POLITECNICO DI TORINO

Master's Degree in Physics of Complex Systems

Master's Degree Thesis

Competition for intracellular resources: from experimental data to parameter estimation

Supervisor:
Dr. Carla BOSIA

Candidate:
Marta Cunial

December 2023

Abstract

In recent years, modelling genetic circuits has emerged as a powerful tool to unravel underlying mechanisms of gene expression. Experimental data of single cells expressing two fluorescent proteins (mCherry and eYFP) over time are available: they present various shapes and behaviours which are not predictable by standard models of gene regulation assuming infinite resources. Inspired by many studies suggesting that competition for molecular resources can modify and shape the response of a genetic circuit, this Master's Thesis aims at understanding if a resource-aware model can be used to correctly predict gene expression in environments with finite pools of intracellular resources. To do this, a minimal stochastic model is built to study the simple case where only the ribosomes are present in finite number, we investigate the effects of this finite pool on gene expression levels, and infer model parameters. In order to fix as many parameters as possible, we performed wet-lab experiments with HEK293 Tet-off cells transfected with plasmids containing sequences for mCherry and eYFP fluorescent proteins. Degradation rates of the proteins are then estimated using fluorescence microscopy, while the degradation rates of mRNAs are estimated via quantitative PCR techniques. Experimental data of the two fluorescent proteins are used to validate the model and to infer other parameters that are difficult to determine experimentally.

Table of Contents

List of Figures

Introduction

Understanding genetic regulation is crucial for unravelling the underlying mechanism of our organism. Over the past decades, significant efforts have been dedicated to discover in depth how cells work. The primary objective has been to unravel the origins and progression of diseases aiming at understanding how they start and develop, with the hope of finding effective treatments. This pursuit is intricately linked to a thorough comprehension of gene expression mechanisms [1]. This Master's thesis plans to contribute to the ongoing attempt to unravel cell behaviour by studying gene expression via mathematical modelling and statistical analysis.

Collaborative efforts across various disciplines, including molecular biology, genomics, bioinformatics, and physics, have encouraged research promoting the translation of scientific discoveries into clinical applications. The inherent complexity of our organism along with its ability to adapt, evolve and overcome difficulties that can emerge at cellular level, can be traced back to how DNA is expressed. Gene expression is a fundamental process that regulates the behaviour and functionality of cells, and it plays an essential role in coordinating various biological processes, including cell growth, differentiation, and response to environmental changes. By controlling the synthesis of specific proteins, gene expression determines the unique characteristics and functions of different cell types within an organism [2].

The development and progression of many diseases, among which cancer, involves genetic and epigenetic alterations, disrupting the finely tuned regulation of gene expression in cells [3]: understanding the tangled relationship

between gene expression and diseases is critical for unravelling the molecular mechanisms underlying disease's initiation and progression. Over the past few decades, advances in genomic technologies have revolutionized our ability to investigate gene expression patterns. By creating a mathematical model, more and more accurate and complex, it would be possible to understand the basal mechanisms of gene expression and eventually exploit the model to better infer sick cells and their gene regulation.

By now, stochastic simulation and computational biology have been proven as powerful tools for studying gene expression in cells [4]. Traditional deterministic models assume that gene expression follows precise rules and deterministic kinetics. On the other hand, stochastic simulation approaches, such as Gillespie's algorithm [5], allow us to model gene expression at the single-cell level, capturing the inherent variability and randomness observed in biological systems. Stochastic models prove invaluable in enhancing our comprehension of experimental data, especially when dealing with single-cell fluorescence trajectories [6]. Within this Master's thesis, we model gene expression in a competing environment with a basal stochastic model, involving two genes that compete to bind to ribosomes for translation; this model is examined to gather more information about the effects of competition for molecular resources on gene expression. The principal aim is to test model's ability to accurately predict behaviours observed in experimental data, which are not explainable by simplistic models assuming infinite abundance of resources. To this end, we analyze a dataset at our disposal from previous experiments, containing results from single-cell fluorescence signals. This dataset not only serves as a valuable resource for validating our model, but also plays a pivotal role in the inference of model's parameters. The available dataset was obtained following cells in time, after they were transfected with plasmids encoding for two fluorescence proteins, and the fluorescence signals were measured.

Each cell expresses two fluorescent proteins (mCherry and eYFP), fluorescence signals are followed in time from 6 hours after cell transfection, and fluorescence intensities are measured every 20 minutes up to 45 hours.
Fluorescence levels of four different cells are depicted in Figure 1: it is clearly visible how gene expression varies enormously in shape, behaviour and intensity between cells. Understanding what causes different conducts in cells is often impossible to determine from experiments. As an example, in the transfection experiment conducted to obtain the data in Figure 1, the quantity of plasmids each cell has uptaken is not known, and this can of course determine diverse gene expression levels. A possible educated guess we can make is that if a cell has taken in a high number of plasmids, a greater demand for molecular resource is expected inside the cell. In this competing scenario, the available resources become limited and are distributed among plasmids and endogenous genes; consequently, we would expect a reduced gene expression level since limited resources represent a bottleneck point, reducing efficiency of gene expression. Studying the characteristics of a system in which genes compete for resources (represented in our basal model by finite ribosomes pool) with stochastic simulation methods gives us a tool to distinguish a cell in a competing environment from one in which the resources do not represent a limiting factor.

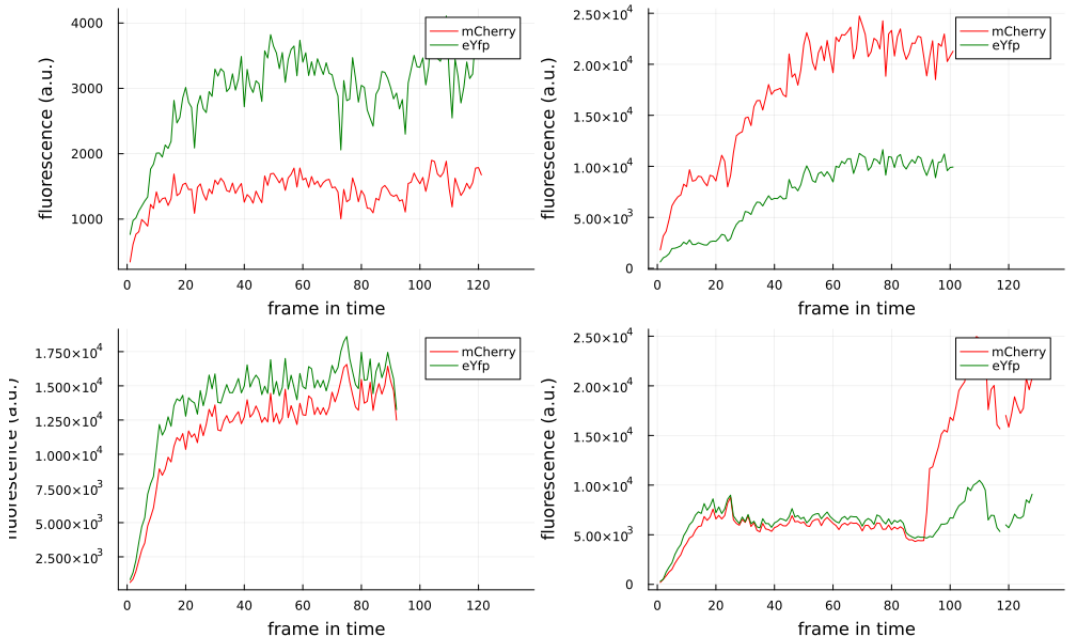

Figure 1: Different experimental fluorescence trajectories for four different cells .
Four representative trajectories over 19.001 cells available are displayed, with mCherry (red) and eYFP (green) signals. The saturation level, the behaviour and the shape differ from cell to cell.

Changes in concentration of biochemical species can be formulated as a system of differential equations which can be solved and/or simulated. First, all possible reactions in the system are to be identified and each reaction is associated to a rate. After this, the differential equations for species concentration can be written. To clarify, consider the rate of change of a molecule of mRNA m, transcribed with a rate α_m and degraded with a rate β_m; similarly, a protein p is translated from a molecule of mRNA with a rate α_p and it degrades with a rate β_p: the associated differential equations for the two species are

$$\frac{dm}{dt} = \alpha_m - \beta_m\, m; \qquad \frac{dp}{dt} = \alpha_p\, m - \beta_p\, p. \tag{1}$$

The equations 1 are a representation of the mass conservation law, where

the change of a species concentration is the sum of *creation* contributions, transcription and translation which increase the corresponding species concentration, and *consumption* contributions leading to a decrease in the species concentrations. It is common to find the solution of the system of equations at the steady-state, when the system has stopped evolving, setting the derivative to zero, i.e. $\frac{dm}{dt} = \frac{dp}{dt} = 0$. The steady-state solution of the above system of differential equations leads to a system of algebraic equations,

$$0 = \alpha_m - \beta_m \, m \qquad 0 = \alpha_p \, m - \beta_p \, p. \tag{2}$$

The straightforward solutions of this systems are $m^* = \frac{\alpha_m}{\beta_m}$ and $p^* = \frac{\alpha_p}{\beta_p} m^* = \frac{\alpha_p \, \alpha_m}{\beta_p \, \beta_p}$.

This representation of gene expression with a system of ordinary differential equations (ODE) provides a deterministic approach able to predict average values of concentrations of species involved in the system. However, the ODE approach does not take into account the randomness and discrete nature of the process; when the aforementioned aspects play an important role, another approach is necessary. Stochastic simulations methods provide a probabilistic representation considering single random events and correlations between species: these methods are capable of capturing effect of stochastic fluctuations on system dynamics.
Once the model has been defined, the following step necessary to completely determine the system is estimating reactions parameters, which can be determined experimentally or inferred from data and simulations. In this work, both procedures are exploited (see Chapter 3).

This Master 's Thesis is organized in five parts. First of all, a brief introduction about gene expression from a biological point of view is given in Chapter 1 in Section 1.1.1; section 1.1.2 focuses on regulation of gene expression in general and the most important findings available in literature on competition's effects on gene expression regulation are presented in Section 1.1.3.

Chapter 2 offers an overview about modelling genetic circuits in Section 2.1; in Section 2.1.1 the functioning of the stochastic simulation method utilised in this Master's Thesis is explained, before actually depicting the used model, in Section 2.2. Results obtained from simulation are discussed and compared to experimental findings in Chapter 4, in Chapter 5 conclusions and suggestions for future research are presented.

Chapter 1

State of the Art

1.1 Gene expression

1.1.1 How genes are expressed from DNA.

Genetic information in eukaryotes are contained in the DNA located inside the nucleus of cells. Genes encoded into DNA are firstly transcribed into messenger RNAs (mRNAs), which are then translated into proteins [7].

The principal enzyme engaged in transcription is RNA polymerase, that transcribes a molecule of RNA from one strand of DNA via base pairing but substituting thymine (T) with uracil (U) [8]; transcription starts when RNA polymerase binds to a specific region of DNA called promoter, where, with the help of transcription factors (TFs) proteins, the transcription process can start. Binding to specific regions of the DNA with various affinities, TFs can enhance or reduce the binding of RNA polymerase, tuning transcription [7]. During the elongation phase of transcription, for each unit in the DNA strand, a nucleotide is added via base pairing to the mRNA molecule, thereby increasing its length. This progression halts when the RNA polymerase complex encounters a specific site on the DNA strand known as the terminator [9]. The obtained molecule of RNA is called pre-mRNA and it undergoes

various steps ensuring the production of a functional transcript (mRNA). During this maturation process, the 5' end is capped adding a 7-methyl guanosine residue, and a poly-A tail is added to the 3' end (polyadenylation) to stabilize the mRNA. After this, a further step called splicing occurs, mediated by the spliceosome: non-coding regions (introns) are removed, and exons (coding regions) are joined together, forming the mRNA mature molecule [10]. The mature mRNA is transported from the cell nucleus to the cytoplasm through nuclear pores, where it is available for translation by ribosomes into a functional protein.

Translation is the process of assembling a protein from the mRNA and it is performed by an organelle called ribosome. The idea behind the translation is simple: the nucleotide sequence is read in series of three nucleotides (a codon), and each triplet is associated to an amino acid. The ribosome is composed by two subunits that combine when translation occurs, binding together at a specific region of the mRNA, which is called Kozak sequence in eukaryotes [11]; Kozak sequence is located near the initiation codon methionine (AUG), and translation can start [10]. The larger unit of the ribosome reads three nucleotides of the mRNA at a time, and each codon is associated to a specific amino acid; the elongation of the protein chain continues until the ribosome reaches a terminator codon (UUA, UGA or UAG), that symbolises the end of the coding region.

Apart from mRNA and the two ribosome units, another molecule of RNA named transfer RNA (tRNA) is necessary for translation to take place; tRNA functions as a physical link between the nucleotides and the amino acid. Each tRNA matches a codon of the mRNA with a complementary sequence of base pairs (anticodon), located at one end of the molecule; the amino acid corresponding to the anticodon is attached to the opposite end of the tRNA (see Figure 1.1) [7] [10] .

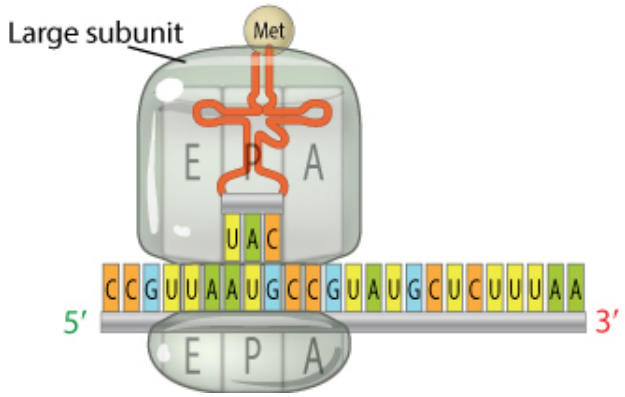

Figure 1.1: Ribosome units bind together to allow a translation of a codon.
The large ribosome unit binds to the smaller units and the initial codon AUG is aligned with the P site of the ribosome, where the tRNA transporting the corresponding methionine amino acid (Met) attaches; the next codon to be translated is positioned in correspondence of the A site of the ribosome.
(Adapted from: Clancy S, Brown W. (2008) "*Translation: DNA to mRNA to Protein*". Nature Education 1:101, https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/.)

1.1.2 Regulation of gene expression

The wide variety observed in nature and among various type of cells of the same organism is the result of regulation of gene expression, a fundamental process that determines the timing and amount of proteins production in living organisms. Gene regulation involves a complex web of molecular mechanisms that controls which genes are activated or repressed in response to various internal and external signals and stimuli. Regulation of gene expression occurs at various levels: epigenetic control, transcriptional regulation, post-transcriptional regulation, translational and post-translational regulation [7].

Epigenetic regulation refers to inheritable changes in the structure of DNA, which do not modify the underlying sequence of nucleotides; these changes in the DNA structure, named histone modifications (see Figure 1.2), can lead to gene silencing or enhance gene transcription into mRNA.

Histone proteins change their configuration depending on signals (called "tags") mainly represented by phosphate, methyl, or acetyl groups located on the histone; tags can be removed or added depending on whether the translation of a gene is necessary or not [7].

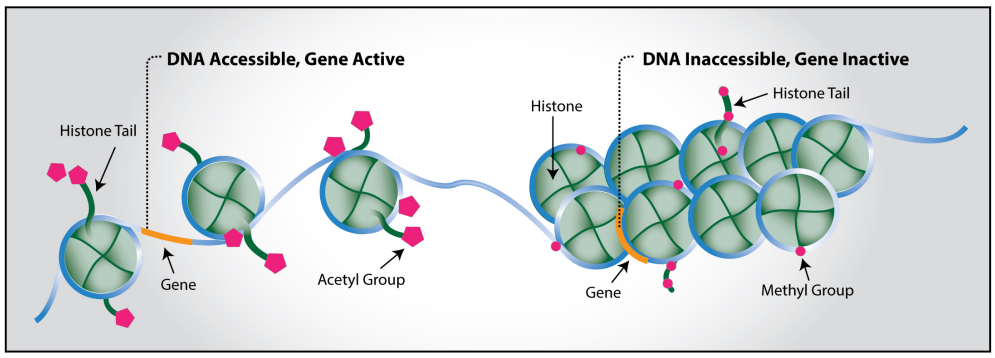

Figure 1.2: Different conformations of DNA and histone proteins. Histone proteins can modify their configuration allowing transcription of a part of DNA (left) or forbidding it (right). In the left configuratiom the RNA polymerase can have access to the gene and initiate the transcription process.
(Adapted from: https://www.epigentek.com/catalog/antibodies-for-histone-modifications-lp-49.html).

Transcriptional regulation determines whether a gene is transcribed into mRNA and its frequency of transcription. In order to initiate transcription of a gene, it is not sufficient that RNA polymerase binds to the promoter region, but it also requires the presence of other transcription factors, helping in the formation of the transcription complex; a longer promoter region is linked to a higher probability for transcription factors to bind, boosting transcription [10]. In eukaryotes, there exists some particular DNA regions called enhancers able to favour transcription: a type of TFs (activators) binds to the enhancer and creates interactions with the transcription machine, promoting its formation and advance on the mRNA strand [12]. Once the RNA polymerase is linked to the promoter, an additional transcription regulation step requires a set of TFs allowing RNA polymerase to move on to the next triplet and proceed in the transcription [13].

Post-transcriptional regulation modulates the processing, stability, and transport of RNA molecules; this step of gene regulation occurs between the transcription and translation into protein. The most important post-transcriptional regulation processes are mediated by microRNA (miRNA), non-coding RNA molecules of $\sim$ 21 nucleotides [14], [15]. Various studies showed that miRNAs are able to enhance degradation of mRNA , binding at the 3' untranslated region (UTR) of the mRNA molecule inducing deadenylation and decapping [16]. In addition to this, miRNA can inhibit translation, by interfering with the recruitment of the translation machinery [16] - [17]. RNA-binding proteins (RBPs) also act as post-transcriptional regulators: binding to the untranslated regions of mRNA strand, they enhance or diminish its stability, depending on the specific RBP that bind to the mRNA [7], [18].

Translation and post-translation regulation orchestrate the conversion of genetic information encoded in mRNA into functional proteins. Post-translation steps play a crucial role in refining and modifying the nascent protein. Processes like protein folding, post-translational modifications (e.g., phosphorylation, glycosylation), and subcellular trafficking contribute to the protein's final structure and function. After translation, proteins may undergo ubiquitination, a process where ubiquitin molecules are covalently attached to specific target proteins. This modification serves various cellular functions, primarily in protein degradation. Ubiquitination marks proteins for recognition and subsequent degradation by the proteasome, a cellular complex responsible for breaking down and recycling proteins [19]- [20].

1.1.3 Regulation by competition

Competition for natural resource plays a critical role at all levels in biology. Different species sharing resources compete, and this competition can lead to changes in population distributions [21]. Natural evolution is a result of competition between individuals of the same species, but also identical cells

can compete to regulate growth and activity, promoting the domination of the most fitted cell [22].
Within cells, competition plays a key role in the regulation of gene expression. Many cellular components such as ribosomes, DNA polymerase, transcription factors, RNA polymerase and various organelles can represent pools of resources that are subjected to competition in the intracellular environment.

TFs, for example, represent a shared resource among different genes, each associated with different binding affinities for transcription factors; the abundance of transcription factors in *Escherichia coli* and its relevance in gene expression regulation was examined by Brewster and colleagues in 2014 [23]. The study addresses how the fold-change of a repressor depends on the presence of different TF copy numbers: the Lac repressor (LacI) is the specific TF analyzed and the effects on its target genes are examined. The fold-change in gene expression is largely unaffected when the copy number of TFs is significantly higher than the copy number of genes. On the contrary, when the TF copy number is lower than the gene copy number, the effects of LacI on gene expression are muted, resulting in a poorer repression of the target genes [23].
Competition is also able to regulate cell growth: circular RNA (circRNAs) can bind to multiple miRNAs competitively, limiting their activity and preventing them from binding to their target mRNAs. This competition between circRNAs and miRNAs plays a crucial role in gene expression regulation and can have significant implications for cell growth, as proved by Zheng and colleagues in 2016 [24]. MiRNAs are a crucial targets for competition since their binding with other RNAs can inhibit their function; this effect is called *sponge effect* and consist in other molecules of RNA that sequester miRNAs from binding to their target, regulating their activity. Sponge effect has been proved to be a basal mechanism for cell differential [25] and tumor suppression [26].
Another study conducted in 2020 by Frei et al. [27] reveals that endogenous

and exogenous genes can compete for transcriptional and translational resources. Researchers transfected HEK293t cells with two fluorescent proteins (mCitrine and mRuby3) driven by the same promoter. When the exogenous gene is increased in quantity, the levels of the analysed gene decreased, showing that shared resources constitute a limiting factor in gene expression; this genetic burden results in a correlation between genes that are normally independent. Different ratio of the two proteins were used in the transfection mix and two amounts of encoding plasmids were compared (50 ng and 500 ng) in the study. As expected in a competing environment, higher levels of encoding plasmids increase the demand for cellular resources, reducing gene expression levels (Figure 1.3). Furthermore, the fluorescence levels of mCitrine and mRuby3 were negatively correlated (Figure 1.3 right panels) and this correlation was exacerbated with high values of encoding plasmids, showing that competition for molecular resources was more intense when the genes copy number to translate increases.

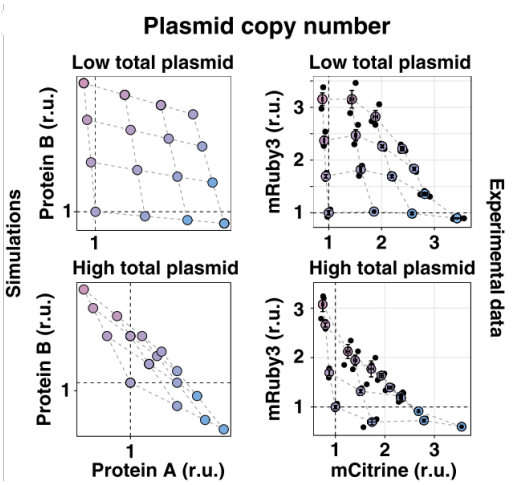

Figure 1.3: Effects of competition in genetic circuits.
(Left) The total gene expression is flattened by an higher amount of encoding plasmids.
(Right) Levels of mCitrine and mRuby are negatively correlated and this correlation is more pronounced with higher amount of transfected plasmids (lower panel).
(Adapted from: Frei T. et al., *Characterization and mitigation of gene expression burden in mammalian cell*, Nature Communications, 2020 [27]).

Frei et al. also claim that heterologous genetic loads take parts in the competing environment; to prove this, they co-transfected H1299 cells with two fluorescent proteins, EGPF and mKate, which are driven by the same promoter, and measure expression levels of heterologous and endogenous genes (CyCA2, eIF4E, GAPDH). When cells are transfected with high or intermediate levels of EGFP and mKate, the expression of endogenous genes decreases compared to cell that were not transfected [27].
Another comprehensive study carried out by Wei and colleagues in 2019 [28] examined in details what can be the outcomes of competition in gene expression. Different cellular resources which are involved in competition are addressed in the study: transcription factors (Figure 1.4 B), competitions for miRNAs (Figure 1.4 C), ribosomes (Figure 1.4 D), and degradation enzymes (Figure 1.4 E). A general mathematical model was studied where two target molecules T_1 and T_2 are in competition to bind with a shared regulatory molecule species R (Figure 1.4 F).

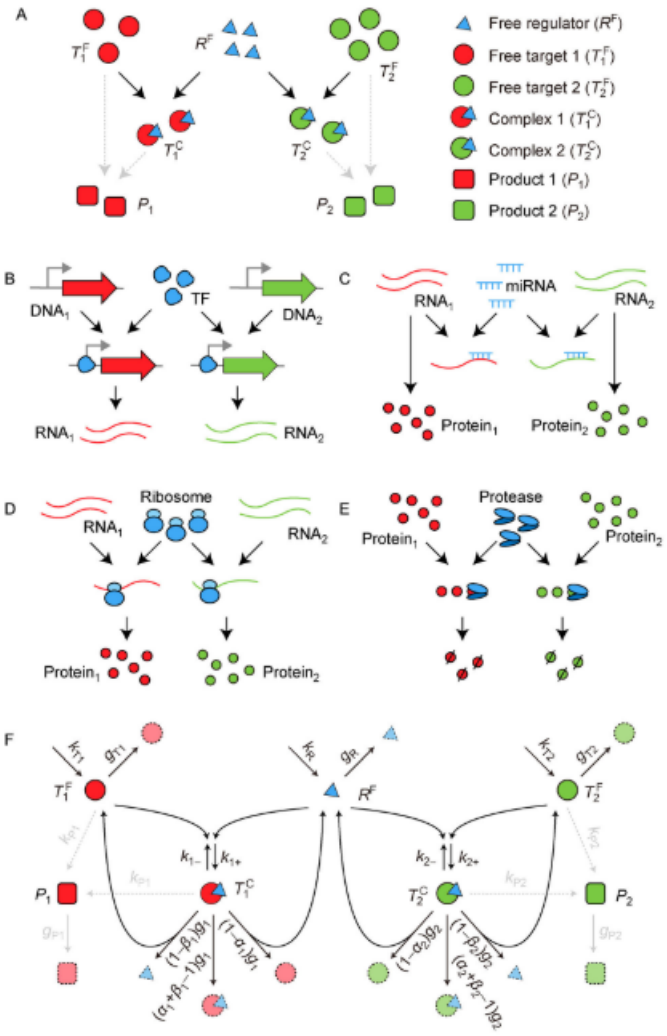

Figure 1.4: Genetic circuits with various competition resources.
(A) Basic structure of a competing model.
(B - E) Competing models with different resources that can cause competition (TFs, miRNAs, ribosomes and proteases in order).
(F) Kinetic model of a competing system.
(Adapted from: Wei L. et al. *Regulation by competition: a hidden layer of gene regulatory network.* Quantitative Biology, 2019 [28]).

The study showed how competition can affect the shape of the dose-response curve (which analyzes the relationship between the concentration of the regulator and the response of a biological system), and can generate a

threshold behaviour in the concentration of the complex T_{1C} as a function of the concentration of the complex T_{2C} when the regulator (R) is present in a scarce pool; on the contrary, when the regulator (R) is abundant the two species are not influenced one by the other (see Figure 1.5). Competition can also modify the dose-response curve's dynamics, delaying or accelerating its edges [28].

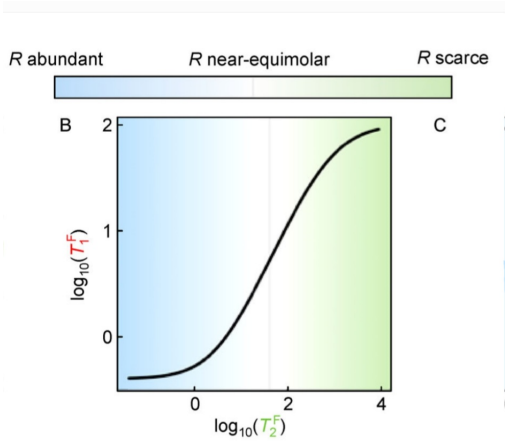

Figure 1.5: Abundance of T_1^F as a function of T_2^F, in different regimes.
In the "scarce R" regime (*blue region*), level of free T_1^F is not sensitive to abundance of T_2^F, whereas when the system enters the "near-equimolar" regime (*white region*), levels of free T_1^F and T_2^F depend on each other [28].
(Adapted from: Wei L. et al. *Regulation by competition: a hidden layer of gene regulatory network.* Quantitative Biology, 2019 [28]).

Inspired by numerous studies on gene expression in a competing environment and its consequences, the main objective of this work is to build a mathematical model able to predict the effects of regulation by competition for a finite pool of ribosomes. An introduction of mathematical modeling of genetic circuits and simulations are given in Chapter 2, where the used model is also described.

Chapter 2

Modeling

2.1 Modeling gene expression

Understanding the dynamics underlying biological processes allows to derive differential equations that represent changes in species concentration mathematically, as we showed in equation 1 for the simple production and degradation of a molecule of mRNA or protein. Solving these equations provides a valuable tool to unravel the behaviour of biological systems, not only granting the prediction of changes in species concentrations, but it also leads to a better understanding on how external conditions will affect biological outcomes.

As pointed out by Scott M. [29] three crucial assumptions are at the root of mathematical modelling of biological processes; first of all, in order to be able to predict the evolution of a biological system with differential equations, one must assume that species concentrations evolve continuously in time and that they are differentiable. A continuous and differentiable process emerges when many reactions among a great number of particles pile over time, and each reaction leads to a finite change in the molecules numbers.
Next, biochemical reactions are usually influenced by the system's past states and there exist many feedback mechanisms: to include this feature in a

system of differential equations, it is necessary to use rates that depend on the past states of the systems. In this way it is possible to merge together the assumption of instantaneous reactions necessary for the differential equations representation, and the more complex and time-dependent nature of biological systems.
The third and last assumption needed to adapt differential equations to biological reactions is to presume that reactants are distributed in space in a homogeneous way; to mathematically account for spatial density of reactants, it is necessary to adopt partial differential equation formalism instead of ordinary differential equations [29].

Solution of ODE systems are particularly useful for systems with large populations, where individual variations are less significant and they can be averaged out, thus providing a valuable tool to describe the average behaviour of the system. It is however necessary to point out that ODE models assume a continuous and deterministic view of processes, excluding randomness and discrete nature of processes. As a consequence, situations where stochastic behaviours play a significant role, such as small populations or single events, require a more appropriate approach: probabilistic models and stochastic simulations allow us to capture the discreteness of these events.

Since biochemical reactions are the results of discrete reaction events, such as molecules binding, unbinding or transformation, it is almost immediate to associate each reaction to a probabilistic event depending on the probability of encounter and reaction between two molecules. To provide a more accurate representation of biological systems, we switch to a probability representation, where $P(\mathbf{n}, t)$ represents the probability for the system to be in the state $\mathbf{n}$ at time t, and the vector $\mathbf{n} = (n_1, n_2 ... n_N)$ represents the number of molecules for species $i = 1, ...N$ present in the state at that time. To describe how $P(\mathbf{n}, t)$ evolves in time, we can write the so called Master equation (Eqn. 2.1).

$$\frac{dP(\mathbf{n},t)}{dt} = \sum_{\mathbf{n'}\neq\mathbf{n}} W(\mathbf{n'} \rightarrow \mathbf{n})P(\mathbf{n'},t) - \sum_{\mathbf{n'}\neq\mathbf{n}} W\mathbf{n} \rightarrow \mathbf{n'}P(\mathbf{n},t) \tag{2.1}$$

The first term in RHS represents the transition from another state **n'** to state **n** with a transition rate $W(\mathbf{n'} \rightarrow \mathbf{n})$, while the second term is a transition directed outward from state **n** to any other state **n'**, with associated rate $W(\mathbf{n} \rightarrow \mathbf{n'})$. Writing the Master equation of a process corresponds to describe it as a kind of random-walk process which is governed by a single differential-difference equation [30]. Note that writing the Master equation means that the process describing the jump from one state to another is Markovian, i.e. the transition probability from state **n** to a state **n'** depends only on the present state **n**, no matter what are the past states of the process. The Master equation describing the time evolution of a system of discrete states is often challenging to solve analytically, especially for complex systems. Stochastic simulation methods, such as Gillespie algorithm, offer a practical and efficient mean to capture the inherent randomness and discrete nature of molecular interactions within biological systems described by the Master equation.

Moreover, stochastic simulation methods provide a significant advantage over deterministic approaches by precisely capturing the random nature of molecular interactions. In contrast to deterministic methods, which are based on solving differential equations using average reaction rates, stochastic simulation deals with the discrete and stochastic behaviour of individual molecules.

In a biological context, especially at the molecular level, randomness is omnipresent. Molecules collide, react, and diffuse in a manner that is fundamentally probabilistic. Stochastic simulations acknowledge this fundamental idea, and this allows to model individual reaction events and the time intervals between those events, reflecting real-world scenarios where reactions

occur sporadically due to chance encounters between reactants. By incorporating this discrete and probabilistic perspective, stochastic simulations provide a powerful tool to explore the full range of possibilities that arise from random interactions, representing a richer understanding of the system's dynamics compared to deterministic models, which tend to smooth out these fluctuations. To simulate a stochastic process with inherited randomness like biochemical reactions, we can use the famous stochastic simulation method called Gillespie algorithm.

2.1.1 Gillespie algorithm

Gillespie algorithm is one of the most famous stochastic simulation algorithms, created by Joseph L. Doob in 1945 but published by Dan T. Gillespie in 1977 [5], where he applied the algorithm to simulate biochemical reactions. Consider a system with two species X_1 and X_2, characterized by the following chemical reactions:

$$X_1 + X_2 \xrightarrow{c_1} 2X_2 \tag{2.2}$$

$$X_1 \xrightarrow{c_2} 0 \tag{2.3}$$

The probability that the first reaction happens in a time interval dt depends of course on the reaction rate per unit time c_1, but also on the concentration of the reactants that have to combine to make the reaction happen; more specifically, the probability that the reaction 2.2 happens in the time interval dt is $P = c_1 n_1 n_2 dt$, where n_1 and n_2 are the number of molecules of type X_1 and X_2, and the value $h = n_1 n_2$ is the number of combinations of reagents yielding to the reaction 2.2. In a system of M reactions, it is necessary to understand which of the reactions is going to happen first.

We can write the probability that the system is in state $(n_1, n_2...)$ at time t, that the next reaction to happen is of type R_i $(i = 1....M)$ and it happens

in the interval $[t+\tau, t+\tau+d\tau]$ as

$$P(\tau, i)d\tau \tag{2.4}$$

This probability can be seen as a product of two parts, i.e. $P(\tau, i)d\tau = P_0(\tau) \cdot P_i d\tau$: the probability that nothing happens in the time interval $(t, t+\tau)$, here called $P_0(\tau)$, multiplied by the probability that the reaction R_i happens in the time interval $(t+\tau, t+\tau+d\tau)$, with an associated probability $P_i = c_i h_i d\tau$, where h_i is the number of combinations of reagents molecules that can give origin to a reaction of type i.
To evaluate the probability that no reaction happens in the time interval τ ($P_0(\tau)$), we can divide the interval in K sub-intervals, each of size $\epsilon = \frac{\tau}{K}$. The probability that nothing happens in the first interval $(t, t+\epsilon)$ is simply:

$$\prod_{j=1}^{M} [1 - c_j h_j \epsilon] = 1 - \sum_{j=1}^{M} c_j h_j \epsilon + O(\epsilon). \tag{2.5}$$

Since equation 2.5 depends only on the size of the interval, and we have K consecutive intervals of length ϵ, we obtain

$$P_0(\tau) = [1 - \sum_{j=1}^{M} c_j h_j \epsilon + O(\epsilon)]^{K} = [1 - \sum_{j=1}^{M} c_j h_j \frac{\tau}{K} + O(K^{-1})]^{K} \tag{2.6}$$

Performing the limit $K \to \infty$, the probability that no reaction occurs in the time interval of size τ becomes

$$P_0(\tau) = \exp\left(-\sum_{j=1}^{M} c_j h_j \tau\right) \tag{2.7}$$

Recalling that $P(\tau, i)d\tau = P_0(\tau) \cdot P_i d\tau$ we arrive at

$$P(\tau, i) = a_i \exp\left(-a_0 \tau\right) \tag{2.8}$$

where we have defined the new quantity $a_i = c_i h_i$, the propensity function for reaction i; the propensity function measures the instantaneous rate at which the i-th reactions happens. The sum of the propensity functions is $a_0 = \sum_{j=1}^{M} c_j h_j$ [5], [30]. To summarize, we obtain the famous result stated by Gillespie in his paper in 1977 [5]:

$$P(\tau, i) = \begin{cases} a_i \exp(-a_0 \tau) & \text{if } 0 \leq \tau < \infty \text{ and } i = 1, ..., M \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

It is necessary to sample the random pair τ, i from distribution 2.9; as shown by Gillespie [5], to sample the time jump τ we can use the inverse cumulative function method that allows us to sample from a known distribution (in this case exponential distribution) [31] yielding to

$$\tau = \frac{1}{a_0} \log\left(\frac{1}{r_1}\right) \tag{2.10}$$

where r_1 is a random number generated according to a uniform distribution in [0,1]. The reaction i must be extracted according to its weight $\frac{a_i}{a_0}$ and this can be achieved again using inverse cumulative distribution method, sampling from a distribution with associated weights $\left[\frac{a_1}{a_0}, \frac{a_2}{a_0}, ... \frac{a_M}{a_0}\right]$.
Gillespie algorithm consists in the following steps [5]:

Algorithm 1 Gillespie Algorithm

Step 0: Define the M reaction constants $c_1, c_2, ...c_M$ and the N initial population numbers $X_1, X_2, ..., X_N$. Initialize the time variable $t = 0$, and the reaction counter $n = 0$. Initialize uniform random number generator.

Step 1: Calculate and store the M quantities $a_1 = h_1 c_1$, $a_2 = h_2 c_2, \ldots, a_M = h_M c_M$ for the current population numbers. Calculate and store the value $a_0 = a_1 + a_2 + \ldots + a_M$.

Step 2: Generate random number r_1 using the uniform random generator, calculate τ according to eqn 2.10 and extract the reaction i from a distribution with associated weights $\left[\frac{a_1}{a_0}, \frac{a_2}{a_0}, ...\frac{a_M}{a_0}\right]$.

Step 3: Update the time and the state according to the values extracted in Step 2. Increase $t = t + \tau$ and perform reaction R_i (adjust population levels accordingly). Update the reaction counter $n = n + 1$. Go to **Step 1**.

In this Master Thesis the Gillespie algorithm is implemented in Julia program using the package *Gillespie.jl* [32].

2.2 Building the model to reproduce experimental data.

Many previous studies have modeled genetic circuits taking into account finite shared resource pools. Frei T. and colleagues [27] created a generic genetic model accounting for limited transcriptional and translational resources and showed how competition shape gene expression levels, resulting in negative correlation between different genes. A recent study carried out by Cella F. et al. [33] suggested a resource-aware ODE model to capture post-transcriptional events and resource reallocation caused by miRNA-mediated downregulation process.
The experimental data we want to study are obtained from HEK293t cells transfected with encoding plasmids for mCherry and eYFP fluorescent proteins. Trying to explain the observed discrepancies between theoretical models and experimental results, we implement a model in which genes compete for a finite pool of ribosomes.
A schematic but illustrative scheme is displayed in Figure 2.1, containing all reactions and processes involved in the model used in this work. We want to model a genetic circuit in which two molecules of mRNAs of two different genes, m_1 and m_2, are transcribed together: this correlated transcription is defined to create a model as similar as possible to our experimental data, where mCherry and eYFP are driven by a common promoter. Each mRNA molecule can bind to a free ribosome R and form a translation complex, b_1 or b_2, that is in time translated into a protein molecule, one for each gene (p_1 and p_2). Proteins and mRNAs are subjected to degradation: more specifically, the protein can be degraded once it is translated, while the mRNA molecule can be degraded when it is in its "free" state, but also when it is bound in the complex with the ribosome.
The most important feature in this model is that the pool of ribosome is finite: if the number of mRNA molecules is sufficiently higher with respect

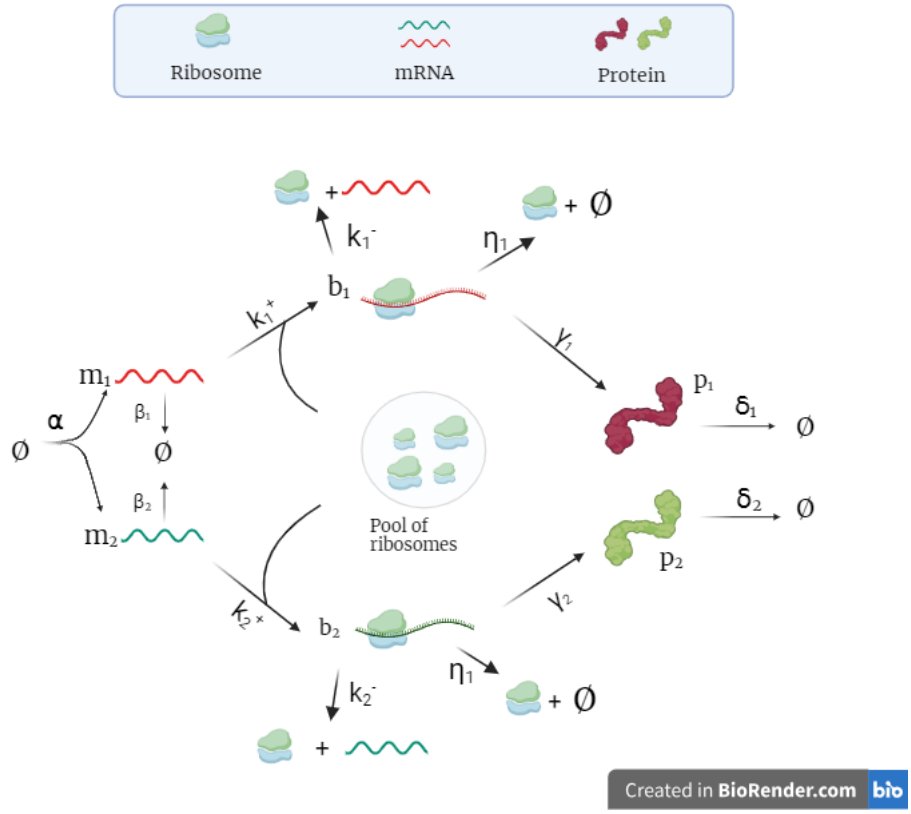

Figure 2.1: Illustrative scheme of processes and reactions studied in the model used.
(Created in Biorenders.com)

to the size of ribosome's pool, mRNAs have to compete to bind and form complexes with the free ribosomes, i.e. ribosomes which are not already involved in other complexes.

We imagine that during our transfection procedure, some cells may have assimilated an higher number of encoding plasmids with respect to other cells; an elevated number of plasmids corresponds to more molecules of mRNA competing to bind to ribosomes, establishing a competing scenario, not observed in cell with fewer plasmids where the numbers of mRNA molecules is smaller with respect to available ribosomes.

Following the scheme depicted in the Figure 2.1, the analyzed system is identified in terms of the following chemical reactions (2.11-2.23):

$$0 \xrightarrow{\alpha} m_1 + m_2 \qquad \text{transcription of correlated mRNAs} \tag{2.11}$$

$$m_1 + R \xrightarrow{k_1^+} b_1 \qquad \text{association of a molecule of } m_1 \text{ and a ribosome} \tag{2.12}$$

$$m_2 + R \xrightarrow{k_2^+} b_2 \qquad \text{association of a molecule of } m_2 \text{ and a ribosome} \tag{2.13}$$

$$b_1 \xrightarrow{k_1^-} m_1 + R \qquad \text{dissociation of the complex } b_1 \tag{2.14}$$

$$b_2 \xrightarrow{k_2^-} m_2 + R \qquad \text{dissociation of the complex } b_2 \tag{2.15}$$

$$b_1 \xrightarrow{\gamma_1} m_1 + R + p_1 \qquad \text{translation of } b_1 \text{ into a protein } p_1 \tag{2.16}$$

$$b_2 \xrightarrow{\gamma_2} m_2 + R + p_2 \qquad \text{translation of } b_1 \text{ into a protein } p_1 \tag{2.17}$$

$$b_1 \xrightarrow{\eta_1} R \qquad \text{degradation of } m_1 \text{ when bound in } b_1 \tag{2.18}$$

$$b_2 \xrightarrow{\eta_2} R \qquad \text{degradation of } m_2 \text{ when bound in } b_2 \tag{2.19}$$

$$m_1 \xrightarrow{\beta_1} 0 \qquad \text{degradation of a free } m_1 \tag{2.20}$$

$$m_2 \xrightarrow{\beta_2} 0 \qquad \text{degradation of a free } m_2 \tag{2.21}$$

$$p_1 \xrightarrow{\delta_1} 0 \qquad \text{degradation of a protein } p_1 \tag{2.22}$$

$$p_2 \xrightarrow{\delta_2} 0 \qquad \text{degradation of a protein } p_2 \tag{2.23}$$

It is now more visible and understandable what are the reactions considered while building the model and the rate at which each of them happens. The model is able to describe transcription of two correlated mRNAs at rate α (reaction 2.11); each of these mRNAs can create a translation complex binding to a free ribosome (reactions 2.12, 2.13) with rates k_1^+ or k_2^+ respectively, but these complexes can also undergo dissociation with rate k_1^- or k_2^-, freeing the mRNA and a ribosome each (reactions 2.14, 2.15). When the translation is complete, the mRNA and the ribosome are again freed and a protein p_1 (p_2) is produced with rate γ_1 (γ_2) (reactions 2.16, 2.17). The

degradation processes described before for the mRNA happen with rates β_1 and β_2 for the free mRNA (reactions 2.18, 2.19), η_1 and η_2 when the mRNA is in a complex (reactions 2.20, 2.21). Finally, the degradation of the proteins are presented in reactions 2.22 and 2.23, with associated rates δ_1 and δ_2. Inspired by the set of ordinary differential equations obtained in the study of Cella F. [33], available in the supplementary material of the article [34], we write a ODE system represented by 2.24-2.30. The system of differential equations describing the changes in concentration of species in time, associated to the above mentioned system of reactions, is:

$$\frac{dm_1}{dt} = \alpha - k_1^+ m_1 R + k_1^- b_1 + \gamma_1 b_1 - \beta_1 m_1 \tag{2.24}$$

$$\frac{dm_2}{dt} = \alpha - k_2^+ m_2 R + k_2^- b_2 + \gamma_2 b_2 - \beta_2 m_2 \tag{2.25}$$

$$\frac{db_1}{dt} = k_1^+ m_1 R - k_1^- b_1 - \gamma_1 b_1 - \eta_1 b_1 \tag{2.26}$$

$$\frac{db_2}{dt} = k_2^+ m_2 R - k_2^- b_2 - \gamma_2 b_2 - \eta_2 b_2 \tag{2.27}$$

$$\frac{dR}{dt} = -k_1^+ m_1 R - k_2^+ m_2 R + k_1^- b_1 + k_2^- b_2 + \gamma_1 b_1 + \gamma_2 b_2 + \eta_1 b_1 + \eta_2 b_2 \tag{2.28}$$

$$\frac{dp_1}{dt} = \gamma_1 b_1 - \delta_1 p_1 \tag{2.29}$$

$$\frac{dp_2}{dt} = \gamma_2 b_2 - \delta_2 p_2 \tag{2.30}$$

To take into account randomness and correlations between species as explained in section 2.1, we switch to a probabilistic point of view of the system, for which we can write the associated Master equation (referring to equation 2.1). Suppose a system characterised by m_1 molecules of mRNA m1, m_2 molecules of mRNA m2, R free ribosomes, b_1 (b_2) molecules of complexes of m1 (m2), and p_1 (p_2) molecules of protein p1 (p2). Equation 2.31 accounts for the transitions into and out of the state $X = (m_1, m_2, b_1, b_2, R, p_1, p_2)$

and is the Master equation for the studied system.

$$\begin{aligned}
\frac{dP(m_1,m_2,b_1,b_2,R,p_1,p_2)}{dt} &= \alpha P(m_1-1,m_2-1,b_1,b_2,R,p_1,p_2) + \\
&+ \eta_1(b_1+1)P(m_1,m_2,b_1+1,b_2,R-1,p_1,p_2) + \\
&+ \eta_2(b_2+1)P(m_1,m_2,b_1,b_2+1,R-1,p_1,p_2) + \\
&+ \beta_1(m_1+1)P(m_1+1,m_2,b_1,b_2,R,p_1,p_2) + \\
&+ \beta_2(m_2+1)P(m_1,m_2+1,b_1,b_2,R,p_1,p_2) + \\
&+ \delta_1(p_1+1)P(m_1,m_2,b_1,b_2,R,p_1+1,p_2) + \\
&+ \delta_2(p_2+1)P(m_1,m_2,b_1,b_2,R,p_1,p_2+1) + \\
&+ k_1^+(m_1+1)(R+1)P(m_1+1,m_2,b_1-1,b_2,R+1,p_1,p_2) + \\
&+ k_2^+(m_2+1)(R+1)P(m_1,m_2+1,b_1,b_2-1,R+1,p_1,p_2) + \\
&+ k_1^-(b_1+1)P(m_1-1,m_2,b_1+1,b_2,R-1,p_1,p_2) + \\
&+ k_2^-(b_2+1)P(m_1,m_2-1,b_1,b_2+1,R-1,p_1,p_2) + \\
&+ \gamma_1(b_1+1)P(m_1-1,m_2,b_1+1,b_2,R-1,p_1-1,p_2) + \\
&+ \gamma_2(b_2+1)P(m_1,m_2-1,b_1,b_2+1,R-1,p_1,p_2-1) - \\
&- P(m_1,m_2,b_1,b_2,R,p_1,p_2)[(\eta_1 + \gamma_1 + k_1^-)b_1 + (\eta_2 + \gamma_2 + k_2^-)b_2 + \\
&+ k_1^+ m_1 R + k_2^+ m_2 R + \beta_1 m_1 + \beta_2 m_2 + \delta_1 p_1 + \delta_2 p_2] \quad (2.31)
\end{aligned}$$

The Master equation 2.31 is highly intricate and poses a formidable challenge in terms of finding a solution through conventional methods such as generating function or numerical methods. Acknowledged the complexity of the equation, the most practical and efficient way to obtain results is to resort to simulation techniques such as Gillespie algorithm [5], and its implementation in Julia programming language [32].

Chapter 3

Parameters Estimation in Gene Regulation Model

Recalling the reactions based on which the model is built, i.e. reactions (2.11-2.23), the next necessary step to define the model is to determine parameters.
Since our experimental data are obtained with HEK293t cells, the same cell line is used to perform wet-lab experiments to determine model's parameters. The most immediate rates we can estimate are the translation rates (γ_1 and γ_2) for the two fluorescent proteins, mCherry and eYFP. Knowing the length of the mRNAs coding for the proteins (data available on SnapGene [35]), and knowing the elongation rate, i.e. the rate at which a ribosome translates a codon in cells used in experiments (elongation rate for HEK293t is available on the database Bionumbers [36]), it is immediate to calculate the translation rate.

Ribosomes in HEK293 cells translate ~ 3 codons/s [36] and the molecule of mRNA of mCherry is 711 bp (base-pairs) long, while eYFP mRNA is 720 bp long [35]. As a consequence, a molecule of mCherry is translated in nearly ≈ 79 s, whereas a molecule of eYFP is translated in approximately

$\approx 80\ s$; the rates at which a protein molecule of each species is produced are evaluated to be $\gamma_1 \approx 0.01266\ s^{-1}$ and $\gamma_2 \approx 0.0125\ s^{-1}$. The evaluation of the translation rate for a mRNA based solely on the speed of ribosomes and the length of the mRNA strand might oversimplify the complex process of protein synthesis. It is crucial to acknowledge the intricate journey that mRNA undergoes before reaching its final translated state into a fluorescent protein. The mRNA strand not only goes through refinement processes such as splicing, but it also needs to traverse from the nucleus to the cytoplasm for translation. Additionally, the subsequent steps involving protein folding and maturation are integral to the complete development of a functional protein with observable fluorescence. Ignoring these complex operations in the translation process may lead to an inaccurate estimation of the translation rate. Therefore, it becomes imperative to reconsider the initial parameter estimation and incorporate a more comprehensive understanding of the mRNA strand's journey from the transcribed mRNA to the creation of its fully matured and fluorescent protein state. In line with existing literature, we estimate the parameter to be determined by the longest time observed across the various processes involved in protein synthesis from mRNA strand, encompassing mRNA splicing, translocation from nucleus to cytoplasm, and the subsequent steps of protein folding and maturation. The maturation process requires the most time, making the rate equivalent to the duration of this particular step. Maturation time for mCherry and eYFP are estimated as $\simeq 50 - 60\ h$ [37], corresponding to translation rates $\gamma_1 = \gamma_2 \simeq 0.0002\ s^{-1}$.

3.1 Parameters calculated from experiments.

3.1.1 Experimental Setup

Experiments are performed at IIGM - Italian Institute of Genomic Medicine, whose laboratories are located at IRCCS in Candiolo (TO).

HEK293 Tet-off cells are employed during the experiments aimed at estimating degradation rates; these cells are derived from human embryonic kidney and are commonly used in biological research. Cells are maintained in DMEM (Gibco) supplemented with 10 % FBS serum.
To insert genes encoding for fluorescent proteins in the cell, transfection is needed: transfection consists in introducing inside the cell exogenous material, in our case plasmids that contain genes for mCherry and eYFP, expressed by the same promoter. For our experiments, Effectene® (Qiagen) is employed as transfection agent. Transfections were carried out on 6-well plates for both experiments, adding 2 μg of encoding plasmids in each plate, according to manufacturer's instructions.

3.1.2 Degradation rate of mRNAs

To determine the degradation rates of the two mRNAs molecules, quantitative PCR experiments are carried out. The aim of the experiment is to determine how fast the two mRNAs are degraded inside cells: to achieve this, it is necessary to block transcription and quantify the mRNA over time.
In a *Tet-off* cell line as HEK293t, blocking transcription is easily achieved inserting in the medium the antibiotic doxycycline in a concentration of 1 $\mu g/ml$ of medium. *Tet-Off* stands for "tetracycline off", indicating the possibility to regulate gene expression in response to the presence or absence of tetracycline or its derivative, doxycycline.
In the *Tet-Off* system, a key component is the tetracycline-controlled transactivator (tTA) protein. The tTA protein binds to a specific DNA sequence, which is placed upstream of the gene of interest in the cell's genome. When tetracycline or doxycycline is absent, the tTA protein can bind to the specific DNA sequence and activate transcription of the gene. However, in the presence of tetracycline (or doxycycline), these molecules bind to the tTA protein, causing a conformational change that prevents it from binding to the DNA sequence, inhibiting gene transcription [10].

Two 6-well plates are transfected 24 hours after plating; the encoding plasmids for mCherry and eYFP are transfected using Effectene® and 2 μg of DNA per well. To study the degradation in time of mRNA, doxycycline is administered to eleven of the twelve transfected wells at different times, the twelfth well is kept as control. Right after the last administration, RNAs are extracted from all wells using *RNeasy Mini Kit* (Qiagen) and quantified: only samples with sufficient concentration of RNA are considered.
To quantify how the mRNA varies over time, a further step is necessary before performing quantitative PCR. Reverse transcription converts RNA molecules into complementary DNA (cDNA) strands: the resulting cDNA can be then amplified using qPCR.
Quantitative Polymerase Chain Reaction (qPCR) is a technique used to quantify the amount of specific RNA sequences (mCherry and eYFP in our case) in the biological sample extracted over time.
The amplification process in qPCR is identical to that of traditional PCR. In a single tube, components necessary for the amplification reaction are mixed with a segment of target DNA that acts as a template (in our experiment cDNA obtained with reverse transcription of extracted RNA): these include fluorescent dye probe BrightGreen, forward and backward primers of the target genes, and RNase/DNase free water. Following this, the reaction's contents go through a number of temperature and time-dependent processes, including primer annealing, denaturation, and extension (see scheme in Figure 3.1). The DNA template is amplified exponentially once this set of procedures is carried out various times [38].

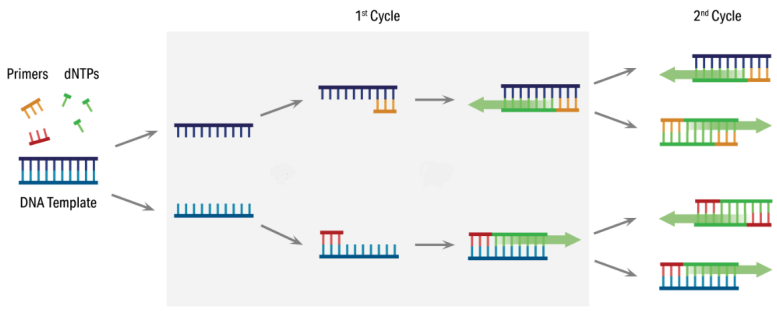

Figure 3.1: Illustrative scheme of three-step amplification procedure of qPCR: denaturation, annealing, and extension.
(Adapted from https://www.stratech.co.uk/aat-bioquest/real-time-pcr-qpcr/ [38]).

qPCR is performed for each time sample, for mCherry, eYFP and housekeeping gene (GAPDH), used as control measure and necessary to evaluate the fold-change. Data obtained from qPCR experiments consist in quantification cycles (Ct), the number of PCR cycles required for the fluorescence signal coming from the dye probe to achieve a threshold level for each sample. Lower Ct values imply that the target DNA is more abundant in the sample. Obtained data are analyzed using the ΔΔCt ("*delta delta Ct*") method:

1. **Calculate ΔCt:** ΔCt represents the difference in Ct values between the target gene (mCherry or eYFP) and the housekeeping gene (GAPDH), for each sample: $\Delta Ct = Ct_{target} - Ct_{housekeeping}$.

2. **Calculate ΔΔCt:** ΔΔCt is used to compare the ΔCt values between the sample result and the control result (the control well not treated with doxycycline).
   It is calculated as $\Delta\Delta Ct = \Delta Ct_{target} - \Delta Ct_{control}$

3. **Calculate Fold Change**: The fold change in gene expression between the experimental and control samples is calculated using the formula:

   $$\text{Fold Change} = 2^{-\Delta\Delta Ct}$$

   A fold change greater than 1 indicates upregulation, while a fold change

smaller than 1 indicates downregulation of the target gene in the experimental sample compared to the control.

Data obtained with this procedure for both mRNAs are used to interpolate an exponential decreasing function $f(t) = e^{-rt}$, since we expect that mRNA degrades exponentially, using *LsqFit* package in Julia.
The degradation rate is calculated from the half-life time with equation 3.1:

$$rate = \frac{\log(2)}{r_{\text{opt}}} \tag{3.1}$$

where r_{opt} is the parameter obtained from the fit.
For each time step, three samples are examined and before applying the $\Delta\Delta$Ct method, and the mean of the samples is calculated. In Figure 3.1.2, results for degradation rates of mCherry (left) and eYFP (right) are displayed: black stars represent half-life time evaluated for each fit.

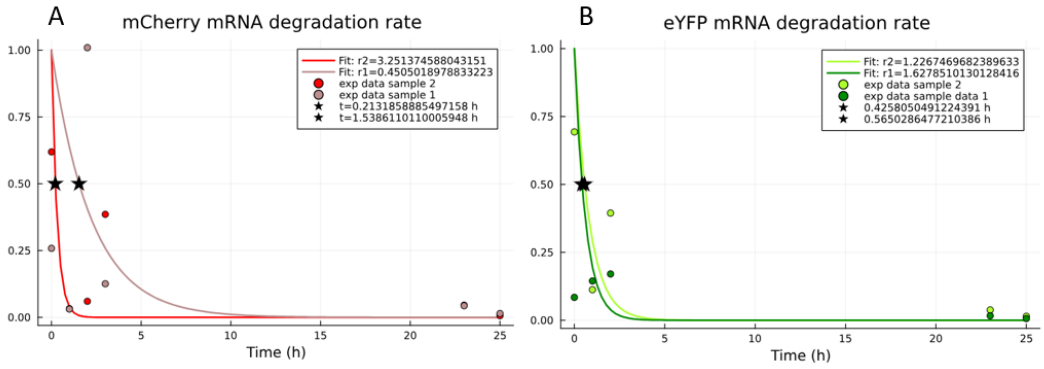

Figure 3.2: Fit and experimental data for mRNAs.
Experimental data (scatter points) of mCherry (A) and eYFP (B) obtained from qPCR, fitted with an exponential decreasing function. For each species, the data named *sample 1* are obtained evaluating the mean Ct over the triplicate, for each time point, whereas data labeled with *sample 2* are obtained averaging Ct, excluding the most different data for each triplicate.
(All data are normalized with respect to the first time point.)
Parameters r represent the best fitting parameters of the function $f(t) = e^{-rt}$ for each dataset. Black stars show the half-life time, the time necessary for normalized fluorescence to drop from 1.0 to 0.5.

From the obtained results we can estimate the degradation rate of mRNA approximately as $\beta_1 \approx 0.0005\ s^{-1}$ and $\beta_2 \approx 0.0003\ s^{-1}$ for the mRNAs associated to mCherry and eYFP, respectively. The obtained results are in agreement with the different lengths of the two genes, which influence not only the translation rates, but also their degradation rates.

3.1.3 Degradation rate of proteins

To study the degradation process of fluorescent proteins mCherry and eYFP, we examine the transfected cells over time with a fluorescence microscope, and measure the fluorescence signal analyzing the photos captured at the microscope. In a 6-well plate, a well is used as control and two wells as samples. Cells are transfected using Effectene® as described in section 3.1.2. After 24 hours from transfection, the medium is changed with medium containing doxycycline in a concentration of 1 $\mu g/ml$ to stop transcription and thus observe only degradation of proteins. Cells are followed for one week, changing medium every two days so that the action of doxycycline does not fade in time; photos are taken at fluorescence microscope twice a day: for each well, ten sectors are examined in red and green channels to capture both fluorescent proteins. Examples of the photos obtained are displayed in Figure 3.3.

Images from microscope are analyzed in MATLAB programming language, opened with Bio-Formats library and the ten sectors are concatenated to form two 3D matrices of integers (one for each channel), for each time point, for the two wells.

First of all, saturated pixels and their surroundings are excluded from analysis. Then, each image is segmented to understand which pixels belong to cells and what should not be taken into account and therefore considered as background. The images undergo a binarization process applying a thresholding technique to convert grayscale images into black and white images: pixels above a certain intensity level are set to white, while the others are set to black

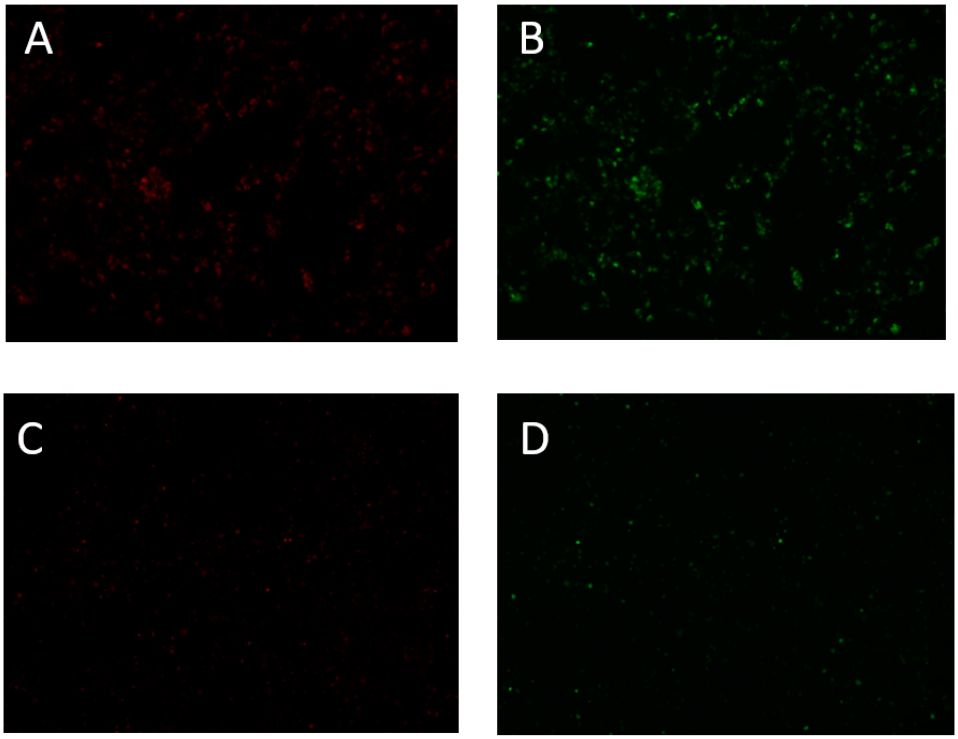

Figure 3.3: Example images obtained from fluorescence microscopy.

(A-B) Images taken right after the first administration of doxycycline.
(C-D) Images taken six days after the first administration of doxycycline.
Obs.: The two images are taken from the second sample well and do not represent the same cells.

and will belong to the background. To determine the optimal value of the threshold, an iteration procedure is implemented, increasing the threshold value at each iterative step; this iterative procedure concludes when the number of objects identified as cells does not increase anymore, ensuring that the optimal segmentation threshold is determined. In addition to this, during the analysis, objects identified as cell are retained only if their area falls within a specif range of areas in term of pixels, in order to exclude too large or too small structures, that would lead to inexact results. Once the segmentation is complete, background value is calculated as the mean of all pixels not segmented as cells plus twice the standard deviation of the

same data; this background is subtracted from each pixel in the images, to remove unwanted noise and evaluate the fluorescence of only relevant objects. To evaluate the mean fluorescence two possible strategies are implemented. The first one consists in considering only pixels that belong to a segmented object and are not saturated, summing their value and dividing by their total number (Figure 3.4 A, B). This process leads to very stable curves and for this reason we also implemented another strategy: summing all pixels that are not saturated, both segmented or not, and divide for the total number (Figure 3.4 C, D).

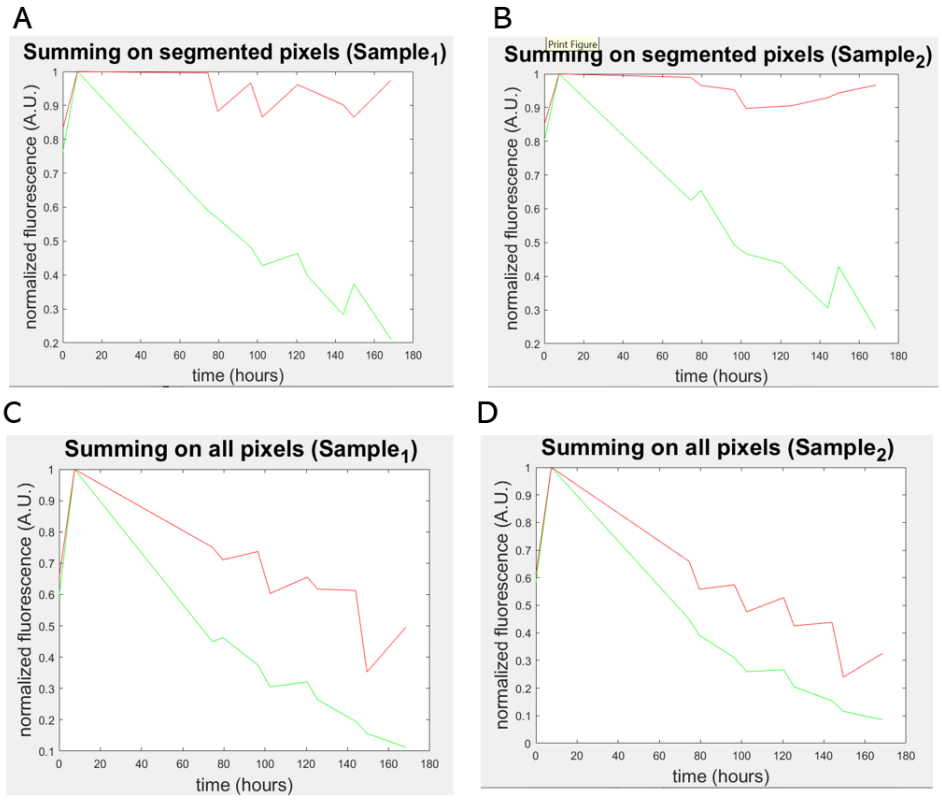

Figure 3.4: Fluorescence data obtained after segmentation of microscopy images.
Transfected cells are followed over a period of 1 week (170 hours) after the first administration of doxycycline.
(A-B) Fluorescence trajectories as a function of time, obtained summing only pixels that belong to a segmented object and are not saturated.
(C-D) Trajectories obtained summing all pixels that are not saturated, both segmented or not.
Red trajectories represent mCherry fluorescence, green trajectories represent eYFP fluorescence signal.

As clearly visible from Figure 3.4, degradation curves obtained summing only on segmented and non-saturated pixels are very stable signals, especially mCherry (Figure 3.4 A,B), but summing all non-saturated pixels produce faster decreasing curves for both proteins (Figure 3.4 C,D); this is due to the fact that at larger time, more and more cells that were fluorescent before are considered background since their signal smothers, but we divide by the total

number of pixels not saturated, which is bigger and bigger in time. In any case, both strategies lead to half-life time results far greater than half-life time found in literature of $24 - 48$ hours for fluorescent proteins [39].
Indeed, fitting results from our experiments using LsqFit package in Julia with an exponential decreasing function $f(t) = e^{-rt}$, we obtain curves reported in Figure 3.5: half-life times are indicated by black stars, $t_{1/2} \sim 60\ h$ for eYFP (Figure 3.5 B) and $t_{1/2} > 100\ h$ for mCherry (Figure 3.5 A); summing over only non saturated pixels leads to even higher half-life time, especially for mCherry where $t_{1/2} >> 100\ h$, as visible also from 3.4 A, B.

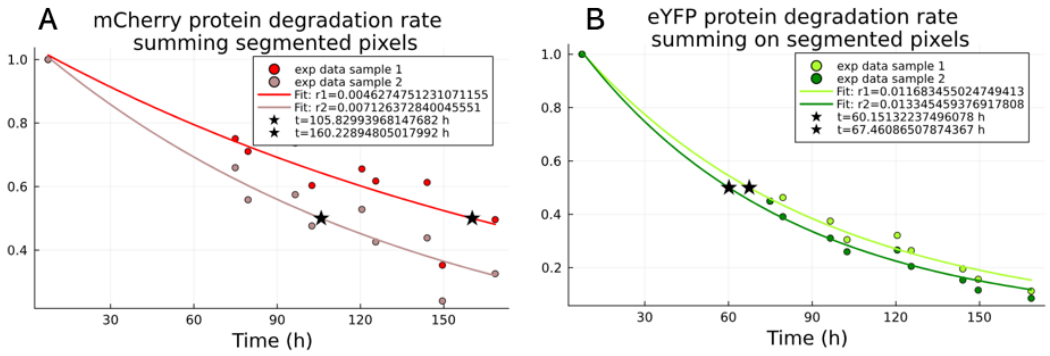

Figure 3.5: Fit and experimental fluorescence data for proteins obtained summing over segmented pixels.
Experimental data (scatter points) of mCherry (A) and eYFP (B) fluorescence for well 1 and well 2, fitted with an exponential decreasing function.
(All data are normalized with respect to the first time point.)
Parameters r represent the best fitting parameters of the $(f(t) = e^{-rt})$ for each dataset. Black stars show the half-life time, the time necessary for normalized fluorescence to drop from 1.0 to 0.5.

To better understand this discrepancy between half-life times obtained from our experiments and those found in literature, we have to analyze conditions in which experiments were carried out. Cells were plated in a 6-well plate, transfected and analyzed at microscope: when the microscopy observation started, cells were at confluence and a process called contact inhibition took place. Contact inhibition is a regulation process that occurs

in cells which stops cells proliferation when all place available is occupied [40]; consequently, the observed half-life time in our experiment reflects the degradation in non-diving cells. On the contrary, data at our disposal represents cells which divide and therefore it becomes evident that the predominant factor influencing the experimental degradation is the cellular division process. Accordingly, the degradation rate for fluorescent proteins, is assigned equal to the cell division rate: HEK293T cells divide every 48 h [41], meaning that the corresponding doubling rate is approximately $5.8 \times 10^{-6}\ s^{-1}$.
Consequently, the degradation rates in our model for mCherry and eYFP are $\delta_1 = \delta_2 \cong 5.8 \times 10^{-6}\ s^{-1}$.

3.2 Parameters inferred from experimental data.

Values of fluorescence for 19.001 HEK293t cells are available and each channel is stored in a .MAT file. First and foremost, experimental trajectories must be validated to ensure accuracy and reliability of the obtained results. Experimental data are indeed prone by nature to imperfection, noise and missing information. The importance of a validation process is highlighted in the context of our experimental data, which entails measuring an experimental variable over time. The nature of the cell cycle and its dynamics bring about new obstacles such as cell division or segmentation operations, which might lead to errors or distortions in the detected fluorescence. Validation procedure is thus essential to account for data gaps, sudden fluctuations or saturation of signals.
A trajectory validation function is designed to assess individual trajectories based on multiple criteria. First of all, trajectories containing a too elevated number of NaN values, corresponding to missing signal, are discharged. After that, we measure and discharge trajectories which saturate in one or both

channels: saturation can be caused by overexposure to excitation light or too high fluorophore concentration. Trajectories presenting sudden downwards jumps are also excluded; however, we acknowledge that during the experiment cell division can occur, to accommodate for this, trajectories showing a descending jump equivalent to approximately one-half of the starting fluorescence level are considered valid. This approach aim at distinguishing a biological event like cell division from other source of noise. If a trajectory presents more than five consecutive NaN values, it is considered valid only if it has a sufficient number (> 80) of non-NaN values after the consecutive NaN sequence. If the signal is not present for more than five time frames, it corresponds to more than one hour during which the cell was not followed and the signal not registered. During this time the cell may have divided or moved, so it is crucial to consider only trajectories that have a sufficient number of valid data after this sequence of NaN.
The validation procedure described above is performed on both channels and a cell is kept only if both trajectories (eYFP and mCherry) passed the validation check. A further adjustment is conducted on valid trajectories, shifting data to obtain trajectories starting with non NaN values.

3.2.1 Transcription rate of mRNAs

To evaluate the transcription rate of the two mRNAs, the analytical distribution of proteins studied by Swain and Shahrezaei in 2008 [42] is used to fit high level experimental trajectories. Following our reasoning and findings, we choose to fit high level trajectories since we expect that they correspond to cells with an infinite ribosome pool, necessary feature needed to estimate the transcription rate from distribution 3.2, obtained in an infinite resource model. They modelled a two-stage model where a single gene is transcribed and translated, which corresponds to the infinite model described by reactions 4.1-4.7, with only one species. By solving the associated Master equation, they found the probability to have n proteins at time τ, when $\gamma = \frac{\beta}{\delta} \gg 1$, is

represented by $P_n(\tau)$:

$$P_n(\tau) = \frac{\Gamma(a+n)}{\Gamma(n+1)\Gamma(a)}\Big(\frac{b}{b+1}\Big)^n\Big(\frac{1+be^{-\tau}}{b+1}\Big)^a \times {}_2F_1\Big(-n,-a,1-a-n;\frac{1+b}{e^{\tau}+b}\Big) \quad (3.2)$$

where $a = \frac{\alpha}{\beta}$, is the ratio between the transcription rate and the protein degradation rate, and $b = \frac{\gamma}{\beta}$, is the protein burst size, i.e. the ratio between the translation rate and the mRNA degradation rate. ${}_2F_1(a,b,c;z)$ is the hypergeometric function, and Γ represents the gamma function [42]. We want to fit experimental data with distribution 3.2 to infer the transcription rate α from the best fitting parameter a.

It is crucial to note a key assumption in our study: while our experimental data are based on fluorescence values, all probability distributions are described as a function of *proteins number*, meaning that we operate under the assumption that fluorescence is directly proportional to the number of proteins, i.e. $f \simeq c \cdot p$. This hypothesis is grounded in the fundamental relationship between protein concentration and fluorescence intensity. The emitted fluorescence signal is generally assumed to correlate with the concentration of fluorescent proteins present in a sample. Consequently, we need to re-scale experimental fluorescence data with the proportionality factor c between fluorescence and number of proteins. Following the study of van Oudenaarden A. and Thatthai M. [43], it is possible to evaluate the conversion factor between fluorescence and proteins number, analyzing experimental high-level trajectories evaluated at frame $k = 100$ (corresponding to about 39 h $\simeq 2360$ s) after transfection. According to the study [43]

$$< p >= a \cdot b \cdot (1 - e^{-\beta t}) \quad \text{and} \quad \frac{\delta p^2}{< p >} = \Big(\frac{1-e^{-2\beta t}}{1-e^{-\beta t}}\Big)(b+1) \quad (3.3)$$

where we recall that β is the degradation rate of the mRNA molecule, $a = \frac{\alpha}{\delta}$ and $b = \frac{\gamma}{\beta}$.

Assuming $< f > \simeq c \ < p >$, we can evaluate the conversion factor c as

$$c = \frac{\delta f^2}{< f >} \cdot \frac{1}{(b+1)\Big(\frac{1-e^{-2\beta t}}{1-e^{-\beta t}}\Big)} = 2886.25 \quad (3.4)$$

where δf^2 is the variance of our data, and $< f >$ is the mean value.
We scale the fluorescence data of high level trajectories of mCherry evaluated at frame $k = 100$ with the calculated factor c from equation 3.4, and fit the obtained number of proteins with the probability distribution function in equation 3.2. Additionally, we could also calculate the theoretical value of parameter a combining equations 3.4 - 3.3, obtaining

$$a = \frac{< f >}{c \cdot b \cdot (1 - e^{-\beta t})} \simeq 19.95 \tag{3.5}$$

Fit is performed using the Julia package *Optim*, minimizing the logarithmic likelihood between experimental data distribution (scaled up by the factor c) and theoretical distribution (equation 3.2). Results of the fitting for mCherry are displayed in Figure 3.6: the graph is obtained fitting experimental scaled data of high level trajectories, optimizing parameter a.
During this procedure, the parameter $b = \frac{\gamma}{\beta} = \frac{0.0002}{0.0005} = 0.4$ is kept constant, since it is already fully determined. Experimental data are displayed with *orange histograms*, *red line* represents the distribution obtained with the fit minimizing the likelihood, while *black dashed line* represents the distribution obtained with parameter a evaluated from equation 3.5.

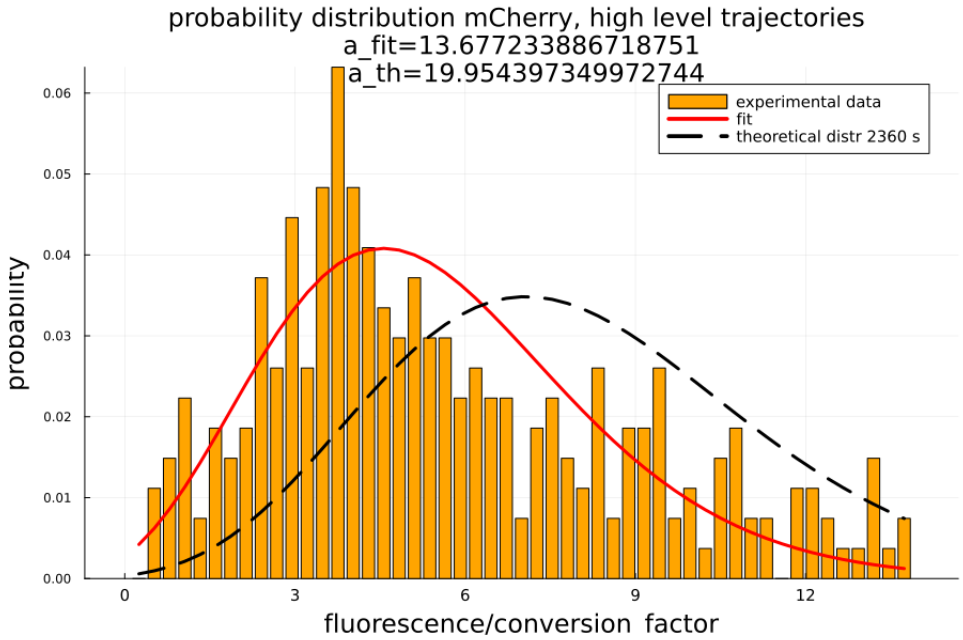

Figure 3.6: Probability distribution of proteins fitted with parameter *b* experimentally determined.
Probability distributions plotted as histogram of high-level trajectories of mCherry (*Orange bars*). Fit with distribution in equation 3.2 is represented by *red line*. Distribution with parameter a obtained from equation 3.5 is displayed with *black dashed line*.

Fitting the scaled experimental data, we obtain the value of parameter that better fits our data ($a = 13.68$), from which we obtain the transcription rate $\alpha = a \cdot \delta_1 = 13.68 \cdot 5.8 \times 10^{-6} \simeq 7.9 \times 10^{-5}\ s^{-1}$.

Chapter 4

Results

As stated in the Introduction, in a competitive environment where resource are distributed among multiple genes, an anticipated outcome is expected: an increase in the amount of genes competing for ribosomes (manifested in experiments by a higher number of transfected plasmids present within a cell), gives rise to a heightened demand for intracellular resources. The shared limited resources, particularly ribosomes in our model, are distributed across the numerous mRNA gene products, resulting in a consequent reduction in the overall level of gene expression. In our model, we replicate the increases competition within the cellular environment by reducing the ribosome pool size, leaving unchanged all other parameters.

As simulations show, reducing the size of the ribosomal pool has a strong impact on translation efficiency. Indeed, in the competing scenario, when ribosomal pool size ranges from 50 to 700, our results suggest how reductions in the ribosomal pool size directly lowers translation efficiency (see Figure 4.1).

We simulated the model conducting 10.000 independent simulations for different ribosomal pool sizes (all other parameters are fixed), measuring expected levels of proteins p_1 and p_2, when the system has reached steady state. Results of protein levels for p_2 are reported in Figure 4.1. Predictably,

the levels of gene expression for both fluorescence proteins diminishes as the ribosomal pool size reduces (Figure 4.1). This is due to the bottleneck effect that limited resources create inside the system, reducing the efficiency of gene expression.
The number of proteins translated diminishes heavily when the pool size is reduced: from more than 400.000 proteins obtained with a ribosomal pool of size $R = 700$ (see purple box in Figure 4.1) to 150.000 proteins reached when $R = 50$ (dark-red box in Figure 4.1).

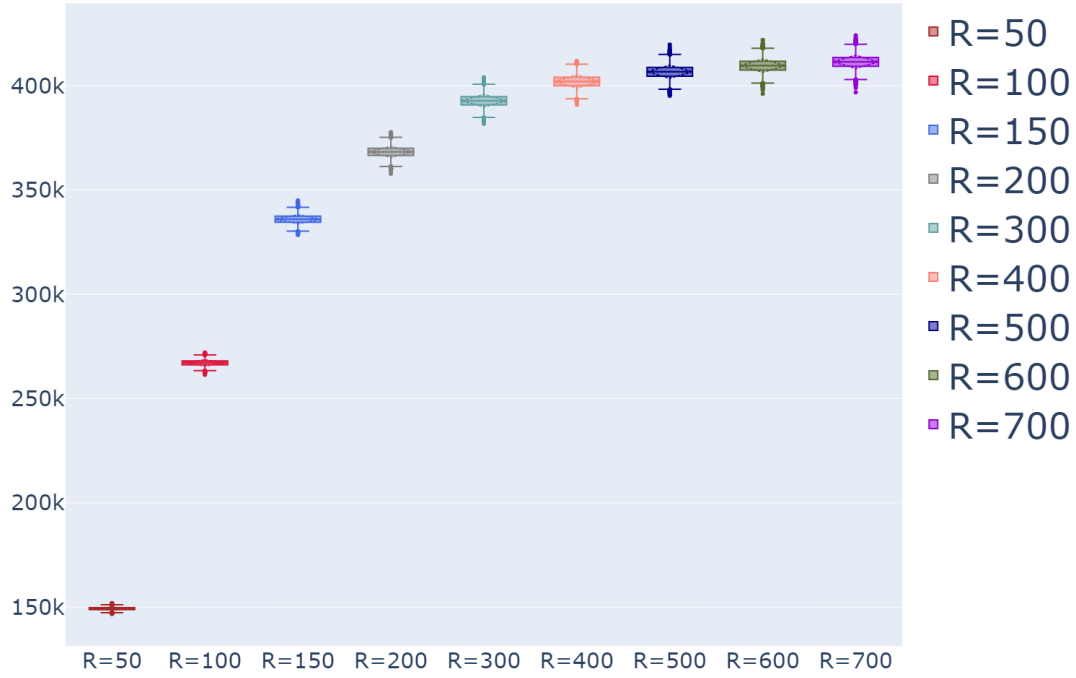

Figure 4.1: Size of ribosome pool shapes protein levels.
Reducing ribosomes pool size increases competition between genes, leading to a decreased translation efficiency.
Data obtained simulating the competing symmetric model with 10.000 simulations per pool size, with parameters *para*=$[0.03, 0.001, 0.001, 0.0009, 0.0009, 0.012, 0.012, 0.0005, 0.0005, 0.0005, 0.0005, 1.7 \cdot 10^{-6}, 1.7 \cdot 10^{-6}]$.

Since the simulated model is symmetric, where translation and degradation rates are equal for both species, we only display results for one protein, as the results depicted for p_2 are equal to those obtained for p_1.
Within our competitive environment, another feature affected by the limiting size of the ribosomal pool is the correlation between the two fluorescent proteins. Initially, their corresponding mRNAs are transcribed together, naturally resulting in a positive correlation between them ($\text{cor}(m_1, m_2) \approx 0.5$). In an ideal scenario characterized by an infinite pool of molecular resources, this correlation persists also between the proteins (black dashed line in Figure 4.2), taking into account that the translation process can lead to a minor reduction in correlation with respect to mRNAs.
What proves astonishing, however, is the rapid decrease observed in protein-protein correlation as we manipulate the ribosomal pool size, ultimately reaching negative value, indicative of anti-correlated proteins (Figure 4.2). When the pool has a non-limiting size (cf. $R = 1000$ in Figure 4.2), correlation results are in accordance with those obtained in a infinite pool model simulated as described in reactions 4.1 - 4.7, depicted with dashed black line in Figure 4.2. The correlation reduction as a function of the pool size is particularly noteworthy when recalling the initial correlation of the mRNAs at transcriptional stage.

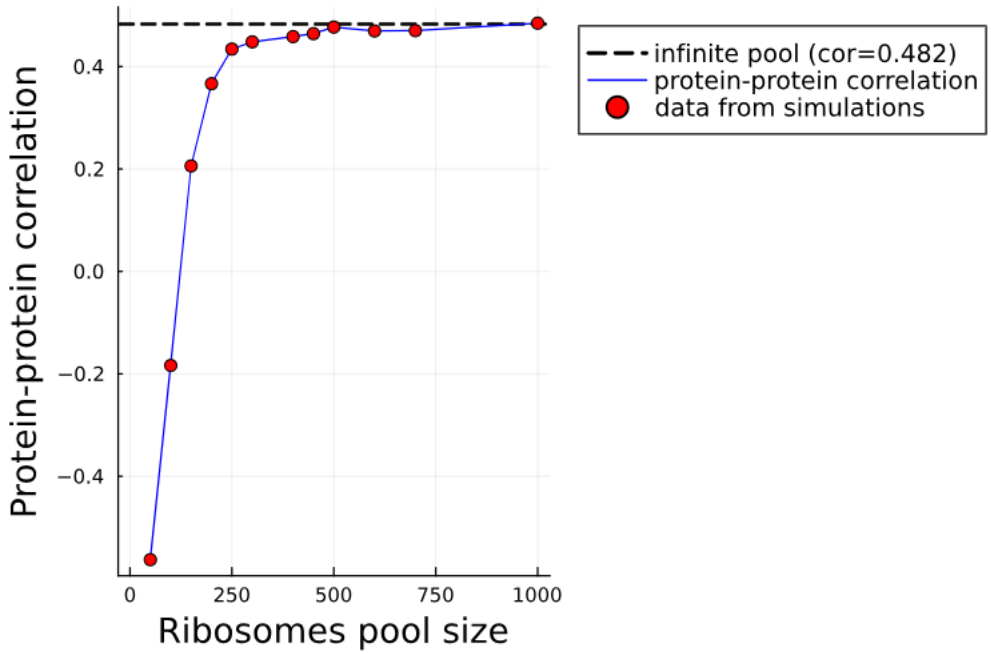

Figure 4.2: Correlation between the two proteins as a function of ribosomal pool size (simulations).
Reducing ribosomal pool size increases competition among genes. This competition produces a reduction of the correlation between the two proteins as the pool shrinks.

What has emerged is that the number of ribosomes imposes a significant constraint on the cellular system. In line with this idea, it is crucial to show that a lower size of ribosomal pool results in a surplus of untranslated mRNA molecules. This proposition indeed arises from the competing scenario where different mRNA molecules compete for limited resource, ultimately leading to a reduction in the overall protein expression levels.
To further support our thesis stating that a higher competition leads to an excess of untranslated mRNA molecules, we can plot the mRNA and protein distributions for different pool sizes and compare them with results obtained from an "infinite ribosomes" model.
For the sake of clarity, an "infinite ribosomes" model simply describes a gene which is first transcribed into mRNA and then translated into protein with constant rates. In this model, we disregard reactions involving ribosomes,

since we assume that they are abundantly available and do not pose a bottleneck to the gene expression process. The basic reactions characterising the model are the following:

$$0 \xrightarrow{\alpha} m_1 + m_2 \quad \text{transcription of correlated mRNAs} \tag{4.1}$$

$$m_1 \xrightarrow{\gamma_1} m_1 + p_1 \quad \text{translation of } m_1 \text{ into a protein } p_1 \tag{4.2}$$

$$m_2 \xrightarrow{\gamma_2} m_2 + p_2 \quad \text{translation of } m_2 \text{ into a protein } p_2 \tag{4.3}$$

$$m_1 \xrightarrow{\beta_1} 0 \quad \text{degradation of a molecule of } m_1 \tag{4.4}$$

$$m_2 \xrightarrow{\beta_2} 0 \quad \text{degradation of a molecule of } m_2 \tag{4.5}$$

$$p_1 \xrightarrow{\delta_1} 0 \quad \text{degradation of a protein } p_1 \tag{4.6}$$

$$p_2 \xrightarrow{\delta_2} 0 \quad \text{degradation of a protein } p_2 \tag{4.7}$$

We performed simulations of the competing model, gradually increasing the size of the ribosomal pool. The aim is to show that, while total mRNA levels (free mRNA and mRNA in complex with ribosomes) between infinite and competing models are not influenced by the pool sizes, heightened competition leads to a proportional reduction in protein expression levels, as displayed in Figure 4.3. When the pool is small ($R = 100$ in Figure 4.3), the difference between protein levels in the competing model (*green histograms*) and the infinite model (*purple histograms*) is elevated; on the other hand, when the pool in not limiting ($R = 1000$ in Figure 4.3), the protein level reached in the competing model approaches the result of infinite model.

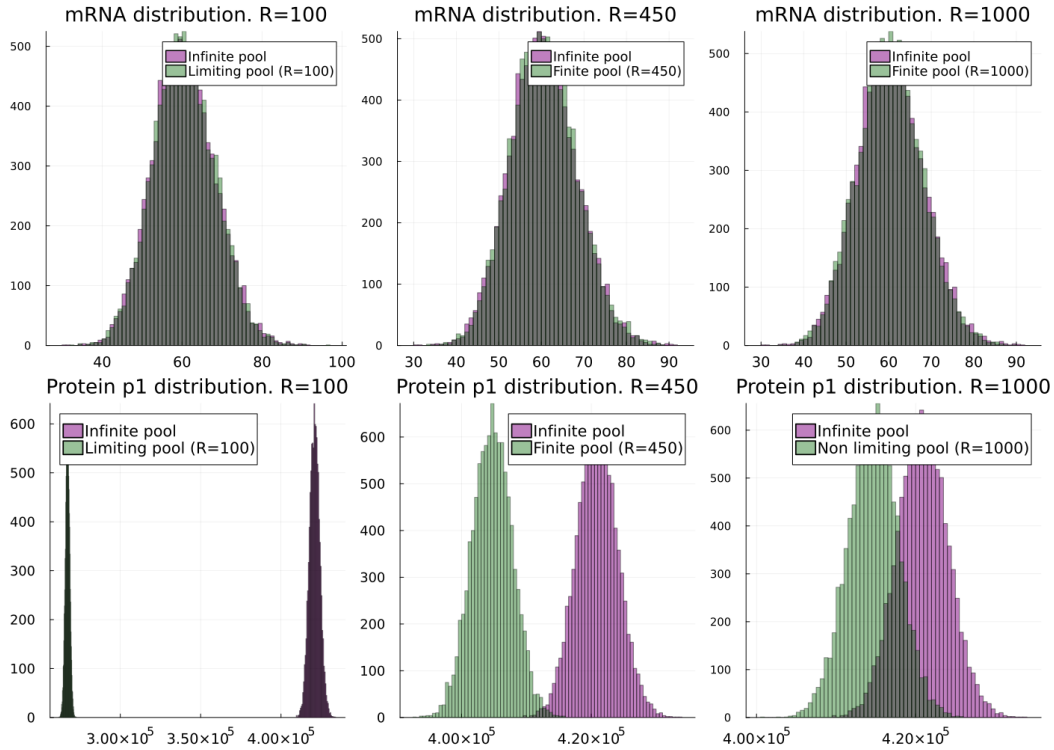

Figure 4.3: Distributions of mRNA and proteins for different pool sizes compared with infinite model results.
(*Upper panel*) Probability distributions of mRNAs in competing models (green) obtained summing free mRNAs and mRNAs in complex for one species evaluated with different pool sizes, compared with probability distributions of mRNA from infinite model (purple). mRNA transcription is not influenced by ribosome pool.
(*Lower panel*) Probability distributions of one protein in competing models (green) evaluated with different pool sizes, compared with protein probability distributions in a infinite model (purple).
Competition between mRNA molecules leads to a surplus of untranslated mRNA molecules, thus resulting in lower levels of protein expression. Increasing the pool size, we recover the infinite model results when the pool becomes non-limiting.

Having outlined our expectations for a competing environment in cells through simulations, the next step consists in seeking empirical validation of the anticipated behaviours: a limited resource pool results in poorer protein translation and leads to a decrease in correlation between two fluorescent proteins. Trying to confirm a similar pattern of correlation in real scenario,

we analyzed experimental trajectories, validated as described in section 3.2. Valid trajectories are then divided into three groups depending on the protein expression level they reached: low, medium, high. Following our hypothesis that a stronger competition for resources among genes reduces translation efficiency, we would expect, on general basis, that trajectories reaching a lower protein level correspond to the competing regime. Furthermore, a lower correlation between the two proteins is expected in a competing environment, as obtained in simulations (cf. Figure 4.1 and 4.2).
Protein-protein correlation at different time frames is calculated for trajectories belonging to the aforementioned three categories. Results obtained are partially in agreement with theoretical expectations and they are depicted in Figure 4.4: trajectories with lower expression level (*orange line*) show a diminished correlation between proteins with respect to trajectories with high protein levels (*blue line*), but correlation values are more elevated with respect to those obtained in simulations.

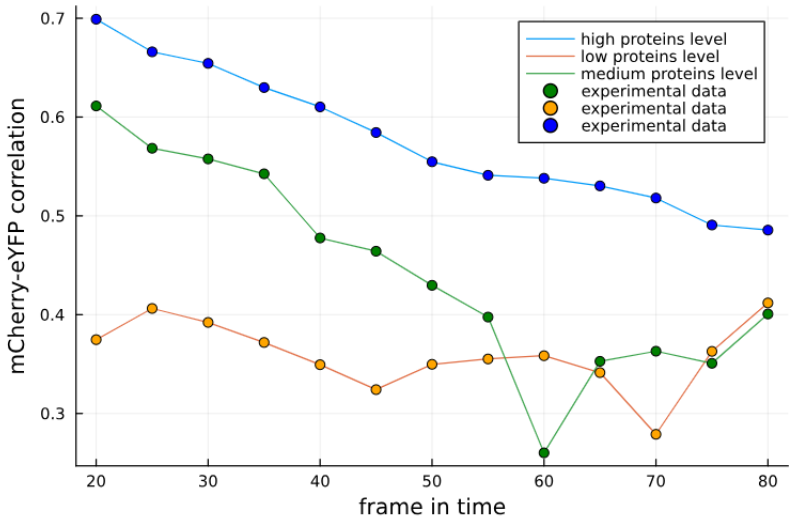

Figure 4.4: Correlation between the two proteins evaluated in experimental data, at different time points.
Correlation between mCherry and eYFP is higher in trajectories reaching high protein levels (blue line); lower protein levels (orange line) are associated to a decreased correlation between proteins with respect to high levels trajectories. Trajectories associated with a medium level of gene expression (green line) exhibit a correlation value between two proteins that falls within the range defined by trajectories with high and low expression levels at nearly all examined time frames.

The obtained results have to be handled with care: a detailed description of problems which arised during the analysis of experimental data is given in the following, and reasons because the obtained results in Figure 4.4 does not completely align with model results are presented in the ensuing discussion.

First of all, one must consider that trajectories were divided based on their protein levels: this criterion, however, does not distinguish a trajectory obtained from a cell which expresses low protein levels because it has assimilated a very low number of encoding plasmids from a low levels trajectory caused by very high competition among genes due to the elevated number of encoding plasmids.
As a matter of fact, this scenario can be also obtained through simulations modifying the transcription rate of correlated mRNAs in the infinite resource

model; the same protein level can be reached simulating a competing model with finite and limiting ribosomal pool (Figure 4.5 *green histograms*) and with an infinite model where the transcription rate has been lowered to achieve the same protein levels at steady state (Figure 4.5 *purple histograms*). Data displayed in Figure 4.5 for the finite pool model show a surplus of untranslated mRNAs, in agreement with our findings about competition lowering gene expression (cf. Figure 4.1).

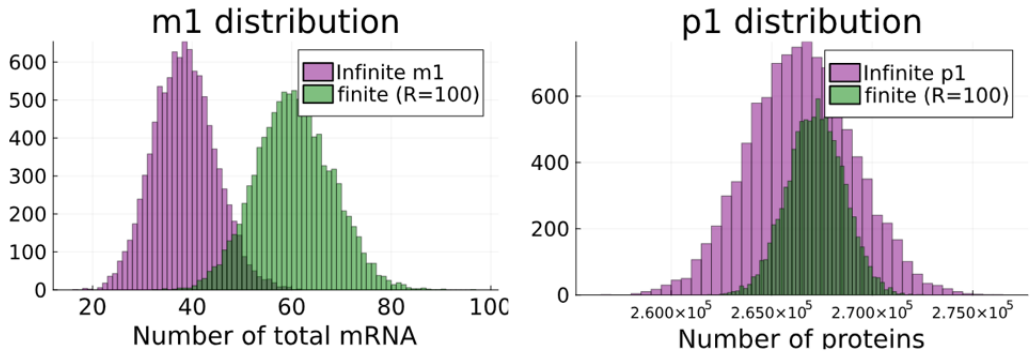

Figure 4.5: Model with limiting pool and infinite pool model can lead to same protein levels.
A low level of protein expression can be determined by two possible situations: an infinite model with a low transcription rate which corresponds to a cell with few plasmids to be expressed (purple), or a cell expressing low protein level because many plasmids are competing for the finite ribosomes pool (green); the latter is modelled with a finite pool model, diminishing the pool size to R=100.

Moreover, high and low level trajectories frequently displayed a different behaviour in their profiles. More specifically, high level trajectories tend to extend for a greater number of time frames, whereas trajectories associated with low protein levels often end early in time. This difference in trajectory lengths introduces a potential inaccuracy in the evaluation of correlations, particularly over the last evaluated frames.

If we compare correlation results from simulation of Figure 4.2 and from experimental data in Figure 4.4, one may argue that even if a decrease in correlation between high and low level trajectories is visible, correlation

values in both cases are higher with respect to those obtained in simulations. To interpret and explain this discrepancy between theoretical and experimental outcomes, one must take into account that theoretical results depicted in Figure 4.2 are derived sampling each simulation when it has reached the steady state, while we evaluated correlation in experimental trajectories at different time frames; experimental data often show increasing trend indicative of the transient state present before reaching steady state. Accordingly to this, we decided to sample simulations at different times in the transient state, to gasp some more information about correlation behaviour far from the steady state. We simulated two scenarios, one with a limiting ribosomal pool $R = 100$ and another where the pool is finite but not limiting $R = 1000$, and evaluated the correlation between proteins p_1 and p_2 at different time points during the transient (see Figure 4.6). It appears evident that during the transient state the two proteins have a correlation value higher with respect to the value reached at the steady state. Correlation diminishes during to transient state, reaching steady state value. The rapid decrease is more enhanced in the limiting pool with $R = 100$ ribosomes (Figure 4.6 A), where correlation drops from $\simeq 0.70$ at 500 s to $\simeq -0.15$ at 50.000 s.

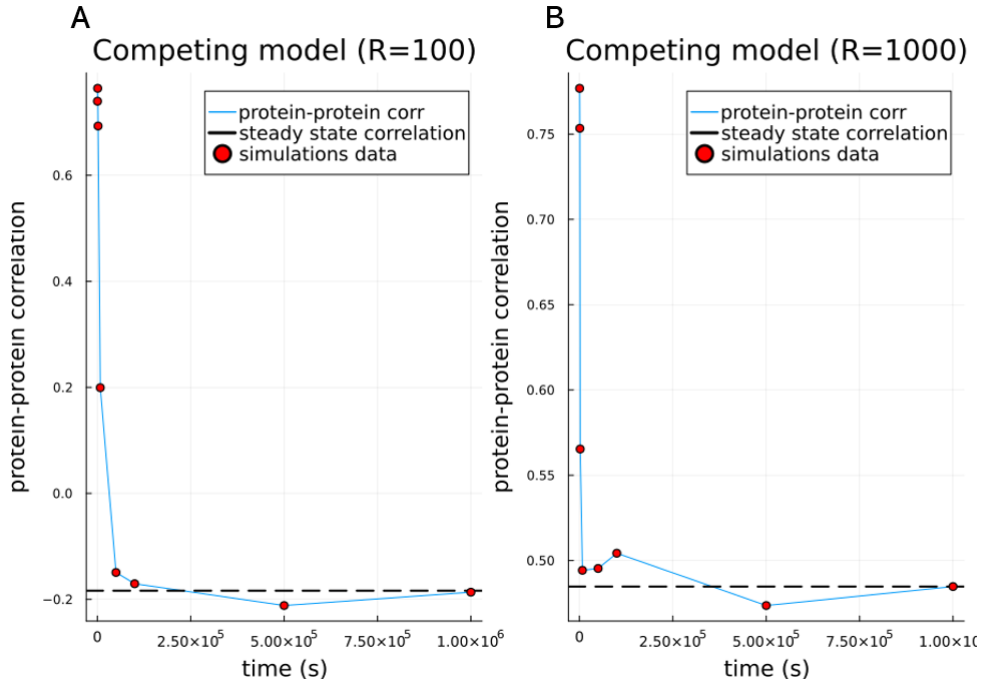

Figure 4.6: Correlation between proteins during the transient state.

Correlation between proteins at different time points during the transient shows that correlation has higher values during the transient state and it reaches the steady state correlation value (black dashed line) with a rapid decrease.
(A) Simulations performed with a limiting ribosome pool of size R=100.
(B) Simulations performed with a non limiting ribosome pool of size R=1000.
Simulations are sampled at times $[1 \times 10^3,\ 2 \times 10^3,\ 8 \cdot 10^3,\ 5 \times 10^4,\ 1 \times 10^5,\ 5 \times 10^5,\ 1 \times 10^6]$ s; steady state is evaluated at time 3×10^6 s.

Since we have proved that in both competing or infinite resource scenario, the correlation evaluated during the transient state is higher with respect to steady state values (see Figure 4.4), the difference in value between theoretical results from simulation in Figure 4.2 and experimental results in Figure 4.4 can be accounted by the presence of transient behaviour in numerous trajectories. This dynamics results in a general augmentation of the addressed correlations. This increase is more pronounced in the first time frames, reflecting the initial transient before reaching steady state of the fluorescence trajectory.

Another issue encountered during the inference of model's parameter concerns the value found for the transcription rate and the proportionality factor between fluorescence and number of proteins c. In section 3.2.1 we derived $\alpha \simeq 7.9 \times 10^{-5}\ s^{-1}$, and $c \simeq 2886$. However, the obtained results do not allow us to interpret experimental results in a correct way. Indeed, if we scale up by the same factor c the experimental data of low level trajectories, assuming that the proportion coefficient between fluorescence and number of protein is a fixed quantity, we obtain the distribution in Figure 4.7. These scaled up data suggest that the mean value of mCherry protein per cell, in cells expressing a low level fluorescence is smaller than 1, meaning that some cell may have not assimilated any encoding plasmid. This result is in contrast with experimental observation, since all trajectories examined express a non-zero value of fluorescence.

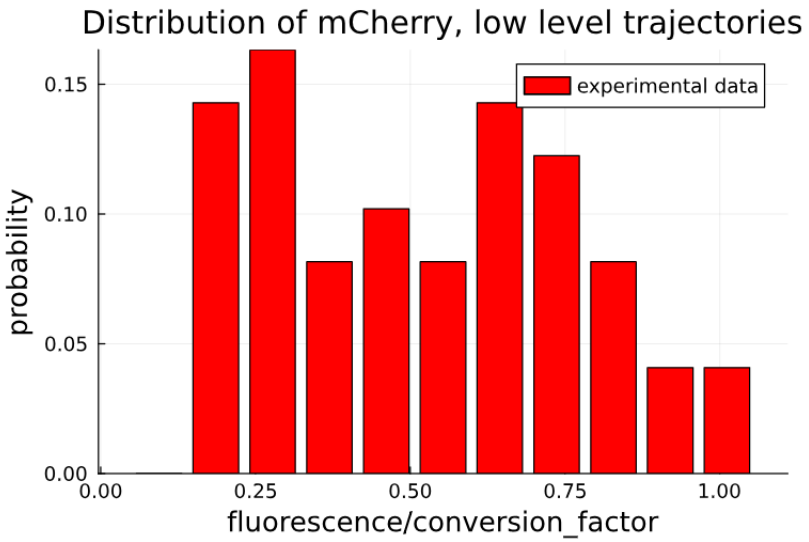

Figure 4.7: Probability distribution obtained scaling up by the obtained conversion factor c.
Scaling the fluorescence low level trajectories by the same factor obtained in section 3.2.1 lead to a maximum number of proteins per cell equal to 1.

This discrepancy leads us to question the efficacy of utilizing fluorescence

levels as the only parameter for distinguishing trajectories. It suggests that relying solely on the fluorescence level may not be a reliable or comprehensive means to categorize trajectory data based on the scenario (limited / infinite ribosomes) inside the cell.

All previous considerations strongly indicate that assessing experimental trajectories solely based on their fluorescence levels results in an imprecise categorization. The assumption linking a finite ribosome pool to reduced translation efficiency and lower protein production holds true only when analyzing cells with equal transcription rates. However, this assumption cannot be applied to our experimental data: since we are not able to assess how many plasmids each cell has assimilated, the effective transcriptional rate may vary from a cell containing many plasmids to one with fewer plasmids. In the category designated as low-level, where, according to our initial assumptions, cells with a restricted ribosomal pool were expected, we can find cells with an infinite pool but low transcription rate (see Figure 4.5), corresponding to cells with few encoding plasmids.

To correctly distinguish cells with a limiting ribosomal pool from those with an infinite pool, it becomes imperative to assess a new criterion to categorize trajectories. In this context, stochastic model simulations play a vital role. One of the key insights obtained during this Thesis is the observed decrease in proteins correlation when the ribosomal pool is limited. Building on this awareness, analyzing the correlation between proteins within the same experimental trajectory emerges as a valuable tool for accurately determining the trajectory category. However, a further difficulty emerges from the study of correlation as a function of time reported in Figure 4.6: in the transient state, proteins exhibit strong correlation both in the limiting and infinite pool, with differentiation in values based on the finite nature of the pool occurring only during the steady state. We model the potential process of measuring protein-protein correlation within an experimental trajectory. Through simulations, we assessed the correlation among proteins

within the same trajectory, from the start of the trajectory to a designated "end-time" (depicted in the Figure 4.8 on the x-axis in logarithmic scale). As evident, analyzing correlation including the transient state does not enable the differentiation between trajectories with finite or infinite pools. The distinction becomes evident only when evaluating correlation up to a significantly later time beyond the start of the steady state (dashed line in Figure 4.8).

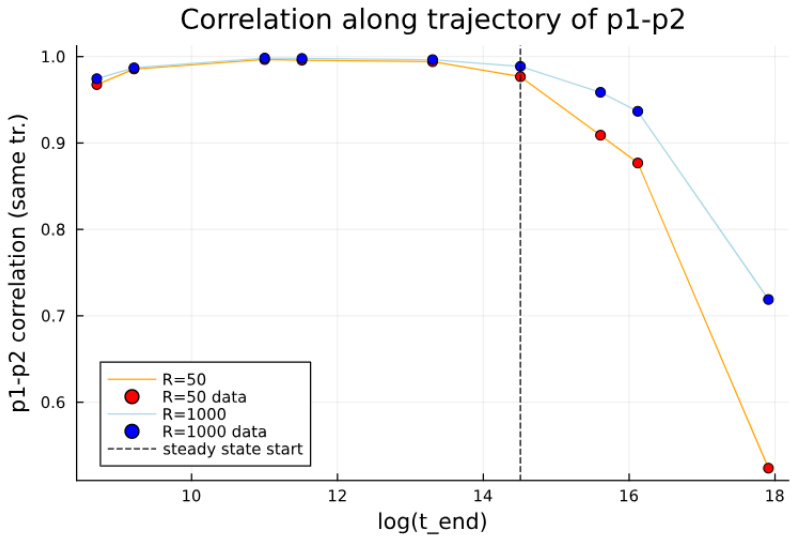

Figure 4.8: Proteins correlation evaluated up to different times, along the same trajectory.
Correlation between proteins along the same trajectory, stopping at increasing time. Correlation is calculated from start to an *end-time*: $[1 \times 10^3,\ 5 \times 10^3,\ 6 \times 10^4,\ 1 \times 10^5,\ 6 \times 10^5,\ ,\ 2 \times 10^6,\ 6 \times 10^6,\ 9 \times 10^6,\ 6 \times 10^7\ s]$.
The steady state is considered to start from $t = 2 \times 10^6\ s$ (*dashed black line*).
Protein- protein correlation diminishes and differentiates only during the steady state, reaching lower values for the limiting pool ($R = 50$ *orange line*). Non-limiting pool ($R = 1000$) is represented by *blue line*.

As depicted in Figure 4.8, it becomes apparent that a very long steady state phase is imperative for effectively discerning between limiting and non-limiting cases. However, experimental trajectories frequently present a

challenging scenario, characterized by extended transients and abbreviated steady states: some trajectories are depicted as example in Figure 4.9.

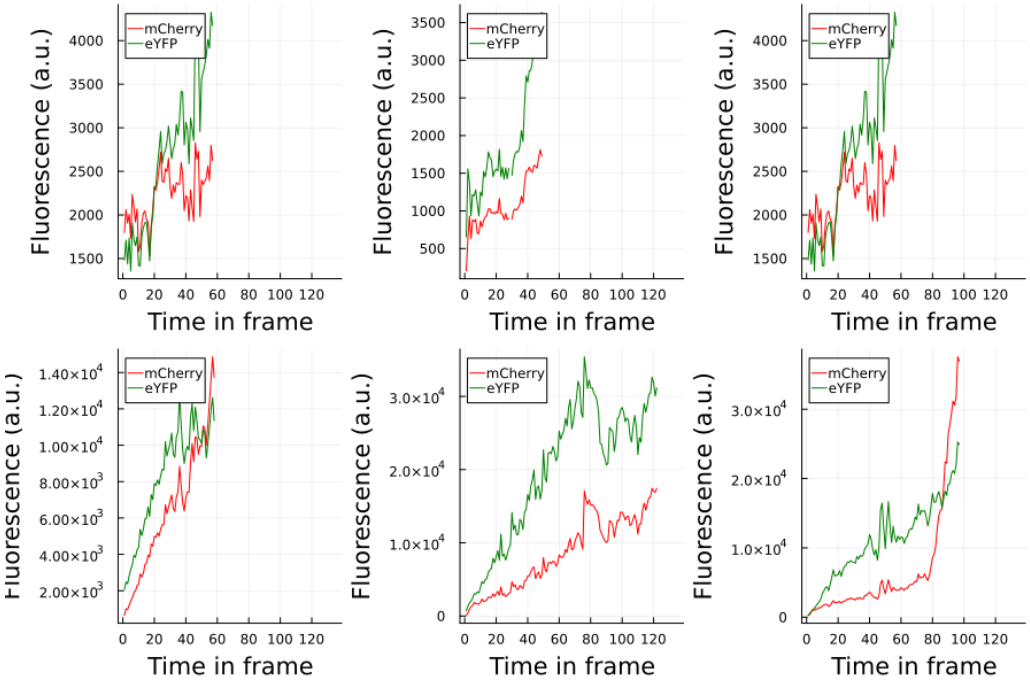

Figure 4.9: Examples of trajectorories showing more transient state than steady state values.
An issue encountered while analyzing experimental trajectories are the complex and various behaviours observed among different cells. Some cells display only transient behaviour or very few data in the steady state.

To address this challenge, we try to assess correlation along a simulated trajectory, but discerning between the transient and steady-state phases. The outcomes are illustrated in Figure 4.10, for a limiting ribosomal pool of size $R = 50$ (upper panel Figure 4.10 A,B) and for a non-limiting pool ($R = 1000$ in lower panel Figure 4.10 C,D). The graph represented the evaluation of the correlation (along the same trajectory) in 100 simulated trajectories for each pool, correlation mean values and standard deviations are displayed. Each point in the graphs A-C in Figure 4.10 corresponds to the mean value of correlations (for 100 simulations) evaluated from the start of the trajectory

up to the time measured on the x-axis.
The steady state is evaluated to start at time $t = 2 \times 10^6$ s from visual inspection of simulated trajectories; each point in graphs B-D in Figure 4.10 represents the mean value of correlations (for 100 simulations) evaluated from $t = 2 \times 10^6$ s up to the time represented on the x-axis. This representation allow us to state that if the correlation is evaluated in the transient state of a trajectory, it does not present any difference between the infinite pool and a limiting pool (see Figure 4.10 A-C). On the other hand, if we evaluate the correlation along the trajectory examining only the steady state, correlation between proteins is lower in the limiting case if compared to correlation evaluated on the same time interval for the infinite mode. It is important to observe that this difference is appreciable provided that time analyzed is sufficiently far from the start of the steady state, i.e. we have a sufficient number of points in the steady state. Figure 4.10 suggests that an effective strategy to differentiate cells with a limited pool from those with an infinite ribosomal pool is to assess protein-protein correlation along the trajectory, specifically during the steady state.

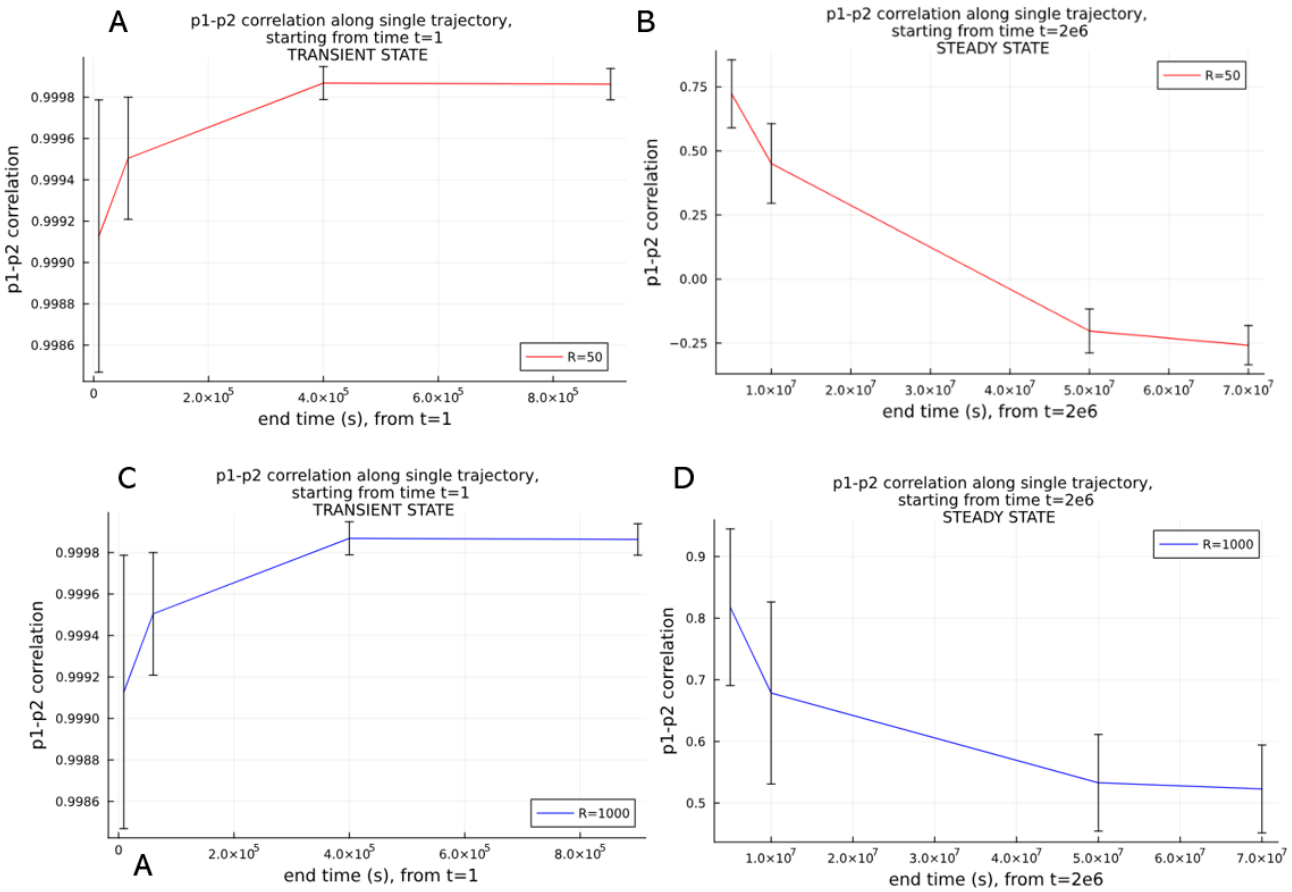

Figure 4.10: Proteins correlation evaluated along the same trajectory, during transient and steady state.
Correlation between proteins along the same trajectory, stopping at increasing times. (A-B) Limiting pool with $R = 50$. (C-D) Non-limiting pool with $R = 1000$.
Data are obtained performing 100 simulations for each case, mean values and standard deviations of the evaluated correlation are depicted.
During the transient state (A-C), correlation is $\simeq 1$ in both cases, difference in correlation between limiting- non limiting cases is appreciable if it is evaluated during the steady state (B-D).
A limiting pool (B) reduces the protein-protein correlation up to negative values, while a non-limiting pool (D) reduces correlation up to $\simeq 0.5$.

Categorizing trajectories based on the proteins correlation evaluated at the steady state for sufficient time can be a good criterion to distinguish cells with a limiting ribosomal pool from those where the pool is non-limiting.

With a better categorization, it would be possible to infer the transcription rate from a fit of trajectories belonging to the non-limiting pool case.

These considerations collectively contribute to a comprehensive understanding of differences between experimental results and simulation expectations. In light of these factors, this Thesis has examined in details the possible outcomes derived from competition for molecular resources among genes with simulation, and we found a partial alignment between theoretical and experimental results. It is however compulsory to acknowledge that the data analysis is to be improved and trajectories further classified on basis of protein-protein correlation evaluated during the steady state.

Chapter 5

Conclusions and Future Work

In conclusion, this Thesis has successfully investigated the effects of competition for molecular resources in gene expression, yielding significant insights into correlation profile and protein expression levels.
Competition for a finite resource pool diminishes protein expression level due to the limited availability of essential components necessary for translation process and it induces decrease in correlation between proteins that are translated from correlated mRNAs.

A more accurate method to analyze and categorize experimental trajectories is to be implemented, since we have proven that dividing trajectories based on their fluorescence level leads to only partially accurate results. We have also demonstrated that distinguishing trajectories based on protein-protein correlation evaluated on the single trajectory at the steady state could be a valuable categorization method to distinguish gene products in a competing environment from those in which the ribosomal pool can be assumed infinite. To achieve this result, an algorithm able to determine where the steady state starts for each experimental trajectory is to be defined, taking into account the noisiness and different behaviours of experimental

trajectories.

While this study has advanced our understanding of competition in gene expression, it is essential to acknowledge its limitations, first of all that the model created is basal, since it only considers one limited resource pool. Gene expression is a complex and multifaceted process influenced by various molecular resources. Looking ahead, future research should explore multiple finite pools scenario, building upon the foundations laid in this work. We have examined translational resource finiteness but many other resources can lead to bottleneck effects as we have proved for ribosomes. For instance, in addition to translation-related resources like ribosomes, degradation and transcription resources can significantly shape the regulatory landscape within a cell. The finite availability of degradation resource, including RNA-degrading enzymes (ribonucleases, exosome...) and degradation machineries, can play a pivotal role in determining lifetime of mRNA transcripts. In a scenario where multiple genes are competing for these degradation resources, an increase in transcription for one gene may lead to higher level of its mRNA, monopolizing the available degradation resource pool and consequently leading to an accumulation of other gene products, contributing to their higher expression levels. This scenario is however context-dependent and should be analyzed in details.
Transcriptional resources can also be considered as a limiting pool of molecular resources. RNA polymerases and transcription factors can influence the initiation and elongation of mRNA synthesis. If a gene is highly active and utilizes a significant portion of the available transcription resources, it can result in a reduced access for other genes seeking to initiate or enhance their transcription. This competition-induced limitation in transcription resources may lead to a decrease in the expression levels of genes that are not as actively engaged in this competitive process.

Competition for multiple resource pools introduces a layered complexity, influencing the rates of transcription, translation, and mRNA degradation

for individual genes. Consequently, the availability and prioritization of different resources can shape the overall gene expression landscape, leading to variations in the abundance of transcripts and proteins. This intricate interplay highlights the need for a more comprehensive understanding of how diverse cellular components collectively contribute to the regulation of gene expression. The presence of multiple finite pools can also help in a deeper understanding of experimental trajectories: each resource pool and the associated competition could contribute to a plateau in the trajectory of protein levels. As the gene encounters different resource limitations throughout its expression process, the protein abundance may reach plateaus at distinct levels, reflecting the availability of the specific resources at each stage.
Future research exploring scenarios involving modelling multiple finite resource pools will illuminate the intricate dynamics of these interactions, providing deeper insights into the orchestration of cellular processes and the fine-tuning of gene expression patterns.

Acknowledgements

First of all, I would like to express my gratitude to my thesis supervisor, Carla Bosia, for her guidance, invaluable insights, and continuous support throughout the research process, especially in difficult moments she has always believed in me and in my abilities. I wouldn't be able to achieve these results without the help of people working at Candiolo's lab, especially Elsi, who taught me, advise me and guided me through all my lab experience.

I am indebted to my parents and sister for providing the necessary resources to conclude my studies, and for creating an understanding and always encouraging environment for during the challenging periods of this academic journey.

A special thanks is due to Andrea, who has always believed in me and encouraged me throughout this journey.
Last but not least, I am grateful to all my friend who contributed in various ways, directly or indirectly, to the completion of this thesis. Your support has been invaluable, and I am deeply grateful to have you in my life.

Prima di tutto, vorrei esprimere la mia gratitudine alla mia relatrice di tesi, Carla Bosia, per la sua guida, per i preziosi suggerimenti e per il continuo supporto durante l'intera ricerca, soprattutto nei momenti difficili, in cui ha sempre creduto in me e nelle mie capacità. Non sarei riuscita a ottenere questi risultati senza l'aiuto delle persone che lavorano nel laboratorio di Candiolo, in particolare Elsi, che mi ha insegnato, consigliata e guidata attraverso tutta la mia esperienza in laboratorio.

Sono grata ai miei genitori ed a mia sorella, per aver fornito tutto il supporto e le risorse necessarie per concludere i miei studi, e per aver creato un ambiente

comprensivo e sempre incoraggiante durante i periodi difficili di questo percorso accademico.

Un ringraziamento speciale va ad Andrea, che ha sempre creduto in me e mi ha incoraggiata lungo questo percorso, supportandomi e sopportandomi.

Infine, ma non meno importante, sono grata a tutti i miei amici che hanno contribuito in molti modi, direttamente o indirettamente, al completamento di questo percorso. Il vostro supporto e la vostra presenza è stato importante per me, e sono sinceramente grata di avervi nella mia vita.

Bibliography

[1] Emilsson V., Thorleifsson G., and Zhang B. et al. «Genetics of gene expression and its effect on disease.» In: *Nature.* (2008). DOI: 10.1038/nature06758 (cit. on p. v).

[2] *3.6 Cellular Differentiation-Anatomy and Physiology.* URL: https://open.oregonstate.education/aandp/chapter/3-6-cellular-differentiation/ (cit. on p. v).

[3] Zoghbi H.Y. and Baudet A.L. «Epigenetics and Human Disease». In: *Cold Spring Harbor perspectives in biology.* (2016). DOI: 10.1101/cshperspect.a019497 (cit. on p. v).

[4] Paulsson J. «Models of stochastic gene expression.» In: *Physics of Life Revies.* (2005). DOI: 10.1016/j.plrev.2005.03.003 (cit. on p. vi).

[5] Gillespie D.T. «Exact simualtion of coupled chemical reactions». In: *Journal of Chemical Physics* 81 (25) (1977), p. 2340. DOI: 10.1021/j100540a008 (cit. on pp. vi, 14, 16, 22).

[6] Hahl S. K. and Kremling A. «A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: on fixed points, means, and modes.» In: *Frontiers in Genetics.* (2016). DOI: 10.3389/fgene.2016.00157 (cit. on p. vi).

[7] Mattaini K. *Introduction to Molecular and Cell Biology.* Roger Williams University PressBooks, 2020 (cit. on pp. 1–5).

[8] Dornell J. «RNA Polymerase: Function and Definition.» In: *Genomics Research From Technology Networks.* (2021). URL: https://www.technologynetworks.com/genomics/articles/rna-polymerase-function-and-definition-346823 (cit. on p. 1).

[9] Yamaguchi Y. «Encyclopedia of Systems Biology». In: Springer New York, 2013, pp. 2221–2224. DOI: 10.1007/978-1-4419-9863-7_1407 (cit. on p. 1).

[10] Reece R. «Analysis of Genes and Genomes». In: John Wiley and sons., 2004. ISBN: 0 470 84380 2 (cit. on pp. 2, 4, 25).

[11] Kozak M. «Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes.» In: *Cell.* 44 (1986). DOI: 10.1016/0092-8674(86)90762-2 (cit. on p. 2).

[12] Ma J. «Transcriptional activators and activation mechanisms.» In: *Protein Cell.* (2011). DOI: 10.1007/s13238-011-1101-7 (cit. on p. 4).

[13] Kwak H and Lis JT. «Control of transcriptional elongation.» In: *Annual Review of Genetics* 47 (2013), pp. 483–508. DOI: 10.1146/annurev-genet-110711-155440 (cit. on p. 4).

[14] O'Brien J., Heyam H., Yara Z., and Chun P. «Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation». In: *Frontiers in Endocrinology* 9 (2018). DOI: 10.3389/fendo.2018.00402 (cit. on p. 5).

[15] Bartel D. P. «MicroRNAs: Genomics, Biogenesis, Mechanism, and Function». In: *Cell* 116 (2023). DOI: 10.1016/s0092-8674(04)00045-5 (cit. on p. 5).

[16] Huntzinger E. and Izaurralde E. «Gene silencing by microRNAs: contributions of translational repression and mRNA decay.» In: *Nat Rev Genet* 12 (2011). DOI: 10.1038/nrg2936 (cit. on p. 5).

[17] Valinezhad Orang A. et al. «Mechanisms of miRNA-Mediated Gene Regulation from Common Downregulation to mRNA-Specific Upregulation.» In: *International journal of genomics.* (2014). DOI: doi:10.1155/2014/970607 (cit. on p. 5).

[18] Corley M., Burns M., and Gene W. «How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms». In: *Molecular Cell.* (2020). DOI: 10.1016/j.molcel.2020.03.011 (cit. on p. 5).

[19] *10.8: Regulation of Translation. Biology LibreTexts.* URL: https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Cells_-_Molecules_and_Mechanisms_(Wong)/10%3A_Translation/10.08%3A_Regulation_of_Translation. (cit. on p. 5).

[20] Hershey J. et al. «Principles of translational control: an overview.» In: *Cold Spring Harbor perspectives in biology* 4 (2012). DOI: 10.1101/cshperspect.a011528 (cit. on p. 5).

[21] Hardin G. «The Competitive Exclusion Principle.» In: *Science* 131.3409 (1960), pp. 1292–97. URL: http://www.jstor.org/stable/1705965 (cit. on p. 5).

[22] Khare A. and Shaulsky G. «First among equals: competition between genetically identical cells.» In: *Nature reviews.Genetics* 7 (2006), pp. 577–83. DOI: 10.1038/nrg1875 (cit. on p. 6).

[23] Brewster R. C., Weinert F. M., García H. G., Song D., Rydenfelt M., and Phillips R. «The Transcription Factor Titration Effect Dictates Level of Gene Expression». In: *Cell* 156 (2014). DOI: 10.1016/J.CELL.2014.02.022 (cit. on p. 6).

[24] Zheng Q., Bao C., Guo W., and et al. «Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs.» In: *Nature Communications.* 7 (2016). DOI: 10.1038/ncomms11215. (cit. on p. 6).

[25] Cesana M., Cacchiarelli D., Legnini I., Santini T., Sthandier O., Chinappi M., Tramontano A., and Bozzoni I. «A Long Noncoding RNA Controls Muscle Differentiation by Functioning as a Competing Endogenous RNA.» In: *Cell.* 147 (2011). DOI: 10.1016/j.cell.2011.09.028 (cit. on p. 6).

[26] Sumazin P., Yang X., Chiu H.S, Guarnieri P., Silva J., Califano A., and et al. «An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma.» In: *Cell.* 147 (2011). DOI: 10.1016/j.cell.2011.09.041 (cit. on p. 6).

[27] Frei T., Cella F., Tedeschi F., Gutiérrez J., Stan G.-B., Khammash M., and Siciliano V. «Characterization and mitigation of gene expression burden in mammalian cells». In: *Nature Communications* (2020). DOI: 10.1038/s41467-020-18392-x (cit. on pp. 6–8, 18).

[28] Wei L., Yuan Y., and Hu T. et al. «Regulation by competition: a hidden layer of gene regulatory network». In: *Quant Biol* 7 (2019), pp. 110–121. DOI: 10.1007/s40484-018-0162-5 (cit. on pp. 8–10).

[29] Scott M. «Tutorial : Genetic circuits and noise». In: 2006. URL: https://api.semanticscholar.org/CorpusID:18799510 (cit. on pp. 11, 12).

[30] *Stochastic systems: The Gillespie algorithm.* University Lecture. 2018. URL: https://e-l.unifi.it/pluginfile.php/642238/mod_resource/content/1/SistemiStocastici4.pdf (cit. on pp. 13, 16).

[31] Bonakdarpour M. *Inverse transform sampling.* 2016. URL: https://stephens999.github.io/fiveMinuteStats/inverse_transform_sampling.html (cit. on p. 16).

[32] Frost S. «Gillespie.jl: Stochastic Simulation Algorithm in Julia». In: *Journal of Open Source Software* 1.3 (2016). DOI: 10.21105/joss.00042 (cit. on pp. 17, 22).

[33] Cella F., Perrino G., Tedeschi F., Viero G., Bosia C., Stan G., and Siciliano V. «MIRELLA: a mathematical model explains the effect of microRNA-mediated synthetic genes regulation on intracellular resource allocation». In: *Nucleic Acids Research* 1 (2023). DOI: 10.1093/nar/gkad151 (cit. on pp. 18, 21).

[34] Cella F., Perrino G., Tedeschi F., Viero G., Bosia C., Stan G., and Siciliano V. «Supplementary Information: MIRELLA: a mathematical model explains the effect of microRNA-mediated synthetic genes regulation on intracellular resource allocation». In: *Nucleic Acids Research* 1 (2023). URL: https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/nar/51/7/10.1093_nar_gkad151/2/gkad151_supplemental_file.pdf?Expires=1701765273&Signature=dR2krXxzXiHVhD4qRERfMezbvONC6SjmfFZGchIHjGCnhPgElDT3PAWyzAPllfk-GDP9AO4NAr-o4a6BQPtRrRXe-hxOs6t1qbg~HA-OCmXTl1r8fAuFJgS8xYJRuZvosHi4L~pN9xWG2BwUFKTOT7w6FN8g3kfYhe5sBES~bkzjkVu~x~BSFkw1uA4iP4LDWEL~giEskUCabbHjmo4-BVf2FSQJqkdyli4C9-i7wHUT2K2vwEZVMMLv1N38X9DFCVcoycp02351qH15EqU7H087jolcxmbXzq4zux62z6CKPmCLUavdVgG~MYmraNHHIPfAs2QGjzfICE~82ecHew__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA (cit. on p. 21).

[35] *SnapGene website*. URL: https://www.snapgene.com/plasmids/fluorescent_protein_genes_and_plasmids (cit. on p. 23).

[36] *BioNumbers website*. URL: https://bionumbers.hms.harvard.edu/bionumber.aspx?id=112744&ver=7&trm=ribosome+translation+rate+Hek293&org= (cit. on p. 23).

[37] Macdonald P.J., Y. Chen, and Mueller J.D. «Chromophore maturation and fluorescence fluctuation spectroscopy of fluorescent proteins in a cell-free expression system.» In: *Analytical biochemistry* 421 (2012). DOI: 10.1016/j.ab.2011.10.040 (cit. on p. 24).

[38] *Real-Time PCR (qPCR).* URL: https://www.stratech.co.uk/aat-bioquest/real-time-pcr-qpcr/ (cit. on pp. 26, 27).

[39] Subach F., Subach O., Gundorov I., Morozova K., Piatkevich K., Cuervo A., and Verkhusha V. «Monomeric fluorescent timers that change color from blue to red report on cellular trafficking.» In: *Nat Chem Biol* (2009). DOI: 10.1038/nchembio.138 (cit. on p. 33).

[40] Pavel M., Renna M., and Park S.J. et al. «Contact inhibition controls cell survival and proliferation via YAP/TAZ-autophagy axis.» In: *Nature Communication* 9 (2018). DOI: 10.1038/s41467-018-05388-x (cit. on p. 34).

[41] Thomas P. and Smart T. «HEK293 cell line: A vehicle for the expressiosmn of recombinant proteins.» In: *Journal of Pharmacological and Toxicological Methods* 51 (2005). DOI: 10.1016/j.vascn.2004.08.014 (cit. on p. 34).

[42] Shahrezaei V. and Swain P. «Analytical distributions for stochastic gene expression.» In: *Proceedings of the National Academy of Sciences* (2008). DOI: 10.1073/pnas.0803850105 (cit. on pp. 35, 36).

[43] Mukund Thattai and Alexander van Oudenaarden. «Intrinsic noise in gene regulatory networks». In: *Proceedings of the National Academy of Sciences* 98.15 (2001), pp. 8614–8619. DOI: 10.1073/pnas.151588598 (cit. on p. 36).