene expression in bacteria, concentrating on the models which have been taken as a first reference for our work and serve as introduction to the classic techniques used.

2.A.1 The Rigney's model

We focus mainly on the model of protein production in a single cell, see [64]. We will discuss briefly the model at population level, which considers new features such as as cell age and cell division.

Introduction

Until recent years, experiments measured *in vivo* protein synthesis as an average over the subpopulations of cells in each sample. However, few experiments in he early seventies observed that the average number of molecules for certain enzymes increases linearly with respect to time, during a cell cycle. When the gene is duplicated this behavior is still present, but it proceeds with twice the previous rate.

These observations served as starting point of the modeling work of Rigney and Schieve, which, up to our knowledge, were among the first to introduce a probabilistic description of the synthesis of each protein specie within a single bacterium. There were in fact physical reasons to believe that chaotic motion of molecules manifests in form of fluctuations and also direct a visual evidence: the times between successive mRNA transcription initiations, by the DNA-directed-RNA polymerase at given promoter, are randomly distributed, as was shown by Miller, Hamkalo and Thomas in 1970. The subsequent transcription and translation events were also thought to be random processes and a stochastic description of the phenomenon seemed natural.

The model

The model described in [64] should predict on one side the linearity in the average production rate, on the other side it should allow individual cells to show fluctuations in their individual rates of synthesis.

The rate at which each protein species is produced in each cell is supposed to be principally determined by the random frequency at which polymerase molecules bind to the corresponding promoter site on the DNA.

The promoter region corresponding to a specific protein may be in one of two possible states: either a polymerase molecule is bound to the promoter (state "on"), or it is not (state "off").

This model does not considered at all the messenger degradation, nor the bacterial dilution. The interest is focused mainly on the transition in which a bound polymerase molecule leaves the promoter region to begin the production of mRNA strand. We denote with N_t the number of times this particular promoter becomes free over an arbitrary period of time t and it is assumed

$$N_t \propto (\text{produced enzymes because of this promoter}),$$

i.e. the number of transitions is proportional to the number of enzyme molecules eventually produced. This allows to transfer directly the analysis on the promoter to the protein level.

The transitions allowed by the model may be described as a Markov jump process. With few calculations is possible to compute the probability distribution for the time between successive mRNA initiations at a specific promoter. This distribution can be used to calculate the probability of certain mRNA initiations over the time period t.

Suppose that cells have been growing under constant conditions for several generations, then the fraction of promoters in each of the states of the model will not change in time, that means that the random process has reached a stationary state (there are changes at level of single promoters but not at population level).

The model is intended to quantify the number of mRNA of a particular protein type initiated on the ensemble of promoters over a period t. The problem can be stated in the context of the *renewal theory*, where the holding times are here the time of detachment plus the time of attachment of the promoter, and where the renewal process is represented by the number of mRNA initiations N_t. The standard renewal theory tells us that the average number of mRNAs initiated on the promoters over a period t is linearly proportional to time. Since the number of proteins is assumed to be proportional to initiated mRNA, the protein number should be linearly proportional to time. These results are based on the assumption that the transition parameters do not change during the cell cycle.

Here we consider the two-states model as in [64], but we make direct calculations and we slightly change notation with respect to the original paper.

Suppose that the promoter region of a particular gene has two possible states: state "0" the promoter is free, state "1" the promoter is associated with a polymerase. The transition from 1 to 0 correspond to a mRNA initiation.

Be $X_t \in \{0, 1\}$ a stochastic process such that the time E_{λ_0} it stays in state "0" is exponentially distributed with parameter λ_0, while the time it stays in state "1" is exponentially distributed with parameter λ_1 and is denoted by E_{λ_1}. The process X_t is a continuous time Markov chain with state space $I = \{0, 1\}$ and with $Q-$matrix

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{pmatrix}.$$

The Kolmogorov forward equations are

$$\begin{aligned} P_t' &= P_t\, Q \\ P_0(i, j) &= \delta_{ij}, \end{aligned}$$

which admit the solutions

$$P_t(1,1) = \frac{\lambda_1}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)t} + \frac{\lambda_0}{\lambda_0+\lambda_1} \tag{2.A.1}$$

$$P_t(0,0) = \frac{\lambda_0}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)t} + \frac{\lambda_1}{\lambda_0+\lambda_1} \tag{2.A.2}$$

$$P_t(0,1) = -\frac{\lambda_0}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)t} + \frac{\lambda_0}{\lambda_0+\lambda_1} \tag{2.A.3}$$

$$P_t(1,0) = -\frac{\lambda_1}{\lambda_0+\lambda_1}e^{-(\lambda_0+\lambda_1)t} + \frac{\lambda_1}{\lambda_0+\lambda_1}. \tag{2.A.4}$$

The stationary distribution is therefore given by $\pi(0) = \frac{\lambda_1}{\lambda_0+\lambda_1}$, $\pi(1) = \frac{\lambda_0}{\lambda_0+\lambda_1}$.

In order to evaluate the number of produced mRNAs we have to compute the time interval between two subsequent messenger initiation. Denote with T the random variable

$$T = E_{\lambda_0} + E_{\lambda_1}. \tag{2.A.5}$$

To compute the probability density function (*p.d.f.*) of T, we use the Laplace transform as in the original paper of Rigney. Since E_{λ_0} and E_{λ_1} are independent, and since the Laplace transform of a random variable exponentially distributed with parameter λ_i is given by

$$\mathcal{L}(E_{\lambda_i})(z) = \mathbb{E}\left[e^{-zE_{\lambda_i}}\right] = \frac{\lambda_i}{\lambda_i+z}, \tag{2.A.6}$$

then

$$\begin{aligned}\mathcal{L}(T)(z) &= \mathbb{E}\left[e^{-zT}\right] = \mathbb{E}\left[e^{-z(E_{\lambda_0}+E_{\lambda_1})}\right] \\ &= \mathbb{E}\left[e^{-zE_{\lambda_0}}\right]\mathbb{E}\left[e^{-zE_{\lambda_1}}\right] = \frac{\lambda_0\lambda_1}{(\lambda_0+z)(\lambda_1+z)}.\end{aligned} \tag{2.A.7}$$

The previous formula can be written as a sum of the type $\mathcal{L}(T)(z) = A_0\frac{\lambda_0}{\lambda_0+z} + A_1\frac{\lambda_1}{\lambda_1+z}$, with

$$A_0 = \frac{\lambda_1}{\lambda_1-\lambda_0} \qquad A_1 = \frac{\lambda_0}{\lambda_0-\lambda_1}.$$

Using formula (2.A.6), we have that the *p.d.f* of the random variable T is given by

$$f_T(x) = A_0\lambda_0 e^{-\lambda_0 x} + A_1\lambda_1 e^{-\lambda_1 x} = \frac{\lambda_0\lambda_1}{\lambda_1-\lambda_0}(e^{-\lambda_0 x} - e^{-\lambda_1 x}). \tag{2.A.8}$$

Using the previous formula we are able to compute the average inter-initiation time, which is

$$\mathbb{E}[T] = \int_0^\infty t f_T(t)dt = \frac{\lambda_0+\lambda_1}{\lambda_0\lambda_1}. \tag{2.A.9}$$

It is possible to relate the distribution of the process N_t with the time T between two occurrences. In fact if we denote with T_i, $i = 1, 2\ldots$, the time between the $(i-1)^{\text{th}}$ and i^{th} event, and with

$$S_r = \sum_{i=1}^{r} T_i \tag{2.A.10}$$

the total time up to the r^{th} occurrence, then we have the identity

$$\{N_t < r\} = \{S_r > t\}.$$

It is possible to derive then the statistics of the process N_t and it results that the average number of initiated mRNAs over the interval $[0, t]$ is

$$\mathbb{E}[N_t] = \frac{t}{\mathbb{E}[T]} = \left[\frac{\lambda_0 + \lambda_1}{\lambda_0 \lambda_1}\right] t. \tag{2.A.11}$$

Moreover the variance of the number of initiated mRNAs gives the formula

$$\mathrm{Var}\,(N_t) = \left(\frac{\lambda_0\lambda_1}{\lambda_0 + \lambda_1} - \frac{2(\lambda_0\lambda_1)^2}{(\lambda_0 + \lambda_1)^3}\right) t + \frac{2(\lambda_0\lambda_1)^2}{(\lambda_0 + \lambda_1)^4}(1 - e^{-(\lambda_0+\lambda_1)t}). \tag{2.A.12}$$

Rigney and Schieve in this article point out how the noise is inherent of the protein production and not a mere consequence of the measurements. The mean mRNA can be derived also by averaging over a population, but the variance gives important insights of the kinetics of the protein production itself.

This model has lots of simplifications and look just on a specific part of the whole protein production. Nevertheless, it shows how the analysis of a simple stochastic model of the protein production can supply important information on the kinetics and, consequently, on the biological and chemical mechanisms that lie behind gene expression.

Towards a population model

In the paper of Rigney 1979 [63] the author analyses few problematics concerning a population of growing cells. The starting point of the author is the fact that the 90% of the DNA of bacteria growing in a glucose are rarely transcribed. Let consider *constitutive genes*, i.e. promoter-controlled structural genes which are rarely transcribed. The latter characteristic make stochastic modeling well-suited for the analysis of the protein production to give an estimation of the variability of the number of produced proteins.

Rigney considers here a population that is able to reach a steady-state, i.e. all the experimental conditions and the characteristics of the bacterial population are constants and are such that the probability that a cell, selected randomly, shows a property is independent of time. Such a population is in the *exponential growth phase*.

The population can be divided in sub-populations, of a specific window of ages $[a, a + \Delta a]$, where a is the "age" with respect to cell cycle. This is fundamental when considering the population perspective, since we may not expect that the cells are synchronous. The aim of the Rigney's population model is to predict distributions of low-level, constitutive proteins.

The main random process under analysis is the number $X(a)$ of a specific protein in a cell in the bacterial population. Since the author deals with a whole asynchronous population of bacteria, he makes further simplifications in order to have a treatable model. The probability λ that transcription initiates is supposed to be constant, independently of the cell age, the number of polymerases, chromosome conformation and other characteristics of the cell itself. Each mRNA is supposed to produce deterministically s proteins during its lifetime, and these s molecules are supposed to be produced simultaneously since the inter-initiation time for constitutive proteins is much greater of the mRNA lifetime, and of the order of the cell lifetime. In order to consider cell growth and division still being able to perform mathematical analysis, Rigney supposes that there is one or two copies of the gene depending on cell age. This is an approximation since the number of gene copies depends on various factors, *in primis* on the position of the gene in the DNA helix. Moreover, the duplication time τ_1 of the gene and the time, or age, of cell division τ_2 are supposed to be deterministic, supposing that the proteins in the mother cell to be equally probabilistically divided between the daughter cells.

The protein degradation is supposed to be exponentially distributed with parameter μ.

The main mathematical tool used here are generating functions. The author starts by studying the probability $P(x_0 \to x)$ that a cell, having x_0 molecules of the specific protein at the time of its division, has a daughter cell having x molecules at the time of its own division. This probability distribution can be used to derive the generating function of $\pi(x, \tau_2)$, which is the probability to have x molecules in a cell just before division, under the hypothesis that the system has reached the equilibrium. This last distribution gives finally the generating function for $\pi(x, a)$, which is the probability to have x molecules of the protein in a cell of age a. The author concludes the technical computations by showing how derive the moments of the probability distribution.

Rigney proposes then a two-variables extension of its model. In this extension he considers two proteins, which have different production characteristics. We point out that he does not consider any competition in the production chain of the two-proteins model. The main difficulty is that the final distribution will depend on the quantities of the two proteins.

The author underlines the importance of the hypothesis for the analytic tractability of the model and he states

> [...] for models in which the initiation or degradation transition probabilities are non-linear functions of the random variables, it will generally be impossible to find exact, analytical solutions.

Despite the simplifications made by Rigney, one of the key point of his population model is that he derives the generating function for the probability distribution at population level.

2.A.2 Paulsson's model survey

Johan Paulsson is one of the most prominent researcher in the field of stochastic gene expression and this is testified by the number of important works he has published on this argument [55, 53, 54].

We focus on the Paulsson's work [54] in which the author focuses on the common characteristics of many stochastic models of gene expression, giving a widespread perspective on the field.

Introduction

Stochastic models are proposed to take into account the fact that cellular events depends on the random collision of molecules. Since many of the cellular events deal with small numbers of molecules and are not independent, a probabilistic description is often the best description of such processes. A stochastic model is fundamental if we want to investigate and understand the process of gene expression, even in cases in which randomness is absent. In fact this lack of randomness should be somehow explained in probabilistic terms.

Most of the stochastic models that can be found in literature focus on genes, mRNAs and proteins, including all the other processes in effective transition rates. This fundamental description may be indefinitely complicated, but in order to better understand the first principles the author chooses to focus just three main processes: gene activation/deactivation, transcription and translation.

The model

Paulsson's supposes to have a constant number of copies of the specific gene, denoted with $n_1^{\max}$, each copy switching independently between two states (active/inactive). The cell growth, cell

division and gene replication are not included in this model. The author denotes with n_1 the number of active genes.

The three main processes modeled, i.e. active genes, mRNAs and proteins, are supposed to chase each other. In particular the active genes will affect the number n_2 of mRNAs, which in turn will affect the number n_3 of proteins. Messengers and proteins are both degraded. The first have short lifetimes, especially in eukaryotes, and are destroyed by the degradative machinery, while proteins are more sensitive to the dilution which they undergo because of cell growth.

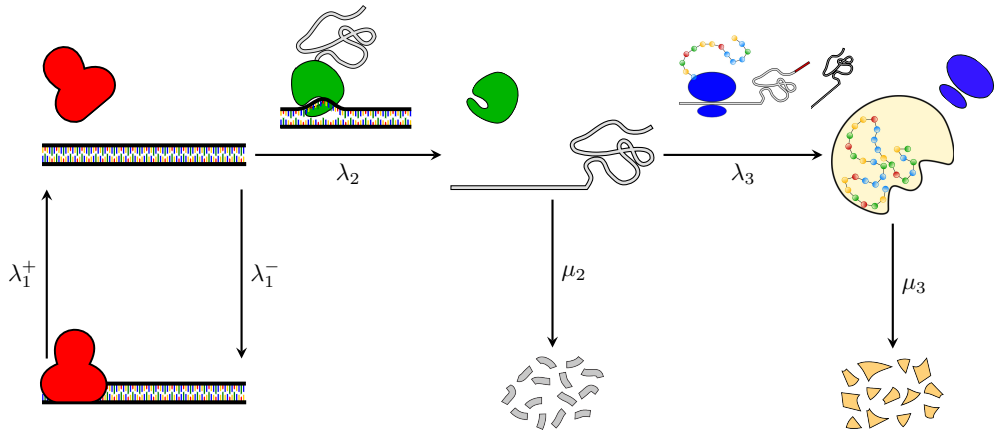


Figure 2.7: Three-Stage model. The gene activation/deactivation occur at rate λ_1 and μ_1 respectively. Transcription and translation occur at rates λ_2 and λ_3 respectively, while the mRNA and protein degradation are supposed to be exponentially distributed with rates μ_2 and μ_3.

The Kolmogorov equations for the Markov process $(n_1(t), n_2(t), n_3(t))$ are then

$$\begin{aligned}\frac{\mathrm{d}P(n_1,n_2,n_3)}{\mathrm{d}t} =&\lambda_1(n_1^{\max}-n_1+1)P(n_1-1,n_2,n_3)-\lambda_1(n_1^{\max}-n_1)P(n_1,n_2,n_3)\\ &+\mu_1(n_1+1)P(n_1+1,n_2,n_3)-\mu_1 n_1 P(n_1,n_2,n_3)\\ &+\lambda_2 n_1 P(n_1,n_2-1,n_3)-\lambda_2 n_1 P(n_1,n_2,n_3)\\ &+\mu_2(n_2+1)P(n_1,n_2+1,n_3)-\mu_2 n_2 P(n_1,n_2,n_3)\\ &+\lambda_3 n_2 P(n_1,n_2,n_3-1)-\lambda_3 n_2 P(n_1,n_2,n_3)\\ &+\mu_3(n_3+1)P(n_1,n_2,n_3+1)-\mu_3 n_3 P(n_1,n_2,n_3),\end{aligned} \tag{2.A.13}$$

where

$$P(n_1,n_2,n_3)=\mathbb{P}\left[n_1(t)=n_1,n_2(t)=n_2,n_3(t)=n_3\right]. \tag{2.A.14}$$

First moments of the modeled processes

We detail now the techniques used to obtain analytic formulas of the first moments in many works found in literature. We may restrain our analysis to the case $n_1^{\max}=1$, since the production relative to each gene copy is independent. Therefore in the general case, because of this independence, then the average and variance of the number of messengers and proteins will result multiplied by the constant $n_1^{\max}$.

Since $n_1 \in \{0,1\}$, we may restate the Chapman-Kolmogorov equations (2.A.13) by conditioning to the random variable n_1. Define

$$f_{(x,y)}(t) = \mathbb{P}\left[n_2(t) = x, n_3(t) = y, n_1(t) = 0\right] \tag{2.A.15}$$
$$g_{(x,y)}(t) = \mathbb{P}\left[n_2(t) = x, n_3(t) = y, n_1(t) = 1\right], \tag{2.A.16}$$

then the Chapman-Kolmogorov equations for $f_{(x,y)}$ and $g_{(x,y)}$ read

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}f_{(x,y)} &= \mu_1 g_{(x,y)} - \lambda_1 f_{(x,y)} + \mu_2(x+1)f_{(x+1,y)} - \mu_2 x f_{(x,y)} + \lambda_3 x f_{(x,y-1)} \\ &\quad - \lambda_3 x f_{(x,y)} + \mu_3(y+1)f_{(x,y+1)} - \mu_3 y f_{(x,y)} \end{aligned} \tag{2.A.17}$$
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}g_{(x,y)} &= \lambda_1 f_{(x,y)} - \mu_1 g_{(x,y)} + \lambda_2 g_{(x-1,y)} - \lambda_2 g_{(x,y)} + \mu_2(x+1)g_{(x+1,y)} \\ &\quad - \mu_2 x g_{(x,y)} + \lambda_3 x g_{(x,y-1)} - \lambda_3 x g_{(x,y)} + \mu_3(y+1)g_{(x,y+1)} - \mu_3 y g_{(x,y)}, \end{aligned} \tag{2.A.18}$$

where, for typographic issues, we have omitted to explicitly report the dependence on the time variable t of $f_{(x,y)}$ and $g_{(x,y)}$.

Define the generating functions

$$F(z,s) = \sum_{x,y} z^x s^y f_{(x,y)} \tag{2.A.19}$$
$$G(z,s) = \sum_{x,y} z^x s^y g_{(x,y)}. \tag{2.A.20}$$

The Kolmogorov equations (2.A.17) and (2.A.17) are finally transformed to the equations

$$\begin{aligned}\frac{\partial F}{\partial t} &= \mu_1 G(z,s) - \lambda_1 F(z,s) - \mu_2(z-1)\frac{\partial F}{\partial z}(z,s) + \lambda_3 z(s-1)\frac{\partial F}{\partial z}(z,s) \\ &\quad - \mu_3(s-1)\frac{\partial F}{\partial s}(z,s) \end{aligned} \tag{2.A.21}$$
$$\begin{aligned}\frac{\partial G}{\partial t} &= \lambda_1 F(z,s) - \mu_1 G(z,s) + \lambda_2(z-1)G(z,s) - \mu_2(z-1)\frac{\partial G}{\partial z}(z,s) \\ &\quad + \lambda_3 z(s-1)\frac{\partial G}{\partial z}(z,s) - \mu_3(s-1)\frac{\partial G}{\partial s}(z,s) \end{aligned} \tag{2.A.22}$$

The aim of this model is to obtain analytic close formula of protein average and variance. Therefore we may obtain from equations (2.A.21) and (2.A.22), simplified equations for the protein average and variance. Observe that

$$\frac{\partial F}{\partial z}(z,s)\bigg|_{z=s=1} = \mathbb{E}\left[n_2, n_1 = 0\right] \qquad \frac{\partial G}{\partial z}(z,s)\bigg|_{z=s=1} = \mathbb{E}\left[n_2, n_1 = 1\right]$$
$$\frac{\partial F}{\partial s}(z,s)\bigg|_{z=s=1} = \mathbb{E}\left[n_3, n_1 = 0\right] \qquad \frac{\partial G}{\partial s}(z,s)\bigg|_{z=s=1} = \mathbb{E}\left[n_3, n_1 = 1\right],$$

where $\mathbb{E}\left[\,\cdot\,, n_1 = 0\right]$ is the expected value corresponding to density function (2.A.17) and $\mathbb{E}\left[\,\cdot\,, n_1 = 1\right]$ using density function (2.A.18). Note that from (2.A.22) we have

$$\frac{\partial G}{\partial t}(z,s)\bigg|_{z=s=1} = \lambda_1\left(1 - G(z,s)\right)\bigg|_{z=s=1} - \mu_1 G(z,s)\bigg|_{z=s=1},$$

since $G(1,1) = \mathbb{P}[n_1 = 1]$ and $F(1,1) = \mathbb{P}[n_1 = 0] = 1 - \mathbb{P}[n_1 = 1]$. We obtain therefore the formulas for the averages of the messengers and of the proteins

$$\mathbb{E}[\text{mRNA}] = \mathbb{E}[n_2] = \mathbb{E}[n_2, n_1 = 0] + \mathbb{E}[n_2, n_1 = 1], \tag{2.A.23}$$
$$\mathbb{E}[\text{P}] = \mathbb{E}[n_3] = \mathbb{E}[n_3, n_1 = 0] + \mathbb{E}[n_3, n_1 = 1]. \tag{2.A.24}$$

In analogy to first derivatives formulas, we compute the second derivatives, which will be used in the following:

$$\frac{\partial^2 F}{\partial z^2}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_2(n_2-1), n_1 = 0] \qquad \frac{\partial^2 G}{\partial z^2}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_2(n_2-1), n_1 = 1]$$
$$\frac{\partial^2 F}{\partial s^2}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_3(n_3-1), n_1 = 0] \qquad \frac{\partial^2 G}{\partial s^2}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_3(n_3-1), n_1 = 1]$$
$$\frac{\partial^2 F}{\partial z \partial s}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_2 n_3, n_1 = 0] \qquad \frac{\partial^2 G}{\partial z \partial s}(z,s)\bigg|_{z=s=1} = \mathbb{E}[n_2 n_3, n_1 = 1].$$

The average number of messengers is described by the differential equations

$$\frac{\partial}{\partial t}\mathbb{E}[n_2, n_1 = 0] = \mu_1 \mathbb{E}[n_2, n_1 = 1] - \lambda_1 \mathbb{E}[n_2, n_1 = 0] - \mu_2 \mathbb{E}[n_2, n_1 = 0] \tag{2.A.25}$$
$$\frac{\partial}{\partial t}\mathbb{E}[n_2, n_1 = 1] = \lambda_1 \mathbb{E}[n_2, n_1 = 0] - \mu_1 \mathbb{E}[n_2, n_1 = 1] + \lambda_2 G(1,1) - \mu_2 \mathbb{E}[n_2, n_1 = 1], \tag{2.A.26}$$

while the equations

$$\frac{\partial}{\partial t}\mathbb{E}[n_3, n_1 = 0] = \mu_1 \mathbb{E}[n_3, n_1 = 1] - \lambda_1 \mathbb{E}[n_3, n_1 = 0] + \lambda_3 \mathbb{E}[n_2, n_1 = 0] - \mu_3 \mathbb{E}[n_3, n_1 = 0] \tag{2.A.27}$$
$$\frac{\partial}{\partial t}\mathbb{E}[n_3, n_1 = 1] = \lambda_1 \mathbb{E}[n_3, n_1 = 0] - \mu_1 \mathbb{E}[n_3, n_1 = 1] + \lambda_3 \mathbb{E}[n_2, n_1 = 1] - \mu_3 \mathbb{E}[n_3, n_1 = 1], \tag{2.A.28}$$

give the average of protein copy number.

To obtain the formula of variance, we have to write down the differential equations for the second moments, which for the messengers read

$$\frac{\partial}{\partial t}\mathbb{E}[n_2(n_2-1), n_1 = 0] = +\mu_1 \mathbb{E}[n_2(n_2-1), n_1 = 1] - \lambda_1 \mathbb{E}[n_2(n_2-1), n_1 = 0] - 2\mu_2 \mathbb{E}[n_2(n_2-1), n_1 = 0] \tag{2.A.29}$$
$$\frac{\partial}{\partial t}\mathbb{E}[n_2(n_2-1), n_1 = 1] = \lambda_1 \mathbb{E}[n_2(n_2-1), n_1 = 0] - \mu_1 \mathbb{E}[n_2(n_2-1), n_1 = 1] + 2\lambda_2 \mathbb{E}[n_2, n_1 = 1] - 2\mu_2 \mathbb{E}[n_2(n_2-1), n_1 = 1]. \tag{2.A.30}$$

In order to obtain a close system of differential equations for the second moments of the number

of proteins, we need also the differential equations for the cross-moments:

$$\frac{\partial}{\partial t}\mathbb{E}\left[n_3(n_3-1),n_1=0\right] = +\mu_1\mathbb{E}\left[n_3(n_3-1),n_1=1\right]-\lambda_1\mathbb{E}\left[n_3(n_3-1),n_1=0\right] \tag{2.A.31}$$
$$+2\lambda_3\mathbb{E}\left[n_2n_3,n_1=0\right]-2\mu_3\mathbb{E}\left[n_3(n_3-1),n_1=0\right]$$
$$\frac{\partial}{\partial t}\mathbb{E}\left[n_3(n_3-1),n_1=1\right] = \lambda_1\mathbb{E}\left[n_3(n_3-1),n_1=0\right]-\mu_1\mathbb{E}\left[n_3(n_3-1),n_1=1\right] \tag{2.A.32}$$
$$+2\lambda_3\mathbb{E}\left[n_2n_3,n_1=1\right]-2\mu_3\mathbb{E}\left[n_3(n_3-1),n_1=1\right]$$
$$\frac{\partial}{\partial t}\mathbb{E}\left[n_2n_3,n_1=0\right] = +\mu_1\mathbb{E}\left[n_2n_3,n_1=1\right]-\lambda_1\mathbb{E}\left[n_2n_3,n_1=0\right]-\mu_2\mathbb{E}\left[n_2n_3,n_1=0\right] \tag{2.A.33}$$
$$-\mu_3\mathbb{E}\left[n_2n_3,n_1=0\right]+\lambda_3\mathbb{E}\left[n_2,n_1=0\right]+\lambda_3\mathbb{E}\left[n_2(n_2-1),n_1=0\right]$$
$$\frac{\partial}{\partial t}\mathbb{E}\left[n_2n_3,n_1=1\right] = \lambda_1\mathbb{E}\left[n_2n_3,n_1=0\right]-\mu_1\mathbb{E}\left[n_2n_3,n_1=1\right]-\mu_2\mathbb{E}\left[n_2n_3,n_1=1\right] \tag{2.A.34}$$
$$-\mu_3\mathbb{E}\left[n_2n_3,n_1=1\right]+\lambda_2\mathbb{E}\left[n_3,n_1=1\right]+\lambda_3\mathbb{E}\left[n_2,n_1=1\right]$$
$$+\lambda_3\mathbb{E}\left[n_2(n_2-1),n_1=1\right].$$

The previous systems of differential equations give the first moments of messengers and proteins. In the following Section we derive the solution of these systems at equilibrium.

Analytic close formulas for first moments

We can obtain the mean and variance of the number of messengers and proteins at equilibrium, by solving the equilibrium differential equations associated to the differential equations presented in the previous section. It results that

$$\mathbb{E}\left[n_1\right] = \frac{\lambda_1 n_1^{\max}}{\lambda_1+\mu_1}, \tag{2.A.35}$$
$$\mathrm{Var}\left(n_1\right) = \frac{\lambda_1\mu_1 n_1^{\max}}{(\lambda_1+\mu_1)^2}, \tag{2.A.36}$$
$$\mathbb{E}\left[n_2\right] = \frac{\lambda_2}{\mu_2}\frac{\lambda_1 n_1^{\max}}{\lambda_1+\mu_1}, \tag{2.A.37}$$
$$\mathrm{Var}\left(n_2\right) = \frac{\lambda_2}{\mu_2}\frac{\lambda_1 n_1^{\max}}{\lambda_1+\mu_1}+\frac{\lambda_2^2}{\mu_2(\mu_2+\lambda_1+\mu_1)}\frac{\lambda_1\mu_1 n_1^{\max}}{(\lambda_1+\mu_1)^2}, \tag{2.A.38}$$

where we have reintroduced the constant $n_1^{\max}$, which accounts for possible copies of the specific gene.

This reasoning can be extended to the *three-stage model*, which lead also to the protein statistics

$$\mathbb{E}\left[n_3\right] = \frac{\lambda_3}{\mu_3}\frac{\lambda_2}{\mu_2}\frac{\lambda_1 n_1^{\max}}{\lambda_1+\mu_1}, \tag{2.A.39}$$
$$\mathrm{Var}\left(n_3\right) = \frac{\lambda_3}{\mu_3}\frac{\lambda_2}{\mu_2}\frac{\lambda_1 n_1^{\max}}{\lambda_1+\mu_1}+\frac{\lambda_2\lambda_3^2}{\mu_2\mu_3}\frac{\lambda_1}{(\lambda_1+\mu_1)}\frac{n_1^{\max}}{(\mu_2+\mu_3)}+ \tag{2.A.40}$$
$$+\left(\frac{\lambda_2\lambda_3}{\mu_2\mu_3}\right)^2\frac{\lambda_1\mu_1 n_1^{\max}}{(\lambda_1+\mu_1)^2}\frac{\mu_2\mu_3[\mu_2+\mu_3+\lambda_1+\mu_1]}{(\mu_2+\mu_3)[\mu_2+\lambda_1+\mu_1][\mu_3+\lambda_1+\mu_1]} \tag{2.A.41}$$

Paulsson gives analytical formulas for both averages and variances of all the three processes that are modeled and gives possible interpretations of the obtained results. Analyzing the resulting variance, he interprets different addends as noises inherent to the studied process itself or deriving from fluctuations in upstream processes in the gene expression.

In a Paulsson's paper of 2004 [53] the author present a possible interpretation of the *linear noise approximation* for gene expression. This concept is restated and explained in [54], where the author shows possible applications of this approximation in order to compute statistics of the *three-stage model* with non linear transition rates, where the only constraint is that each stochastic event add or remove one molecule at a time. Here the *three-stage model* as represented in figure 2.7 is a special case, since in this case the linear approximation results to be exact.

2.A.3 Swain's model

Peter Swain, Michael Elowitz, Eric Siggia, Vahid Shahrezaei, Mukund Thattai and Alexander van Oudenaarden have produced many papers on the gene expression, trying to analyze different aspects of protein production using tools of the theory of probability.

We left aside for the moment the problem of "intrinsic" and "extrinsic" noise, analyzed among others by Swain and Elowitz [69] and by Thattai and van Oudenaarden [71]. We will focus on the so-called "intrinsic noise", which is the noise of the quantity of proteins related inherently to the production chain, as defined in the work [69].

Introduction

The authors consider here the stochasticity which is connected to the protein production: random timing of chemical reactions and stochastic encounter of cellular compounds.

Typically the experimental data are compared with models which predict the mean and possibly the variance of the proteins, since the analytic formula of the probabilistic distribution of the proteins is often hard to derive.

Using few biological data, as the fact that mRNA degradation is faster than protein decay, the authors propose an approximate method for solving the *master equation*, giving the transient and stationary approximated probabilistic distributions of the protein number.

The model

Following the main models in the literature, Swain and Shahrezaei consider the *three-stage model* as the fundamental model to describe gene expression, see Section 2.A.2 for details. The promoter of the gene, corresponding to the specific protein of interest, can be in one of two states: active or inactive. These transitions may occur because of the binding/unbinding of specific proteins or by pausing polymerase. This is also one of the basic mechanisms considered by Rigney in his work. The other two stages considered are the messengers and protein dynamics. The transcription of a new mRNA may occur only if the promoter is active. Transcription, translation and the decay of proteins and messengers are modeled as first-order chemical reactions, i.e. they are supposed to be exponentially distributed.

The fundamental hypothesis of this model is that the messenger lifetime is much shorter than protein lifetime. The protein fluctuations are so determined only by time-averaged properties of the mRNA and not by finer time-dependent ones. This is the key point of Swain approach to simplify the mathematical description in order to give an analytic approximated formula for the probabilistic distribution.

For the two-stage model the probability $P_{m,n}(t)$ of having m mRNAs and n proteins at time t satisfies the *master equation*

$$\begin{aligned}\frac{\partial P_{m,n}}{\partial t} &= \nu_0(P_{m-1,n} - P_{m,n}) + \nu_1 m(P_{m,n-1} - P_{m,n}) + d_0\left[(m+1)P_{m+1,n} - mP_{m,n}\right] \\ &\quad + d_1\left[(n+1)P_{m,n+1} - nP_{m,n}\right],\end{aligned} \tag{2.A.42}$$

where ν_0 and ν_1 are the transcription and translation rates respectively, while d_0 and d_1 are the degradation rates for the messengers and the proteins respectively. Defining the generating function

$$F(z, z') = \sum_{m,n} z'^m z^n P_{m,n},$$

and using equation (2.A.42), we find a first-order partial differential equation for the generating function

$$\frac{\partial F}{\partial \nu} - \gamma\left[b(1+u) - \frac{u}{\nu}\right]\frac{\partial F}{\partial u} + \frac{1}{\nu}\frac{\partial F}{\partial \tau} = a\frac{u}{\nu}F, \tag{2.A.43}$$

where $a = \nu_0/d_1$, $b = \nu_1/d_0$, $\gamma = d_0/d_1$ and where the new variables are $\tau = d_1 t$, $u = z' - 1$ and $\nu = z - 1$.

Solving the previous equation and for $\gamma >> 1$, the generating function is given by

$$F(z, \tau) \approx \left[\frac{1 - b(z-1)e^{-\tau}}{1 + b - bz}\right]^a, \tag{2.A.44}$$

if we assume there are no proteins at $\tau = 0$. So when $\gamma >> 1$ and $\tau >> \gamma^{-1}$, then

$$P_{0,n}(\tau) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^n\left(\frac{1+be^{-\tau}}{1+b}\right)^a {}_2F_1\left(\begin{matrix}-n & -a \\ 1-a-n\end{matrix}; \frac{1+b}{e^{\tau}+b}\right), \tag{2.A.45}$$

where Γ is the gamma function and ${}_2F_1\left(\begin{smallmatrix}-n & -a \\ 1-a-n\end{smallmatrix}; \frac{1+b}{e^{\tau}+b}\right)$ is a hypergeometric function.

Note 2.A.1. *We recall that the* hypergeometric function *is defined as*

$${}_2F_1\left(\begin{matrix}a & b \\ c\end{matrix}; z\right) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!}, \tag{2.A.46}$$

provided that $c \neq 0, -1, -2, \ldots$. *The* $(x)_n$ *notation stand for*

$$(x)_n = \begin{cases} 1 & \text{if } n = 0, \\ x(x+1)\ldots(x+n-1) & \text{if } n > 0. \end{cases}$$

The hypergeometric function may be characterized via a differential equation, the Euler's hypergeometric differential equation

$$z(1-z)\frac{\mathrm{d}^2\omega}{\mathrm{d}z^2} + [c - (a+b+1)z]\frac{\mathrm{d}\omega}{\mathrm{d}z} - ab\omega = 0 \tag{2.A.47}$$

At steady state the previous equation gives

$$P_n = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\left(\frac{b}{1+b}\right)^n\left(1 - \frac{b}{1+b}\right)^a, \tag{2.A.48}$$

which is a negative binomial distribution. Observe that b is the mean number of proteins produced during a mRNA lifetime, this quantity is referred in the literature as "bursts". Protein lifetimes are usually greater than mRNA lifetimes, so, typically, no protein will be degraded during a messenger lifetime. Moreover if we look at the dynamics in the protein time-scale, the proteins produced by a chosen mRNA will appear to be synthesized concurrently, i.e. in a *burst*.

The same analysis can be done in the *three-stage model*, with a greater effort in order to solve the related differential problem and to derive the approximated formulas.

Supposing a difference in the mRNA and protein lifetimes, i.e. $\gamma >> 1$, has allowed simplifications, leading to a complete approximated description of the transient and stationary probability distribution of the protein numbers. This is due mainly to the fact that number of messengers reaches rapidly steady state, reducing protein fluctuations. When γ has large values, then the proteins produced by a mRNA are produced in a burst, which is geometrically distributed. Nevertheless when $\gamma < 1$ then protein degradation will occur during the translation of new proteins, which makes unlikely to have a geometric distribution for the protein production.

Chapter 3

Realistic model of gene expression

The *three-stage model* proposed in Chapter 2 introduces a new description of gene expression using marked Poisson point processes (MPPP). However, if the model presented is a general case of the classic models, the choices of distributions made have no relevant biological meaning and serve as illustration of the mathematical framework. In this chapter, MPPP framework is the fundamental ingredient to obtain a more realistic and appropriate description of gene expression. The impact of such more realistic description will be analyzed together with possible relevant choices of probability distributions, in order to bring light on the existing modeling and provide new insights on gene expression and its characterization.

Mathematical models play a crucial role in gene expression and evolve continuously in order to improve their qualitative and quantitative descriptive capabilities. However, a model is the result of the trade-off between a fine and accurate description of the biological phenomena and the limits imposed by the theoretical tools to characterize and analyze the properties of the model itself. Such a dilemma is even more manifest when explicit analytic formulas are expected.

The classic Markovian description of gene expression [5, 54, 64] captures the main characteristics of the gene expression, but comes with the strong exponential assumption for the duration of every step. In particular, this assumption is not justified for the description of steps such as the protein elongation and the protein decay due to volume dilution, as discussed in Section 1.4.1. The critical analysis of the Markovian description of gene expression before the biological knowledge requires the introduction of an appropriate and more general mathematical framework. In order to include the protein elongation process, we introduce a supplementary step with respect to the model introduced in Chapter 2 by splitting the effective step of translation into the ribosome binding on a messenger and the subsequent elongation of the polypeptide chain. The MPPP framework is then used as the fundamental ingredient to derive a more realistic *four-stage model* of gene expression which includes protein elongation.

The protein dilution is also considered by adopting an appropriate description of the phenomenon, obtained by coupling the stochastic description with a deterministic continuous model of dilution. The resulting model mixes a stochastic description of the dynamics of different components involved in gene expression with a deterministic description of dilution and is referred to as *hybrid four-stage model*.

General results valid for any choice of distributions of various steps are thus derived for both *four-stage* and *hybrid four-stage* models.

Plan of the Chapter. In Section 3.1.1 we describe the main steps of a model of gene expression. Unlike the three-stage model introduced in Chapter 2, we add now a supplementary fourth step

to account for a more realistic description of protein elongation step. The obtained model is referred to as *four-stage model* and we derive the statistics of the main processes under general assumptions of the duration of several steps.

The general model detailed in Section 3.1.1 is specialized with respect to the biological knowledge acquired on the underlying processes. In particular, Section 3.1.2 is devoted to a detailed discussion on the appropriate assumptions for the general steps, focusing on the description of protein elongation step and protein dilution, which are peculiarities of our model with respect to what has been proposed so far.

Section 3.1.3 is devoted to an in-depth analysis of the four-stage model with realistic assumptions. Because of the complexity of the analytic formulas of a realistic protein elongation, we consider in Section 3.2.2 an appropriate approximation of the duration of elongation step, which is shown to be well approximated by a deterministic duration. Given such approximation, a comparison with a classical approach with all exponential steps lead to a counter-intuitive result, detailed in Section 3.2.3.

The accurate description of *protein dilution* lead to precise quantification of this phenomenon and is a major improvement with respect to the classic approach. In particular, the introduction of such realistic description lead to a substantial difference in the resulting formulas: despite the classic formulas capture the qualitative trend of fluctuations, they overestimate fluctuations when dilution is the main decay mechanism. Moreover, the realistic description of protein dilution invalidates a result of the classic approach, i.e. the process describing the number of proteins at equilibrium has a wider distribution than a Poisson process. In fact, using a model with a realistic description of protein dilution we show that this "Poisson limit" can be theoretically trespassed.

Section 3.2.5 is devoted to the analysis of the impact of the different steps on the resulting fluctuations in the four-stage model with realistic assumptions. It is shown that gene activity and protein dilution have the most important effect on fluctuations, while the realistic description of protein elongation have a mild impact, except for specific cases such as unstable proteins. Moreover, is studied the impact of rare codons, whose role is still an open question, on elongation step and, therefore, on protein fluctuations.

3.1 Four-Stage Model of Gene Expression

Protein elongation is usually included as an effective exponential step in classic models. However, this step has proven to be a crucial, and in some cases limiting, step in the production of a protein. For this reason, in order to account more precisely of this process, we introduce now a *Four-Stage Model* of gene expression, where the supplementary step corresponds to the protein elongation step.

3.1.1 Model and general results

We introduce now the Four-Stage model using the mathematical toolkit introduced in Chapter 2. The results obtained in this section rely on general assumptions of the duration of several steps in the model. The Four-Stage Model is described in Figure 3.2, where appears the elongation at step 5.

Notation 3.1.1. *As for Chapter 2, if not specified differently, a process is said to occur at some rate λ, means that the duration of such process is exponentially distributed with parameter λ.*

Gene activation.

The gene regulation is modeled as a two-stage gene activation: the gene is inactivated at rate λ_1^- and is activated by the unbinding of the repressor at rate λ_1^+, see [57, 39, 36, 54, 67] for details. The assumption that the maximum number of active genes $n_{\max}$ is 1 is not restrictive, since the statistics are proportional to $n_{\max}$ by independence and the formulas in the general case can be easily obtained.

The stationary process $(Y(t), t \in \mathbb{R})$ with values in $\{0, 1\}$, represents the gene status, where $Y(t) = 1$ indicates that the gene is active at time t, while $Y(t) = 0$ if it is inactive. The behaviour of the process $(Y(t))$ is well known, see [18] for details, and at equilibrium it results

$$\mathbb{P}(Y = 1) = \delta_+ = 1 - \mathbb{P}(Y = 0), \tag{3.1.1}$$

where $\delta_+ = \lambda_1^+/\Lambda$ and $\Lambda = \lambda_1^+ + \lambda_1^-$. In order to compute the variance of the other processes, we need the quantity

$$\mathbb{P}(Y(t) = 1 | Y(0) = 1) = \delta_+(1 - \delta_+)e^{-\Lambda t}. \tag{3.1.2}$$

At equilibrium, the average of gene activity is given by $\mathbb{E}[Y] = \delta_+$ and the variance by $\mathrm{var}(Y) = \delta_+(1 - \delta_+)$.

mRNA dynamics.

The active gene produces mRNAs at rate λ_2 and we denote with $F_2(\mathrm{d}y)$ the probability distribution of the lifetime of a messenger, i.e. $\mathbb{P}[\sigma_2 \in A] = \int_A F_2(\mathrm{d}y)$. The dynamics of the messengers can be described using the notations and tools of MPPP theory, as shown in Chapter 2. In particular, let $\mathcal{N}_{\lambda_2} = (s_n, \sigma_{2,n})$ be a MPPP on $\mathbb{R} \times \mathbb{R}_+$ with intensity measure $\lambda_2 \,\mathrm{d}u \otimes F_2(\mathrm{d}v)$, where s_n and $\sigma_{2,n}$ represent the instants of messenger creation and the lifetime of a mRNA born at s_n respectively.

The point process $\mathcal{M}$ representing the instants of creation of mRNAs and the associated lifetime can be represented as

$$\mathcal{M} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \delta_{(u,v)} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v), \tag{3.1.3}$$

where δ_z is the Dirac mass at z.

The number of messengers alive at equilibrium can be represented as

$$M = \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{u\le 0\le u+v, Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v). \tag{3.1.4}$$

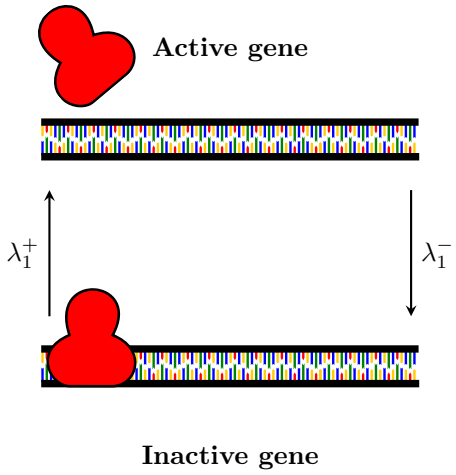


Figure 3.1: Gene activation. The gene activation/deactivation occur at rate λ_1^+ and λ_1^- respectively.

Remark. Here the mRNA is available for translation once a small portion of the growing mRNA chain has been assembled. This assumption is coherent with the prokaryotic dynamics, but should be adapted for the eukaryotic case. In fact in this case we have to wait the completed messenger to be exported to the cytoplasm. If we assume this duration to be deterministic, then the previously defined integral should be shifted of a constant value and we should easily get the corresponding analytic results.

Using the MPPP representation of the number of messengers, we can now get an explicit expression of the average and variance of messengers at equilibrium.

Proposition 3.1.2. *If the distribution of the messenger lifetime σ_2 is $F_2(\mathrm{d}x)$, the average number M of mRNAs at equilibrium is given by*

$$\mathbb{E}(M) = \delta_+ \lambda_2 \mathbb{E}(\sigma_2). \tag{3.1.5}$$

The variance of M is given by

$$\begin{aligned} \mathrm{var}(M) =& \mathbb{E}(M) + 2\lambda_2^2 \delta_+ (1-\delta_+) \cdot \\ & \cdot \int_0^{+\infty} e^{-\Lambda v} \overline{F}_2(u) \overline{F}_2(u+v)\, \mathrm{d}u\, \mathrm{d}v, \end{aligned} \tag{3.1.6}$$

where we use the notations $F_2(x) \overset{\text{def}}{=} F_2([0,x])$ and $\overline{F}_2(x) \overset{\text{def}}{=} (1 - F_2(x))$.

See Chapter 2, section 2.3, or the paper [18] for a complete proof of these results.

Remark. The presented results are still valid in the case in which some random variable have not a probability density function. In such a context, the integrals should then be interpreted in the distribution sense.

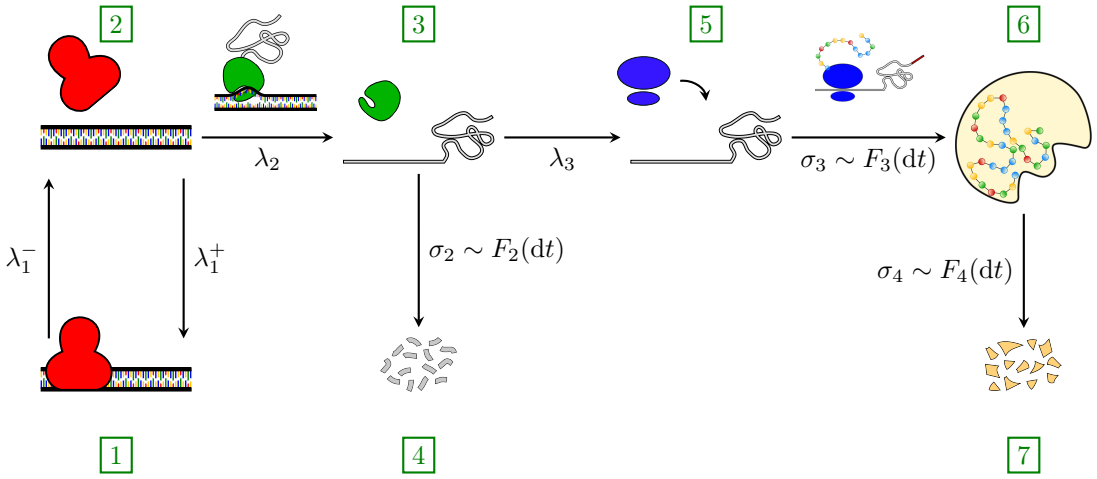


Figure 3.2: **Four-stage model.** *Stage 1 - Gene activation (steps $1-2$).* The initiation of transcription is strongly regulated; schematically the gene is said to be in "inactive state" if a repressor prevents the polymerase binding and is in "active state" otherwise. The gene inactivation (respectively activation) occurs at rate λ_1^- (respectively λ_1^+). *Stage 2 - Transcription (step 3).* When the gene is in active state, the RNA polymerase produces an mRNA at rate λ_2. The probability distribution of an mRNA lifetime is $F_2(\mathrm{d}t)$ and is general (step [4]). *Stage 3 - Translation (steps $5-6$).* A ribosome binds to an active mRNA at rate λ_3 and proceeds to protein elongation, whose duration σ_3 is assumed to have a general probability distribution $F_3(\mathrm{d}t)$. *Stage 4 - Protein decay (step 7).* Protein decay is achieved here by *proteolysis*, the case of protein dilution being treated in Section 3.1.2. The distribution $F_4(\mathrm{d}t)$ of the protein lifetime σ_4 is supposed general for the moment. The parameters λ_1^+, λ_1^-, λ_2, λ_3, indicate that the duration of the relative step has exponential distribution of parameter λ.

Active ribosomes dynamics.

We consider the dynamics of ribosomes, i.e. the binding ribosome-mRNA, which occurs at rate λ_3, and the polypeptide elongation step, whose elongation time σ_3 has a general probability distribution $F_3(\mathrm{d}t)$. Note that multiple ribosomes can be active on the same messenger at a time, resulting in a queue of ribosomes elongating proteins. These ribosomes are referred to as *active ribosomes*. The number of proteins results therefore amplified with respect to the population of mRNAs.

The point process $\mathcal{R}$ representing the instants of ribosome binding on mRNA and the associated protein elongation time can be written as

$$\mathcal{R} = \int_{\mathbb{R}\times\mathbb{R}_+} \mathcal{M}(\mathrm{d}u, \mathrm{d}v) \int_{[u,u+v]\times\mathbb{R}_+} \delta_{(x,y)} \mathcal{N}_{\lambda_3}^{(u)}(\mathrm{d}x, \mathrm{d}y), \tag{3.1.7}$$

where, for $u \in \mathbb{R}$, $\mathcal{N}_{\lambda_3}^{u}(\mathrm{d}x, \mathrm{d}y)$ is a marked Poisson point process on $\mathbb{R} \times \mathbb{R}_+$ with intensity measure $\lambda_3 \,\mathrm{d}x \otimes F_3(\mathrm{d}y)$, created at time u.

The number of active ribosomes at equilibrium can be represented by the following random variable

$$\begin{aligned} R &= \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v) \left(\iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x\le 0\le x+y, u\le x\le u+v\}} \mathcal{N}_{\lambda_3}^{u}(\mathrm{d}x, \mathrm{d}y) \right) \\ &= \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x\le 0\le x+y\}} \mathcal{R}(\mathrm{d}x, \mathrm{d}y) \end{aligned} \tag{3.1.8}$$

where similar arguments than in Chapter 2 apply.

The statistics of the active ribosomes at equilibrium can be obtained using the approach showed in Chapter 2. For this reason we skip the proofs of the following theorem.

Proposition 3.1.3. *If the distribution of the lifetime of an mRNA is $F_2(\mathrm{d}x)$ and the distribution of elongation time is $F_3(\mathrm{d}y)$, then the expected value of the random variable R, which is the number of ribosomes active on mRNAs at equilibrium, is given by*

$$\mathbb{E}(R) = \delta_+ \lambda_2 \lambda_3 \mathbb{E}(\sigma_2)\mathbb{E}(\sigma_3) \tag{3.1.9}$$

and its variance $\mathrm{var}(R)$ *can be expressed as*

$$\begin{aligned} \mathrm{var}(R) = \mathbb{E}(R) &+ \lambda_2 \lambda_3^2 \delta_+ \int_0^{+\infty} \int_{\mathbb{R}_+} \left[\int_{-s}^{(-s+t)\wedge 0} \overline{F}_3(u)\, \mathrm{d}u \right]^2 \mathrm{d}s F_2(\mathrm{d}t) \\ &+ \lambda_2^2 \lambda_3^2 \delta_+ (1-\delta_+) \int_{\mathbb{R}_+^4} e^{-\Lambda|(u_1-u_2)+(v_1-v_2)|} \prod_{i=1}^{2} \overline{F}_2(u_i)\overline{F}_3(v_i)\, \mathrm{d}u_i\, \mathrm{d}v_i, \end{aligned} \tag{3.1.10}$$

where, for $i = 2,\ 3$, $\overline{F}_i(x) = (1 - F_i(x))$ and $x \wedge y \overset{\text{def}}{=} \min\{x, y\}$.

Actually, we can obtain a more general result than in Proposition 3.1.3, allowing to recover any statistic of a function of the point process describing ribosomes. In fact, equation (3.1.7) allows to consider functions depending on the specific point process. More in detail, for any function $f : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ integrable with respect to the point process $\mathcal{R}(\mathrm{d}x, \mathrm{d}y)$ we can consider $\mathcal{R}(f)$, which is a shorthand for

$$\mathcal{R}(f) = \int_{\mathbb{R}\times\mathbb{R}_+} \mathcal{M}(\mathrm{d}u, \mathrm{d}v) \int_{[u,u+v]\times\mathbb{R}_+} f(x, y) \mathcal{N}_{\lambda_3}^{u}(\mathrm{d}x, \mathrm{d}y). \tag{3.1.11}$$

We now derive a theorem analogous of Theorem 3.1.3, in which we obtain the statistics of $\mathcal{R}(f)$ at equilibrium.

Theorem 3.1.4. *Let $f : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ a non negative Borelian function, then if the distribution of the lifetime of an mRNA is $F_2(\mathrm{d}x)$ and the distribution of elongation time is $F_3(\mathrm{d}y)$, the expected value of the random variable $R(f)$ is given by*

$$\mathbb{E}(R(f)) = \delta_+ \lambda_2 \lambda_3 \mathbb{E}(\sigma_2) \int_{\mathbb{R}\times\mathbb{R}_+} f(x,y)\,\mathrm{d}x F_3(\mathrm{d}y) \tag{3.1.12}$$

and its variance $\mathrm{var}(R(f))$ *can be expressed as*

$$\begin{aligned}\mathrm{var}\,(R(f)) =& \delta_+ \lambda_2 \lambda_3 \mathbb{E}(\sigma_2) \int_{\mathbb{R}\times\mathbb{R}_+} f^2(u,v)\,\mathrm{d}u F_3(\mathrm{d}v) \\ &+ \lambda_2 \lambda_3^2 \delta_+ \int_{\mathbb{R}\times\mathbb{R}_+} \left(\int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le u\le s+t\}} f(x,y)\ \mathrm{d}x F_3(\mathrm{d}y) \right)^2 \mathrm{d}s F_2(\mathrm{d}t) \qquad (3.1.13) \\ &+ \lambda_2^2 \lambda_3^2 \delta_+ (1-\delta_+) \int_{\mathbb{R}^2\times\mathbb{R}_+^4} e^{-\Lambda|u-w+z-x|} f(u,v) f(x,y) \overline{F}_2(w) \overline{F}_2(z)\,\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}w\,\mathrm{d}z F_3(\mathrm{d}v) F_3(\mathrm{d}y), \qquad (3.1.14)\end{aligned}$$

where, for $i = 2, 3$, $\overline{F}_i(x) = (1 - F_i(x))$.

Proof. The proof consists in a slight modification of the proof of Proposition 2.3.5. Recall that $R(f)$ can also be represented as

$$R(f) = \sum_{n\in\mathbb{Z}} \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(s_n)=1\}} \mathbb{1}_{\{s_n\le x\le s_n+\sigma_{2,n}\}} f(x,y) \mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y).$$

Denote by $\widehat{\mathbb{E}}$ the conditional expectation $\mathbb{E}[\,\cdot \mid (Y(t)), (s_n, \sigma_{2,n})]$. The conditional generating function $\widehat{\mathbb{E}}\left[e^{-zR(f)}\right]$ can be written as

$$\begin{aligned}&\widehat{\mathbb{E}}\left(\prod_{n\in\mathbb{Z}} \exp\left(-z \mathbb{1}_{\{Y(s_n)=1\}} \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s_n\le x\le s_n+\sigma_{2,n}\}} f(x,y) \mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y) \right) \right) \\ &\qquad = \prod_{n\in\mathbb{Z}} \widehat{\mathbb{E}}\left[\exp\left(-z \mathbb{1}_{\{Y(s_n)=1\}} \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s_n\le x\le s_n+\sigma_{2,n}\}} f(x,y) \mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y) \right) \right],\end{aligned}$$

since the point processes $\mathcal{N}_{\lambda_3}^{s_n}$, $n \in \mathbb{Z}$, are independent. The nth term of this product is, applying Proposition A.0.5 to the marked Poisson point processes $\mathcal{N}_{\lambda_3}^{s_n}$,

$$\widehat{\mathbb{E}}\left[\exp\left(-z \int_{\mathbb{R}\times\mathbb{R}_+} g(x,y) \mathcal{N}_{\lambda_3}^{s_n}(\mathrm{d}x,\mathrm{d}y) \right) \right] \quad = \quad \exp\left(-\lambda_3 \int \left(1 - e^{-zg(x,y)}\right) \mathrm{d}x\, F_3(\mathrm{d}y) \right),$$

where $g(x,y) = \mathbb{1}_{\{Y(s_n)=1\}} \mathbb{1}_{\{s_n\le x\le s_n+\sigma_{2,n}\}} f(x,y)$.

Now,

$$\begin{aligned}\mathbb{E}\left[e^{-z\mathcal{R}(f)} \middle| (Y) \right] &= \mathbb{E}\left[\exp\left\{ -\lambda_3 \sum_n \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s_n\le x\le s_n+\sigma_{2,n}\}} \left(1 - e^{-f(x,y)}\right) \mathbb{1}_{\{Y(s_n)=1\}}\, \mathrm{d}x F_3(\mathrm{d}y) \right\} \right] \\ &= \mathbb{E}\left[e^{-\mathcal{N}_{\lambda_2}(G(s,t))} \middle| (Y) \right] = \exp\left\{ -\lambda_2 \int \left(1 - e^{-G(s,t)}\right) \mathrm{d}s F_2(\mathrm{d}t) \right\}\end{aligned} \tag{3.1.15}$$

where

$$G(s,t) = \lambda_3 \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le x\le s+t\}} \mathbb{1}_{\{Y(s)=1\}} \left(1 - e^{-f(x,y)}\right) \mathrm{d}x F_3(\mathrm{d}y).$$

In order to obtain an expression for $\mathbb{E}(R(f))$ and $\mathbb{E}\left(R(f)(R(f)-1)|(Y(t))\right)$, we have to take the first and second derivative of equation (3.1.15) with respect to z and evaluate it at $z = 0$. The resulting formula should then be integrated with respect to $(Y(t))$ to obtain formulas (3.1.12) and (3.1.13), by using similar arguments as in the proof of Proposition 2.3.3 (with more technical calculations). □

Proposition 3.1.2 as well as Theorem 3.1.5 could be generalized by proceeding to a similar proof as done here. We just give the general result in the case of active ribosomes, since, as we will see later on, we will use this result to derive formulas of mean and variance when we consider dilution as the main decay mechanism for proteins during exponential growth phase.

Proteins dynamics.

In a similar way as done for ribosomes, we can describe the protein dynamics in connection with the dynamics of messengers and active ribosomes. The point process $\mathcal{P}$ representing the instants of translation initiation, duration of protein elongation and protein lifetime can be represented as

$$\mathcal{P} = \iint_{\mathbb{R}\times\mathbb{R}_+} \mathcal{M}(\mathrm{d}u, \mathrm{d}v) \iint_{[u,u+v]\times\mathbb{R}_+} \delta_{(x,y,z)} \mathcal{N}_{\lambda_3}^{(u)}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z), \tag{3.1.16}$$

where, for $u \in \mathbb{R}$, $\mathcal{N}_{\lambda_3}^{u}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ is a marked Poisson point process on $\mathbb{R}\times\mathbb{R}_+\times\mathbb{R}_+$ with intensity measure $\lambda_3\, \mathrm{d}x \otimes F_3(\mathrm{d}y) \otimes F_4(\mathrm{d}z)$, created at time u. Here F_3 is the distribution associated with protein elongation time and F_4 is the distribution associated to protein decay.

The number of proteins at equilibrium can be represented by the following random variable

$$\begin{aligned} P &= \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{Y(u)=1\}} \mathcal{N}_{\lambda_2}(\mathrm{d}u, \mathrm{d}v) \left(\iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x+y\le 0\le x+y+z\}} \mathbb{1}_{\{u\le x\le u+v\}} \mathcal{N}_{\lambda_3}^{u}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \right) \\ &= \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x+y\le 0\le x+y+z\}} \mathcal{P}(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \end{aligned} \tag{3.1.17}$$

where the indicator function $\mathbb{1}_{\{x\le 0\le x+y\}}$ is used to represent all the proteins which have started translation at a negative time and which have not yet finished their task at time $t = 0$.

The statistics of the proteins at equilibrium can be obtained using the approach showed in Proposition 3.1.2 and 3.1.3. We now give the results concerning protein, but we skip proofs since they use similar arguments as in proof of Propositions 3.1.2 and 3.1.3, but with more technical computations.

Theorem 3.1.5. *If the distribution of the lifetime of an mRNA is $F_2(\mathrm{d}x)$, the distribution of elongation time is $F_3(\mathrm{d}y)$ and the distribution of the lifetime of a protein is $F_4(\mathrm{d}z)$, then the expected value of the random variable P, which is the number of proteins at equilibrium, is given by*

$$\mathbb{E}(P) = \delta_+ \lambda_2 \lambda_3 \mathbb{E}(\sigma_2)\mathbb{E}(\sigma_4) \tag{3.1.18}$$

and its variance $\mathrm{var}(P)$ *can be expressed as*

$$\mathrm{var}(P) = \mathbb{E}(P) + \lambda_2\lambda_3^2\delta_+ \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le 0\}} \left[\int_{\mathbb{R}_+^2} \mathbb{1}_{\{-(u+s+t)\le y\le -(u+s)\}} \overline{F}_4(u)\ \mathrm{d}u F_3(\mathrm{d}y)\right]^2 \mathrm{d}s F_2(\mathrm{d}t) \tag{3.1.19}$$

$$+ \lambda_2^2\lambda_3^2\delta_+(1-\delta_+) \int_{\mathbb{R}_+^6} e^{-\lambda|\xi+\eta+\zeta-x-y-z|} \overline{F}_2(z)\overline{F}_2(\zeta)\overline{F}_4(x)\overline{F}_4(\xi)\ \mathrm{d}x\,\mathrm{d}z\,\mathrm{d}\xi\,\mathrm{d}\zeta F_3(\mathrm{d}y)F_3(\mathrm{d}\eta),$$

where, for $i = 2,\ 4$, $F_j(x) = F_j([0,x])$ *and* $\overline{F}_i(x) = (1 - F_i(x))$.

The formulas obtained in this section involve general distributions for several steps of the protein production. However, the averages of the modeled processes at equilibrium depend only on the average duration of the general steps. This means that, for given conditions and given mean duration of the different steps, the choice of specific distributions does not affect the average number of messengers, active ribosomes and proteins at equilibrium. On the other side, the choice of specific distributions has an impact on the fluctuations of the different processes.

3.1.2 Realistic assumptions

In this Section we discuss the appropriate assumptions of the different processes described in the four-stage model. The general formulas presented in Section 3.1.1 are now specialized, by considering appropriate choices for the distributions in order to better describe the underlying process and obtain a more realistic model.

As usual, we refer to *exponential assumption* when the time to produce a particular cellular component or its lifetime is exponentially distributed. We have already discussed and described the appropriateness of such assumption in Section 2.1.3, which is well suited to describe the biochemical reactions which requires primarily the encounter of two reactants. We first recall the main steps of the four-stage model with exponentially distributed duration.

Gene activation. The gene activation/inactivation is mainly driven by repressor, which is a DNA-binding protein that regulates the expression of the specific protein by binding the operator thus preventing polymerase binding. For this reason the activation step is modeled as a two state process, where the switching between states occur in exponentially distributed time. Experimental results of Goldin et al.[22] show that the exponential two-step model is a well suited description of the biological phenomenon.

Transcription and translation initiation. Transcription and translation processes have similarities in their main steps. In particular, during transcription initiation a polymerase binds on the gene, followed by specific processes before the start of messenger elongation. Analogously, during translation initiation a ribosome attaches on a messenger and then a series of processes occur in order to well accommodate the ribosome and proceeds to the next step. We assume that the initiations of both transcription and translation to be exponentially distributed, since the main limiting mechanism is the successful binding of polymerase and ribosome.

mRNA degradation. The degradation of messengers occurs by means of *RNase*, which locates and binds to the target messenger in order to start the degradation process. The degradation step is therefore well described to occur in exponentially distributed time, since, once a *RNase* has bound on the messenger, the translation initiation is not possible anymore (the *RNase* prevents the binding of ribosome and has eventually degraded the starting sequence). Note that there is

a competition between translation initiation and mRNA decay. In particular, the *RNase* and ribosome compete to bind to the *ribosome binding site* (RBS) in first place. If *RNase* succeeds, then it degrades the mRNA in the $5'$ to $3'$ direction and does not interact with ribosomes bound to mRNA, which can complete elongation of the specific protein, see also [37, 74].

mRNA elongation. Messenger elongation, despite shares common characteristics with protein elongation, has not been intentionally considered in the *four-stage model*. In fact, in the presented model, once the gene is active, the messenger is available for translation initiation up to an exponentially distributed time. We show now that this model assumption is a licit, since our analysis is restricted to prokaryotic gene expression and the impact of the messenger elongation step is here negligible.

The mRNA elongation can be described in a similar way of protein elongation and is possible to derive general formulas of the variance of messengers at equilibrium in such more realistic case. In particular, we assume that the gene switches between active and inactive states at rates λ_1^- and λ_1^+ respectively. The polymerase binds on the active gene at rate λ_2 and the duration ζ_2 of the subsequent messenger elongation has a general distribution $G_2(\mathrm{d}t)$, as shown in Figure 3.3. The messenger lifetime ζ_3 is then assumed to have distribution $G_3(\mathrm{d}t)$.

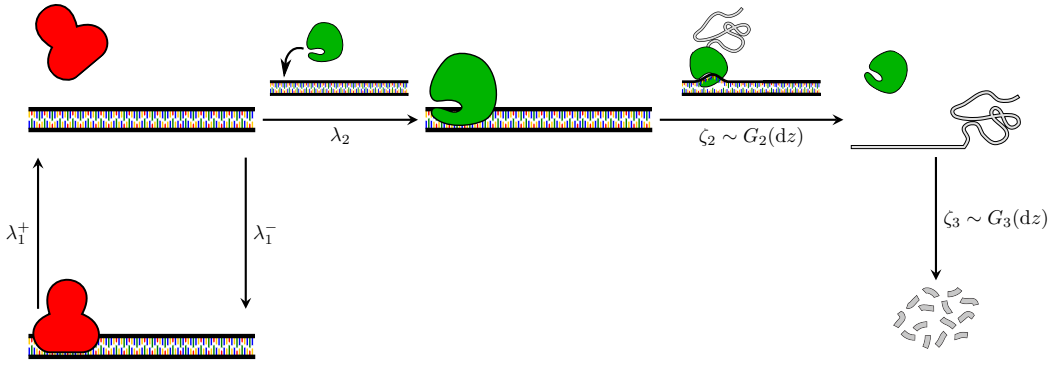


Figure 3.3: **mRNA elongation model.** The gene inactivation (respectively activation) occurs at rate λ_1^- (respectively λ_1^+). When the gene is in active state, the RNA polymerase attaches the gene at rate λ_2 and proceeds to mRNA elongation, whose duration ζ_2 is assumed to have a general probability distribution $G_2(\mathrm{d}t)$. Messenger decay is achieved *RNase* and the distribution $G_3(\mathrm{d}t)$ of mRNA lifetime ζ_3 is supposed general for the moment. The parameters λ_1^+, λ_1^-, λ_2 and λ_3 indicate that the duration of the relative step has exponential distribution of parameter λ.

In order to obtain an estimation of the impact of messenger elongation step on fluctuations, we consider two limit cases: the case of large fluctuations on mRNA elongation, i.e. exponential elongation, and the case with no fluctuations, i.e. deterministic elongation. The messenger decay is supposed from here on to be exponentially distributed with parameter μ_2.

If the messenger elongation is exponentially distributed, then the mRNA fluctuations are described by the formula

$$\mathrm{var}^{(\mathrm{E})}(M) = \mathbb{E}[M]\left[1 + (1-\delta_+)\frac{\lambda_2\mu_2}{\Lambda+\mu_3}\frac{\Lambda+\mu_2+\mu_3}{(\Lambda+\mu_2)(\mu_2+\mu_3)}\right], \tag{3.1.20}$$

where the superscript "E" serves as a reminder of the exponential choice of the elongation time distribution. On the other side, if the elongation time is deterministic and its duration is denoted by τ_2, then the variance of the number of mRNAs at equilibrium is given by

$$\mathrm{var}^{(\mathrm{D})}(M) = \mathbb{E}[M]\left[1 + (1-\delta_+)\frac{\lambda_2\mu_2}{\Lambda+\mu_3}\right], \tag{3.1.21}$$

where now $\mu_2 = 1/\tau_2$ and "D" stands for deterministic elongation.

We can show that $\mathrm{var}^{(\mathrm{E})}(M) < \mathrm{var}^{(\mathrm{D})}(M)$ for any choice of parameters. Moreover, simulations indicate that the variance corresponding to other suitable distribution choices are comprises in between the two previous formulas. For this reason we use those formulas to give an estimation of the impact of the modeling of messenger elongation on its resulting fluctuations.

There is biological evidence, see also [43] and [74], that in prokaryotes ribosomes can bind to the elongating mRNA. In fact, since prokaryotes have no nucleus, the messenger is accessible yet during elongation and a ribosome can bind to it as long as the RBS sequence has been assembled. For this reason, the time the messenger becomes available for translation corresponds to the time required to produce this starting sequence, which varies from 0.6 up to 4 seconds under biologically relevant ranges of parameters.

If we consider a constitutive gene, i.e. $\delta_+ = 1$, then formulas (3.1.20) and (3.1.21) are identical and the elongation distribution has no impact on fluctuations. If $\delta_+ \neq 1$, the gene dynamics have the strongest impact on messenger fluctuations. Analyzing the variances for different rates of activation/repression, with elongation time of the starting sequence in the window $[0.5, 10]$ seconds and with an average lifetime of $1-2$ minutes, the differences of the relative variances of the deterministic and exponential elongation is smaller than $\approx 3\%$. Messenger elongation has a weak impact on fluctuations and the simplification made in the *four-stage model* is reasonable. In conclusion, since we are interested in messenger dynamics just in perspective of the production of proteins, the exponential assumption for the effective step of messenger transcription. Note that in the *four-stage model* we make no distinction between elongating and completed messengers.

Remark. The simplification made in *four-stage model* to not consider messenger elongation is no more valid in the *eukaryotic* case since, because of the presence of a nucleus, transcription and translation occur in separated compartments. However, the MPPP framework can be used to build a model considering the specificity of *eukaryotic* gene expression. In particular, the separation of transcription and translation requires the elongated messenger to be firstly exported to the cytoplasm in order to undergo subsequent processing. The time to completely elongate the messenger as well as the time required to export it to the cytoplasm must then be included into the description, resulting in a slightly more complex, but still treatable, model of gene expression.

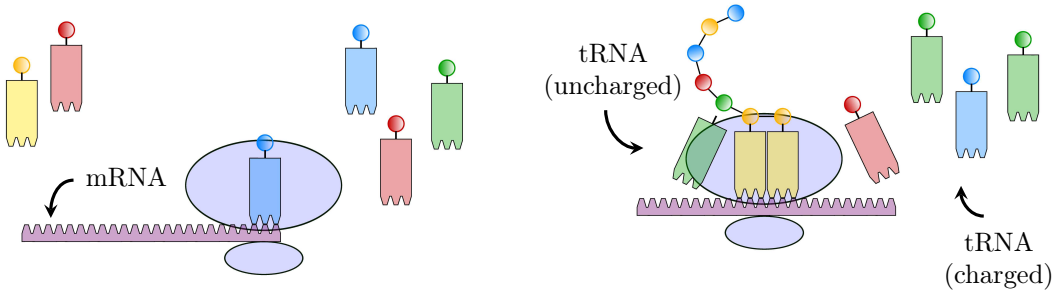
Protein elongation. The protein elongation results in an iterative procedure in which each codon of the messenger chain is coupled with a particular tRNA, which adds a new amino-acid to the growing polypeptide chain by means of ribosome, see Figure 3.4a. The insertion of a new amino-acid requires firstly the encounter of a charged tRNA, corresponding to the codon the ribosome is reading, with the ribosome. The amino acid is then attached to the growing polypeptide chain, while the uncharged tRNA is released in the cytoplasm and the ribosome moves to the next codon. Since the encounter of tRNA with the ribosome proves to be costly in terms of time, then each event of attachment of a new amino acid can be described to occur up to an exponentially distributed random time. If the specific protein is composed of N amino acids, then the duration $T_{\mathrm{el}}^{(N)}$ of the elongation step is the sum of N exponentially distributed random variables.

We can identify K classes of codons, each class including codons with the same characteristics, i.e. each codon that belongs to a class k is assembled in an exponential time of parameter ρ_k. Let

T_k be the time to process the N_k codons of class k, with $k = 1, \dots, K$, hence T_k is the sum of N_k independent exponential random variables of parameter ρ_k. Therefore, T_k is Erlang distributed, $T_k \sim \mathrm{Erl}(N_k, \rho_k)$, where ρ_k is referred to as rate and N_k as shape parameter, and its probability density function is

$$f(x, N_k, \rho_k) = \frac{\rho_k^{N_k} x^{N_k - 1} e^{-\rho_k x}}{(N_k - 1)!},$$

where N_k is such that $\sum_{k=1}^{K} N_k = N$. Consequently, the duration of the protein elongation $T_{\mathrm{el}}^{(N)} = \sum_{k=1}^{K} T_k$ can be described as the sum of independent random variables with Erlang distribution.



(a) Ribosome binding. (b) Chain elongation.

Figure 3.4: **Protein elongation.** Translation starts with the binding of the ribosome on the messenger (see (a)), which assembles the polypeptide chain. Each of the four amino acids is carried by a specific *charged tRNA* (see figure (b)). Once the tRNA has bound to the ribosome, the amino acid is attached to the growing chain and the uncharged tRNA is released in the cytoplasm.

Protein decay: *proteolysis* and volume dilution. There are two distinct mechanisms of protein*proteolysis* and volume dilution.

Protein decay through *proteolysis* is a controlled process and is mainly associated to a quality control of the produced proteins or to particular situations such as stress. Temperature and stress have a deleterious effect on proteins which are eventually denatured. Proteolysis intervenes at this point as maintenance and destroys possibly the damaged proteins. The degradation of protein via *proteolysis* shares a similar description of mRNA decay: a *protease* locates and binds to a protein and breaks down the chain into smaller polypeptides or amino acids. For this reason, the exponential distribution is the most appropriate choice for the distribution $F_4(\mathrm{d}z)$ in the *four-stage model*. In general, *proteolysis* is not the main protein decay process within a cell in normal conditions and the typical protein lifetime is of several cell cycles.

The protein decay via dilution is a process is a completely different process. In general, *prokaryotic* and *eukaryotic* cells increase their internal components in order to give rise to two daughter cells. The volume increase associated to cell growth is the responsible of the dilution of the cellular components. Bacterial protein production occurs mainly in the exponential growth phase, where dilution is the main cause of protein decay. The phenomenon of dilution can be described by a continuous deterministic process, since it is the direct consequence of the (continuous) increase of cell volume and the speed of dilution corresponds to the cell growth rate, which is a well determined parameter for fixed environmental conditions. Consider a (fixed) unit

of volume V_0, which contains one unity of protein production, i.e. it contains a fixed number of genes, the corresponding produced mRNAs and their degradation machinery and a constant pull of free ribosomes. We assume that cell dilution affects only the proteins produced in V_0, in particular this can be visualized in the following way: the production machinery is confined to the volume V_0, except the proteins which can continuously leak out the fixed volume, see figure 3.5. If $P(t)$ denote the number of proteins at time t in the volume V_0, we define the protein concentration $P_c(t)$ as

$$P_c(t) \stackrel{\text{def}}{=} \frac{P(t)}{V_0}, \tag{3.1.22}$$

where $P(t)$ is the protein copy number at time t.

Let $\{l_k\}_{k=1,\dots,\infty}$ be the sequence of the instants of birth of new proteins and denote with t_0 the first time of observations. By definition of the sequence $\{l_k\}_{k=1,\dots,\infty}$, we have

$$\begin{aligned} P(l_k^-) &= P_c(l_k^-)V_0, \\ P(l_k^+) &= P(l_k^-) + 1 = P_c(l_k^-)V_0 + 1. \end{aligned} \tag{3.1.23}$$

During exponential phase of growth, the decay protein of protein is driven by a term of the form $e^{-\nu t}$, where ν is the cell growth rate. In particular, if we interrupt the production of proteins at time l_0, then the protein concentration $P_c(t)$ decreases as follows

$$P_c(t) = P_c(t_0)e^{-\nu(t-l_0)} \qquad t \geq l_0. \tag{3.1.24}$$

Using the previous visualization, ν can be interpreted as the speed of the continuous leakage of proteins out of the volume V_0, while P_c measures the "amount" of proteins inside such volume.

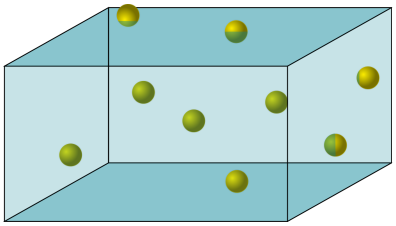


Figure 3.5: **Unit of production.** The dilution affects only the concentration of proteins. This can be visualized as a continuous deterministic leaking of proteins (yellow spheres) out of the volume V_0. The volume V_0 contains the machinery of production of a specific protein.

Remark. In the classic stochastic models of gene expression, protein dilution is described as a stochastic exponentially distributed protein lifetime. This description allows to use the classic tools in order to derive explicit analytic formulas of protein variance, however the use of this description has not been justified. In the following section we see how to couple the continuous deterministic description of protein dilution with a discrete stochastic production process as in the *four-stage model*. The description via MPPP allows then to derive exact formulas.

3.1.3 Explicit formulas under realistic assumptions

In Section 3.1.2 we have analyzed realistic assumptions for the Four-Stage model introduced in Section 3.1.1. We first describe the *hybrid four-stage model* of the gene expression, which includes the protein decay due to dilution. General results on protein fluctuations are derived under general assumptions and the so obtained model is compared to the *four-stage model* detailed previously. Using realistic biological assumptions analyzed in 3.1.2, we then derive an explicit description of protein variance.

Hybrid four-stage model

The modeling of protein dilution introduced in Section 3.1.2 is now coupled with a stochastic description of the other processes involved in gene expression.

From equations (3.1.23) and (3.1.24), we obtain the following linear continuous system describing the protein concentration dynamics

$$\begin{cases} \dfrac{\mathrm{d}P_c}{\mathrm{d}t}(t) = -\nu P_c(t), & t \in [l_{k-1}, l_k) \\ P_c(l_k^+) = P_c(l_k^-) + \dfrac{1}{V_0}, & k = 1, \dots, \infty \\ P_c(l_0) = P_c^0, \end{cases} \tag{3.1.25}$$

where l_0 is the initial time. The birth of new proteins is a discrete phenomenon and results in jumps of protein concentration at instants $\{l_k\}_{k=0,\dots,\infty}$, as expressed by equation in (3.1.23).

The solution of the system (3.1.25) is given by

$$P_c(t) = P_c^0 e^{-\nu t} + \frac{1}{V_0} \sum_{k=1}^{\infty} \mathbb{1}_{\{l_k < t\}} e^{-\nu(t-l_k)}. \tag{3.1.26}$$

Since V_0 is constant, using equation 3.1.22 we can write the previous equation in terms of number of proteins. For this reason, for here on we will write all relations in terms of number of proteins P.

We are interested in the equilibrium statistics of proteins, therefore we can assume $P^0 = 0$, i.e. we forget the initial condition. The system (3.1.25) describe just the protein dynamics in the case of volume dilution given the jump instants $\{l_k\}_{k=0,\dots,\infty}$. It is possible to couple this deterministic description with the stochastic description of the other involved processes by writing the protein birth instants in terms of the dynamics of genes, messengers and ribosomes. More in detail, if we denote with r_k the instant of ribosome binding on a mRNA and with $\sigma_{k,n}$ the time required to translate such messenger, then

$$l_k = r_k + \sigma_{3,k}, \tag{3.1.27}$$

since a new protein is released into the cytoplasm once a ribosome has completed chain assembly.

Since the volume dilution affects only protein concentration, we can reuse the description of the *four-stage model* for gene activation/inactivation, mRNAs and active ribosomes dynamics. In particular, the point process describing active ribosomes can be written as $\mathcal{R} = (r_k + \sigma_{3,k}, k \in \mathbb{Z})$ and, using the functional description (3.1.11), we can rewrite equation (3.1.26) as follows

$$P = \sum_{k=1}^{\infty} \mathbb{1}_{\{l_k<0\}} e^{\nu l_k} = \iint_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{x+y<0\}} e^{\nu(x+y)} \mathcal{R}(\mathrm{d}x, \mathrm{d}y) = \mathcal{R}(f), \tag{3.1.28}$$

where

$$f(x, y) = \mathbb{1}_{\{x+y<0\}} e^{\nu(x+y)}. \tag{3.1.29}$$

The resulting model is depicted in Figure 3.6 and is referred to as *hybrid four-stage model*.

The statistics of proteins in the *hybrid four-stage model* can be obtained by using the results of Section 3.1.1, as detailed in the following theorem.

Theorem 3.1.6. *If the distribution of the lifetime of an mRNA is $F_2(\mathrm{d}x)$ and the distribution of elongation time is $F_3(\mathrm{d}y)$, then the expected value of the variable P, which is the concentration of proteins at equilibrium, is given by*

$$\mathbb{E}(P) = \frac{\delta_+ \lambda_2 \lambda_3}{\nu} \mathbb{E}(\sigma_2) \tag{3.1.30}$$

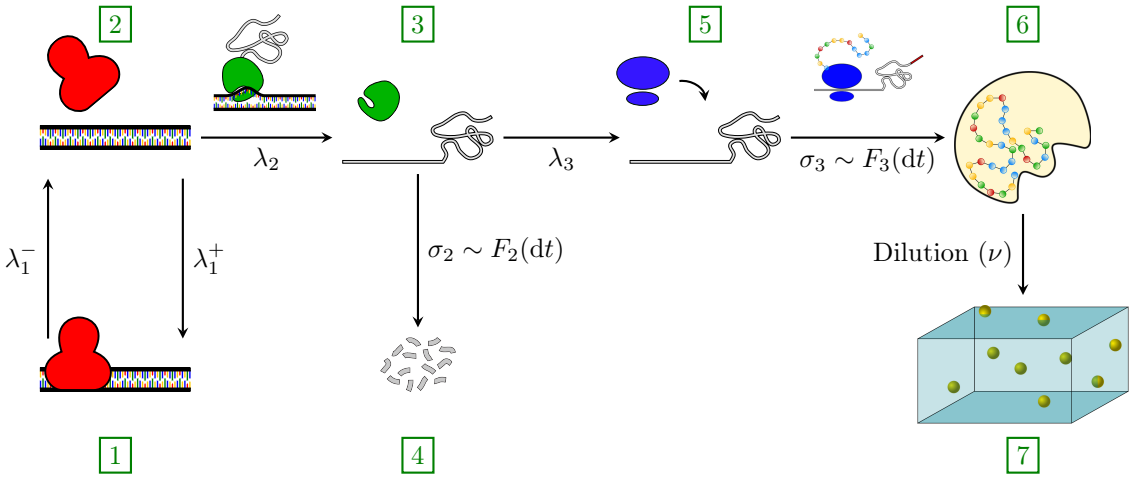


Figure 3.6: **Four-stage model with protein dilution – Hybrid model.** This model is a slight modification of the *four-stage model*, see Figure 3.2. Here protein dilution is the main protein decay mechanism (step7) and is described as a continuous deterministic process. By coupling the continuous model of protein dilution with the point process view of Section 3.1.1 for all other processes we obtain the *hybrid four-stage model*. The volume dilution proceeds at rate ν and is strictly connected to the rate of cell volume growth. The parameters λ_1^+, λ_1^-, λ_2 and λ_3 indicate that the duration of the relative step has exponential distribution. The distributions F_2 and F_3 are general.

and its variance $\mathrm{var}(P)$ *can be expressed as*

$$\begin{aligned}
&\mathrm{var}_{Hybrid}(P) = \qquad (3.1.31)\\
&= \frac{1}{2}\mathbb{E}(P) + \lambda_2\lambda_3^2\delta_+ \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le 0\}} \left[\int_{\mathbb{R}_+^2} \mathbb{1}_{\{-(u+s+t)\le y\le -(u+s)\}} e^{-\nu u}\, \mathrm{d}u F_3(\mathrm{d}y)\right]^2 \mathrm{d}s F_2(\mathrm{d}t)\\
&\quad + \lambda_2^2\lambda_3^2\delta_+(1-\delta_+)\int_{\mathbb{R}_+^6} e^{-\lambda|\xi+\eta+\zeta-x-y-z|}\overline{F}_2(z)\overline{F}_2(\zeta)e^{-\nu x}e^{-\nu\xi}\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}\xi\,\mathrm{d}\zeta F_3(\mathrm{d}y)F_3(\mathrm{d}\eta),
\end{aligned}$$

where, for $i = 2, 3$, $\overline{F}_i(x) = (1 - F_i(x))$ *and* ν *is the rate of volume dilution.*

Proof. By applying Theorem 3.1.4 with $f(u, v)$ given by (3.1.29), we obtain the results. □

Notation 3.1.7. *From here on we use the subscript "Hybrid 4-Stage" to distinguish the formulas obtained from the* hybrid four-stage model *from the formulas of the* four-stage model*, tagged now with "4-Stage".*

Four-stage and hybrid four-stage models: comparison of general formulas. Assume that dilution is the main protein decay mechanism; we want now to compare the fluctuations resulting of the description of gene expression in the case of dilution obtained by the *four-stage model* and the *hybrid four-stag model*. In particular, we will use Theorems 3.1.5 and 3.1.6.

The *four-stage model* describe the protein decay as a discrete stochastic process, which occurs up to a time σ_4 with distribution $F_4(\mathrm{d}z)$. Classically it is assumed that this step is exponentially

distributed, i.e. $\sigma_4 \sim F_4(\mathrm{d}z) = \mu_4 e^{-\nu z}\,\mathrm{d}z$. In this case, formula (3.1.19) read

$$\begin{aligned}
&\mathrm{var}_{\text{4-Stage}}(P) = \\
&= \mathbb{E}(P) + \lambda_2\lambda_3^2\delta_+ \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}_{\{s\le 0\}} \left[\int_{\mathbb{R}_+^2} \mathbb{1}_{\{-(u+s+t)\le y\le -(u+s)\}} e^{-\nu u}\,\mathrm{d}uF_3(\mathrm{d}y)\right]^2 \mathrm{d}sF_2(\mathrm{d}t) \\
&\quad + \lambda_2^2\lambda_3^2\delta_+(1-\delta_+) \int_{\mathbb{R}_+^6} e^{-\lambda|\xi+\eta+\zeta-x-y-z|}\overline{F}_2(z)\overline{F}_2(\zeta)e^{-\nu x}e^{-\nu\xi}\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}\xi\,\mathrm{d}\zeta F_3(\mathrm{d}y)F_3(\mathrm{d}\eta).
\end{aligned}$$

A comparison of the previous formula with formula (3.1.31) shows that the only difference reflects in the first addend, no matter the choices of distributions $F_2(\mathrm{d}t)$ and $F_3(\mathrm{d}t)$.

This difference is strictly connected to the modeling of volume dilution and not to the more realistic description of protein elongation. Indeed, if we consider the three-stage model presented in Chapter 2 with the description of dilution given, we obtain a formula of variance as in formula (2.3.10) where the first addend is replaced by $\frac{1}{2}\mathbb{E}[P]$. In Section 3.2.4 we will analyze more in detail the impact of the more realistic description of dilution on protein fluctuations.

The previous comparison of formulas (3.1.19) and (3.1.31) has an important consequence, which will be analyzed hereafter. The development of stochastic models of gene expression has relied in part on the comparison of the distribution of proteins at equilibrium with a simple Poisson model. More in detail, we can obtain a first simplistic stochastic description of gene expression using a birth and death process, as shown in Figure 3.7: a new protein is synthesized at rate λ and degraded at rate μ, each event occurring in exponential random time. This

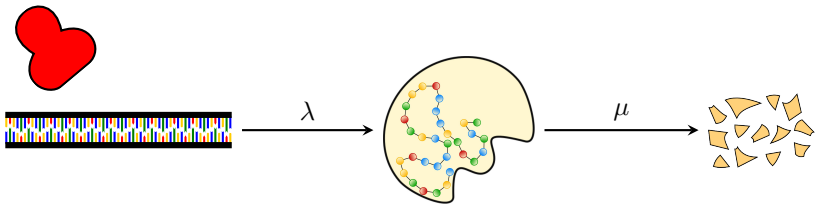


Figure 3.7: Poisson model. Gene expression is described as a simple birth and death process, in which the birth of a new protein occurs at rate λ, while it is degraded at rate μ.

simple description leads to well-known formulas for protein mean and variance at equilibrium, i.e. $\mathbb{E}(P) = \lambda/\mu$ and $\mathrm{var}(P) = \mathbb{E}(P)$, since the resulting process describing proteins dynamics at equilibrium has Poisson distribution.

The stochastic models proposed in the last thirty years have marked a breaking point with respect to this naive modeling by showing that the distribution of proteins resulting by a more realistic description of the underlying processes is no more Poissonian. In particular, if the mRNA dynamics are also considered, the variance of protein number is bigger than the Poisson case and, therefore, their distribution is not-Poissonian. This argument is already present in the early work of Berg [5], in which the author showed that the variance of proteins at equilibrium could be wider than the Poisson case, which is recovered in the limit when each messenger produces exactly one protein.

The comparison with Poisson process has led to the introduction of a measure of the deviation from Poissonian behavior: the Fano factor. The Fano factor, defined as $v = \mathrm{var}(P)/\mathbb{E}(P)$, is equal to 1 in the case of Poisson dynamics. Many other works have investigated the non-Poisson nature

of protein dynamics both from theoretical and experimental point of view, see [54, 70, 47, 2, 48] to cite few.

Using general approach described in Section 3.1.1 for the *four-stage model*, the variance of proteins given by formula (3.1.19) has the form

$$\mathrm{var}_{\text{4-Stage}}(P) = \mathbb{E}(P)\,[1 + K]$$

where $K > 0$, which allows to conclude that, even for general distributions, the protein fluctuations are wider than a corresponding Poissonian description.

On the other side, if we consider the *hybrid four-stage model*, the formula of variance (3.1.31) has the form

$$\mathrm{var}_{\text{Hybrid}}(P) = \mathbb{E}(P)\left[\frac{1}{2} + K\right],$$

where $K > 0$. The previous formula tells us that the fluctuations deriving from the realistic description of protein dilution given in Section 3.1.2 may in principle be smaller than the Poissonian description of figure 3.7, depending on the value of K.

This is not in contrast with the results obtained experimentally showing fluctuations larger than the Poisson case. In fact, the last inequality just implies that the lower bound for fluctuations in protein number is lower than Poisson. In particular, if we consider a constitutive protein, i.e. $\delta_+ = 1$, and we assume protein elongation to be deterministic and mRNA degradation to occur in exponentially distributed time, then

$$\mathrm{var}_{\text{Hybrid}}(P) = \mathbb{E}(P)\left[\frac{1}{2} + \frac{\lambda_3}{\mu_2 + \mu_4}\right].$$

Therefore, if we consider for example a mRNA average lifetime of 2 minutes, a doubling average time of 20 minutes and we consider a weakly affinity of the ribosome, i.e. $\lambda_3 < 0.3$, then we obtain that the second term within brackets in the previous formula is smaller than 0.5 and the resulting noise is smaller than the Poisson case, which breaks down a common result on the subject. The parameters have been chosen in biologically relevant ranges, however no experiments has been performed to investigate this specific aspect up to present. It would be worth to set up experiments in order to explore more in detail fluctuations in this particular direction.

In conclusion, even if for the majority of proteins we possibly have a variance larger than a Poisson, this could not be the case for a small bunch of them and would be worth to investigate experimentally. Moreover, note that if on one side we may experience a variance smaller than the Poisson, on the other side this does not change the qualitative behavior of the whole system, since the variance still results proportional to the mean value and scales with it, as has been also shown experimentally [2, 48].

Proteolysis & Dilution: user manual. The volume dilution is in general the main mechanism of protein decay, however in specific cases, such as in the case of unstable proteins see Appendix B.1.5, *proteolysis* can play a central role and this should be considered in the modeling choices. In this chapter we have presented two models of gene expression: the *four-stage model* and the *hybrid four-stage model*. We are now going to describe the cases in which we should use one or the other model, depending on the specific conditions under analysis.

If volume dilution is the main mechanism of protein decay and the effect of *proteolysis* can be neglected, then the *hybrid four-stage model* is the more adapted description of gene expression, since it includes a more realistic description of protein dilution. From here on we use the subscript "DIL." to refer to the results obtained from the *hybrid four-stage model* under the assumption that dilution is the main protein decay mechanism.

On the other side, if *proteolysis* is the main protein decay mechanism, then the description of the *four-stage model* is more appropriate, since the decay phenomenon occurs at a specific

(random) point in time. As discussed in Section 3.1.2, since the limiting step of *proteolysis* is the encounter of the *protease* with the protein, the lifetime of a protein is well described to be exponentially distributed, i.e. $\sigma_4 \sim F_4(\mathrm{d}z) = \mu_4 e^{-\mu_4 z}\,\mathrm{d}z$. For these reasons, from here on we tag with the subscript "LYSIS" the results obtained using the *four-stage model* with exponentially distributed protein lifetime under the assumption that *proteolysis* is the main decay mechanism.

Notation 3.1.8. *From here on, the results obtained from the* hybrid four-stage model *in case dilution is the main protein decay mechanism are tagged with the subscript "DIL.", for example the variance of proteins in this case is denoted with* $\mathrm{var}_{DIL.}(P)$.

Similarly, the four-stage model *will be used only if* proteolysis *is the main protein decay mechanism and the the variance of proteins will be denoted with* $\mathrm{var}_{LYSIS}(P)$ *in this case.*

From here on we will just use the parameter μ_4 *to denote the rate of the protein decay, this parameter being intended to be the rate of the exponential protein lifetime in the case of* proteolysis *or the rate of volume dilution otherwise.*

Protein elongation and explicit formulas

After the detailed discussion of the previous sections, we derive now a formula for the variance of proteins in the general case in which the dilution is the main protein decay mechanism under realistic assumptions for the duration of elongation discussed in Section 3.1.2. If each codon requires the same amount of time in order to be processed, then the polypeptide elongation time $T_{\mathrm{el}}^{(N)}$ exhibits an Erlang distribution $\mathrm{Erl}(N,\rho)$, see Figure 3.8.

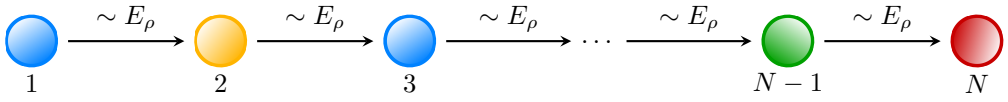


Figure 3.8: **Scheme of protein elongation.** The time required by each amino acid in order to be attached to the growing polypeptide chain is a random variable E_ρ exponentially distributed with parameter $\rho > 0$. N is the total number of amino acids composing the specific protein.

If we consider a constitutive protein, i.e. the gene is always active ($\delta_+ = 1$), then, by integrating the assumptions previously discussed and applying formula (3.1.31), the protein variance is given by

$$\begin{aligned}\mathrm{var}_{\mathrm{DIL.}}(P) = \mathbb{E}[P]\Bigg[\frac{1}{2} + \frac{2\lambda_3\mu_2}{\mu_2^2-\mu_4^2}\frac{\rho^{2N}}{(N-1)!} \\ \times\left(\frac{\mu_2}{(\rho^2-\mu_4^2)^N}\int_{\mathbb{R}_+} s^{N-1}e^{-s}Q\left(N,\frac{s}{\rho^2-\mu_4^2}\right)\mathrm{d}s\right. \\ \left.-\frac{\mu_4}{(\rho^2-\mu_2^2)^N}\int_{\mathbb{R}_+} s^{N-1}e^{-s}Q\left(N,\frac{s}{\rho^2-\mu_2^2}\right)\mathrm{d}s\right)\Bigg], \quad (3.1.32)\end{aligned}$$

where $Q(n,s)$, $n \in \mathbb{N}$, is the cumulative distribution function of a Poisson random variable of parameter s, i.e. if $X \sim \mathrm{Poi}(s)$ then $\mathbb{P}[X < n] = Q(n,s)$.

Note 3.1.9. *The function* $Q(h,s)$ *is also called* regularized Gamma function *and is defined as* $Q(h,s) = \Gamma(h,s)/\Gamma(s)$*, where* $\Gamma(h,s)$ *is the* upper incomplete gamma function *and* $\Gamma(s)$ *is the* gamma function.

If we consider the more realistic description $T_{\text{el}}^{(N)} = \sum_{k=1}^{K} T_k$, where $T_k \sim \text{Erl}(N_k, \rho_k)$, $\rho_k \neq \rho_j$ for any $k \neq j$ and $\sum_{k=1}^{K} N_k = N$, the distribution characterizing the elongation step is no more Erlang, but it can still be characterized. In fact, in this case the distribution of the duration of elongation is the convolution of Erlang random variables and the density function is characterized by the formula

$$f_T(t; N, \rho_1, \dots, \rho_K) = \sum_{k=1}^{K} \rho_k^{N_k} e^{-\rho_k t} \left[\sum_{j=1}^{N_k} \frac{(-1)^{N_k - j}}{(j-1)!} t^{j-1} \cdot \right. \tag{3.1.33}$$
$$\left. \cdot \left(\sum_{\substack{m_1 + \dots + m_K = N_k - j \\ m_i = 0}} \prod_{\substack{l=1 \\ l \neq i}}^{K} \binom{N_l + m_l - 1}{m_l} \frac{\rho_l^{N_l}}{(\rho_l - \rho_i)^{N_l + m_l}} \right) \right],$$

for $t \geq 0$, while $f_T(t; N, \rho_1, \dots, \rho_K) = 0$ for $t \leq 0$, see [32] for further details.

In first approach, we can split codons in two different classes: normal and rare codons. The main difference between normal and rare codons is the time required to assemble them, the latter being present at low concentrations which cause the ribosome to wait long time, blocking eventually the elongation of proteins. The presence of rare codons has been proven in experiments, but their role is still under study. Thus, using the formalism introduced, if we have N_1 normal codons, each needing an exponential random time of parameter ρ_1 to be processed, then the time T_1 to assemble the normal codons has an Erlang distribution $\text{Erl}(N_1, \rho_1)$. Similarly, the time T_2 needed to process the rare codons is an an Erlang random variable $\text{Erl}(N_2, \rho_2)$, with $\rho_2 << \rho_1$. The probability density function of the protein elongation is now given by

$$\begin{aligned} f_T(t) = & \\ = \; & \rho_1^N e^{-\rho_1 t} \sum_{j=1}^{N} \binom{N+M-j}{N-j} \frac{\rho_2^M}{(\rho_2 - \rho_1)^{N+M-j}} \frac{(-1)^{N-j}}{(j-1)!} t^{j-1} + \\ & + \rho_2^M e^{-\rho_2 t} \sum_{j=1}^{M} \binom{N+M-j}{M-j} \frac{\rho_1^N}{(\rho_1 - \rho_2)^{N+M-j}} \frac{(-1)^{M-j}}{(j-1)!} t^{j-1}. \end{aligned}$$

Unfortunately, the previous formula, and therefore formula (3.1.33), cannot be easily used to obtain a simple analytic formula of variance such as equation (3.1.32).

3.2 Qualitative and quantitative analysis of realistic model

In Section 3.1.3 we have described the model that results when we consider realistic description and assumptions of the underlying processes. We now investigate the resulting formulas and quantify the impact of different choices on protein fluctuations using biologically relevant ranges for the parameters. The aim of the following analysis is to obtain insights on the underlying mechanisms and to obtain some intuition on the protein fluctuations through the use of the models presented. This is even more important in the present context of gene expression, because of the lack of reliable experimental data at present days. In fact, despite the sophisticated techniques developed in recent years and the expertise of biologists, the experimental outcomes are often not sufficiently reliable to estimate precisely the protein variance.

3.2.1 Biological data and model parameters

Before analyzing more in detail the results of the modeling introduced in the previous sections, we devote this section to relate the available biological data to the model parameters. These relations we will be used in the following sections in order to analyze the models outcomes with respect to biological reality.

The protein dilution is caused by cell volume growth and is therefore strictly connected to the cell doubling time τ, i.e. the time required to the cell to double its volume. In particular, if we denote with ν the rate of the exponential growth, then the protein concentration $P_c(t)$ satisfies the relation $P_c(t) = P_c^0 e^{-\nu t}$, where $t \in [0, \tau)$ and P_c^0 is the protein concentration at the beginning of cell cycle. Therefore,

$$\nu = \frac{\ln(2)}{\tau}, \tag{3.2.1}$$

where ν corresponds also to the protein dilution rate.

Biologists commonly use the concept of *half-life* ($t_{1/2}$), i.e. the amount of time required for a quantity to fall to half its value measured at the beginning of the initial time of observation. In particular, the messengers and protein lifetimes are usually expressed in terms of "half-lives". Consider now the messenger decay, the case of protein decay via *proteolysis* being analogous. The mRNA degradation has been described as a first-order reaction, i.e. a biochemical reaction which depends on the concentration of only one reactant, the number of mRNAs in the specific case. The (average) rate law of the mRNA decay reaction is therefore $\frac{\mathrm{d}}{\mathrm{d}t} M(t) = -\mu_2 M(t)$, where μ_2 is the decay rate. If we denote with τ_{hl} the mRNA half-life, then we can obtain the parameter μ_2, characterizing the exponentially distributed messenger degradation, thanks to formula

$$\mu_2 = \frac{\ln(2)}{\tau_{hl}}. \tag{3.2.2}$$

The *proteolysis* rate μ_4 can be obtained using a similar formula, where instead of τ_{hl} we consider the protein half-life.

The parameters λ_1^+ and λ_1^- can be related to the dynamics of the switching of gene state (active/inactive). In particular, the average time of a change of gene state is given by $\Lambda^{-1} = 1/(\lambda_1^+ + \lambda_1^-)$. The parameter δ_+, defined as $\delta_+ = \lambda_1^+/(\lambda_1^+ + \lambda_1^-)$, can be then tuned by observing the residence time of gene in active state with respect to the residence time in inactive state.

We fix now the average number of proteins, $\mathbb{E}[P]$, and the average number of mRNAs, $\mathbb{E}[M]$. In particular, we may derive the parameters λ_2 and λ_3 as functions of the degradation parameters and of the fixed averages. The transcription initiation parameter λ_2 is obtained by formula

$$\lambda_2 = \frac{\mu_2 \mathbb{E}[M]}{\delta_+}. \tag{3.2.3}$$

The translation initiation can be now be obtained as

$$\lambda_3 = \frac{\mu_2 \nu \mathbb{E}[P]}{\lambda_2 \delta_+} = \nu \frac{\mathbb{E}[P]}{\mathbb{E}[M]}. \tag{3.2.4}$$

We conclude this section by giving reference parameters that will be used in the following sections in order to analyze the derived close analytic formulas. These parameters are resumed in table 3.3, where we have considered the parameters corresponding to different growth regimes. Section 3.A shows a more detailed list of biological parameters and should be kept in mind while reading the following sections.

The speed of polypeptide chain elongation changes with respect to the growth regime, see table 3.2. We whose then the value of 18aa/s as average protein elongation speed. The average

protein length is of about 300 amino acids, see [7], and of about 360 amino acids for *E. Coli*, using table [46]. We choose then 300 as reference value for simulations. Therefore, if we took the reference value of 18 aa/s as average protein elongation speed, the corresponding time of protein elongation is of about 16 seconds. The average messenger half life considered in simulation is $\tau_{hl} = 120$ seconds.

3.2.2 Estimation of fluctuations: deterministic elongation

The analytic formula (3.1.32) can be studied with the classical tools of analysis and allows to precisely account for the fluctuations of the number of proteins. Nevertheless, there is no hope to further simplify it, since the integrals which appears in (3.1.32) have no explicit analytic form. For this reason, we analyze the two following limiting cases, in order to obtain an analytic estimation of the possible fluctuations on the protein copy number: deterministic and exponential protein elongation.

Suppose that the protein elongation is deterministic, i.e. the elongation time of every protein is $\tau_3 \stackrel{\text{def}}{=} 1/\mu_3$ $\mathbb{P}-$almost surely, with $\mu_3 > 0$. Leaving unchanged every other assumption as discussed in previous section, we obtain the formula

$$\text{var}^{(\text{D})}_{\text{DIL.}}(P) = \mathbb{E}(P)\left[\frac{1}{2} + \frac{\lambda_3}{\mu_2+\nu} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\nu)}{(\mu_2+\nu)(\Lambda+\mu_2)(\Lambda+\nu)}\right] \tag{3.2.5}$$

where the subscript "D" stands for deterministic protein elongation. Note that the previous formula is independent on parameter μ_3 describing the elongation time. Intuitively, since the assumption of a deterministic protein elongation results in a delay of the protein production, the fluctuations are not affected by this delay but rather "transferred" from the ribosomes to the proteins. This independence is strictly connected with the assumption of the elongation step: by considering any distribution law different from deterministic, this will impact the protein fluctuations.

Suppose now that the protein elongation is exponentially distributed. This assumption, despite not realistic, will serve to give an estimation of fluctuations and allow for a comparison with the formulas available in literature. Therefore, assuming $\sigma_3 \stackrel{\text{dist.}}{=} E_{\mu_3}$, the variance of the number of proteins in the case of protein dilution is

$$\begin{aligned}\text{var}^{(\text{E})}_{\text{DIL.}}(P) = \mathbb{E}(P)\Bigg[&\frac{1}{2} + \frac{\lambda_3\mu_3(\mu_2+\mu_3+\nu)}{(\mu_2+\mu_3)(\mu_2+\nu)(\mu_3+\nu)} + \\ &+ \frac{\lambda_2\lambda_3(1-\delta_+)\mu_3\nu^2}{(\Lambda+\mu_2)(\nu^2-\mu_3^2)} \times \left(\frac{\Lambda+\mu_2+\mu_3}{\mu_3(\mu_2+\mu_3)(\Lambda+\mu_3)} - \frac{\Lambda+\mu_2+\nu}{\nu(\mu_2+\nu)(\Lambda+\nu)}\right)\Bigg],\end{aligned} \tag{3.2.6}$$

where we used formula (3.1.30) and the subscript "E" refers to the exponential assumption of protein elongation is used to distinguish with formula (3.2.5).

It is possible to prove that $\text{var}_{\text{E}}(P) \leq \text{var}_{\text{D}}(P)$ for any choice of parameters. More in general, numerical results have shown that $\text{var}_{\text{E}}(P) \leq \text{var}_{\text{Erlang}}(P) \leq \text{var}_{\text{D}}(P)$ for any set of chosen parameter, as shown in Figure 3.9. In particular, formulas (3.2.6) and (3.2.5) seem to represent a lower and upper bound for different choices of parameters of Erlang distribution. Therefore, formulas (3.2.6) and (3.2.5) can be used to estimate the impact of a realistic description of protein elongation with respect to a classic approach.

The assumption of a deterministic duration of protein elongation is actually a good approximation of the description via Erlang distribution. In fact, assume that the duration of each elementary process of the assembly of a new amino acid is an independent and identically distributed random variable T_i, with $i = 1, \ldots, N$ and N is the number of amino acids in the specific

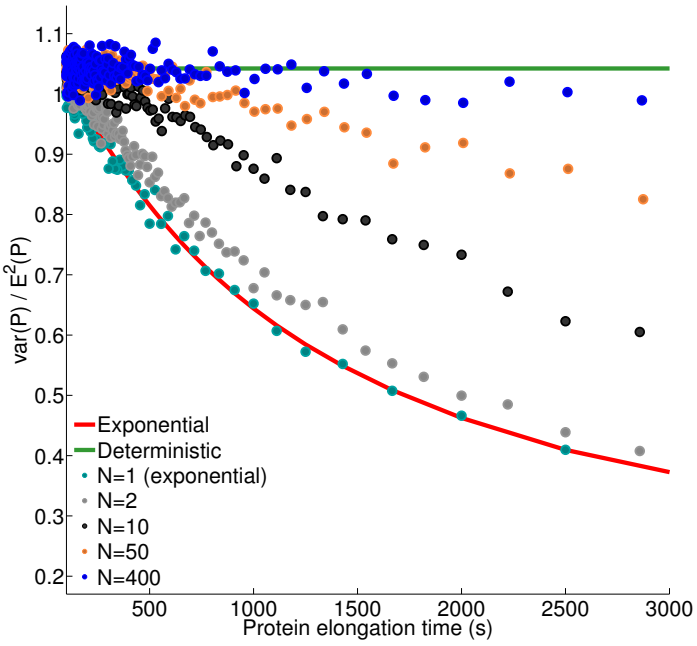


Figure 3.9: **Relative variance with respect to protein elongation time.** The protein elongation time is denoted with $T_{\text{el}}^{(N)}$. The red continuous curve is the relative variance, i.e. $\text{var}(P)/\mathbb{E}^2(P)$, in the case of exponential protein elongation as given by formula (3.2.6). The green continuous curve is the relative variance in the case of deterministic protein elongation as given by formula (3.2.5). Observe that fluctuations associated with deterministic elongation are independent of protein elongation time. Colored circles are obtained via simulations under the assumption of Erlang duration of elongation for different values of the shape parameter N, i.e. the number of steps of the Erlang distribution (the number of amino acids in our description). In particular, the simulator is used to obtain the empirical mean and variance for the specific case of Erlang distributed elongation. The case $N = 1$ corresponds to the case of a single exponential step and we therefore recover the exponential profile (red curve). This case would corresponds to the case of a protein made of a single amino acid. The case $N = 400$ corresponds to average case in *prokaryotes*, since the average number of amino-acids in *E. Coli* is in between 360 and 400. In order to compare different curves, we have changed the time to process each amino acid in such a way that the protein elongation time is consistent for each vertical cut of the curves. Note that as long as N increases, the variance of the duration of elongation step diminishes, but the overall protein variance increases. In particular, the profile corresponding to the more realistic case $N = 400$ is close to the profile corresponding to the deterministic elongation.

protein. The average length of proteins in *E. Coli*, expressed in terms of number of amino acids, is $N = 360$, see table [46]. Consequently, since N is quite large, the central limit theorem gives the following approximation

$$T^{(N)} \sim mN + G\sqrt{N}$$

where m is the average elongation time for one amino-acid and G is a centered Gaussian random

variable. If the variance of the random variable G is not too large, then $T^{(N)}$ is close to the deterministic constant mN. Formulas (3.1.19) and (3.1.31) show that the variance is a continuous functional of the probability distributions, since it is expressed in integral form depending on the specific probability measures. In particular, as long as the distribution of the protein elongation converges continuously towards a normal distribution with narrower standard deviation, the variance obtained by the formulas in the deterministic case will be close to the exact value.

3.2.3 Four-Stage Model: a counter-intuitive result

The general description of the gene expression via MPPP approach as detailed in Sections 3.1.1 and 3.1.2 allows to consider any distribution for few steps of the model.

The classic approach requires each step to have exponentially distributed duration $T \sim E_\lambda$, whose variance is given by

$$\mathrm{var}(T) = \mathbb{E}^2[T] = \frac{1}{\lambda^2},$$

and thus its fluctuations are large compared to its the average value.

As in Section 2.4, we consider the "opposite" case of deterministic duration of some step described in order to see the impact of distribution choice on the resulting fluctuations in protein number.

If we assume each step to have exponentially distributed duration except for protein elongation and take deterministic/exponential distributions for this step, we obtain the counter-intuitive result $\mathrm{var}_{\mathrm{EXP}} \leq \mathrm{var}_{\mathrm{DET}}$. Figure 3.10 illustrates this results, here the square root of relative variance is plotted with respect to the parameter describing the RBS affinity between ribosomes and messengers.

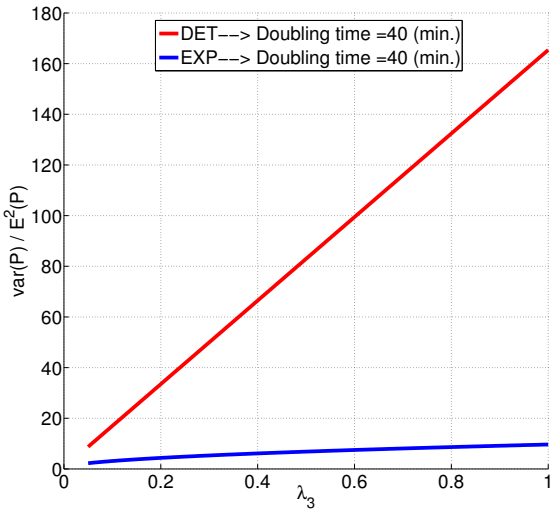


Figure 3.10: Impact of elongation distribution on protein relative variance.

This result holds also when replacing other exponential steps with deterministic ones. We do not investigate here the appropriateness of these choices, but we want rather to extract general information about the response of the system with respect to deeply different choices. Moreover, simulations indicate that this result holds also for other distributions with intermediate fluctuations. In particular, the obtained results seem to indicate a distribution ordering with respect to the fluctuations of duration distribution. If we consider Erlang (or Gamma) distributions and Gaussian distributions, cfr Chapter 2, then by tuning distribution parameters we can change the resulting statistics of the duration: decrease variance of duration corresponds to increase in the resulting fluctuations in protein number.

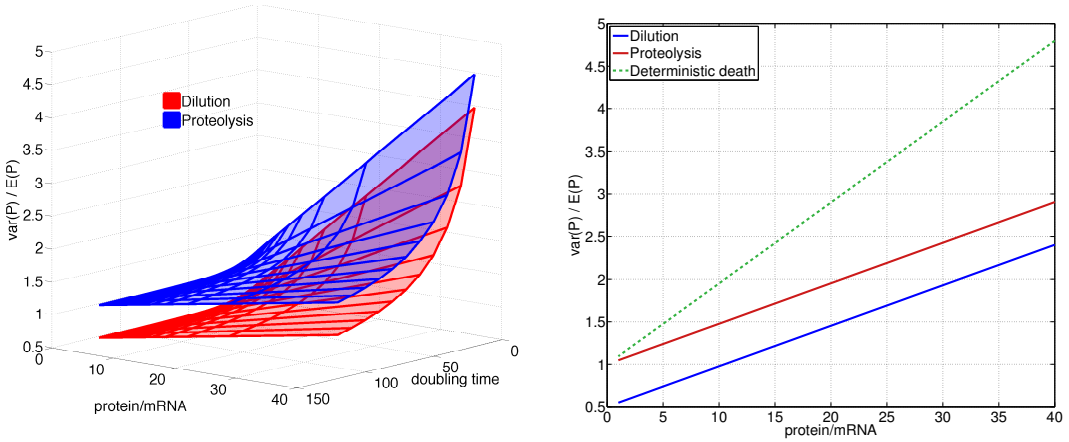
3.2.4 Proteolysis vs. dilution

Consider now a four-stage model with deterministic protein elongation and $\sigma_2 \sim E_{\mu_2}$ as discussed in Section 3.1.2. If we assume that the protein decay occurs by *proteolysis*, then using formula

(3.1.19) and the previous assumptions we obtain

$$\mathrm{var}_{\mathrm{LYSIS}}(P) = \mathbb{E}(P)\left[1 + \frac{\lambda_3}{\mu_2+\mu_4} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\lambda+\mu_2)(\Lambda+\mu_4)}\right], \tag{3.2.7}$$

where $\mathbb{E}[P] = \delta_+ \frac{\lambda_2\lambda_3}{\mu_2\mu_4}$ and $\Lambda = \lambda_1^+ + \lambda_1^-$.



(a) Proteolysis and volume dilution

(b) Protein decay: proteolysis, dilution and deterministic (punctual)

Figure 3.11: In (a) the Fano factor, i.e. $\mathrm{var}(P)/\mathbb{E}[P]$, is plotted with respect to the ratio $\mathbb{E}[P]/\mathbb{E}[M]$ and the cell doubling time varying from 20 minutes up to 120 minutes. The mRNA number, which affects the ratio $\mathbb{E}[P]/\mathbb{E}[M]$, varies now from 1 up to the average number of proteins $\mathbb{E}[P] = 40$. The protein elongation speed is 18 amino acids per second and the protein length is 300 amino acids. The mRNA lifetime is fixed at 2 minutes. The red-coloured graph represents the Fano factor for the four-stage model with proteolysis, while the blue one corresponds to the dilution case as described in Section 3.1.2. In (b) the Fano factor is plotted as function of $\mathbb{E}[P]/\mathbb{E}[M]$, with cell doubling time fixed to 20 minutes. Here, with respect to plot (a), we have added the curve relative to the case of discrete deterministic protein death, green dashed line, corresponding to the *four-stage model* with deterministic death. The plots are obtained using the analytic formulas (3.2.7) and (3.2.8).

If on the other side we consider volume dilution as the major mechanism of protein decay, then, using Theorem 3.1.6 and the formula (3.1.31), we obtain

$$\mathrm{var}_{\mathrm{DIL}}(P) = \mathbb{E}(P)\left[\frac{1}{2} + \frac{\lambda_3}{\mu_2+\mu_4} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\lambda+\mu_2)(\Lambda+\mu_4)}\right], \tag{3.2.8}$$

where μ_4 represents now the cell growth rate, i.e. $\mu_4 = \nu$. The comparison of the two scenarios is represented in Figure 3.11, where we plot the relative variance for both dilution and *proteolysis* case, as function of cell doubling time τ and of the average number of proteins produced per mRNA. As shown in Figure 3.11a, the variance associated with *proteolysis* (red surface) overestimates the fluctuations with respect to the surface relative to dilution (blue surface). Therefore, the description of protein dilution in classic models to have an exponentially distributed duration overestimates the variance.

Remark. In Figure 3.11 we have plotted the Fano factor, defined as $\mathrm{var}(P)/\mathbb{E}[P]$, since both formulas (3.2.7) and (3.2.8) are linear with respect to the average number of proteins. For this reason, in the limit case of small values of the average number of proteins per messenger we recover the first constant terms of equations (3.2.7) and (3.2.8), 1 and 0.5 respectively.

The difference between the two mechanisms of decay is just in the first term and, in both formulas protein noise $\mathrm{var}(P)/\mathbb{E}^2(P)$ decreases as the level of expression of the protein increases. However, if such scaling behavior is captured by both descriptions, there is a difference, because of the scaling constant is now replaced with $1/2$. This explains the reason of the qualitative agreement of experimental results, such as [2], with the noise scaling, but shows that if qualitatively classic models capture this phenomenon, the formulas might not be quantitatively accurate.

The difference of these two approaches is not noticeable as long as the first constant term of equations (3.2.7) and (3.2.8) is smaller than the other two terms, which seems to be the case in many situations. However, when we consider an active promoter, i.e. $\delta_+ = 1$, the last term of the previous formulas disappear and the difference between the two descriptions become remarkable as long as we consider weak affinity between messengers and ribosomes, i.e. $\lambda_3 << 1$, in the case of an unstable protein, i.e. $1/\mu_4$ smaller than cell cycle. In this case, still in a parameter window of biological interest, the impact of a more realistic description is substantial, especially for highly-abundant proteins. It would be worth verifying experimentally such conclusions obtained just by the theoretical model even if in parameter ranges of biological interest.

Conclusions on protein decay step

We have analyzed the impact of different descriptions of protein decay on protein fluctuations. In particular, we have shown that in the case in which dilution is the main decay mechanism, then the impact of a more appropriate model description, the *hybrid four-stage model*, reflects in a change in the first addend of the variance of the number of proteins. Comparing the *hybrid four-stage model* with the *four-stage model* we have seen that there could be significant differences between the two descriptions in terms of protein variance under biologically relevant sets of parameters in the case of non-regulated genes, i.e. $\delta_+ = 1$, and it would be worth to verify experimentally the results obtained.

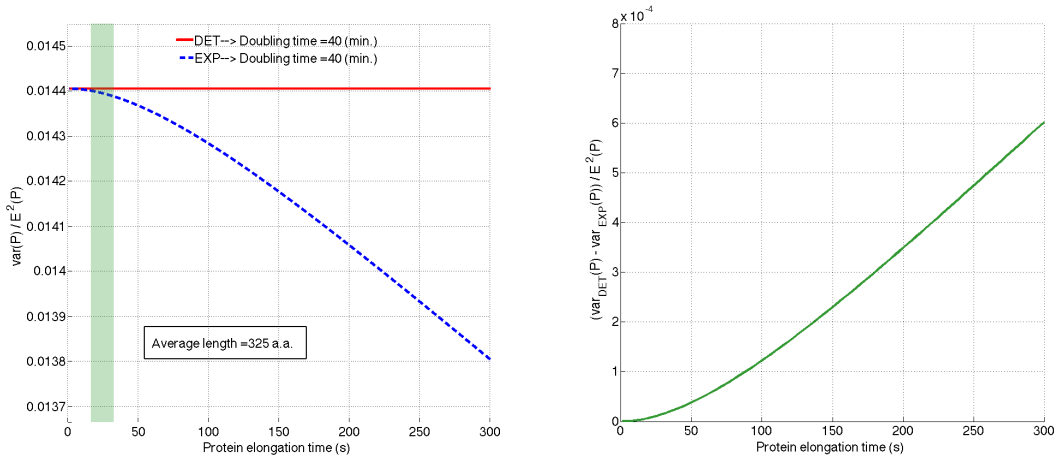
However, in the case of regulated proteins, i.e. $\delta_+ \neq 1$, the third term in the formulas (3.2.7) and (3.2.8) dominates and the results obtained by the two models are very close.

3.2.5 Impact of different steps on protein fluctuations.

We analyze now the four-stage model with realistic assumptions discussed in Section 3.1.2, i.e. all processes are supposed to have an exponentially distributed duration except the protein elongation which is supposed to be deterministic and the protein decay, which is supposed to be driven by dilution. Recall that, in this case, the formula of variance is (see equation (3.2.5))

$$\begin{aligned}\mathrm{var}^{(\mathrm{D})}(P) =& \mathbb{E}[P]\left[\frac{1}{2} + \frac{\lambda_3}{\mu_2+\mu_4} + \frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\Lambda+\mu_2)(\Lambda+\mu_4)}\right] \\ =& \mathbb{E}[P]\left[\frac{1}{2} + \frac{\mathbb{E}[P]}{\mathbb{E}[M]}\frac{\mu_4}{\mu_2+\mu_4} + \mathbb{E}[P]\frac{\mu_2\mu_4}{\mu_2+\mu_4}\frac{(1-\delta_+)}{\delta_+}\frac{(\Lambda+\mu_2+\mu_4)}{(\Lambda+\mu_2)(\Lambda+\mu_4)}\right].\end{aligned}$$

In order to analyze the impact of the choice of deterministic protein elongation, we compare this model with a similar model where the elongation duration is supposed to be exponentially distributed and the variance is given by formula (3.2.6).



(a) Active gene. Deterministic vs Exponential elongation.

(b) Active gene. Difference in relative variances.

Figure 3.12: **Impact of elongation distribution on fluctuations of stable protein with no promoter regulation.** In figure (a), the relative variance for the four-stage model with deterministic (red curve) and exponential (blue curve) protein elongation is plotted with respect to the protein elongation time. The average protein number is $\mathbb{E}[P] = 400$, average mRNAs number $\mathbb{E}[M] = 4$ with average lifetime of 2 minutes, while the cell doubling time is 40 minutes. The green rectangle corresponds to the elongation time window of a reference protein of 325 amino acids, the elongation time depending on the speed of elongation varying between 10 and 20 amino acids per second ($a.a./s$). In figure (b) we plot the difference between the relative variances with the two choices of distribution.

General remarks

The deterministic protein elongation has been chosen as a limiting case and seems to act as an upperbound for the variance of proteins for any other chosen distribution of protein elongation. Formula (3.2.5) shows clearly that the variance scales with respect to the protein level or, equivalently, the relative variance decreases as $1/\mathbb{E}[P]$.

The second term of the previous formula shows also the dependence of the noise on the amplification factor of messengers. In particular, this term is proportional to the ratio $\mathbb{E}[P]/\mathbb{E}[M]$, i.e. the number of proteins produced per messenger, and we expect to increase the noise by increasing this ratio, as confirmed by the Figure 3.13 and 3.18b. Therefore, a strategy to reduce noise at fixed protein expression level would be to increase the number of messengers, reducing the impact of this addend. This is in agreement with the work of Ozbudak et al.[51], where the authors studied the sensibility of noise with respect to changes in the transcription or translation rates by modifying genetically strains of *Bacillus Subtilis*. An increased transcription rate resulted in a sensible decrease in the noise of the corresponding protein.

The third term in the previous equation corresponds to the noise deriving from the dynamics of gene regulation. This term is proportional to the ratio $(1 - \delta_+)/\delta_+ = \lambda_1^-/\lambda_1^+$, which is the ratio of the probability of inactive gene over the probability of active one. This term seems to dominates the others in several cases and the introduction of a gene regulation is associated with noisier profiles of the protein dynamics. Since the introduction of the first model considering

gene activation, see Peccoud and Ycart [57], a debate has occurred in the community because the gene activation gives a good explanation of bursty profiles of messenger and protein numbers, see [22, 61]. The observation of high intensity production followed by long quite periods, also called *bursts*, is consistent with the results obtained by the introduction of the gene regulation.

Stable proteins

We first analyze the case of stable proteins, i.e. $\delta_+ = 1$. The main mechanism of protein decay is now represented by volume dilution, whose typical time scale is the cell cycle time, which varies between ~ 20 and ~ 60 minutes. In the case of non-regulated genes, the variance of the four-stage model with realistic assumptions (3.2.5) is greater than formula (3.2.6), in fact, if $\delta_+ = 1$,

$$\begin{aligned}\mathrm{var}^{(\mathrm{D})}(P) =& \mathbb{E}[P]\left[\frac{1}{2} + \frac{\mathbb{E}[P]}{\mathbb{E}[M]}\frac{\mu_4}{\mu_2+\mu_4}\right] \\ \geq& \mathbb{E}[P]\left[\frac{1}{2} + \frac{\mathbb{E}[P]}{\mathbb{E}[M]}\frac{\mu_4}{\mu_2+\mu_4}\left(1 - \frac{\mu_2\mu_4}{(\mu_2+\mu_3)(\mu_3+\mu_4)}\right)\right] = \mathrm{var}^{(\mathrm{E})}(P)\end{aligned}$$

In order to analyze the impact of this step under realistic assumptions with respect to classic approach, we consider realistic parameter ranges.

In Figure 3.12a we plot the relative variance of protein number with deterministic (red) and exponential (blue) elongation step. For short elongation times, the two descriptions give similar fluctuations, while the deterministic description differs more and more as long as the elongation time increases. Despite the difference might seem important, the plot of the difference in relative variances shows that the differences is of the order of 10^{-4} and seems not to be as crucial as other steps, as we will see later in this Section.

We focus on proteins corresponding to regulated genes, i.e. $\delta_+ \neq 1$. Because of the previous result, we have to study the third addend in formula (3.2.5) which can be rewritten as

$$\frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\Lambda+\mu_2)(\Lambda+\mu_4)} = \mathbb{E}[P]\frac{\lambda_1^-}{\lambda_1^+}\frac{\mu_2\mu_4}{(\mu_2+\mu_4)}\frac{\Lambda+\mu_2+\mu_4}{(\Lambda+\mu_4)(\Lambda+\mu_2)}. \tag{3.2.9}$$

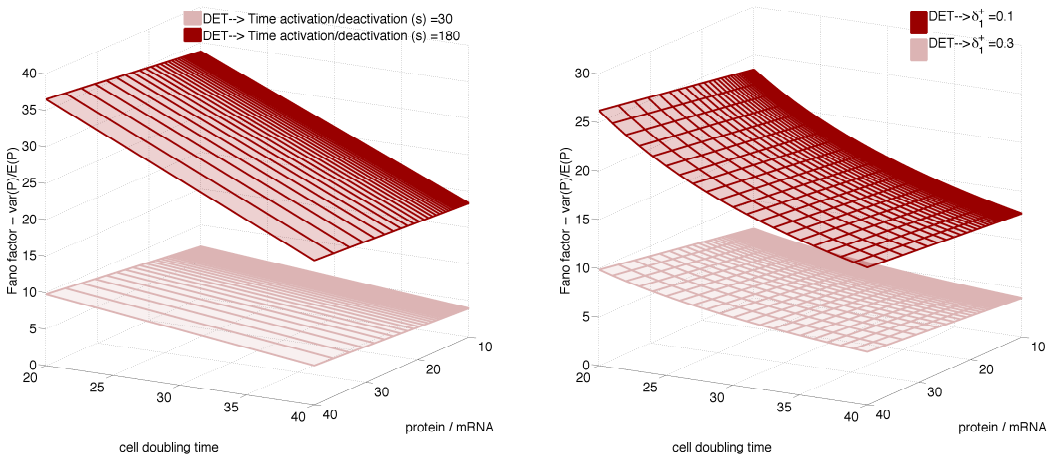
We restrict our analysis to biologically relevant ranges of the model parameters. The parameter $\Lambda = \lambda_1^+ + \lambda_1^-$ gives the order of magnitude of the time to wait in order to observe changes in the gene regulation activity, while the term δ_+ represents the percentage of activation of the gene. The gene regulation is obtained by means of transcript factors (*activators* or *repressors*); experiments show that in average each transcript factor visit the corresponding gene once each ~ 45 minutes. Moreover each gene exhibits in average ~ 100 transcript factors, which makes the visit time of transcript factors on the DNA of the order of at most few minutes. For example, as showed by Elf et al.[15], the *lac* repressor spends almost 90% of time non-bounded and diffusing, with a residence time of < 5 milliseconds. The time for a single *lac* repressor dimer in one cell to find a specific operator has been estimated to be ~ 354 seconds.

We suppose therefore for stable proteins and under physiologically meaningful conditions that $\Lambda >> \mu_4$ and $\mu_2 > \mu_4$. In this case, (3.2.9) reduces to the following approximated formula

$$\frac{\lambda_2\lambda_3(1-\delta_+)(\Lambda+\mu_2+\mu_4)}{(\mu_2+\mu_4)(\Lambda+\mu_2)(\Lambda+\mu_4)} \approx \mathbb{E}[P]\frac{(1-\delta_+)}{\lambda_1^+}\frac{\mu_2\mu_4}{\mu_2+\mu_4},$$

which gives the following approximated formula for the protein variance

$$\mathrm{var}^{(\mathrm{D})}(P) \approx \mathbb{E}[P]\left[\frac{1}{2} + \frac{\mathbb{E}[P]}{\mathbb{E}[M]}\frac{\mu_4}{\mu_2+\mu_4} + \mathbb{E}[P]\frac{\lambda_1^-}{\lambda_1^+}\frac{1}{\Lambda}\frac{\mu_2\mu_4}{\mu_2+\mu_4}\right], \tag{3.2.10}$$



(a) Impact of gene timescale.

(b) Impact of regulation activity.

Figure 3.13: **Impact of gene regulation on protein variance.** The Fano factor for the four-stage model with deterministic protein elongation is plotted with respect to the ratio $\mathbb{E}[P]/\mathbb{E}[M]$ and with different cell doubling times, from 20 up to 40 minutes. Here the average protein number is fixed to $\mathbb{E}[P] = 40$. The average number of mRNAs varies between 1 and $\mathbb{E}[P]/10$ and their lifetime is fixed to 2 minutes.
The two surfaces in figure (a) correspond to two different typical times for gene regulation changes, i.e. the typical activity time τ_{act} takes the value 30 and 180 seconds and we have the relation $\Lambda = 1/(60 * \tau_{\text{act}})$. In this plot, the activation rate is $\delta_+ = 0.1$.
The two surfaces in figure (b) correspond to two different rates of gene regulation activity, i.e. we suppose that the gene probability of activation δ_+ takes the values 0.1 and 0.3, where the lower surface corresponds to the higher gene activity.

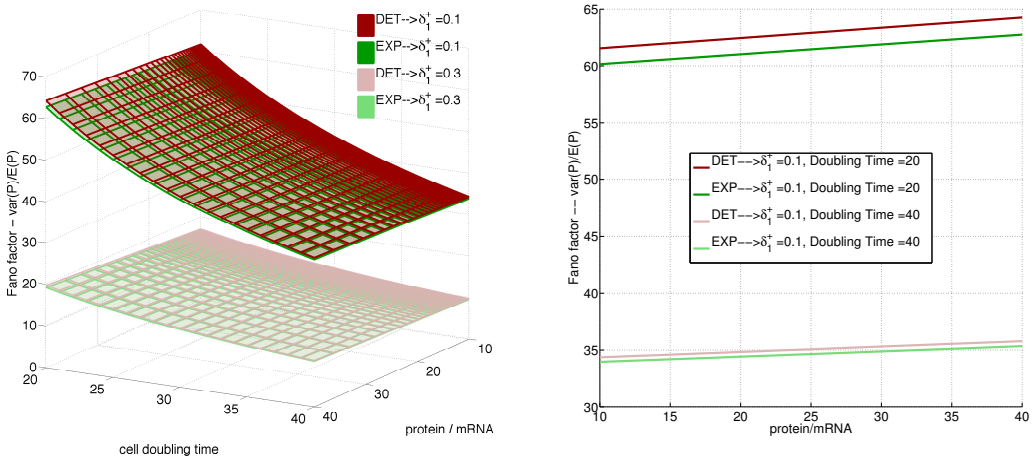
where the approximation concerns only the third term inside brackets.

The third term of formula (3.2.10) is proportional to the protein average $\mathbb{E}[P]$ and no more on the ratio $\mathbb{E}[P]/\mathbb{E}[M]$. As consequence, the impact of this term is increased as long as we consider highly expressed proteins, no matter the number of proteins produced per transcript. This term depends also on gene activation: by lowering the probability of the repressor unbinding δ_+ (resp. increasing the probability of repressor binding $1 - \delta_+$) or by increasing the time $1/\Lambda$ between gene activation/deactivation events we increase the protein variance. The results relative to changes in the gene activity are showed in figure 3.13. In particular, figure 3.13a shows the impact of changes in the typical time of gene regulation: when regulation events become rare, the corresponding fluctuations increase considerably (here $\delta_+ = 0.1$). On the other side, figure 3.13b analyzes the impact of changes in the activity of the gene, i.e. the influence of changes of the probability to have activation/deactivation. If we consider low probability of gene activation, $\delta_+ = 0.1$, then we experience larger fluctuations than for a case where the gene is more active. Figure 3.13 shows also that the cell doubling time has an important role in the resulting protein fluctuations, shorter doubling time corresponds to higher fluctuations.

We now compare the impact of gene activity and of different choices of protein chain elongation, in order to identify the most important steps and test the impact of more realistic assumptions of protein elongation on protein fluctuations. In Figure 3.14, we plot the Fano factor of protein number with respect to different values of gene activity, i.e. $\delta_+ = 0.1$ and $\delta_+ = 0.3$, for

the exponential and deterministic assumption of protein elongation. We consider here a "standard" prokaryotic protein, i.e. it is composed of 300 amino acids and the speed of translation is fixed at $18\,a.a./s$. Red and green surfaces in figure 3.14a represent the protein Fano factor with deterministic and exponential chain elongation. It is clear that a change on gene activity has a major impact on protein fluctuations, while the red and green surfaces are close for identical parameters.

In order to analyze the impact of cell doubling time, we now fix the gene regulation at $\delta_+ = 0.1$ and we consider the curves relative to 20 and 40 minutes of cell doubling time, see Figure 3.14b. As for gene regulation, different growth conditions have a major impact on protein fluctuations with respect to different choices for the protein elongation distribution. In particular, while changes between deterministic and exponential elongation account for less than 3%, the change in cell doubling time accounts here for almost 50%. Thus, also in the case of switching gene status, the impact of protein elongation on the overall protein fluctuations seems to have a smaller impact than other steps. In particular, the results suggest that protein fluctuations are extremely sensitive in changes in the gene regulation and cell doubling time, which seem to account for a large part of the observed fluctuations.



(a) Impact of gene activity.

(b) Impact of cell doubling ($\delta_+ = 0.1$), 2D plot.

Figure 3.14: **Impact of gene regulation vs protein elongation.** The Fano factor for the four-stage model with **deterministic** and **exponential** protein elongation is plotted with respect to the ratio $\mathbb{E}[P]/\mathbb{E}[M]$ and with different cell doubling times, from 20 up to 40 minutes, where $\mathbb{E}[P] = 40$. The average number of mRNAs varies between 1 and $\mathbb{E}[P]/10$ and their average lifetime is fixed to 2 minutes. In particular, gene activity takes the values $\delta_+ = 0.1$ and $\delta_+ = 0.3$, while gene regulation typical time is of $360\,s$ [15].
The red and green surfaces in figure (a) correspond to deterministic and exponential protein elongation.
In figure (b) we plot two cuts at $\tau_{\text{doubling}} = 20$ and $\tau_{\text{doubling}} = 40$ of the surfaces corresponding to $\delta_+ = 0.1$.

We conclude this section by giving approximated formulas of protein variance, which allow to highlight the dependence of fluctuations on global characteristics such as protein and messenger

averages. The last term in equation (3.2.10) can be written as

$$\frac{\mu_2\mu_4}{\mu_2+\mu_4} = \frac{1}{\frac{1}{\mu_2}+\frac{1}{\mu_4}},$$

so we can reduce the third term if we have a slow cell dilution or long-lasting mRNAs, i.e. for small values of μ_4 and μ_2. Now since the second term in (3.2.10) requires large values of μ_2 in order to be minimized, then a good strategy is to consider unstable messengers coupled with slow cell growth.

From formula (3.2.10) we obtain the following formula for the Fano factor

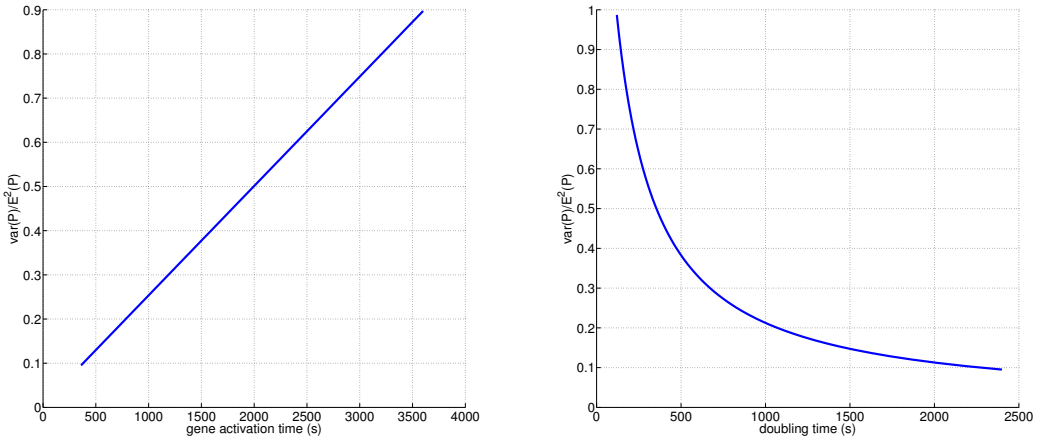
$$\frac{\mathrm{var}^{(\mathrm{D})}(P)}{\mathbb{E}[P]} \approx \frac{1}{2} + \frac{\mathbb{E}_{\max}[P]}{\mathbb{E}_{\max}[M]}\frac{\mu_4}{\mu_2+\mu_4} + \mathbb{E}_{\max}[P]\frac{1-\delta_+}{\Lambda}\frac{\mu_2\mu_4}{\mu_2+\mu_4}, \tag{3.2.11}$$

where

$$\mathbb{E}_{\max}[M] \overset{\text{def}}{=} \frac{\lambda_2}{\mu_2}, \tag{3.2.12}$$

$$\mathbb{E}_{\max}[P] \overset{\text{def}}{=} \frac{\lambda_2\lambda_3}{\mu_2\mu_4} \tag{3.2.13}$$

are the average number of mRNAs (resp. proteins) in the case of non-regulated gene, i.e. $\delta_+ = 1$. Note that, once we have fixed the parameters except gene activation, $\mathbb{E}_{\max}[M]$ and $\mathbb{E}_{\max}[P]$ represent the maximum amount of messenger and proteins respectively, therefore the label "max".



(a) Impact of gene activity.

(b) Impact of cell doubling time.

Figure 3.15: **Approximated relative variance for stable proteins.** Plot of the formula (3.2.14) for different values of gene typical activation/deactivation time ($\Lambda = \lambda_1^+ + \lambda_1^-$), figure (a), and for changes in cell doubling time, figure (b). We consider here a protein with mean $\mathbb{E}_{\max}[P] = 400$, associated with $\mathbb{E}_{\max}[M] = 10$, the average messenger lifetime is 2 minutes and the gene is mostly inactive, $\delta_+ = 0.1$.

Equivalently, we can compute the relative variance

$$\frac{\mathrm{var}^{(\mathrm{D})}(P)}{\mathbb{E}^2[P]} \approx \frac{1}{2\mathbb{E}_{\max}[P]} + \frac{1}{\mathbb{E}_{\max}[M]}\frac{\mu_4}{\mu_2+\mu_4} + \frac{1-\delta_+}{\Lambda}\frac{\mu_2\mu_4}{\mu_2+\mu_4}. \tag{3.2.14}$$

Remark. There is little information of the time of regulation of genes for large part of the genome and the activity of each gene seems to have specific characteristics. However, if $\Lambda \approx \mu_4$, then the gene regulation would be a rare event during cell cycle, causing abruptly changes in protein quantities from one generation to the other, resulting in a quite faulty regulation of the protein production. For this reason, we do not expect expressed proteins to show too long gene regulation cycles. A particular case is represented by stress-connected proteins, which are activated only in emergency cases. In this case, it is reasonable that the repressor is bound to the gene for really long times and, once activated, the corresponding protein would show a peak in production. This would be consistent with the results of Newman et al.[48], where stress proteins show high levels of noise.

Rare codons

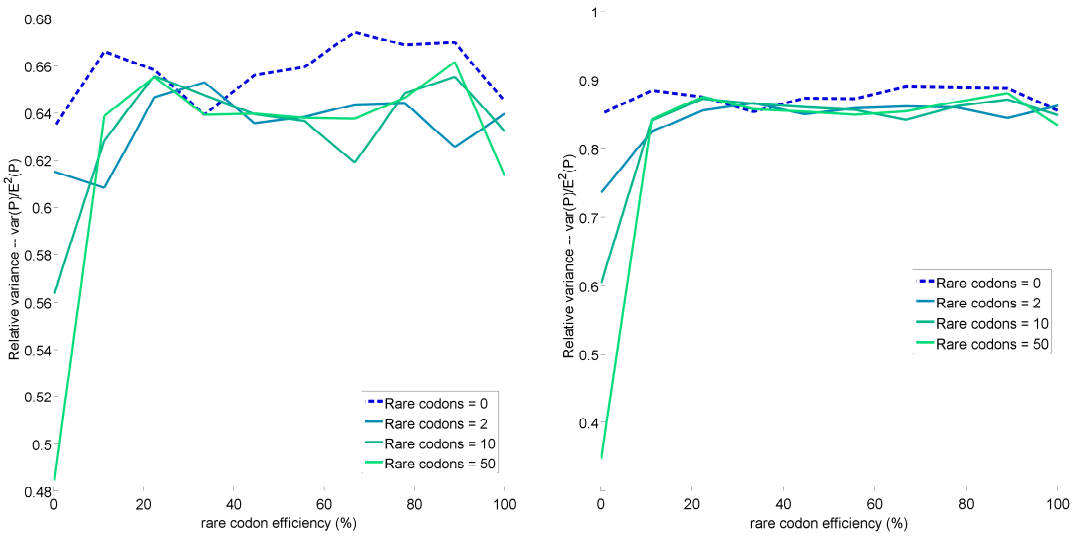
We briefly discuss the case of rare codons and their impact on protein fluctuations in the present context.

Any cell uses 61 different codons, i.e. triplets of nucleotides, to encode for 20 different amino acids and three termination codons. As a consequence, there is a redundancy in the possible choices of a genetic sequence and each codon is encoded by one up to six different sequences. These sequences are read by *tRNAs* which have been charged with the corresponding amino-acid. It results that the same protein can be therefore be produced by many alternative sequences of nucleic acids. The large variability on the frequency of the use of codons in different organisms is the result of evolutionary forces [24]. It turns out that there is a tight connection between condons and tRNA concentrations, since the system has evolved in order to reach a balance between these two fundamental components in order to optimize the efficiency of the translation apparatus.

The codon biases between different organisms have become of crucial interest since the intensive use in biotechnology of heterologous hosts to produce a desired protein. This technique consists to force a host organism to express, possibly at high rates, a gene which is not part of its genome mainly via recombinant DNA technology. However, it happens often that the protein is not expressed or it is expressed at very low levels. A plausible explanation of this abrupt change in expression level of the specific protein in hosts is that some codons used in the codification are rare codons in the host organism. Improvements have been done since the first use of this technique in 1977 [30], but there are still open questions on the subject and we refer to Gustafsson et al.[25] for further details.

The presence of rare codons also in wild types has arisen the question on the utility of such codons and their possible effect on the resulting gene expression. Using our modeling approach we study the impact of the presence of rare codons on the resulting fluctuations of different processes. In particular, we consider a 400 amino-acids protein with "standard" biological parameters, which is elongated in an average time of 100 seconds if no rare codons are present (dashed curves in figures 3.16a and 3.16b). We proceed by increasing the portion of rare codons, studying the impact with respect to the efficiency of the rare codons, i.e. we study the impact when the efficiency to process a rare codon is reduced to a fraction of the normal codon, low percentages meaning a long time to be processed. The duration of elongation step is described as the sum of two independent Erlang random variables, one accounting for the normal codons, the other for the rare codons. In accord with the results of the previous section, when the efficiency of codon processing drops down, fluctuations diminish. However, when the efficiency is above the 20% of the normal codons, even if the fraction of rare codons if larger than 10% of the total codons, the fluctuations align with those of a protein with no rare codons, see figure 3.16.

Rare codons are addressed in literature to have a negative impact on the expression of a



(a) Impact on protein fluctuations.

(b) Impact on ribosome fluctuations.

Figure 3.16: **Rare codons impact.** Relative variance of protein and ribosome concentration as function of the efficiency of the rare codons with respect to normal ones, i.e. 10% meaning that the rare codon rate of translation is the 10% of the translation rate of a normal codon. Given the speed of elongation and the average number of amino acids in a protein, we can estimate the time required to process each codon and obtain the rate of assembly of a "normal" codon. The rare codons are then modeled as to have less efficient assembly rates and the curves are plot with respect to the efficiency of the rare codons with respect to normal ones. Different curves represent the results for different concentrations of rare codons within a 400 amino-acids protein The protein is elongated in 100 seconds if no rare codons are present (dashed blue line).

specific gene, by showing low expression levels and possibly being responsible of bad folding of proteins. The result obtained is in apparent contrast with these results. In fact, as long as ribosomes are stuck during translation because of rare codons, all other protein types take advantage and can exploit common resources and rare codons seem here to create a disadvantage in this competition. The previous result says that, if all other conditions are fixed, the increase in protein translation time associated to the presence of rare codons would result in a less noisier protein production.

Unstable proteins

In the previous section we have seen that the protein elongation step has little impact for stable proteins. We now analyze the impact of the protein elongation step in the case of unstable proteins, i.e. proteins that are degraded at high rates through *proteolysis*. In particular, we assume protein lifetime is of the order of messenger's lifetime, i.e. $1-2$ minutes.

If $\mu_4 \approx \mu_2$, then *proteolysis* is the main protein decay mechanism while dilution is negligible. In this case, we consider non-regulated protein production, then the formulas reduce now to

$$\mathrm{var}^{(\mathrm{D})}_{\mathrm{LYSIS}}(P) \approx \mathbb{E}[P]\left[1+\frac{\mathbb{E}[P]}{\mathbb{E}[M]}\right] \geq \mathbb{E}[P]\left[1+\frac{\mathbb{E}[P]}{\mathbb{E}[M]}\left(1-\left(\frac{\mu_2}{\mu_2+\mu_3}\right)^2\right)\right] \approx \mathrm{var}^{(\mathrm{E})}_{\mathrm{LYSIS}}(P).$$

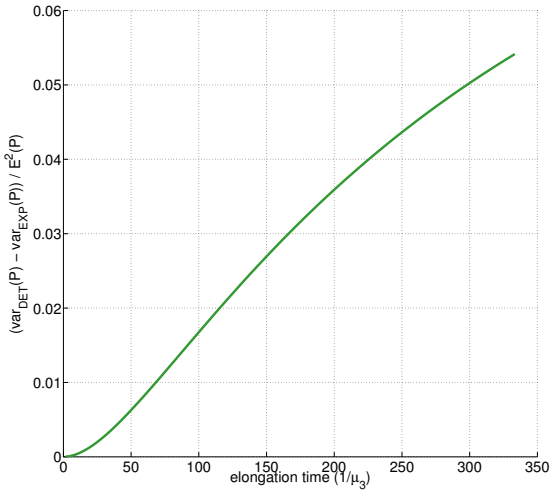


Figure 3.17: **Impact of elongation distribution on fluctuations of unstable protein with no promoter regulation.** Difference between relative variances with deterministic/exponential elongation for unstable proteins with non regulated promoter, i.e. $\delta_+ = 1$.

Note 3.2.1. *Observe that here we have used the general formula* (3.1.19)*, since the main mechanism of decay is* proteolysis*, while in the previous section we used the formula relative to protein dilution.*

The impact of the description of the elongation step is big as we consider long protein elongation time, i.e. small values of the parameter μ_3. Figure 3.17 shows that the impact of a more realistic description of protein elongation is significantly more important than for stable proteins, see Figure 3.12b.

The difference becomes more relevant if we consider also the promoter regulation. Similarly to the previous section, we may derive simplified formulas for the variance of proteins, given the assumption $\mu_4 \approx \mu_2$. The relative variance can be determined as

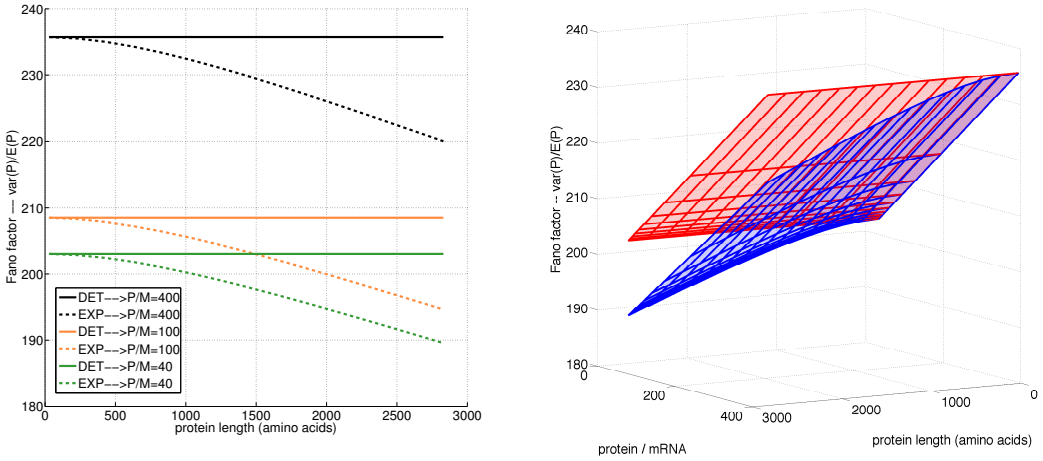
$$\frac{\mathrm{var}^{(\mathrm{D})}(P)}{\mathbb{E}^2[P]} \approx \frac{1}{\mathbb{E}_{\max}[P]} + \frac{1}{2}\frac{1}{\mathbb{E}_{\max}[M]} + \frac{1-\delta_+}{\delta_+}\frac{\mu_2(\Lambda+2\mu_2)}{(\Lambda+\mu_2)^2}, \tag{3.2.15}$$

where $\mathbb{E}_{\max}[M]$ and $\mathbb{E}_{\max}[P]$ are the average number of mRNAs (resp. proteins) in the case of non-regulated gene, i.e. $\delta_+ = 1$.

In Figure 3.18 we plot results relative to this situation. The plot 3.18b show the surfaces relative to deterministic (red) and exponential (blue) as function of the number of amino acids composing the protein and with respect to the average number of proteins produced per mRNA transcript. The elongation speed is fixed to $18\,a.a./s$ (amino acids per second). The discrepancy of the two surfaces is bigger when we consider longer proteins or when each transcript is translated dozens/hundreds of times. If we consider the three cuts shown in figure 3.18a it results that changes in the amplification factor of mRNAs/proteins change the level of noise, but there is little change in the relative difference between the exponential and the deterministic approach.

Despite the impact of elongation is more relevant in the case of unstable proteins, however, it seems to show a sensible difference just for very long proteins, while the exponential modelisation

seems to be a good approximation for short proteins.



(a) Fano factor as function of protein length for three different values of the ratio $\mathbb{E}[P]/\mathbb{E}[M]$.

(b) Fano factor as function of protein length and of the ratio $\mathbb{E}[P]/\mathbb{E}[M]$.

Figure 3.18: **Impact of protein elongation - deterministic vs. exponential.** Behavior of an unstable protein for the four-stage model under the assumptions of deterministic and exponentially distributed protein elongation. The protein lifetime is supposed to be identical to the mRNA lifetime, which is fixed to 2 minutes and the protein elongation speed is 18 amino acids per second. The average number of proteins is 400 and the gene deactivation rate $\lambda_1^- = 0.02$ is twenty times greater than the activation rate. In figure 3.18a is represented the Fano factor of protein number with respect to protein length, varying between 30 and 3000 amino acids. The three curves represent the Fano factor for three different values of the quantity number of proteins per mRNA ($P/M = 40, 100, 400$). Fano factor of the four-stage model under deterministic assumption, solid line in figure (a), is insensitive to changes in the protein length and is always larger than for the four-stage model with exponentially distributed protein elongation time, dashed line in figure (a). In figure 3.18b are shown the 3D graph, as function also of the ratio $\mathbb{E}[P]/\mathbb{E}[M]$, the red surface is relative to deterministic elongation, while the blue one to exponential elongation.

Conclusions on protein elongation step

In this section we have analyzed the impact of the description of protein elongation step on resulting fluctuations on the protein number. In Section 3.1.3 we have seen that the more realistic description of elongation step to have an Erlang distributed duration is close to assume the step to occur in a deterministic fashion, see also Figure 3.9. For this reason, we have focused on the deterministic and exponential choices for the elongation step.

Under deterministic assumption, protein shows higher fluctuations, which confirm that the common exponential assumption is not adapted to describe this step. However, if we look at the quantitative results for stable proteins in a range of biologically relevant parameters, then the difference between the two descriptions is small if compared to the impact of gene regulation and protein decay rate by dilution or *proteolysis*. This is slightly different in the case of unstable proteins, where we assumed that the decay time of proteins and of messengers is comparable.

In this case the deterministic elongation shows most marked difference for proteins composed of many amino acids.

Observe that the analysis conducted here assumes that the time to process each codon is identical. This brings to have a completely equivalent description in terms of time of elongation or protein length. If we consider also rare codons, the impact of elongation can be more significant. In fact, rare codons cause the ribosome to stop elongation and wait for the specific amino acid, slowing down the production of the specific protein.

In conclusion, the description of the elongation step seems to have a minor impact on the second moment of protein number except for the case of long unstable proteins, when we consider biologically meaningful model parameters.

3.A Reference parameters

In this section we resume the parameters of reference for *prokaryotes*, which has been used as reference parameters for our model.

The specific data considered here are referred to *E. Coli* bacterium and are mainly obtained from the classical reference *Modulation of chemical composition and other parameters of the cell by growth rate* by Bremer and Dennis [26]. We will further specify the reference for the parameters obtained from different papers.

Table 3.1: **Global parameters.** We list now few global parameters for *E.Coli* bacteria.

Parameters		Units	Value	Ref.
Proteins per transcript		unitless	10	[44]
Average percent of a given transcript species occupied by a ribosome		%	70	[60]
Average mRNA half-life		min.	5 to 10	[70]
Average repressor binding time	$\lambda_1^+ + \lambda_1^-$	s	~ 360	[15]
Number of proteins / mRNA	$\mathbb{E}[P]/\mathbb{E}[M]$	unitless	100 to 10000	[70]
Number of proteins / mRNA	$\mathbb{E}[P]/\mathbb{E}[M]$	unitless	10	[44]
Unstable protein half life		min.	5	[59]

In table 3.1 we have reported few global parameters of *E. Coli* bacterium. Using the data of Bremer [26] we have built table 3.2, which gives different cell parameters with respect to different growth regimes. These parameters have been used to built reference parameters to study the results of our model in different biologically meaningful situations.

Table 3.2: **Biological parameters.** Parameters pertaining to the macromolecular synthesis rates in exponentially growing *E. Coli* as function of growing rates at 37°C. Reference Bremer [26].

Parameters	Units	Division time		
		τ, 24 min.	τ, 60 min.	τ, 100 min.
RNA polymerases per cell	10^3 RNAP/cell	11.4	2.8	1.5
RNA polymerase activity	%	30	20	17
mRNA chain elongation	Nucl/s	55	45	39
Peptide chain elongation	aa/s	21	16	12
Ribosomes per cell	10^3 ribosomes/cell	72	13.5	6.8
Ribosomes activity	%	80	80	80
Distance of ribo. on mRNA	nucleotides	41	85	79

We conclude this section by giving few insights about the calibration of the model parameters using the biological data of tables 3.1, and 3.2.

Protein decay may occur because of cell dilution effect or because of *proteolysis*, as studied in Section 3.1.2. The protein dilution is caused by cell volume growth; for this reason it is strictly connected to the cell doubling time τ and the rate ν of protein dilution satisfies the relation

$$\nu = \frac{\ln(2)}{\tau}. \tag{3.A.1}$$

Proteolysis is a rare event for which we have no precise characterisation. For this reason we do not specify here any relation for the rate of proteolysis μ_4.

Bernstein *et alii* [6] showed that the 80% of all mRNAs has half-lives τ_{hl} between 3 and 8 minutes, while Yu *et alii* showed in [75] that the *tsr-venus* mRNA has a degradation time of 1.5 ± 0.2 minutes. Biologists commonly use the concept of *half-life* ($t_{1/2}$), i.e. the amount of time required for a quantity to fall to half its value measured at the beginning of the initial time of observation. In particular, the messengers and protein lifetimes are usually expressed in terms of "half-lives". Consider now the messenger decay, the case of protein decay via *proteolysis* being analogous. The mRNA degradation has been described as a first-order reaction, i.e. a biochemical reaction which depends on the concentration of only one reactant, the number of mRNAs in the specific case. The (average) rate law of the mRNA decay reaction is therefore $\frac{\mathrm{d}}{\mathrm{d}t}M(t) = -\mu_2 M(t)$, where μ_2 is the decay rate. If we denote with τ_{hl} the mRNA half-life, then we can obtain the parameter μ_2, characterizing the exponentially distributed messenger degradation, thanks to formula

$$\mu_2 = \frac{\ln(2)}{\tau_{hl}}. \tag{3.A.2}$$

The *proteolysis* rate μ_4 can be obtained using a similar formula, where instead of τ_{hl} we consider the protein half-life.

Table 3.3: **Model parameters.** Parameters of reference used in simulations. Here the parameters for a protein with average number of $\mathbb{E}[P] = 100$. We consider three different mRNAs average, $\mathbb{E}[M] = 1, 10, 100$, which affect the parameter λ_2. The average ribosome translation speed considered is 18 amino acids per second and the average protein length taken is 300 amino acids.

Parameters		Units	Division time		
			τ, 20 min.	τ, 60 min.	τ, 120 min.
Gene activation	λ_1^+	s^{-1}	1	1	1
Gene deactivation	λ_1^-	s^{-1}	0	0	0
Transcription init.	λ_2	$10^{-2}\mathrm{s}^{-1}$	0.58, 5.8, 58	0.58, 5.8, 58	0.58, 5.8, 58
mRNA half-life	$\ln(2)/\mu_2$	10^2s	1.73	1.73	1.73
Ribosome binding	λ_3	$10^{-3}\mathrm{s}^{-1}$	58,5.8, 0.58	19.3, 1.93, 0.19	9.6, 0.96, 0.096
Protein elongation	$1/\mu_3$	s	16.66	16.66	16.66
Dilution	ν	$10^{-4}\mathrm{s}^{-1}$	5.77	1.92	0.96

We fix now the average number of proteins, $\mathbb{E}[P]$, and the average number of mRNAs, $\mathbb{E}[M]$. In particular, we may derive the parameters λ_2 and λ_3 as functions of the degradation parameters and of the fixed averages. The transcription initiation parameter λ_2 is obtained by formula

$$\lambda_2 = \frac{\mu_2 \mathbb{E}[M]}{\delta_+}, \tag{3.A.3}$$

where we recall $\delta_+ = \lambda_1^+/(\lambda_1^+ + \lambda_1^-)$.

The translation initiation can be now be obtained as

$$\lambda_3 = \frac{\mu_2 \nu \mathbb{E}[P]}{\lambda_2 \delta_+} = \nu \frac{\mathbb{E}[P]}{\mathbb{E}[M]}. \tag{3.A.4}$$

We conclude this section by giving reference parameters that have been used to run simulations and to study the close analytic formulas of the Four-Stage Model for different probability distributions and parameter choices.

The speed of polypeptide chain elongation changes with respect to the growth regime, see table 3.2. We chose then the value of 18aa/s as average protein elongation speed.

The average protein length is of about 300 amino acids, see [7], and of about 360 amino acids for *E. Coli*, using table [46]. We choose then 300 as reference value for simulations. Therefore, if we took the reference value of 18 aa/s as average protein elongation speed, the time for protein elongation is of about 16 seconds.

The average messenger half life considered in simulation is $\tau_{hl} = 120$ seconds.

All these parameters are resumed in table 3.3. Here are considered parameters for different growth regimes, which have been taken to be 20, 60 and 120 minutes as cell doubling time.

Chapter 4

Multi-protein model

In the previous two chapters we have investigated the production of a specific protein, focusing on the characterization of the fluctuations of a single protein, which is, to our knowledge, the case of all the mathematical models of gene expression proposed up to now. When several classes of proteins are considered, we have to take into account the sharing of resources, since each class requires a fraction of the common and limited resources of the cell. Therefore, the allocation of the resources within the cell has to be understood in order to improve our knowledge of the gene expression. Large scale experiments have been performed in recent years in order to analyze the fluctuations of many different types of proteins within organisms [2, 48], but these global experimental results have not been accompanied by stochastic models considering the production of many types of proteins. In this chapter, a first step in this direction is done by investigating the allocation of a limited number of ribosomes in order to produce different types of proteins by means of a Markovian representation.

In particular, asymptotic results for equilibrium and transient behavior are obtained through a scaling procedure, under the reasonable biological assumption of *saturation*, i.e. the resources of the cell are limited. It is shown in particular that the number of free ribosomes, i.e. the ribosomes which are not elongating any polypeptide chain, converges in distribution to a Poisson distribution whose parameter satisfies a fixed point equation.

Plan of the Chapter. In Section 4.1 we describe the mathematical model of production of multiple-proteins and we derive the equilibrium results. Section 4.2 is devoted to the derivation of the asymptotic behavior of equilibrium and of the probability distribution of free ribosomes in the limiting regime. In Section 4.3 we analyze the underloaded and overloaded case for ribosome charge and analyze the resulting distributions of free and active ribosomes.

Introduction

The characterization of the protein production have focused in the previous chapters on the description of the main steps which lead the genetic information to be transformed into a functional product for a single protein. However, as long as we look at the whole picture and we consider the protein production in its globality, interactions between different protein production chains take place and this significantly complexifies the resulting scenario. Many studies have focused on the gene regulation pathways, studying the mutual interactions of a set of proteins. These models, mainly deterministic, are characterized by the introduction of *feedbacks*. Here we tackle the problem from a different point of view. More precisely, experimental studies focusing on exploration of noise have pointed out how part of the resulting noise is not protein specific

but seems to have a global effect on all proteins under analysis. This global effect has been associated to fluctuations in cellular components such as *polymerases*, *metabolites* or *ribosomes*, as suggested by Elowitz et al.[16, 69]. Therefore, we focus here on the impact of limited global resources, ribosomes, on the production of different types of proteins.

Competition for cell resources

Ribosomes are a crucial player in the expression of all genes and turn out to be extremely expensive in terms of resources. For this reason cells seem to strictly regulate their number in order to optimize the expression of genes and do not waste resources. We will obtain results on the resulting fluctuations in protein production in both cases of under and over-exploitation of ribosomes, the underloaded and overloaded cases in the following.

To each protein type is associated a specific gene and a specific mRNA type. Nevertheless, there are agents, *polymerases* and *ribosomes*, in charge of steps of gene expression, which are common to all protein types. These macromolecules can be seen as common resources in the protein production at cellular level and, because of their complexity, the cost for their production is quite high in term of resources. For this reason the number of polymerases and ribosomes is kept limited and strictly controlled. Protein production can thus be seen, in first approximation, as the result of competition between genes (resp. mRNAs) to associate with polymerases (resp. ribosomes). Most of the mathematical models of the literature consider exclusively the production of a given protein, without taking into account the limited resources due to competition with the simultaneous production of other types of proteins.

In this chapter, we consider a first step in considering the competition for resources resulting from the simultaneous production of many protein types. In particular, we analyze a stochastic model of the competition for ribosomes, which will act as limiting factor. We assume that there are P classes of proteins, each class being characterized to have fixed concentrations $A_1, \dots, A_P$. The total number N of ribosomes in the cell is assumed to be constant. The ribosomes which are not bound to any mRNA are referred to as *free ribosomes*, while the ribosomes which are attached on messengers are referred to as *active ribosomes*. To each class of proteins of concentration A_p, for $1 \leq p \leq P$, are associated K_p messengers. Each messenger associated to a protein of class A_p is said of being of class p and can accomodate a maximum of C_p ribosomes. Each ribosome bound to a messenger of class p elongates a new polypeptide chain at rate μ_p and is then released, thus entering in the class of free ribosomes. To stick with biological evidence, we assume that a *saturation condition* holds, i.e. the total number N of ribosomes is significantly smaller than the number of ribosomes that all mRNAs would bind to at equilibrium in an unconstrained system.

If we denote with y_p the number of active ribosomes bound to a messengers of class p, then the total number of free ribosomes is given by $N - y_1 - \cdots - y_P$. A free ribosome binds at rate λ_p to a messenger of the class p with less than C_p ribosomes already attached (recall that any messenger of class p can accommodate at most C_p ribosomes). The rate λ_p is an effective parameter including different physical characteristics, such as the affinity of a mRNA of class p with ribosomes and is therefore crucial in the description of the competition for free ribosomes. The analysis is complicated, since the vector (y_p) of active ribosomes varies over time because of the bindings of free ribosomes and, once the production of a protein is completed, to the unbinding of active ribosomes.

The main simplification in the model presented in this chapter, is the assumption that the number of mRNAs associated to each class is fixed and we adopt the exponential assumption for each step throughout this chapter. The model does not consider the decay of proteins, since it focus on the dynamics of ribosomes. These simplifications have been necessary in order to obtain a first description of the competition for common resources and to move towards a

systemic description of the protein production.

4.1 Stochastic model

We introduce now the model of production of proteins and the associated notations. The total number N of ribosomes in the cell is supposed to be constant. We denote with $A_1, \ldots, A_P$, the P distinct values of concentration of proteins, for each concentration A_p being associated K_p different types of proteins, for $1 \leq p \leq P$. For any $1 \leq p \leq P$ and $1 \leq k \leq K_p$, the corresponding protein (resp. mRNA) is referred to as being of class (p, k). The rate at which a ribosome binds on an mRNA of class (p, k) is denoted with λ_p, in particular this parameter is assumed to depend exclusively on protein concentration. Similarly, once a ribosome is bound to a mRNA of class (p, k), it produces a new protein at rate μ_p. We define the *load* ρ as

$$\rho_p \stackrel{\text{def}}{=} \frac{\lambda_p}{\mu_p},$$

where $1 \leq p \leq P$. An mRNA of class (p, k), being of finite length, can accommodate a maximum of C_p ribosomes, where this threshold depends only on the respective protein concentration.

Let $X^N_{p,k}(t)$, for $1 \leq p \leq P$ and $1 \leq k \leq K_p$, be the number of ribosomes attached on the messenger of class (p, k). In particular, the quantity

$$N - \sum_{p=1}^{N} \sum_{k=1}^{K_p} X_{p,k}(t)$$

is therefore the number of free ribosomes at time t, i.e. the ribosomes which are not attached on any messenger.

Remark. Note that in the present model we suppose that there is a fixed number of mRNAs for each class of proteins, each messenger having C_p slots for ribosomes. Each class of protein of concentration A_p contains only one type of protein, i.e. the product of a specific gene. In order to consider the case L different types of proteins, it sufficed to consider L classes $(p, m_1), \ldots, (p, m_L)$, hence the assumption of one type of protein per class is not restrictive.

State space and Q-matrix

Since we have supposed that the duration of each event is exponentially distributed, then the protein production is Markovian. More in detail, the Markovian process

$$(X(t)) = \{(X_{p,k}(t)) : 1 \leq p \leq P, 1 \leq k \leq K_p\}$$

describes the protein production and is an irreducible Markov process with values in the state space

$$\mathcal{S} = \left\{ \left(x_{p,k}, \substack{1 \leq k \leq K_p, \\ 1 \leq p \leq P} \right) : x_{p,k} \in \mathbb{N}, 0 \leq x_{p,k} \leq C_p \text{ and } \sum_{p=1}^{P} \sum_{k=1}^{K_p} x_{p,k} \leq N \right\}.$$

Because of our assumptions, the Q-matrix $Q = \{q(x, y) : x, y \in \mathcal{S}\}$ is given by, for $1 \leq k \leq K$ and $x = (x_{p,k}) \in \mathcal{S}$,

$$\begin{cases} q(x, x + e_{p,k}) = \lambda_p r(x), & \text{if } x_p + 1 \leq C_p, \\ q(x, x - e_{p,k}) = \mu_p, & \text{if } x_p > 0, \end{cases} \tag{4.1.1}$$

where

$$r(x) = N - \sum_{p=1}^{P} \sum_{k=1}^{K_p} x_{p,k}$$

is the total number of ribosomes and $e_{p,k}$ is the unit vector associated to the coordinate (p,k), i.e.

$$e_{p,k}(\tilde{p},\tilde{k}) = \delta_{(p,k)(\tilde{p},\tilde{k})} = \begin{cases} 1 & \text{if } \tilde{p} = p \text{ and } \tilde{k} = k \\ 0 & \text{otherwise.} \end{cases}$$

This class of models is related to stochastic models of communication networks like loss networks. See Kelly [34] for example. The difference here is that the analogue of the "input rate", i.e. the rate at which free ribosomes binds to mRNAs, depends on the state of the process through the variable $r(x)$. Another class of related models are the Gordon-Newel networks where a fixed number of customers/ribosomes travel through the different nodes of the network with some routing mechanism. See Gordon and Newel [23].

Invariant distribution

We derive now few results concerning the invariant distribution.

Proposition 4.1.1 (Stationary distribution)**.** *The Markov process $(X(t))$ is reversible and its invariant distribution π is given, for $x = (x_{p,k}) \in \mathcal{S}$, by*

$$\pi(x) = \frac{1}{Z} \frac{1}{r(x)!} \prod_{p=1}^{P} \prod_{k=1}^{K_p} \rho_p^{x_{p,k}} \tag{4.1.2}$$

where, for $1 \le p \le P$, $\rho_p = \lambda_p/\mu_p$,

$$r(x) = N - \sum_{p=1}^{P} \sum_{k=1}^{K_p} x_{p,k}$$

is the number of free ribosomes and Z is the normalization constant.

Remark. Note that, when $(X(t))$ is in state $x = (x_{p,k})$, the quantity

$$\mu_p(\mathbb{1}_{\{x_{p,1}>0\}} + \mathbb{1}_{\{x_{p,2}>0\}} + \cdots + \mathbb{1}_{\{x_{p,K_p}>0\}})$$

is the total rate of production of proteins with concentration A_p.

Proof. One has only to check that, see [34, Theorem 1.3, page 6], for any $1 \le p \le P$ and $1 \le k \le K_p$, the relation

$$\pi(x) q(x, x + e_{p,k}) = \pi(x + e_{p,k}) q(x + e_{p,k}, x),$$

holds, i.e. that

$$\pi(x) \lambda_p r(x) = \pi(x + e_{p,k}) \mu_p$$

holds for any $x \in \mathcal{S}$ such that $x + e_{p,k} \in \mathcal{S}$. In fact, the previous relation is equivalent to

$$\rho_p r(x) \prod_{\tilde{p}=1}^{P} \prod_{\tilde{k}=1}^{K_p} \rho_{\tilde{p}}^{x_{\tilde{p},\tilde{k}}} = \frac{r(x)!}{r(x + e_{p,k})} \prod_{\tilde{p}=1}^{P} \prod_{\tilde{k}=1}^{K_p} \rho_{\tilde{p}}^{x_{\tilde{p},\tilde{k}} + e_{p,k}(\tilde{p},\tilde{k})}.$$

Now,

$$r(x+e_{p,k}) = N - \sum_{\tilde{p}=1}^{P}\sum_{\tilde{k}=1}^{K_p}\left(x_{\tilde{p},\tilde{k}} + e_{p,k}(\tilde{p},\tilde{k})\right) = r(x) - 1,$$

and

$$\prod_{\tilde{p}=1}^{P}\prod_{\tilde{k}=1}^{K_p}\rho_{\tilde{p}}^{x_{\tilde{p},\tilde{k}}+e_{p,k}(\tilde{p},\tilde{k})} = \rho_p^{x_{p,k}+1}\prod_{\substack{\tilde{p}=1\\ \tilde{p}\neq p}}^{P}\prod_{\substack{\tilde{k}=1\\ \tilde{k}\neq k}}^{K_p}\rho_{\tilde{p}}^{x_{\tilde{p},\tilde{k}}} = \rho_p\prod_{\tilde{p}=1}^{P}\prod_{\tilde{k}=1}^{K_p}\rho_{\tilde{p}}^{x_{\tilde{p},\tilde{k}}},$$

and the proposition is proved. □

A direct consequence of the previous proposition is the following proposition, which expresses the invariant distribution in terms of geometric random variables. This representation will be useful to derive asymptotic results in Section 4.2.

Proposition 4.1.2. *If, for $1 \leq p \leq P$ and $1 \leq k \leq K_p$, $(G_{p,k})$ is an i.i.d. sequence of geometric random variables with parameter ρ_p on the state space $\{0,\dots,C_p\}$, then the distribution of the number R of free ribosomes at equilibrium is given by, for $0 \leq n \leq N$,*

$$\mathbb{P}(R=n) = \frac{1}{Z_R}\frac{1}{n!}\mathbb{P}\left(\sum_{p=1}^{P}\sum_{k=1}^{K_p} G_{p,k} = N-n\right), \tag{4.1.3}$$

where Z_R is the convenient normalization constant.

Proof. Note that

$$R = N - \sum_{p=1}^{P}\sum_{k=1}^{K_p} X_{p,k}.$$

If $x = (x_{p,k})$ we have $\pi(x) = \mathbb{P}\left[(X_{p,k} = x_{p,k}, 1 \leq p \leq P, 1 \leq k \leq K_p)\right]$ and

$$\begin{aligned}\mathbb{P}\left[R=n\right] &= \mathbb{P}\left[\sum_{p=1}^{P}\sum_{k=1}^{K_p} X_{p,k} = N-n\right]\\ &= \mathbb{P}\left[(X_{p,k} = x_{p,k}, 1 \leq p \leq P, 1 \leq k \leq K_p), \sum_{p,k} x_{p,k} = N-n\right].\end{aligned}$$

Now if $G_{p,k}$ is a geometric random variable on $\{0,\dots,C_p\}$ of parameter $1-\rho_p$, for $p = 1, \dots, P$ and $k = 1, \dots, K_p$, then $\mathbb{P}\left[G_{p,k} = m\right] = \frac{(1-\rho_p)}{(1-\rho_p^{C_p+1})}\rho_p^m$, which, together with formula (4.1.2), gives

$$\mathbb{P}(R=n) = \frac{1}{Z_R}\frac{1}{n!}\mathbb{P}\left(\sum_{p=1}^{P}\sum_{k=1}^{K_p} G_{p,k} = N-n\right),$$

since $r(x) = n$. □

The expression of the invariant distribution seems satisfactory since it gives an explicit representation of the equilibrium of the system. Nevertheless, because of the large size of the state space $\mathcal{S}$, the partition function Z turns out to be hard to estimate and, consequently, it is difficult

to obtain the interesting characteristics of this system. In the following section we consider a scaling approach for this model by letting both the number of free ribosomes N and the number of messengers go to infinity with specific relation. Through this scaling method, we will obtain a qualitative description of the system and, in particular, the rate of production of proteins.

4.2 Asymptotic Behavior of Equilibrium

From now on, we assume that the number N of ribosomes is large and that the number of messengers K_p associated to class p of proteins with a fixed concentration A_p, $1 \leq p \leq P$, scales with N in the following way,

$$K_p^N = N\beta_p + o(\sqrt{N}), \tag{4.2.1}$$

in particular the sequence (K_p^N/N) converges to β_p. A superscript N is added to the various quantities related to the production of proteins when the cell contains N ribosomes. Note that for $1 \leq p \leq P$ and $1 \leq k \leq K_p^N$, the number of ribosomes that can be attached to an mRNA of class (p, k) is still bounded by the quantity C_p.

Saturation Regime

Proposition 4.1.1 shows that, at equilibrium, given that the number of free ribosomes is n, the state of the process $(X_{p,k}(t))$ has the same distribution as $(G_{p,k})$ where $G_{p,k}$ is a geometric random variable with parameter ρ_p restricted to the state space $\{0, 1, \ldots, C_p\}$,

$$\mathbb{P}(G_{p,k} = m) = \rho_p^m \frac{1-\rho_p}{1-\rho_p^{C_p+1}}, \quad 0 \leq m \leq C_p,$$

conditioned on the event

$$\mathcal{A}_n = \left\{ \sum_{p=1}^{P} \sum_{k=1}^{K_p^N} G_{p,k} = N - n \right\}.$$

Since ribosomes are costly macromolecules, their number is kept as small as possible provided the performance of the cell is maximum. In fact, the total number of ribosomes is the result of the trade-off between the efficiency of translation, obtained with a larger number of ribosomes, and the impact in terms of resources which derives from the production of ribosomes themselves. For this reason, a typical biological assumption is that the total number of ribosomes is much smaller than the mean number of places available on mRNAs. Therefore, it is reasonable to assume that the average number of ribosomes that would be active on the messengers in an unconstrained case is bigger than the total number of ribosomes N, i.e.

$$\sum_{p=1}^{P} K_p^N \mathbb{E}(G_{p,1}) > N$$

holds, i.e.

$$\sum_{p=1}^{P} K_p^N \rho_p \frac{C_p \rho_p^{C_p+1} - (C_P+1)\rho_p^{C_p} + 1}{(1-\rho_p)(1-\rho_p^{C_p+1})} > N$$

so that the probability of event $\mathcal{A}_n$ is small for all $n \geq 1$.

This saturation condition will hold for N sufficiently large if the following relation is satisfied

$$\sum_{p=1}^{P} \beta_p \rho_p \frac{C_p \rho_p^{C_p+1} - (C_p+1)\rho_p^{C_p} + 1}{(1-\rho_p)(1-\rho_p^{C_p+1})} > 1, \tag{4.2.2}$$

from relation (4.2.1).

Note 4.2.1. *Note that, if we consider the unconstrained case for the a available slots on mRNAs for ribosome bindings, i.e. we remove the constraint that each mRNA of class p can bind to at most C_p ribosomes, then, with similar arguments that lead to relation* (4.2.2)*, we obtained the saturation condition*

$$\sum_{p=1}^{P} \beta_p \frac{\rho_p}{1-\rho_p} > 1. \tag{4.2.3}$$

In the case $\rho < 1$, then the previous condition is satisfied whenever (4.2.2) *holds.*

One derives the asymptotic distribution of the number of free ribosomes at equilibrium in this limiting regime.

Theorem 4.2.2 (Free ribosomes limiting regime)**.** *Under the limiting regime* (4.2.1) *with the saturation condition* (4.2.2)*, as N goes to infinity, at equilibrium the number of free ribosomes converges in distribution to a Poisson random variable with parameter $\gamma(\underline{C})$ where $\gamma(\underline{C})$ is the unique non-negative solution y of the equation*

$$\sum_{p=1}^{P} \beta_p \rho_p y \frac{C_p \rho_p^{C_p+1} y^{C_p+1} - (C_p+1)\rho_p^{C_p} y^{C_p} + 1}{(1-\rho_p y)(1-\rho_p^{C_p+1} y^{C_p+1})} = 1, \tag{4.2.4}$$

$\underline{C} = (C_p)$ and $\rho_p = \lambda_p/\mu_p$.

The proof of Theorem 4.2.2 is given below.

We first introduce a new representation allowing to rewrite the number of free ribosomes and few connected results. For $N \geq 1$, define S_N as

$$S_N = N - \sum_{p=1}^{P} \sum_{k=1}^{K_p^N} G_{p,k},$$

where $G_{p,k} \sim \text{Geo}(\rho_p)$, with space state restricted to $\{0, 1, \ldots, C_p\}$.

From Equation (4.1.3) of Proposition 4.1.2, the distribution of R_N the number of free ribosomes can be expressed as

$$\mathbb{P}(R_N = n) = \frac{1}{Z_{R_N}} \frac{1}{n!} \mathbb{P}(S_N = n), \tag{4.2.5}$$

where Z_{R_N} is the normalization constant.

For a fixed $n \geq 0$, we first derive the asymptotic behavior of the sequence $(\mathbb{P}(S_N = n))$ by using a change of probability distribution similar to the one used in the proof of Bahadur-Rao's Theorem, see Dembo and Zeitouni [13]. A strengthened version of the local central limit theorem is also an important ingredient. This can be seen in fact as the probabilistic analogue of a saddle point method applied to the series

$$\sum_{\substack{x=(x_{p,k}) \in \mathcal{S} \\ r(x)=n}} \prod_{p=1}^{P} \prod_{k=1}^{K_p} \rho_p^{x_{p,k}},$$

see Flajolet and Sedgewick [17, chapter VIII] for example.

The advantage of the probabilistic formulation is that the technical framework is significantly lighter, due to use of a powerful central limit theorem in fact.

For $\theta > 0$, define S_N^θ the random variable whose distribution is defined by, for a non-negative function f on $\mathbb{N}$,

$$\mathbb{E}\left(f\left(S_N^\theta\right)\right) = \frac{1}{\phi_N(\theta)}\mathbb{E}\left(f\left(S_N\right)e^{\theta S_N}\right), \tag{4.2.6}$$

with $\phi_N(\theta) = \mathbb{E}\left(\exp(\theta S_N)\right)$. By taking for the function f the generating function, i.e. $f(n) = e^{n\eta}$ for $\eta \geq 0$, we get that S_N^θ can be represented as a sum of independent random variables

$$S_N^\theta = N - \sum_{p=1}^{P}\sum_{k=1}^{K_p^N} G_{p,k}^\theta, \tag{4.2.7}$$

where, for $1 \leq p \leq P$, $(G_{p,k}^\theta, k \geq 0)$ is a sequence of geometric random variables on $\{0, 1, \ldots, C_p\}$ with parameter $\rho_p e^{-\theta}$. In fact, since the moment generating function of the r.v. $G_{p,k}$ is

$$\mathbb{E}\left[e^{-\eta G_{p,k}}\right] = \frac{1-\rho_p}{1-\rho_p^{C_p+1}}\frac{1-(\rho_p e^{-\eta})^{C_p+1}}{1-\rho_p e^{-\eta}},$$

and since $G_{p,k}$ are i.i.d., then

$$\begin{aligned}\mathbb{E}\left[e^{\eta S_N^\theta}\right] &= \frac{1}{\phi_N(\theta)}\mathbb{E}\left[e^{\eta S_N}e^{\theta S_N}\right] = \frac{e^{N(\eta+\theta)}}{\phi_N(\theta)}\prod_{p,k}\mathbb{E}\left[e^{-(\eta+\theta)G_{p,k}}\right] \\ &= e^{\eta N}\prod_{p,k}\frac{1-\rho_p e^{-\theta}}{1-(\rho_p e^{-\theta})^{C_p+1}}\frac{1-(\rho_p e^{-(\eta+\theta)})^{C_p+1}}{1-\rho_p e^{-(\eta+\theta)}},\end{aligned}$$

which is the moment generating function of the random variable $N - \sum_{p=1}^{P}\sum_{k=1}^{K_p^N} G_{p,k}^\theta$.

From (4.2.7), with $f(\eta) = \mathbb{1}_{\{\eta \geq n\}}e^{-\theta\eta}$, for $n \geq 0$, we obtain

$$\mathbb{E}\left[\mathbb{1}_{\left\{S_N^\theta \geq n\right\}}e^{-\theta S_N^\theta}\right] = \frac{1}{\phi_N(\theta)}\mathbb{E}\left[\mathbb{1}_{\{S_N \geq n\}}\right],$$

therefore

$$\begin{aligned}\mathbb{P}\left[S_N \geq n\right] = \mathbb{E}\left[\mathbb{1}_{\{S_N \geq n\}}\right] &= \phi_N(\theta)\mathbb{E}\left(\mathbb{1}_{\left\{S_N^\theta \geq n\right\}}e^{-\theta S_N^\theta}\right) \\ &= \phi_N(\theta)\int_0^{+\infty}\theta e^{-\theta u}\mathbb{P}\left(n \leq S_N^\theta \leq u\right)\mathrm{d}u.\end{aligned} \tag{4.2.8}$$

In fact, let X denote a positive random variable with p.d.f. f_X and let $C > 0$, then

$$\begin{aligned}\int_0^C \mathbb{1}_{\{n \leq u\}}e^{-\theta u}f_X(u)\,\mathrm{d}u &= g(u)\Big|_{u=0}^{u=C} + \int_0^C \theta e^{-\theta u}\left(\int_0^u \mathbb{1}_{\{n \leq \xi\}}f_X(\xi)\,\mathrm{d}\xi\right)\mathrm{d}u \\ &= g(u)\Big|_{u=0}^{u=C} + \int_0^C \theta e^{-\theta u}\mathbb{P}\left[n \leq X \leq u\right]\mathrm{d}u,\end{aligned}$$

where $g(u) = e^{-\theta u}\int_0^u \mathbb{1}_{\{n \leq \xi\}}f_X(\xi)\,\mathrm{d}\xi$. Now, since $g(0) = 0$ and $\lim_{u\to\infty} g(u) = 0$, by letting $C \to +\infty$ we obtain the relation (4.2.8).

From relation (4.2.8), the asymptotic behavior of $(\mathbb{P}(S_N \geq n))$ can be obtained through limit results concerning S_N^θ. We will show in Lemma 4.2.3 that θ can be chosen so that the average of S_N^θ is 0 and then, due to the representation in terms of independent variables, a central limit theorem gives the desired asymptotics.

Lemma 4.2.3. *For any $N \geq 1$, there exists a unique θ_N such that $\mathbb{E}(S_N^{\theta_N}) = 0$ and the sequence (θ_N) converges to $-\log\gamma(\underline{C})$, where $\gamma(\underline{C})$ is defined by Equation (4.2.4), furthermore $\mathbb{E}((S_N^{\theta_N})^2) = \sigma^2 N + o(N)$, where*

$$\sigma = \sqrt{\sum_{p=1}^{P} \beta_p \frac{y_p^{C_p+1}(C_p(C_p-1)y_p^2 - 2(C_p+1)(C_p-1)y_p + C_p(C_p+1)) - 2y_p^2}{(1-y_p)^2(1-y_p^{C_p+1})}},$$

with, for $1 \leq p \leq P$, $y_p = \rho_p\gamma(\underline{C})$.

Proof. The function $\theta \mapsto \phi_N(\theta) = \mathbb{E}\left[e^{\theta S_N}\right]$ is strictly convex with $\phi_N(0) = 1$ and

$$\begin{aligned}\phi_N'(\theta) = \mathbb{E}\left[S_N^\theta\right] &= N - \sum_p K_p^N \mathbb{E}\left[G_{p,1}^\theta\right] = N\left(1 - \sum_p \frac{K_p^N}{N}\mathbb{E}\left[G_{p,1}^\theta\right]\right) \\ &= N\left(1 - \sum_{p=1}^{P} \beta_p\rho_p e^{-\theta}\frac{C_p\rho_p^{C_p+1}e^{-\theta(C_p+1)} - (C_p+1)\rho_p^{C_p}e^{-\theta C_p} + 1}{(1-\rho_p e^{-\theta})(1-\rho_p^{C_p+1}e^{-\theta(C_p+1)})}\right) + o(\sqrt{N}).\end{aligned}$$

The saturation condition (4.2.2) shows that, for N sufficiently large, $\phi_N'(0) < 0$. Now, since $\phi_N(0) = 1$ and $\lim_{\theta\to+\infty}\phi_N(\theta) = \lim_{\theta\to+\infty} e^{\theta N}\mathbb{E}\left[e^{-\theta\sum_{p,k}G_{p,k}}\right] = +\infty$, there exists a unique θ_N such that $\phi'(\theta_N) = 0$, where the subscript stresses the dependence over N, such that $\phi'(\theta_N) = \mathbb{E}\left[S_N e^{\theta_N S_N}\right] = 0$. Therefore, from (4.2.6) with f the identity, $\mathbb{E}(S_N^{\theta_N}) = 0$, which gives

$$\mathbb{E}(S_N^{\theta_N}) = N - \sum_{p=1}^{P} K_P^N\rho_p e^{-\theta_N}\frac{C_p\rho_p^{C_p+1}e^{-\theta_N(C_p+1)} - (C_P+1)\rho_p^{C_p}e^{-\theta_N C_p} + 1}{(1-\rho_p e^{-\theta_N})(1-\rho_P^{C_p+1}e^{-\theta_N(C_p+1)})} = 0,$$

hence, with the above expansion, one gets

$$1 - \sum_{p=1}^{P} \beta_p\rho_p e^{-\theta_N}\frac{C_p\rho_p^{C_p+1}e^{-\theta_N(C_p+1)} - (C_P+1)\rho_p^{C_p}e^{-\theta_N C_p} + 1}{(1-\rho_p e^{-\theta_N})(1-\rho_P^{C_p+1}e^{-\theta_N(C_p+1)})} = o(\frac{1}{\sqrt{N}}).$$

The previous result shows that, for N sufficiently large, $\exp(-\theta_N)$ is in any arbitrary small neighborhood of $\gamma(C)$. This gives the desired convergence of (θ_N). The other expansion is obtained with direct, straightforward, calculations. □

Proof of Theorem 4.2.2. From the expansions related to the central limit theorem, see Gnedenko and Kolmogorov [20, Theorem 1, page 213], one gets

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{S_N^{\theta_N}}{\sigma\sqrt{N}} \leq x\right) - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-v^2/2}\,\mathrm{d}v \right. \\ \left. - \frac{e^{-x^2/2}}{\sigma\sqrt{2\pi N}}\left(Q(x) + L\left(x\sigma\sqrt{N}\right)\right)\right| = o\left(\frac{1}{\sqrt{N}}\right),$$

where $Q(x) = \alpha(1-x^2)$ for some fixed constant $\alpha > 0$ and L is the periodic function, $L(x) = \lceil x \rceil - x + 1/2$.

By plugging this estimation in the integral of the right hand side of Equation (4.2.8) with $\theta = \theta_N$, one gets that

$$\begin{aligned}
&\int_0^{+\infty} \theta_N e^{-\theta_N u} \mathbb{P}\left(n \leq S_N^{\theta_N} \leq u\right) \mathrm{d}u \\
&\qquad = \frac{1}{\sqrt{2\pi}} \int_a^{+\infty} \int_{n/(\sigma\sqrt{N})}^{u/(\sigma\sqrt{N})} e^{-v^2/2} \,\mathrm{d}v\, \theta_N e^{-\theta_N u} \,\mathrm{d}u + o\left(\frac{1}{\sqrt{N}}\right) \\
&\qquad = \frac{1}{\sigma\sqrt{2\pi N}} \int_a^{+\infty} \int_n^u e^{-v^2/(2\sigma^2 N)} \,\mathrm{d}v\, \theta_N e^{-\theta_N u} \,\mathrm{d}u + o\left(\frac{1}{\sqrt{N}}\right) \\
&\qquad = \frac{1}{\theta_N \sigma\sqrt{2\pi N}} e^{-\theta_N n} + o\left(\frac{1}{\sqrt{N}}\right), \qquad (4.2.9)
\end{aligned}$$

and this estimation is *uniform* with respect to $n \geq 0$.

Now, by using equation (4.2.8) and the previous estimation we obtain

$$\frac{\mathbb{P}\left[S_N \geq n\right]}{\mathbb{P}\left[S_N \geq 0\right]} = \frac{\frac{e^{-\theta_N n}}{\theta_N\sigma\sqrt{2\pi N}} + o\left(\frac{1}{\sqrt{N}}\right)}{\frac{1}{\theta_N\sigma\sqrt{2\pi N}} + o\left(\frac{1}{\sqrt{N}}\right)} = \frac{e^{-\theta_N n} + \frac{o(1/\sqrt{N})}{1/\sqrt{N}}}{1 + \frac{o(1/\sqrt{N})}{1/\sqrt{N}}}$$

which gives, therefore,

$$\lim_{N\to+\infty} \sup_{n\geq 0} \left| \frac{\mathbb{P}(S_N \geq n)}{\mathbb{P}(S_N \geq 0)} - e^{-n\theta_N} \right| = 0$$

and, consequently,

$$\lim_{N\to+\infty} \sup_{n\geq 0} \left| \frac{\mathbb{P}(S_N = n)}{\mathbb{P}(S_N \geq 0)} - e^{-n\theta_N}\left(1 - e^{-\theta_N}\right) \right| = 0.$$

If we plug this uniform asymptotic result into the expression (4.2.5) of the distribution of the number of free ribosomes, given that the normalization constant Z_{R_N} is

$$\sum_k \frac{1}{k!} \mathbb{P}(S_N = k),$$

we obtain

$$\mathbb{P}\left[R_N = n\right] = \frac{\frac{1}{n!}\frac{\mathbb{P}[S_N = n]}{\mathbb{P}[S_N \geq 0]}}{\sum_{k\in\mathbb{N}} \frac{1}{k!}\frac{\mathbb{P}[S_N = k]}{\mathbb{P}[S_N \geq 0]}}.$$

The limit result of Lemma 4.2.3 concludes the proof of theorem. □

The fixed point equation (4.2.4) has a simple intuitive explanation. First one notices that the number of free ribosomes evolves on a very rapid time scale of the order of N, since the number of attached ribosomes and the number of free places for ribosomes on mRNAs are of this order. In fact, because of equation (4.2.2) we have that the total number of available places for ribosomes grows as N and, from Theorem 4.2.2, since the free ribosomes are Poisson distributed, we derive that the total number of attached ribosomes is of this order as well. The number of free ribosomes is therefore quickly at equilibrium, let $\gamma(\underline{C})$ be its average. As an approximation, for a given class (p, k) of proteins, provided it is less than C_p and positive, the number of ribosomes attached

to an mRNA increases by 1 at rate $\gamma(\underline{C})\lambda_p$ and decreases by 1 at rate μ_p. At equilibrium, one gets that the average number of ribosomes attached is

$$\mathbb{E}(X_{p,k}) = \rho_p\gamma(\underline{C})\frac{C_p\rho_p^{C_p+1}\gamma(\underline{C})^{C_p+1} - (C_P+1)\rho_p^{C_p}\gamma(\underline{C})^{C_p}+1}{(1-\rho_p\gamma(\underline{C}))(1-\rho_P^{C_p+1}\gamma(\underline{C})^{C_p+1})},$$

since, by supposing that the free ribosomes have reached equilibrium, the resulting process is a birth-death process and, therefore, the stationary distribution is geometric with parameter $\frac{\lambda_p\gamma(\underline{C})}{\mu_p} = \rho_p\gamma(\underline{C})$, where the space is restricted to $\{0, 1, \ldots, C_p\}$. Therefore, the previous formula corresponds to the average value of a geometric random variable $G_{p,k}^{\gamma}$ of parameter $\rho_p\gamma(\underline{C})$ with values in $\{0, 1, \ldots, C_p\}$.

As consequence, the average total number of attached ribosomes is given by

$$\sum_{p=1}^{P} K_p^N \mathbb{E}(X_{p,k}).$$

This quantity must be of the order of N (few free ribosomes) which gives the desired fixed point equation for $\gamma(\underline{C})$.

We now look at more detailed characteristics of the protein production process, namely the number of ribosomes attached to mRNAs of class (p, k), with $1 \leq p \leq P$ and $1 \leq k \leq K_p^N$, at equilibrium. Note that, since all the proteins with specific concentration have the same parameters, for $1 \leq p \leq P$ at equilibrium the random variables $X_{p,k}$ for any $1 \leq k \leq K_p^N$ have the same distribution. The following theorem gives a limit result for the distribution of these variables as N gets large.

Theorem 4.2.4. *Under the limiting regime with the saturation condition* (4.2.2)*, as N goes to infinity, at equilibrium*

$$\lim_{N\to+\infty}(X_{p,1}^N, 1 \leq p \leq P) = (Y_p, 1 \leq p \leq P)$$

for the convergence in distribution, where Y_p, $1 \leq p \leq P$ are independent random variables and Y_p has a truncated geometric distribution on $\{0, 1, \ldots, C_p\}$ with parameter $\rho_p\gamma(\underline{C})$ where $\gamma(\underline{C})$ is the unique non-negative solution of Equation (4.2.4)*.*

Proof. The notations of the previous proof will be used. From Equation (4.1.2) for the equilibrium distribution, we obtain, for any non-negative function f on $\mathbb{N}^P$,

$$\mathbb{E}\left(f\left(X_{p,1}^N, 1 \leq p \leq P\right)\right) = \frac{1}{Z_N}\sum_{n=0}^{+\infty}\frac{1}{n!}\mathbb{E}\left(f\left(G_{1,1}, \ldots, G_{p,1}\right)\mathbb{1}_{\{r((G_{p,k}))=n\}}\right), \tag{4.2.10}$$

where, for $1 \leq p \leq P$ and $1 \leq k \leq K_p^N$, $G_{p,k}$ is a geometric random variable with parameter ρ_p restricted to the state space $\{0, 1, \ldots, C_p\}$, and Z_N is the normalization constant. If $(m_p) \in \mathbb{N}^P$, we have that

$$\mathbb{P}\left[(G_{p,1}) = (m_p), r((G_{p,k})) = N\right] = \phi_N(\theta_N)\mathbb{E}\left(\mathbb{1}_{\left\{(G_{p,1}^{\theta_N})=(m_p), r\left((G_{p,k}^{\theta_N})\right)=n\right\}}e^{-\theta_N S_N^{\theta_N}}\right),$$

where, as in the proof of Theorem 4.2.2, $(G_{p,k}^{\theta_N})$ are geometric random variables with parameter $\rho_p\exp(-\theta_N)$.

As a consequence we obtain

$$\mathbb{E}\left[\mathbb{1}_{\left\{(G_{p,1}^{\theta_N})=(m_p),r((G_{p,k}^{\theta_N}))=n\right\}}e^{-\theta_N S_N^{\theta_N}}\right] = \mathbb{P}\left((Y_p)=(m_p)\right)\mathbb{E}\left(\mathbb{1}_{\left\{r((\overline{G}_{p,k}))=n+\|m\|\right\}}e^{-\theta_N(\overline{S}_N^{\theta_N}-\|m\|}\right),$$

where $\|m\| = m_1 + \cdots + m_P$ and the bar $\overline{S}$ and $\overline{G}$ is a short notation to specify that all the components with an index $k = 1$ in $S_N^{\theta_N}$ and $(G_{p,k}^{\theta_N})$ are removed (recall that $S_N = N - \sum_{p,k} G_{p,k}^{\theta_N}$). In fact,

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\left\{(G_{p,1}^{\theta_N})=(m_p),r((G_{p,k}^{\theta_N}))=n\right\}}e^{-\theta_N S_N^{\theta_N}}\right] &= \\ &= \mathbb{P}\left[\left(G_{p,k}^{\theta_N}\right)=(m_p)\right]\mathbb{E}\left[\mathbb{1}_{\left\{r((G_{p,k}^{\theta_N}))=n\right\}}e^{-\theta_N S_N^{\theta_N}}\middle|(G_{p,1})=(m_p)\right] \\ &= \mathbb{P}\left[\left(G_{p,k}^{\theta_N}\right)=(m_p)\right]\mathbb{E}\left[\mathbb{1}_{\left\{r((\overline{G}_{p,k}^{\theta_N}))=n+\|m\|\right\}}e^{-\theta_N(\overline{S}_N^{\theta_N}-\|m\|)}\right].\end{aligned}$$

where the last equality comes from the fact that $r\left(\overline{G}_{p,k}^{\theta_N}\right) = n + \|m\|$ and $S_N^{\theta_N} = \overline{S}_N^{\theta_N} - \|m\|$.

From Estimation (4.2.9), one gets that

$$\mathbb{E}\left(\mathbb{1}_{\left\{r((\overline{G}_{p,k}^{\theta_N}))=n+\|m\|\right\}}e^{-\theta_N(\overline{S}_N-\|m\|}\right) = \frac{1}{\theta_N\sigma\sqrt{2\pi\overline{N}}}e^{-\theta_N n}(1-e^{\theta_N}) + o\left(\frac{1}{\sqrt{N}}\right), \qquad (4.2.11)$$

holds uniformly in n and m.

By plugging relation (4.2.11) in equation (4.2.10) with $f(x) = \mathbb{1}_{\{(x_p)=(m_p)\}}$, we obtain

$$\lim_{N\to+\infty}\mathbb{P}((X_{p,1}^N)=(m_p)) = \mathbb{P}\left((Y_p)=(m_p)\right). \qquad (4.2.12)$$

The theorem is proved. □

The above theorem is actually a mean-field result. In the limit any finite subset of components of $(X_{p,k}^N)$ converges in distribution to independent random variables.

4.3 Analysis of fixed point equation

We focus now on the possible biological consequences of the Theorem 4.2.4. The key parameter $\gamma(\underline{C})$ is the solution of the fixed point equation (4.2.4), which is characterized by complicated expressions involving vectors of the loads (ρ_p) and of the capacities (C_p). A simple and tight approximation of $\gamma(\underline{C})$ is now presented. In particular, it is shown that, for a wide range of values of the ρ_p's, the solution $\gamma(\underline{C})$ actually depends on the vector of capacities $\underline{C} = (C_p)$ in a simple way .

The main ingredient of the estimations for $\gamma(\underline{C})$ is the simple following fact. If G is a random variable whose distribution is a truncated geometric distribution with parameter ρ on $\{0, 1, \ldots, C\}$, if $\rho < 1$ then, provided that C is sufficiently large, G is close in distribution to a plain geometric distribution with parameter ρ and, in particular, its expected value is close to $\rho/(1-\rho)$. In the case $\rho > 1$, then G can be expressed as $G = C - \tilde{G}$, where $\tilde{G}$ is a truncated geometric distribution with parameter $1/\rho < 1$. Explicit bounds on the accuracy of the approximations of $\gamma(\underline{C})$ are also provided.

4.3.1 The underloaded case

In this section, without loss of generality, we assume that $\rho_p < 1$, for all $1 \leq p \leq P$. The main result of the following Lemma is the fact that the quantity $\gamma(\underline{C})$ is independent of $\underline{C}$.

Lemma 4.3.1. *Under the conditions $\rho_p < 1$, for all $1 \leq p \leq P$, and the saturation condition (4.2.2) if γ_0 is the unique solution of the equation*

$$\sum_{p=1}^{P} \beta_p \frac{\rho_p \gamma_0}{1 - \rho_p \gamma_0} = 1, \tag{4.3.1}$$

then the solution $\gamma(\underline{C})$ of the fixed point Equation (4.2.4) is such that $\gamma_0 \leq \gamma(\underline{C})$ and

$$|\gamma(\underline{C}) - \gamma_0| \leq \frac{1}{\sum_{p=1}^{P} \beta_p \rho_p} \sum_{p=1}^{P} \beta_p \frac{(C_p + 1)\rho_p^{C_p+1}}{1 - \rho_p^{C_p}} \tag{4.3.2}$$

with $\underline{C} = (C_p)$.

Proof. For $x \geq 0$, define the implicit function $\phi(x)$ as the unique solution of the equation

$$\sum_{p=1}^{P} \frac{\beta_p \rho_p \phi(x)}{1 - \phi(x)\rho_p} = x, \tag{4.3.3}$$

then γ_0 is simply $\phi(1)$. Note that, since we restrict to $x \geq 0$, then $\phi(x) \geq 0$. In fact, if for some $\tilde{x} \geq 0$ we had $\phi(\tilde{x}) < 0$, then we would obtain $0 > \sum_{p=1}^{P} \frac{\beta_p \rho_p \phi(\tilde{x})}{1 - \phi(\tilde{x})\rho_p} = \tilde{x} \geq 0$, which proves our statement. Moreover, we have that $\phi(x) < 1/\rho_p$ for all $1 \leq p \leq P$, since, as long as $\phi(x)$ approaches $1/\rho_p$ for any p, then the left had side of (4.3.3) diverges to $+\infty$. In summary, for any $x \geq 0$, we have a unique solution $\phi(x)$ of equation (4.3.3), such that $0 \leq \phi(x) \leq 1/\rho_p$

It is easy to show that, by Implicit Function Theorem, the function ϕ is differentiable and

$$\phi'(x) \sum_{p=1}^{P} \frac{\beta_p \rho_p}{(1 - \phi(x)\rho_p)^2} = 1. \tag{4.3.4}$$

We have, in particular, that

$$0 \leq \phi'(x) \leq 1/(\beta_1 \rho_1 + \beta_2 \rho_2 + \cdots + \beta_P \rho_P).$$

In fact, since $\phi(x) \geq 0$, then $1 - \phi(x)\rho_p \leq 1$ and, therefore $\phi'(x) \sum_p \beta_p \rho_p \leq 1$. The other inequality is straightforward from the formula (4.3.4).

For $0 \leq y < 1$ and $c > 0$, the relation

$$\frac{cy^{c+1} - (c+1)y^c + 1}{1 - y^{c+1}} = 1 - (c+1)y^c \frac{1 - y}{1 - y^{c+1}} \tag{4.3.5}$$

holds.

Since (4.2.2) holds, then the existence of $\gamma(\underline{C})$ is assured. Equation (4.2.4) can then be rewritten as

$$\gamma(\underline{C}) = \phi\left(1 + \sum_{p=1}^{P} \beta_p (C_p + 1) \frac{(\rho_p \gamma(\underline{C}))^{C_p+1}}{1 - (\rho_p \gamma(\underline{C}))^{C_p+1}}\right),$$

by using relation (4.3.5) together with the definition (4.3.3). This proves, in particular, that $\gamma_0 \leq \gamma(\underline{C})$. Since $\gamma_0 = \phi(1)$ and the previous relation, we obtain by the mean value theorem

$$|\gamma(\underline{C}) - \gamma_0| \leq \sum_{p=1}^{P} \beta_p (C_p + 1) \frac{(\rho_p \gamma(\underline{C}))^{C_p+1}}{1 - (\rho_p \gamma(\underline{C}))^{C_p+1}} \Bigg/ \sum_{p=1}^{P} \beta_p \rho_p$$

and therefore Relation (4.3.2), since $0 \leq \phi(x) \leq 1/\rho_p$, for any $x \geq 0$ and for all $1 \leq p \leq P$. □

The above lemma states that if the right hand side of Relation (4.3.2) is sufficiently small then the simple constant γ_0 is an accurate estimation of the parameter $\gamma(\underline{C})$. Since the ρ_p's are < 1, this approximation will hold in the case where the values of the C_p's are not too small which is a reasonable biological assumption. One gets therefore the following proposition.

Proposition 4.3.2 (System with only Underloaded Classes of mRNAs)**.** *Under the conditions $\rho_p < 1$, for all $1 \leq p \leq P$, and the saturation condition* (4.2.2)*, then, under the limiting regime* (4.2.1)*, at equilibrium as N goes to infinity,*

1. *the distribution of the number of free ribosomes is Poisson with parameter $\gamma_0 + o(h(\underline{C}))$,*
2. *the distribution of the number of ribosomes attached to an mRNA of class (p, k) is geometric with parameter $\rho_p \gamma_0 + o(h(\underline{C}))$.*

where γ_0 is the unique solution of the equation

$$\sum_{p=1}^{P} \beta_p \frac{\rho_p \gamma_0}{1 - \rho_p \gamma_0} = 1$$

and

$$h(\underline{C}) = \sum_{p=1}^{P} \beta_p \frac{(C_p + 1)\rho_p^{C_p+1}}{1 - \rho_p^{C_p+1}}.$$

As a consequence, under the assumptions of the above proposition, the production rate of proteins of class (p, k) is given by

$$\frac{\lambda_p \gamma_0}{1 - \rho_p \gamma_0} + o(h(\underline{C})).$$

4.3.2 The Case of Overloaded mRNAs

The case where at least one of the ρ_p is strictly greater than 1 is now investigated. Without loss of generality, we assume the ordering $\rho_1 \leq \rho_2 \leq \cdots \leq \rho_P$ and we suppose that $\rho_P > 1$ holds. Let

$$L = \sup\{p \geq 1 : \rho_p < 1\},$$

with the convention that $\sup\{\emptyset\} = 0$. For simplicity, it is assumed that none of ρ_p is equal to 1.

With similar arguments then in Section 4.3.1 concerning the truncated geometric random variables which describes the system at equilibrium, if $\rho_p < 1$, then the truncation of the geometric random variable G at C_p has little impact in the analytic expressions with respect to a plain geometric random variable. If conversely $\rho_p > 1$, then we may express the random variable G as $C_p - \tilde{G}$, where $\tilde{G}$ has a truncated geometric distribution with parameter $1/\rho_p$.

The saturation condition can now be written as

$$\sum_{p=1}^{L} \beta_p \frac{\rho_p}{1-\rho_p} + \sum_{p=L+1}^{P} \beta_p \left(C_p - \frac{1}{\rho_p - 1} \right) + \phi(\underline{C}) > 1, \tag{4.3.6}$$

where

$$\phi(\underline{C}) = -\sum_{p=1}^{L} \beta_p \frac{(C_p+1)\rho_p^{C_p+1}}{1-\rho_p^{C_p+1}} + \sum_{p=L+1}^{P} \beta_p \frac{C_p+1}{\rho_p^{C_p+1}-1}.$$

The following proposition gives a simplified version of the results of Section 4.2 in this case.

Proposition 4.3.3 (System with Overloaded Classes of mRNAs)**.** *If, for some* $1 \leq L < P$*, one has* $\rho_1 \leq \rho_2 \leq \cdots \rho_L < 1 < \rho_{L+1} \leq \rho_P$*, then if*

$$\sum_{p=1}^{L} \beta_p \frac{\rho_p}{1-\rho_p} + \sum_{p=L+1}^{P} \beta_p \left(C_p - \frac{1}{\rho_p - 1} \right) > 1 \tag{4.3.7}$$

then, under the limiting regime (4.2.1)*, at equilibrium as* N *goes to infinity,*

1. *the distribution of the number of free ribosomes is Poisson with parameter* $\gamma_0 + o(h(\underline{C}))$*.*
2. *For* $1 \leq p \leq P$ *and* $k \geq 1$ *the number* $X_{p,k}$ *of ribosomes attached to an mRNA of class* (p,k) *is such that*
   - *if* $p \leq L$
     $X_{p,k}$ *is a geometric r.v. with parameter* $\rho_p\gamma_0 + o(h(\underline{C}))$*;*
   - *if* $L+1 \leq p \leq P$
     $C_p - X_{p,k}$ *is a geometric r.v. with parameter* $\gamma_0/\rho_p + o(h(\underline{C}))$*,*

where $1/\rho_{L+1} \leq \gamma_0 \leq 1$ *is the unique solution of the equation*

$$\sum_{p=1}^{L} \beta_p \frac{\rho_p\gamma_0}{1-\rho_p\gamma_0} + \sum_{p=L+1}^{P} \beta_p \left(C_p - \frac{1}{\rho_p\gamma_0 - 1} \right) = 1 \tag{4.3.8}$$

and

$$h(\underline{C}) = \sum_{p=1}^{L} \beta_p \frac{(C_p+1)\rho_p^{C_p+1}}{1-\rho_p^{C_p+1}} + \sum_{p=L+1}^{P} \beta_p \frac{C_p+1}{\rho_p^{C_p+1}-1}$$

Proof. The existence of a solution $1/\rho_p < \gamma_0 < 1$ to Equation (4.3.8) is a consequence of Condition (4.3.7) and therefore that $\gamma_0\rho_p > 1$ for $L+1 \leq p \leq P$. The rest of the proof follows the same arguments as in the proof of the above lemma. □

Coming back to the proof of Theorem (4.2.2) where a change of probability was used, for $1 \leq p \leq P$, a geometric random variable with parameter ρ_p is transformed with a reduced parameter $\gamma(\underline{C})\rho_p$, it should be noted that this change of distribution does not affect the set of overloaded classes under the assumptions of the above proposition, i.e. the $1 \leq p \leq P$ such that $\gamma(\underline{C})\rho_p > 1$.

Appendix A

Mathematical tools

A.0.3 Marked Poisson Point Processes

The main results concerning marked Poisson point processes (MPPP) are briefly recalled in the present section. For more details we refer to Kingman [38] and Chapter 1 of Robert [65]. Throughout this section H is the space $\mathbb{R}^d$ for some $d \geq 1$.

Definition A.0.4. *If $\lambda > 0$, $\mu(\mathrm{d}x)$ is a probability distribution on H, a marked Poisson process on $\mathbb{R}_+ \times H$ with intensity $\lambda\,\mathrm{d}u \otimes \mu(\mathrm{d}x)$ is a sequence $\mathcal{N}_\lambda = (t_n, X_n)$ of elements of $\mathbb{R}_+ \times H$ where*

- *(t_n) is a (classical) Poisson process on $\mathbb{R}_+$ with rate λ.*
- *(X_n) is an i.i.d. sequence with values in H and whose distribution is $\mu(\mathrm{d}x)$.*

The sequence $\mathcal{N}_\lambda$ can also be seen as a marked point process on $\mathbb{R}_+ \times H$, i.e. if $f : \mathbb{R}_+ \times H \to \mathbb{R}_+$ is a continuous function then

$$\mathcal{N}_\lambda(f) = \int_{\mathbb{R}_+ \times H} f(u,x)\mathcal{N}_\lambda(\mathrm{d}u, \mathrm{d}x) = \sum_{n \geq 1} f(t_n, X_n).$$

In other words $\mathcal{N}_\lambda$ can also be seen as a sum of Dirac masses at the points (t_n, X_n).

Laplace's transform of MPPP

The following important proposition characterizes marked Poisson point processes.

Proposition A.0.5. *The point process $\mathcal{N}_\lambda = (t_n, X_n)$ is a marked Poisson point process with intensity $\lambda\,\mathrm{d}u \otimes \mu(\mathrm{d}x)$ if and only if the relation*

$$\mathbb{E}\left[\exp\left(-\mathcal{N}_\lambda(f)\right)\right] = \exp\left(-\lambda \int_0^{+\infty} \left(1 - e^{-f(u,x)}\right)\,\mathrm{d}u\,\mu(\mathrm{d}x)\right) \tag{A.0.1}$$

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

The left-hand-side of Equation (A.0.1) is usually defined as the Laplace transform of $\mathcal{N}_\lambda$ at f. This quantity determines completely the distribution of any marked point process.

For $\xi > 0$, by replacing f by ξf in Relation (A.0.1), one gets an expression for

$$\mathbb{E}\left[\exp\left(-\xi\mathcal{N}_\lambda(f)\right)\right]. \tag{A.0.2}$$

If one differentiates formula (A.0.2) with respect to ξ and sets $\xi = 0$, from (A.0.1) we obtain

$$\mathbb{E}\left[\mathcal{N}_\lambda(f)\right] = \mathbb{E}\left[\int_{\mathbb{R}_+\times H} f(u,x)\mathcal{N}_\lambda(\mathrm{d}u,\mathrm{d}x)\right] = \lambda \int_{\mathbb{R}_+\times H} f(u,x)\ \mathrm{d}u\,\mu(\mathrm{d}x). \tag{A.0.3}$$

MPPP and Martingales

Be $\mathcal{N}_\lambda$ a Poisson marked point process, *i.e.* $\mathcal{N}_\lambda$ is a Poisson point process on $\mathbb{R}\times H$ with intensity measure $\lambda\,\mathrm{d}x \otimes \nu(\mathrm{d}y)$, where $\nu(\mathrm{d}y)$ is a probability measure on H, locally compact metric space. Then we denote with $\mathcal{F}_t$ the σ-field generated by the random variables $\mathcal{A}\times\mathcal{U}$, where A is a Borelian subset of $(\infty, t]$ and U is a Borelian subset of H, *i.e.*

$$\mathcal{F}_t = \sigma\left(\mathcal{N}_\lambda((\infty,t]\times U) : a,b \le t, U \in \mathcal{B}(H).\right)$$

We have the following result, see Robert [65, Proposition A.9, page 353].

Theorem A.0.6. *If $X(t,x)$ is a* cádlág *process adapted to filtration $\mathcal{F}_t$, measurable and square integrable with respect to to $\nu(\mathrm{d}x)$, then the process*

$$\mathcal{G}(t) = \iint\limits_{(0,t]\times H} X(s-,y)\left[\mathcal{N}_\lambda(\mathrm{d}s,\mathrm{d}y) - \lambda\,\mathrm{d}s\nu(\mathrm{d}y)\right] \tag{A.0.4}$$

is a square integrable martingale, whose increasing process is

$$I(t) = \iint\limits_{(0,t]\times H} X^2(s,y)\lambda\,\mathrm{d}s\nu(\mathrm{d}y). \tag{A.0.5}$$

Appendix B

Biology

B.1 Biological Mechanisms

B.1.1 Gene activation

Gene activation is the process which allows a gene to be expressed at a specific time. The way this activation may occur varies a lot from gene to gene and from organism to organism. The main mechanisms causing gene activation are the dissociation of a repressor, the association of an activator and the chromatin remodeling.

Repressor

A *repressor* is a DNA-binding protein that regulates the expression of a specific gene by binding the operator, which is a segment of DNA that a regulator binds to. The binding of the repressor blocks the attachment of RNA polymerase to the *promoter* and prevents the transcription of the genes. If an *inducer* molecule is present, it can interact with the repressor and inhibit its action by detaching it or preventing its binding to the operator.

Activator

An *activator* is a DNA-binding protein that regulates one or more genes by increasing the transcription rate. RNA polymerase binds to the *promoter region* of the gene, forming a complex which sometimes proceeds to gene transcription. An activator recruits the RNA polymerase to its promoter region.

Chromatin remodeling

The *chromatin remodeling* is another way to regulate the expression of a specific gene in eukaryotes. Chromatin is the complex of DNA and histone proteins with which it associates. Histones are highly alkaline proteins found in eukaryotic cell nuclei that package and order the DNA into structural units called *nucleosomes*. Chromatin on one side serves as a way to condense DNA within the cellular nucleus and, on the other side, as a control of gene expression. In eukaryotes specific genes are not expressed if they are not accessible by RNA polymerases and transcription factors. Therefore, chromatin must open to let transcription to take place. More in detail, when the surrounding chromatin is in an acetylated state, i.e. it is open, the gene can be transcribed,

while the transcription is repressed when the chromatin is in a condensed state. The process of "opening" the chromatin complex is called chromatin remodeling.

At nowadays there are no direct proofs of the role of chromatin remodeling to stochastic activation/deactivation of gene expression in eukaryotes. Nevertheless, a number of studies have tried to find such a connection by studying different effects by indirect means. Becskei et al. [2005] [4] and Raj et al. [2006] [61] have studied the correlations between proximally located genes, focusing therefore on positional effects. Raser and O'Shea [2004] [62] and of Xu et al. [2006] [73] have altered, on the other side, the behavior of chromatin-remodeling agents and then analyzed the effects. Raser and O'Shea [62] hypothesize that chromatin remodeling is the key regulation mechanism for certain eukaryotic promoters. In particular, the noise strength profile of promoter *PHO5* of *Saccharomyces cerevisiae* is approximated by the noise profile of a three-stage model, introduced in Section 2.A.2. In this specific case gene activation is an infrequent process with respect to transcription and the active promoter is stable. The *PHO5* is then supposed to depend on a slow promoter transition. In the inactive state the *PHO5* promoter exhibits positioned nucleosomes. Chromatin complexes contribute then to the nucleosome removal, causing gene activation.

The *DNA methylation* is another way to control transcription. In this process, *cytosine bases* of eukaryotic DNA is converted in 5-*methylcytosine*, resulting in the repression of the transcription.

We have here tried to show that the control of gene expression is complex and it results to be much more complicated when looking more in detail the mechanisms. The details vary between different genes and different species. Genes may show different states, but in first approximation we will suppose that each gene may show only two possible states: active or inactive.

The number of copies of a gene within bacteria is a fundamental factor and should be considered in a model describing the expression of a specific gene. When bacteria are growing, they duplicate their DNA, that leads to a number of at least two copies per gene, since the genetic information has to be split between daughter cells.

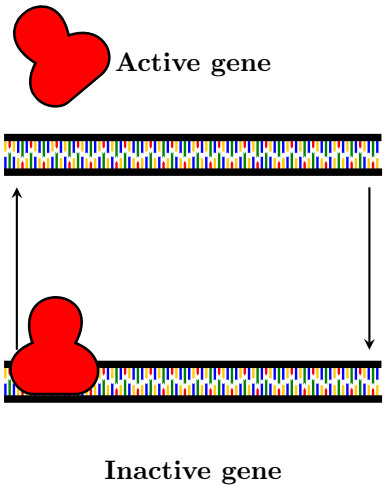


Figure B.1: Gene activation. Schema of gene activation/deactivation. When the *repressor*, red component, binds to the gene it inhibits the mRNA transcription, while the gene is activated when the *repressor* unbinds.

Nevertheless bacteria like *Escherichia coli* are characterized by cell division cycles of 20 minutes, while the DNA replication requires about 40 minutes. So those bacteria are obliged to have more than two copies of DNA and have parallelized the tasks in order to increase the number of genes copies per cell.

So in case of "normal" growth, only one copy of a gene is present, which is doubled in about 40 minutes. Moreover, since the DNA replication always starts at same point, the *origin of replication* is easier to find this part of DNA with respect to the *terminating region of replication*. In "normal" growth regime the division time is smaller then one hour and we have regions with one, two or four copies of genes.

In "regeneration" regime the cell division cycle takes about 20 minutes and we have up to eight copies of genes that are localized closer to the origin of replication.

B.1.2 Transcription

The first step of gene expression is the *transcription* of the information of a gene into RNA. RNA polymerase is one of the central processes in transcription process. Bacteria contain a single type of polymerases, while eukaryotes contain three dinstinct types. Nevertheless, transcriptional mechanisms are quite similar in both species.

Transcription begins with the binding of the RNA polymerase complex to the promoter, which is a particular DNA sequence at the beginning of the gene. Activation of the RNA polymerase complex enables transcription initiation.

In prokaryotes transcription takes place in cytoplasm. Transcription consists of three phases: initiation, elongation and termination. Initiation takes place at promoter, see Section B.2.7 for details, immediately upstream the gene is going to be transcribed. DNA polymerase is made by five subunits: 2α, β, β' and σ. The σ subunit recognizes the "consensus sequence" in the gene promoter, it binds to the DNA and recruits other components of the enzymatic complex. Once the polymerase is bound to DNA, the σ-subunit detaches and β' is responsible to the binding with the DNA. The polymerase unrolls the DNA double-helix and elongation phase starts. The transcription ends at terminating codon, a specific sequence downstream the transcribed gene. The mRNA messenger is then released in the cytoplasm due to termination protein *Rho*. Since transcription and translation both take place in cytoplasm, it may happens that translation starts before completion of transcription.

The transcription process can be described through the following fundamental steps:

1. **initiation:** the polymerase binds to one of the specificity factors σ to form a "holoenzyme" in order to attach to a specific promoter in the DNA. The more similar is a sequence to a "consensus sequence" the stronger is the binding to the DNA. After the first bond is synthesized, the RNA polymerase must clear the promoter (this phase is called *promoter clearance*). During this time it may occurs that a truncated transcript, called *abortive initiation*, is released;

2. **elongation:** after the promoter clearance, the polymerase assembles in a controlled fashion the mRNA chain;

3. **termination:** two kind of termination processes exist: the *ρ-independent transcription termination* or the *ρ-dependent transcription termination*. The first involves terminator sequences within the RNA that signals the RNA polymerase to stop. The latter uses the ρ terminator factor to stop RNA synthesis.

In eukaryotes transcription takes place in a more complex manner. In particular, there are involved three types of RNA polymerases: type I, type II and type III. These polymerases differ in the number and type of sub-units they are made of, as well as the class of RNAs they transcribe. Type I transcribes ribosomal RNAs (*rRNA*), type II transcribes messenger RNAs (*mRNAs*) and most of the small nuclear RNAs, type III transcribes transfer RNAs (*tRNAs*) and other small RNAs. Since RNA polymerase II transcribes protein-encoding genes, it has been deeply studied and many experiments and models on stochastic gene expression in eukaryotes study this type of polymerase.

The transcription regulation controls the frequency and the number of produced messengers. The gene transcription is subject to many control mechanisms; we just recall the most common. The *specificity factors* alter the specificity of RNA polymerase for a given promoter or set of promoters, making it more or less likely to bind to them, *i.e.* sigma factors in prokaryotic transcription. *Repressors* bind to non-coding sequences on the DNA strand that are close to or overlapping the promoter region, impeding RNA polymerase's progress along the strand, thus

impeding the expression of the gene. *General transcription factors* position RNA polymerase at the start of a protein-coding sequence and then release the polymerase to transcribe the mRNA. *Activators* encourage the expression of the gene by enhancing the interaction between RNA polymerase and a particular promoter. More in detail, they do this by increasing the attraction of RNA polymerase for the promoter through interactions with subunits of the RNA polymerase or, indirectly, by changing the structure of the DNA. Finally, *enhancers* are sites on the DNA helix that are bound to by activators in order to loop the DNA bringing a specific promoter to the initiation complex. Enhancers are much more common in eukaryotes than prokaryotes.

In post-transcriptional phase the regulatory machine controls the number of mRNAs that are translated into proteins. The stability and distribution of the different transcripts are regulated (*post-transcriptional regulation*) by means of RNA binding protein (RBP) that controls the various steps and rates of the transcripts. These proteins achieve these events thanks to a RNA recognition motif (RRM) that binds a specific sequence or secondary structure of the transcripts.

The typical hypothesis in stochastic models for gene expression is to consider transcription exponentially distributed. This can be linked to the fact that, in order to produce a new messenger, the polymerase must be attached to the gene. The encounter of particles is usually modeled as an exponentially distributed process. Nevertheless, this modelisation does not take into account the fact that, once polymerase is attached, several processes have to occur in order to effectively start the elongation of the new mRNA. In the elongation step the polymerases combine the basis reading the information contained into the gene, until it reaches the terminator codon and a new mRNA is released into the cytoplasm. We just point out that we have given a coarse-grained description of the transcription process not taking into account possible control mechanisms of the process itself and the other cellular components or other external factors which may play a role.

Using our approximate description it is clear that the exponential hypothesis is a brute simplification of the real situation. Nevertheless, this assumption may be justified in specific cases and allows to make computations. In fact, it has been observed in prokaryotes that ribosomes attache to the growing chain of mRNA, so the exponential hypothesis is equivalent of supposing that the ribosome may attach to the messenger once transcription is just started. Moreover, mRNAs dynamics are much faster with respect to the protein dynamics. This may lead to consider finer modelisation for the proteins and to accept a compromise of simplicity for the mRNAs dynamics.

On the other side, this assumption may not be appropriate in all cases and for all purposes. In fact, the binding of RNA polymerase may change the native structure of the gene, blocking or facilitating transcription. In the same manner the DNA replication can change the chromatin structure, which can remove both activators and repressors thus changing the expression rate of certain genes. In eukaryotes transcription and translation take place in different parts of the cell, the nucleus and cytoplasm respectively. Messengers have to be transported out the nuclei in order to be translated, which may affect the dynamics.

B.1.3 Translation

Translation consists of the following steps:

1. **initiation:** it involves the assemblage of components ribosomal subunits (50S and 30S), mRNA, the first aminoacyl tRNA, GTP and initiation factors (IF1, IF2, IF3). The ribosome has three sites (A, P and E). The A site is the entry-point for aminoacyl tRNA, except for the first that binds directly on the P site. In the P site the peptidyl tRNA is formed and in the E site the uncharged tRNA detaches from the ribosome;

2. **elongation:** it is a controlled process in which the polypeptide chain is elongated with the addition of amino acids to the carboxyl end of the growing end. Elongation involves several elongation factors, a conformal change, bond formations, etc. The aminoacyl tRNA attaches in the A site, then moves to the P site where the polypeptide is attached to the growing chain and the uncharged tRNA is moved to the E site where exits from the complex;

3. **termination:** it occurs when one of three terminating codons moves to the A site. These codons are not recognized by any tRNA but by the so called release factors. These factors trigger hydrolysis of the ester bond and release the newly produced protein in the cytoplasm. The ribosome recycling step is responsible of ribosome disassembly in such a way to be ready to start translation of other messengers.

Translation is carried out by more than one ribosome simultaneously. Ribosomes are rather closely spaced on mRNAs, with their average distance not exceeding a few ribosomes diameters, see [28]. This has been confirmed by measurement on individual *E. Coli* genes such as *lacZ* [35]. Small ribosome spacing means a high frequency translation initiation (FTI), i.e. hence the majority of bacterial mRNAs seems endowed with FTIs that fall only slightly short of the maximum that would result in ribosome queueing.

B.1.4 mRNA degradation

The process of messenger degradation is an essential function for recycling nucleotides and for regulating the level of gene expression and is performed by *RNase*. The decay process occurs in short time scales, i.e. the typical half-life of a messenger is of about two minutes at 37°C in most cases.

The instability of mRNAs was first discovered in the early 60's in *E. Coli* and set this type of RNA apart from the previously known as stable RNA. It was then assumed a conservation of mRNA decay mechanism among different prokaryotes, but recent works [3] have showed a specificity in different prokaryotes. In *E. Coli* mRNA turnover, the major ribonucleases known to participate are *RNase E* and several $3' \to 5'$ *exoribonucleases* (the hydrolytic enzymes *RNase II* and *RNase R* and the phosphorolytic enzyme *PNPase*). In *B. Subtilis* the ribonucleases *RNase* J_1 and *RNase Y* seem to play a major role in the initiation of mRNA decay, however much has to be understood to clarify the primary role of these ribonucleases.

In many bacterial organisms, the decay mechanism is unilateral, as for instance in the *lacZ* gene, which decays in a net $5'$ to $3'$ direction. This has been proven in other organisms thanks to the observation of lag. More precisely, it has been observed that the degradation started with few minutes lag, corresponding to the time needed to complete transcription of the specific gene. This evidence brings to conclude that the degradation follows a specific direction.

B.1.5 Protein degradation

Protein degradation is regulated and selective process within cell: proteins are continuously degraded to amino acids, with protein-specific rates. The degradation machinery differs between eukaryotes and prokaryotes and we try to give few key insight on this complex process, using the review article of Goldberg [21] as reference.

Misfolded proteins can originate by mutations or post-synthetic events and can be highly toxic for cells. Abnormal proteins may result also from errors in transcription or translation. The accumulation of abnormal proteins has been shown to cause *JNK* kinases and *apoptosis*. It has also been hypothesized that the ability of unfolded proteins to activate cell-death program could be associated with neurodegenerative diseases and with certain cancers.

Prokaryotes have developed an elaborate proteolytic machinery to quickly destroy misfolded proteins. This machinery is way more complex than the *proteases*, which is an enzyme that conducts proteolysis. In fact, if such enzymes were allowed into the cytosol, "they would quickly convert the cell into a bag of amino acids" [21]. Evolution had to solve the problem to provide a machinery capable of degrade damaged or misfolded proteins without destroying essential constituents.

The ability of cells to degrade misfolded proteins was found in early 70s, which has allowed to understand that protein structure determines not only its specific cellular function, but also its intracellular stability. A conformal change in *hemoglobin* leads their lifespan to decrease from about 110 days to an half-life of about 10 minutes.

In 80s it was found that bacteria were able to degrade external proteins which are considered abnormal, explaining why medical proteins such as *insulin* cannot accumulate in *E. Coli*.

Incomplete proteins, which can arise for mRNA mistakes or mutations, are quickly degraded. Bacteria are endowed with a mechanism which is capable to target for degradation incomplete polypeptide while they are still on stalled ribosomes.

The folding of newly synthesized proteins is a key process, since the protein conformation is crucial for it to accomplish its function. This process can take several minutes and the failures of folding are not negligible: in fact, about 30% of proteins in eukaryotes might undergo degradation, possibly for failures in the folding process or of multimer assembly. However, it is difficult to identify anomalous proteins, because of the short lifetime of some regulatory proteins. Degradation has been hypothesize to have also a control mechanism, in order to assure that multimeric complexes are present in the proper stoichiometry. In fact, it has been observed that free ribosomal subunit are really unstable, unless they are assembled in ribosomes.

Mature proteins are exposed to the risk of denaturation or chemical damage, which can be caused by different means in a cell, like high temperature, high salt concentrations, other unfolded proteins, etc. Most part of proteins show long lifetimes: in mammals about $1 - 2\%$ of these proteins are degraded each hour.

The *ubiquitin-proteasome* pathway is one of the most important selective degradation machinery for abnormal proteins in eukaryotes. *Ubiquitin* is a tiny protein that have earned the nickname of "the kiss of death", since it tags the protein to be degraded, which occurs by means of the 26S *proteasome*. All this machinery requires ATP, which is used to activate ubiquitin. It is still unclear whether the ubiquitin tag is essential for the proteasome degradation for all abnormal proteins, since this process occurs efficiently in bacteria, which lack ubiquitin and eukaryotic 26S proteasomes are able degrade unfolded proteins *in vitro*, without the ubiquitin bond.

In eukaryotes the 26S proteasome is the only enzyme used to degrade various types of substrate. This enormous protein contains a central cavity where the lysis occurs, ensuring protein digestion to be isolated from the cytosol and other cellular components. The entry channel of the 26s is narrow, in order to exclude normal globular cells to be degraded. Moreover an ATP-dependent mechanism is used to recognize and linearize proteins inside the enzyme in order to enter in the 20S subunit and be degraded. Proteolysis in prokaryotes involves several proteases. Nevertheless the proteasome of *Archea* works together other cellular compounds which select, unfold and translocate abnormal proteins. We may think that the case of eukaryotes is the result emerged by evolution process.

Stable and unstable proteins

Proteins are known to degrade rapidly when their conformations are altered by mutations or denaturation, ... since all these modifications are likely to hinder folding or to disrupt tertiary structures, making these proteins predisposed to be degraded.

However, normal cells could show a high variability in their half-lives and several hypothesis have been proposed, see [**?**] for further details. In particular, the studies of stable and unstable proteins has led to find common characteristics or sub-structures that are supposed to be responsible of the instability of the associated polypeptides. In the manuscript we refer to *stable* proteins those proteins whose lifetime is comparable (or bigger) to cell doubling time, while we refer to *unstable* proteins to those who are highly likely to be degraded rapidly.

B.2 Biological glossary

Here we give short descriptions of few biological components and mechanisms whose we refer to during the manuscript. Our aim is not to be exhaustive but to facilitate the reading.

B.2.1 16S ribosomal RNA

16S ribosomal RNA (16S rRNA) is a component of the 30S subunit of prokaryotic ribosomes. Multiple sequences of 16S rRNA can exist within a single bacterium. The 16S rRNA has several functions

- it has a structural role, acting as a scaffold defining the positions of the ribosomal protein;
- the 3' end contains the anti-Shine-Dalgarno sequence (see B.2.14), which binds upstream to the AUG start codon on the mRNA;
- it interacts with 23S, aiding in the binding of the two ribosomal subunits (50S and 30S);
- it stabilizes the correct codon-anticodon pairing in the A-site.

B.2.2 β-galactosidase

β-galactosidase, or *beta-gal* or *β-gal*, is the enzyme resulting from the expression of *lacZ* gene, which is under the control of the *lac* promoter. The *β-galactosidase* serves as *hydrolase enzime* and catalyses the hydrolisis of *β-galactosides* into *monosaccharides*. Moreover, *β-gal* converts *lactose* into *allolactose*, by *transglycosylation*. The *allolactose* is an inducer of the *lac operon* in *E. Coli*.

β-galactosidase has been the standard reporter for gene expression in both prokaryotes and eukaryotes [5], [8], [50]. A single molecule of *beta-gal* can produce a large number of fluorescent product molecules, leading to amplification of the fluorescent signal.

B.2.3 DNA

The *deoxyribonucleic acid* (DNA) is a nucleic acid that contains the genetic instructions used in development and functioning of all known living organisms, with except of some viruses. The DNA segments that carry this genetic information are called *genes*, but other DNA sequences have structural or regulatory purposes.

It consists of two long polymers of *nucleotides*, with backbones made of sugars and phosphate groups joined by ester bonds. Attached to each sugar is one of four types of molecules, called *bases*: the sequence of these bases encodes the DNA information.

DNA is organized into long structures called *chromosomes*. The chromosomes are duplicated before cells divide in a process called DNA replication. Prokaryotes organisms store their DNA in the cytoplasm, while in eukaryotic it is stored in cell nucleus and in organelles.

DNA usually exists as a pair of molecules that are held tightly together and have the shape of a double helix. The nucleotide repeats contain both the segment of the backbone of the molecule and a base which interacts with the other DNA strand in the helix. A base linked to a sugar is called *nucleoside*, while a base linked to a sugar and one, or more, phosphate group is called a *nucleotide*.

The backbone of the DNA is made of alternating phosphate and sugar (2-deoxyribose i.e. a 5 carbon sugar) residues. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds give to the DNA strand a "direction". In the double helix the direction of nucleotides in one strand is opposite to their direction in the other strand i.e. the strands are *antiparallel*. The asymmetric ends of DNA strands are called *5'* "five prime", which has a terminal phosphate group, and *3'* "three prime", which has a terminal hydroxyl group.

The four bases found in DNA are: *adenine* (A), *cytosine* (C), *guanine* (G) and *thymine* (T). These bases are attached to the sugar/phosphate to form a complete nucleotide.

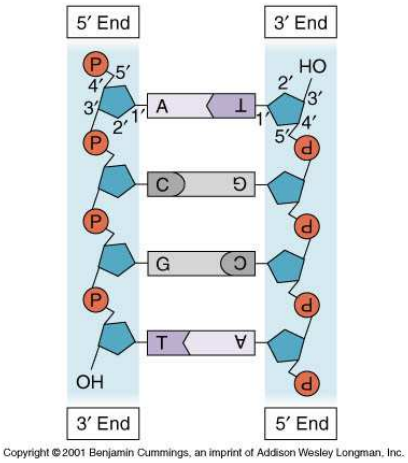


Figure B.2: The 5' end designates the end of DNA or RNA strand that has the fifth carbon of the sugar-ring of the deoxyribose or ribose at its terminus. The 3' end of a strand is so named due to it terminating at the hydroxyl group of the third carbon in sugar-ring.

B.2.4 Gene

The gene is a molecular unit of heredity of a living organism, each gene corresponding to various biological traits. More specifically it has been defined by Pearson et al.[56] as follows

> A gene is a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions.

A *constitutive gene* is a gene that is transcribed continually, i.e. there is no control over its expression. A *housekeeping gene* is typically a constitutive gene that is transcribed at a relatively constant level. The housekeeping gene's products are typically needed for maintenance of the cell. It is generally assumed that their expression is unaffected by experimental conditions.

A *facultative gene*, as opposite to a constitutive gene, is transcribed only when needed. In particular, an *inducible gene* is a gene whose expression is either responsive to environmental change or dependent on the position in the cell cycle.

B.2.5 Inducer

Different proteins can affect gene expression by promoting or preventing mRNA transcription. In *prokaryotes* these cells often intervene at the beginning of a specific gene, called *operator*, which is the DNA sequence to which polymerases attach to synthesize mRNAs.

Inducers cause gene transcription to start, by disabling repressor proteins. More in detail inducers bind to repressors, change their shape and prevent them to bind to DNA, allowing gene transcription. Inducers can be sometimes modulated by *activators* which bind to inducers and allow them to attach to the DNA strand.

B.2.6 Operon

An *operon* is a functioning unit of genomic DNA containing a cluster of genes under the control of a single promoter (see Section B.2.7 for details). This concept was introduced by Monod, Jacob *et alii* in 1960 [31]. The authors proposed a model in which genes are controlled via the operons through a single feedback regulatory mechanism: repression. It was then found that the regulatory mechanisms may be much more complicate. Nevertheless, this concept has revealed fundamental for molecular biology.

The operon is composed of three basic structures:

1. the *promoter*, see Section B.2.7 for details, to which the RNA polymerase binds to initiate transcription,

2. the *operator*, which is a small sequence between the promoter and the genes, to which regulators bind.

3. the *structural genes*, which are the genes that are regulated by the operon.

Operons occur in both prokaryotes and eukaryotes. The genes contained in the operon are transcribed together into a messenger. This mRNA can be translated or cut into several mRNAs, each corresponding to a specific protein. The result is that genes contained in the same operon are expressed or repressed all together.

B.2.7 Promoter

In order for the transcription to take place, the enzyme that synthesizes RNA, known as *RNA polymerase*, must attach to the DNA near a gene. Promoters contain specific DNA sequences and response elements which provide a secure initial binding site for RNA polymerase and for proteins called *transcription factors* that recruit RNA polymerase.

In bacteria, the promoter is recognized by RNA polymerase and an associated *sigma factor* σ, which in turn is often brought to the promoter DNA by an activator protein binding to its own DNA binding site nearby.

B.2.8 Ribosomal Binding Site (RBS)

A *ribosomal binding site* (RBS) is a sequence on mRNA that is bound by the ribosome when initiating protein translation. The sequence is complementary to the 3' end of the rRNA. The ribosome searches for this site and binds to it through base-pairing of nucleotides and then begins the translation process.

B.2.9 Ribosome

The *ribosome* is a large and complex molecular machine, that serves as the primary site of biological protein synthesis (translation). More in detail, ribosomes link together the amino acids as specified by the nucleotide sequence of mRNAs.

Ribosomes consist of two subunits: the large subunit (LSU) and small subunit (SSU). Prokaryotic ribosomes are around $20nm$ in diameter and are composed of 65% ribosomal RNA and 35% ribosomal proteins. For prokaryotes the LSU is called 50S and the SSU is called 30S, while the ribosome is referred to as 70S. Note that the S units of the subunits cannot simply be added because they represent measures of sedimentation rate, that is affected by the component shape and by its mass.

The 30S ribosomal subunit contains the 16S rRNA, see B.2.1, while the 50S contains both the 5S and 23S rRNA. The 3' end of the 16S rRNA (in a ribosome) binds to a sequence on the 5' end of mRNA called the Shine-Dalgarno sequence (see Section B.2.14).

B.2.10 RNA

Ribonucleic acid or RNA is a nucleic acid polymer consisting of nucleotide monomers and perform multiple vital roles in the coding, decoding, regulation and expression of genes. RNA comprises the nucleic acids and, together with DNA and proteins, constitute the three major macromolecules essential for all cells. RNA acts as a messenger (mRNA) between DNA and the protein synthesis complexes (ribosomes), forms portions of ribosomes and acts as an essential carrier molecule for amino acids to be used in protein synthesis (tRNA). RNA is assembled as a chain of nucleotides but, unlike the DNA, it is single-stranded.

B.2.11 Messenger RNA (mRNA)

The *messenger RNA* (mRNA) is a chemical "blueprint" for a specific protein. More specifically, a messenger conveys the genetic information from the DNA to the ribosome, specifying the amino acid sequence of the protein products.

The genetic information in mRNA is encoded in the sequence of nucleotides, which are arranged into *codons*, a triplet of nucleotides. Each codon encodes for a specific amino acid, except the *stop codons* which terminate the protein synthesis.

Because eukaryotic transcription and translation occur in different compartments of the cell, eukaryotic messengers must be exported from the nucleus to the cytoplasm and require a post-transcriptional processing before translation can take place. On the other hand, no processing is required for prokaryotic messengers and translation can initiate even before the mRNA elongation has been completed.

Just before the AUG start codon, prokaryotic mRNAs have a common sequence called the *Shine-Dalgarno sequence*, see B.2.14. The Shine-Dalgarno sequences are complementary to a sequence contained in the 16S ribosomal RNA (rRNA), see B.2.12 and help align the mRNA with the ribosome to properly orient the molecules for translation initiation.

B.2.12 Ribosomal RNA (rRNA)

Ribosomal ribonucleic acid (rRNA) is the RNA component of the ribosome, the protein manufacturing machinery of all living cells. Ribosomal RNA provides a mechanism for decoding mRNA into amino acids and interacts with tRNAs during translation by providing peptidyl transferase activity. The tRNAs bring the necessary amino acids corresponding to the appropriate mRNA codon.

B.2.13 Transfer RNA (tRNA)

A *transfer RNA* (tRNA) is a type of RNA molecule that helps decode a messenger RNA (mRNA) sequence into a protein. More specifically, each amino acid is specified by a three-nucleotide

mRNA sequence (*codon*) and each codon is recognized by a specific tRNA, which contains a specific anticodon triplet sequence. When the tRNA is associated with the corresponding amino acid, we refer to it as *aminoacyl-tRNA* or charged tRNA. The charged tRNA delivers the amino acid to the ribosome for incorporation into the polypeptide chain that is being produced. The uncharged tRNA is then released and ribosome proceeds to the next codon and continues translation step.

B.2.14 Shine-Dalgarno sequence

The Shine-Dalgarno sequence (SD), AGGAGG, is a ribosomal binding site in the mRNA, generally located 8 basepairs upstream of the start codon AUG and exists only in prokaryotes. This sequence helps recruit the ribosome to the mRNA to initiate protein synthesis by aligning it with the start codon. The complementary sequence (UCCUCC), is called the anti-Shine-Dalgarno sequence and is located at the 3' end of the 16S rRNA in the ribosome.

Mutations in the Shine-Dalgarno sequence can reduce translation, due to a reduced mRNA-ribosome pairing efficiency. Anyway, what principally attracts ribosome to mRNA initiation region is apparently *ribosomal protein S1*, which binds to AU-rich sequences found in many prokaryotic mRNAs 15-30 nucleotides upstream of start-codon.

Bibliography

[1] Murat Acar, Jerome T Mettetal, and Alexander van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature genetics*, 40(4):471–475, 2008.

[2] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6):636–643, 2006.

[3] David H Bechhofer. Bacillus subtilis mrna decay: new parts in the toolkit. *Wiley Interdisciplinary Reviews: RNA*, 2(3):387–394, 2011.

[4] Attila Becskei, Benjamin B. Kaufmann, and Alexander van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature genetics*, 37(9):937–944, 2005.

[5] Otto G Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of theoretical biology*, 71(4):587–603, 1978.

[6] Jonathan A Bernstein, Arkady B Khodursky, Pei-Hsun Lin, Sue Lin-Chao, and Stanley N Cohen. Global analysis of mrna decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays. *Proceedings of the National Academy of Sciences*, 99(15):9697–9702, 2002.

[7] Florian Brandt, Stephanie A Etchells, Julio O Ortiz, Adrian H Elcock, F Ulrich Hartl, and Wolfgang Baumeister. The native 3d organization of bacterial polysomes. *Cell*, 136(2): 261–271, 2009.

[8] Long Cai, Nir Friedman, and X Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[9] Ciarán Condon and David H Bechhofer. Regulated rna stability in the gram positives. *Current opinion in microbiology*, 14(2):148–154, 2011.

[10] D. R. Cox. *Renewal theory*. Methuen, London [etc.], 1962.

[11] F. H. C. Crick. On Protein Synthesis. *The Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[12] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[13] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Jones and Bartlett, 1993.

[14] Murray P Deutscher. Degradation of rna in bacteria: comparison of mrna and stable rna. *Nucleic acids research*, 34(2):659–666, 2006.

[15] Johan Elf, Gene-Wei Li, and X Sunney Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–1194, 2007.

[16] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[17] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[18] V. Fromion, E. Leoncini, and P. Robert. Stochastic gene expression in cells: A point process approach. *SIAM Journal on Applied Mathematics*, 73(1):195–211, 2013.

[19] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[20] B. V. Gnedenko and A. N. Kolmogorov. *Limit distributions for sums of independent random variables*. Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills., Ont., 1968.

[21] Alfred L Goldberg. Protein degradation and protection against misfolded or damaged proteins. *Nature*, 426(6968):895–899, 2003.

[22] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[23] W. J. Gordon and G. F Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967.

[24] Henri Grosjean and Walter Fiers. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*, 18(3):199–209, 1982.

[25] Claes Gustafsson, Sridhar Govindarajan, and Jeremy Minshull. Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7):346–353, 2004.

[26] Bremer H. and Dennis P.P. Modulation of chemical composition and other parameters of the cell by growth rate. In Frederick C. Neidhardt and Roy Curtiss, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, second edition, 1996.

[27] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29): 12167–12172, 2011.

[28] John L Ingraham, Ole Maaløe, and Frederick Carl Neidhardt. *Growth of the bacterial cell*, volume 3, chapter Growth rate as a variable, pages 267–315. Sinauer Associates Sunderland, Massachusetts, 1983.

[29] Farren J Isaacs, Jeff Hasty, Charles R Cantor, and James J Collins. Prediction and measurement of an autoregulatory genetic module. *Proceedings of the National Academy of Sciences*, 100(13):7714–7719, 2003.

[30] Keiichi Itakura, Tadaaki Hirose, Roberto Crea, Arthur D Riggs, Herbert L Heyneker, Francisco Bolivar, and Herbert W Boyer. Expression in escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science*, 198(4321):1056–1063, 1977.

[31] F. Jacob, D. Perrin, C. Sanchez, and J. Monod. The operon: a group of genes whose expression is coordinated by an operator. *Comptes Rendus*, pages 250–1727, 1960.

[32] H. Jasiulewicz and W. Kordecki. Convolutions of Erlang and of Pascal distributions with applications to reliability. *Demonstratio Mathematica. Warsaw Technical University Institute of Mathematics*, 36:231–238, 2003.

[33] Benjamin B Kaufmann, Qiong Yang, Jerome T Mettetal, and Alexander van Oudenaarden. Heritable stochastic switching revealed by single-cell genealogy. *PLoS biology*, 5(9):e239, 2007.

[34] Frank P. Kelly. *Reversibility and Stochastic Networks*. John Wiley & Sons Ltd., Chichester, 1979. ISBN 0-471-27601-4. Wiley Series in Probability and Mathematical Statistics.

[35] David Kennell and Howard Riezman. Transcription and translation initiation frequencies of the *Escherichia coli lac* operon. *Journal of molecular biology*, 114(1):1–21, 1977.

[36] Thomas B Kepler and Timothy C Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical Journal*, 81(6):3116–3136, 2001.

[37] Andrzej M Kierzek, Jolanta Zaim, and Piotr Zielenkiewicz. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *Journal of Biological Chemistry*, 276(11):8165–8172, 2001.

[38] J. F. C. Kingman. *Poisson processes*. Oxford studies in probability, 1993.

[39] Minoru SH Ko. A stochastic model for gene induction. *Journal of theoretical biology*, 153 (2):181–194, 1991.

[40] M.S. Ko, H. Nakauchi, and N. Takahashi. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *EMBO J.*, 9:2835–2842, 1990.

[41] Sidney R Kushner. mrna decay in escherichia coli comes of age. *Journal of bacteriology*, 184 (17):4658–4665, 2002.

[42] Hédia Maamar, Arjun Raj, and David Dubnau. Noise in gene expression determines cell fate in bacillus subtilis. *Science*, 317(5837):526–529, 2007.

[43] Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[44] Harley H McAdams and Adam Arkin. Simulation of prokaryotic genetic circuits. *Annual Review of Biophysics and Biomolecular Structure*, 27(1):199–224, 1998.

[45] Jerome T Mettetal, Dale Muzzey, Juan M Pedraza, Ertugrul M Ozbudak, and Alexander van Oudenaarden. Predicting stochastic gene expression dynamics in single cells. *Proceedings of the National Academy of Sciences*, 103(19):7304–7309, 2006.

[46] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research*, 38 (suppl 1):D750–D753, 2010.

[47] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[48] John RS Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[49] J. R. Norris. *Markov chains*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-48181-3. Reprint of 1997 original.

[50] Aaron Novick and Milton Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences of the United States of America*, 43(7):553, 1957.

[51] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):69–73, 2002.

[52] Ertugrul M Ozbudak, Mukund Thattai, Han N Lim, Boris I Shraiman, and Alexander Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737–740, 2004.

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

[54] Johan Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2):157–175, 2005.

[55] Johan Paulsson, Otto G Berg, and Måns Ehrenberg. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13):7148–7153, 2000.

[56] Helen Pearson. Genetics: what is a gene? *Nature*, 441(7092):398–401, 2006.

[57] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234, 1995.

[58] Margit Pedersen, Søren Nissen, Namiko Mitarai, Sine Lo Svenningsen, Kim Sneppen, and Steen Pedersen. The functional half-life of an mrna depends on the ribosome spacing in an early coding region. *Journal of molecular biology*, 407(1):35–44, 2011.

[59] Rob Phillips, Jane Kondev, and Julie Theriot. *Physical biology of the cell*. Garland Science, 2009.

[60] Maria Piques, Waltraud X Schulze, Melanie Höhne, Björn Usadel, Yves Gibon, Johann Rohwer, and Mark Stitt. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in arabidopsis. *Molecular systems biology*, 5(1), 2009.

[61] Arjun Raj, Charles S. Peskin, Daniel Tranchina, Diana Y. Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS biology*, 4(10):e309, 2006.

[62] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004.

[63] David R Rigney. Stochastic model of constitutive protein levels in growing and dividing bacterial cells. *Journal of Theoretical Biology*, 76(4):453–480, 1979.

[64] David R Rigney and William C Schieve. Stochastic model of linear, continuous protein synthesis in bacterial populations. *Journal of theoretical biology*, 69(4):761–766, 1977.

[65] Ph. Robert. *Stochastic Networks and Queues*, volume 52 of *Applications of Mathematics*. Springer-Verlag, Berlin, first edition, 2003.

[66] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain, and Michael B Elowitz. Gene regulation at the single-cell level. *Science Signalling*, 307(5717):1962, 2005.

[67] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[68] Gürol M Süel, Rajan P Kulkarni, Jonathan Dworkin, Jordi Garcia-Ojalvo, and Michael B Elowitz. Tunability and noise dependence in differentiation dynamics. *Science*, 315(5819): 1716–1719, 2007.

[69] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99 (20):12795–12800, 2002.

[70] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[71] Mukund Thattai and Alexander Van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[72] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171 (4356):737–738, 1953.

[73] Eugenia Y Xu, Karl A Zawadzki, and James R Broach. Single-cell observations reveal intermediate transcriptional silencing states. *Molecular cell*, 23(2):219–229, 2006.

[74] Oleg Yarchuk, Nathalie Jacques, Jean Guillerez, and Marc Dreyfus. Interdependence of translation, transcription and mrna degradation in the *lacZ* gene. *Journal of molecular biology*, 226(3):581–596, 1992.

[75] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao, and X Sunney Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–1603, 2006.