TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

CRISTINA PALMA

DISSECTION OF THE RATE CONSTANTS OF A TRANSCRIPTION REPRESSION MECHANISM FROM LIVE SINGLE CELL, SINGLE MOLECULE MICROSCOPY DATA

Master of Science Thesis

Examiner: Associate Prof. Andre S. Ribeiro.
Examiner and topic approved by the Faculty Council of the Faculty of Computing and Electrical Engineering on 29th March 2017

ABSTRACT

CRISTINA PALMA: TUT Thesis Template
Tampere University of technology
Master of Science Thesis, 56 pages
May 2017
Master's Degree Programme: Erasmus Exchange Student
Examiner: Associate Prof. Andre S. Ribeiro.

Keywords: Transcription, Induction, τ Plot, OFF state.

Transcription is a critical process in cells, as it allows to transform the information stored in the DNA and shaped by evolution, into RNA molecules that, once translated into proteins, are capable of performing a multitude of tasks that are necessary for maintaining the cell alive.

Aside from identifying the main molecules involved in transcription, to fully comprehend this process, we need to characterize its dynamics. This will allow a better understanding of the mechanisms regulating gene expression.

The regulatory mechanisms of gene expression are the means by which cells activate or repress, fully or to some extent, a gene's transcriptional activity. It is this regulation that makes possible the response to environmental changes, as well as the establishment of critical internal cycles, such as the cycle responsible for cell replication.

Here, we investigated, at the single cell, single gene level, the dynamics of the process of transcriptional regulation the promoter LacO3O1 by gene-specific regulatory molecules, namely, inducers. Our goal was to, from live, single cell, single molecule data, obtain the values of the rate constants associated with the repression mechanism of transcription of this promoter.

Based on direct measurements of RNA production kinetics at different induction levels, and by estimating the RNA production rate at infinite induction we inferred that, under full induction, the LacO3O1 promoter, on average, spends 12% of the time between consecutive RNA productions in the OFF state.

PREFACE

First I would like to express my sincere gratitude to both my supervisors:

To Prof. Dr. José Manuel Fonseca I thanks for the enthusiasm, believe in me and thoughtful advices given since the first day.

To Prof. Dr. André Sanches Ribeiro for welcoming in his group, for transferring me new knowledge every day and for the contagious science passion.

The LBD lab. Samuel Oliveira, for the logic paths, for the conversations, for having the patient and will to teach me and also for the induction curve provided; Leonardo, for all the friendship; Vinodh, for always keeping up the spirit; Ramakanth, for all the time spent in the lab collecting data; Sofia, for all the problems found and all the solutions presented; Nadia, for all the peace and tranquility transmitted;

FCT family and friends. Ana, for all the 'agrees', for the past trips and for the upcoming ones; Margarida, for the neighborhood during the past 4 years; Jessica, road washer, because we will have more stories to tell; Ricardo, for all the kitchens we will achieve; Laura, free-spirit, for all the inconveniences, all the stories, and for being my crime partner in Funland; Lander, for the deep interest in kites; Henning, for the conversation level; Haresh, for being my first friend in Finland; Laura Garcia, sancha, for being always there during the last 10 years.

"#4ragazzeinMilano". Inês, for the farmyard, for the trips, for the 'exercitatione' and for all the amazing moments during the past 4 years; Mariana, for the tasteless soup, for the tiredness of the pizzas and for the floor cleaning; Sofia, for the experience shared and for the cheap and fun Wednesdays; Martelli, for always keeping the noise down.

"Eights". For the endless number of messages that I hope will keep growing and for all the quality and deep interest of our posts.

Family. To my Parents for supporting me in every step, for being the ones always there, for the guidance and values transmitted. To my Grandmother, for all the love and pride that will never die.

Tampere, 10th of May, 2017

Cristina Palma

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

bp	base pair
CAP	Catabolite Activator Protein
CDF	Cumulative Distribution Function
CFP	Cyan Fluorescent *Protein*
CV^2	Squared Coefficient of Variance
DNA	Deoxyribonucleic Acid
E. coli	*Escherichia coli*
FRET	Fluorescence Resonance Energy Transfer
GFP	Green Fluorescence Protein
GRN	Genetic Regulatory Network
HILO	Highly Inclined and Laminated Optical
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KDE	Kernel Density Estimation
MS2	bacteriophage MS2 coat protein
MS2d	MS2 dimer
mRNA	messenger RNA
NTP	Nucleoside Triphosphate
OD	Optical Density
ppGpp	guanosine 3', 5', bisphosphate
qPCR	quantitative Polymerase Chain Reaction
RBS	Ribosome Binding Site
RNA	Ribonucleic Acid
RNAp	RNA polymerase

SEM	Standard Error of the Mean
TEC	Transcription Elongation Complex
tRNA	transfer RNA
TSS	Transcription Start Site
YFP	Yellow Fluorescent Protein

1. INTRODUCTION

In a constantly changing world, the survival of organisms is determined by their degree of adaptability, i.e. their capability to adjust when facing environmental changes.

Several studies have shown that the regulation of gene expression plays a central role when it comes to cellular adaptation [1]. Regulation of genes' expression levels ensures that the right genes are expressed at the right time. Different mechanisms are responsible for carrying out this process. In order to fully characterize gene regulation one needs to fully characterize these mechanisms.

The process of gene expression begins with transcription, where a specific region of the DNA is transcribed into a messenger RNA. Next, translation occurs and proteins are synthesized. In order to successfully form an mRNA, transcription goes through three different steps: initiation, elongation and termination. In transcription initiation, highly specific interactions occur, which cause this event to be the main regulation point of gene expression in prokaryotes. These highly specific interactions control the kinetics of the open and of the closed complex formation, which are rate-limiting steps, thus controlling the overall transcription rate.

In vitro studies have estimated the time-length of the rate-limiting steps of transcription initiation [2, 3], however characterizing the same rate-constants *in vivo* remains challenging. Importantly, results from *in vivo* and *in vitro* measurements can differ widely, as many interactions occurring in the cell may not be included in the *in vitro* conditions.

Recently, novel experimental techniques of microscopy, molecular probing and computational tools for data analysis, have allowed studies using data from live, individual cells, which showed that transcription is a stochastic process [4] that gives rise to the exhibition of different phenotypes in a population of cells. This diversity is believed to improve the overall adaptation capability of a population of cells [5].

We make use of the particularly valuable new experimental technique of MS2-GFP tagging of RNAs with multiple, specific MS2 binding site sequences. By using this method, we are able to detect and track single RNAs as these are produced in living cells. To apply this method, we need two genetic constructs: a fluorescent protein fused to the RNA bacteriophage MS2 coat protein and a reporter RNA containing multiple MS2 binding sites.

Based on the data that this empirical method provides, a new technique for dissecting the dynamics of transcription initiation was developed [6]. From measurements of the time-intervals between RNA production events in cells with differing RNAP concentrations, it

was shown that, within a certain range of conditions, the inverse of the rate of RNA production changes linearly with the inverse of the concentration of RNAP. Importantly, this change in the rate of RNA production is due solely to changes in the time it takes for the closed complex formation, as this is the step in initiation that depends on RNAP numbers in the cell. As such, it is possible to estimate the duration of the open complex formation, by estimating, from the data, the rate of transcription in cells with infinite RNAP concentration.

Here, based on a stochastic model of transcription that includes the repression mechanism and the closed and open complex formation, we propose a novel, similar methodology that, from microscopy measurements of RNA production at the single molecule level in individual cells subject to different inducer concentrations, allows extracting the rate constants associated to the process of turning OFF and ON the ability of the promoter to be bound by an RNA polymerase.

2. BIOLOGICAL BACKGROUND

This chapter provides an overview of the biological concepts related to this thesis. It begins with a brief summary of the findings on the DNA composition and structure, and then proceeds with a description of the mechanisms of transcription and translation in *Escherichia coli*.

2.1 Brief history about the findings related to the DNA structure

2.1.1 Watson and Crick discoveries

DNA stands for deoxyribonucleic acid. In 1952, Pauling and Corey proposed a structure for the DNA that consisted in three coiled chains, where the bases were on the outside and the phosphates were oriented to the inside, near the axis of the coiled formation [7]. Watson and Crick (1953), considered that this structure did not suit entirely what was already well established regarding the forces between atoms that allow for the molecules to be stable: first, it was unclear what kept the structure stable. Also, some of the van der Waals distances were too small [8]. Thus, they proposed a new structure, based on two helical chains coiled around the same axis. In this model, the basis are located on the inside of the helix and the phosphates on the outside (Fig. 1). While some changes have been made to the original model overtime, the main features proposed remain the same.

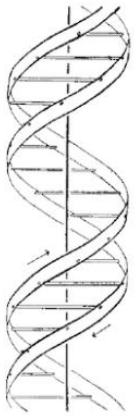

***Fig. 1**: DNA structure. The two strips represent the two phosphates chain. The horizontal lines represent the bases that hold the chains together in a helix format. Image taken from [8].*

2.1.2 Leven's "polynucleotide model"

Long before Watson and Crick discovered the DNA structure, Phoebus Levene proposed that the DNA was composed of a series of nucleotides and that each nucleotide had three major components: phosphate, sugar and a nitrogen-containing base, out of four possible bases (Fig. 2) [9].

Over the years, additional knowledge was acquired and some alterations were made to Levene's original proposal. However, his theory has been proven to be accurate in most aspects. As an example, we now know that the four nitrogenous bases are divided into two categories: purines (adenine and guanine) and pyrimidines (cytosine, thymine and uracil). Meanwhile, the DNA contains thymine while the RNA, instead, contains uracil.

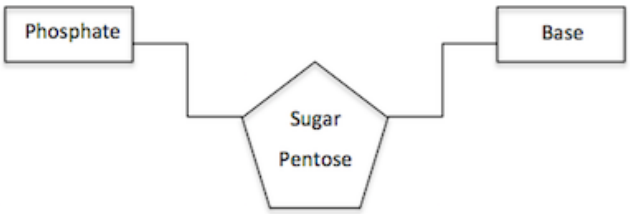

Fig. 2: *Nucleotide diagram. The base block represents one out of the four possible bases: adenine, guanine, cytosine, thymine or uracil. The pentose block is a five–carbon sugar that can represent either a deoxyribose (in the case of DNA) or a ribose (in the case of RNA).*

Erwin Chargaff continued developing Leven's work and presented two major conclusions regarding the DNA structure. First, he noticed that nucleotides do not have a specific order of appearance as Leven suggested, and that there are certain properties which are maintained between species: namely, the amount of adenine is similar to the amount of guanine and the amount of cytosine is similar to the amount of uracil [10]. The latter observation led to what is known as Chargaff Rule, which states that the amount of pyrimidines is equal to the amount of purines.

2.1.3 Double helical structure of the DNA and base pairing

It is known that the DNA double helix is anti-parallel, meaning that the 5' end of one strand is paired with the 3' end of the other strand (Fig. 3). It is possible to see in Fig. 3 that the nucleotides are linked to each other through the phosphate groups, while the DNA base pairs are connected by hydrogen bonds. Due to the hydrogen bonds weak stability, molecules can interact easily with the DNA, being able to perform the tasks of DNA expression and replication.

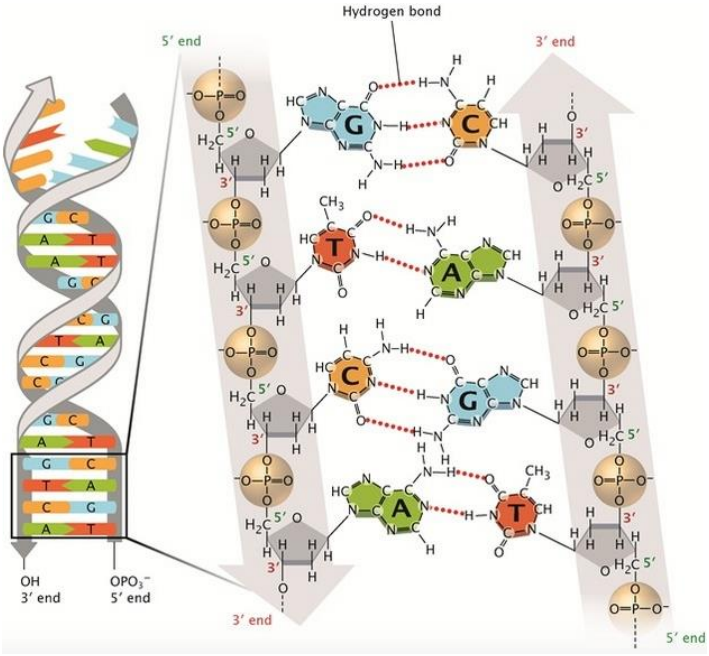

Fig. 3: *DNA helical structure and base pairing. The DNA structure consists of two anti-parallel chains (5'end of one chain is paired with the 3' end of the other). Each chain is composed by a series of nucleotides. Each nucleotide has three components: phosphate, sugar and a nitrogen-containing base. In the case of DNA, the base can be adenine, guanine, cytosine or thymine. The DNA base pairs are connected by hydrogen bonds. The Chargaff rule can be noticed in that thymine only links to adenine, and cytosine to guanine. Image taken from* [11].

2.2 Gene Expression in *Escherichia coli*

Escherichia coli is an important prokaryotic organism of the biosphere. It is typically present in the lower gut of animals but, given that it is also a facultative anaerobe, it can also survive in the natural environment.

E. coli is considered a "model organism" to investigate most of the basic intracellular processes such as: DNA replication, gene expression and protein synthesis. The genome of *E. coli* has approximately 4.6 million base pairs and encodes about 4000 different proteins, whereas the human genome encodes about 100,000 different proteins [12], [13]. Its comparative simplicity, together with the fact that it is easily grown in a laboratory setting provides obvious advantages for genetic analysis.

The process by which the genetic information encoded in the DNA is expressed is called gene expression. The central dogma of molecular biology, enunciated by Francis Crick in 1958, explains how this process occurs [14]. The process begins with transcription, where a section of DNA is transcribed into a messenger RNA (mRNA). This is followed by translation. In translation, the mRNA is used as a template to synthesize an amino-acid sequence that will form a protein (Fig. 4).

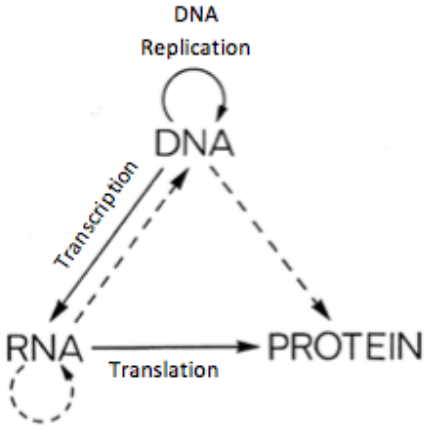

Fig. 4: *Diagram of permitted information flow used by Crick. In the process of DNA replication both strands can work as a template to generate a daughter DNA molecule. In transcription, information of the DNA sequence is used to make an RNA. In translation, information in the RNA sequence is used to make a protein. Due to reverse transcription, information in the RNA can also be used to make DNA. So far, there is no known process capable of synthesizing proteins directly from DNA. Adapted from* [14] .

The genome of *E. coli* consists in a single, circular, double stranded DNA. The cell also contains extra-chromosomal DNA, in the form of plasmids, with additional genes that code mainly for antibiotic resistance [15].

Prokaryote organisms do not have a nucleus, and therefore there is no physical barrier between transcription and translation. This is certainly related to the fact that transcription and translation are coupled [16].

In prokaryotes, the genes are organized in units called operons. An operon consists mainly in three components: a promoter, an operator and a structural gene. The promoter is a specific DNA sequence recognized by the RNAp. Meanwhile, the regulatory molecules that control the expression level of the promoter recognize the operator sites. In 1960, Jacob et al. described the first operon of *E. coli*: the lac operon (Fig. 10) [17].

In prokaryotes, it is possible for a set of genes to be controlled by a single promoter [18]. In this case, the operon is transcribed into a single mRNA molecule that contains information for the expression of multiple proteins. Such mRNA molecules are called polycistronic.

2.3 Mechanisms of Transcription and Translation

Transcription has three main steps: initiation, elongation and termination [19]. In initiation, the RNAp enzyme binds to the promoter and unwinds the DNA. The RNAp is composed of five polypeptide subunits (Fig. 5). Four of these subunits (α, α', β, and β') form the core RNAp. The fifth subunit, σ, confers specificity, and is equally necessary for the polymerase to start synthesizing mRNA. The polymerase with the five subunits is called 'holoenzyme'.

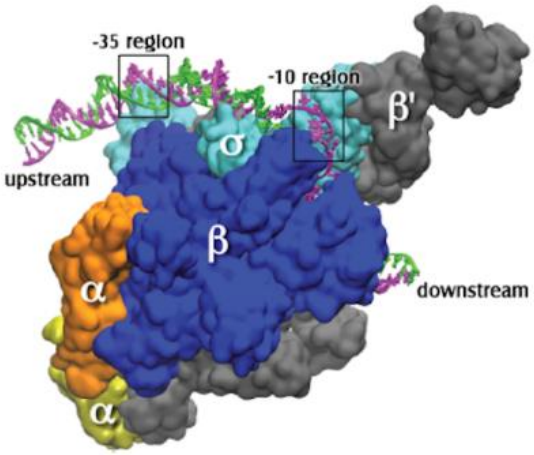

Fig. 5: *Structure of RNA polymerase holoenzyme while interacting with a promoter region during open complex formation in transcription initiation. The open complex formation corresponds to the unwinding of the DNA. Image taken from*[20].

In order to initiate transcription, the holoenzyme must first bind to the promoter. A promoter region in *E. coli* is defined by a consensus region at -35 (TTGACA) and -10 (TATAAT) positions upstream from the transcription start site (TSS) [21]. Without these consensus regions, the RNAp would not be able to recognize the transcription initiation site. Following the binding of the holoenzyme to the promoter region, the holoenzyme unwinds the double stranded DNA (~12bp), forming the transcription bubble.

After the unwinding of DNA and clearance of the RNAp from the promoter, elongation begins. During this process, the RNAp slides on the template strand (3' to 5' direction), while specific nucleotides are added to the 3' end of the growing polynucleotide chain in accordance with the DNA sequence. To ensure the fidelity of transcription (i.e. the specificity of the nucleotides), this process is capable of pausing and backtracking [22]. When the RNAp reaches the termination site, the newly formed mRNA and the RNAp are released from their binding to the DNA, in the process called termination.

Translation of the mRNA is also a three steps process, with initiation, elongation and termination. Ribosomes perform this task. These are complex molecules made of ribosomal RNA and proteins, and composed of a large and a small subunits, which, in the case of *E. coli*, are 50S and 30S, respectively (the "S" stands for svedbergs, a unit used to measure how fast molecules move in a centrifuge). Ribosomes are responsible for translating each codon (set of three nucleotides from the mRNA) into a specific amino acid, which is carried by the transfer RNA molecule.

In translation initiation, the small subunit of the ribosome (30S) binds to the mRNA's ribosome binding site (RBS), forming a 30S-RNA complex. Next, the large subunit (50S) binds to the 30S-RNA complex, in order to initiate translation elongation. In translation elongation, tRNAs bind to the appropriated codon and an amino-acid is added to the growing polypeptide until the stop codon is reached [23].

The above events are dynamic processes, thus, in order model them there is the need to handle them as physical processes and know the underlying rate constant values of this process.

2.3.1 Transcription Initiation steps

Transcription initiation is an essential step of gene expression, in that most of its regulation occurs at this stage. This is a sequential process that includes three steps before RNA chain initiation: binding, isomerization and promoter clearance [3], [24].

The RNAp core enzyme is able to produce an RNA from a DNA template. However, it is incapable of initiating transcription elongation. For this, it is necessary that a specific σ subunit binds to the RNAp core enzyme, forming the RNAp holoenzyme, which is capable of recognizing the promoter and initiate transcription [25].

The *E. coli* genome contains seven distinct σ factors, with σ^{70} being the one that is present in larger numbers [24], [26]. σ^{70} is part of the family of σ factors that are capable of forming RNAp holoenzymes that recognize promoters and form transcriptional promoter complexes, without the need for other factors or energy sources [25]. Using crystallography techniques, the structure of σ^{70} and the interactions between the core enzyme and the promoter DNA have been explicitly described [27].

In order to successfully complete the binding step, first the holoenzyme slides rapidly across the DNA until it finds the TSS [3]. Promoters recognized by σ^{70} contain two main consensus sequences at -35 and -10, counting from the TSS position. The region in between these two consensus sequences is known as a 'spacer' and its sequence differs

between promoters, although its length has approximately ~ . 17 bp in all promoters [24], [27].

After the holoenzyme finds the promoter, it recognizes the promoter site by making specific concts with the -35 and -10 boxes. In this step, the DNA maintains the double stranded structure, thus its named 'closed complex form'. Following the binding of the holoenzyme to the promoter, the σ^{70} factor triggers the destabilization of the DNA double helix, forming the transcription bubble [27]. In σ^{70} promoters, this step does not require ATP energy, as it is achieved by a structural change of the RNAP holoenzyme that is more energetically favorable than the previous state [19].

The following isomerization, which forms the stable open complex, was found to include at least three steps: DNA loading, DNA unwinding, and assembly of the polymerase clamp [24]. Once the promoter DNA sequence is open, NTPs can bind and transcription can begin. Several studies suggest that, before elongation, the RNAp goes through an abortive initiation cycle, where it synthesizes a few small transcripts no longer than 17 nucleotides [28], [29]. The detection of this abortive transcripts *in vivo* suggests that they might play a functional role, e.g. work as primers [28].

Promoter clearance is the last stage of transcription initiation. It consists of the RNA polymerase releasing its contacts with the core promoter and entering in the elongation phase.

The moment the σ^{70} is released is still unclear. It was first thought to coincide with the formation of the elongation complex [29], although recent studies argue the possibility that it may remain bounded to the promoter, be released in the beginning of elongation, or remain bound to the TEC throughout the elongation process [30]–[33].

William McClure [3] identified the steps of initiation by two methods: the abortive initiation assay and the *in vitro* transcription assay [2], [3], [34]. The abortive initiation assay consists of the binding of the two first triphosphates in an RNA sequence in the presence of a saturating amount of RNAp. In the specific experiments carried out by W. McClure, the two triphosphates were ATP and UTP, where ATP is always the first nucleotide, followed by UTP. When both these nucleotides link to each other, a phosphodiester bond is created and both pppApU and PPi are produced. In the absence of more nucleotides, the bond between them is broken (aborting initiation). After a short time, a steady-state production of the abortive product is reached. By measuring the delay in reaching this steady-state in various conditions (differing in the concentrations of RNAp), it is possible to estimate the rate of open-complex formation by estimating how long the process would take when having an infinite concentration of RNAp in the system (Fig. 6).

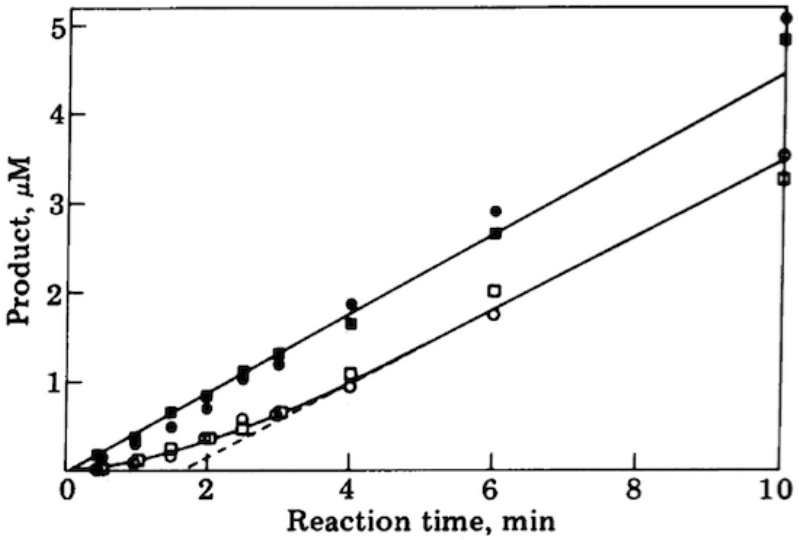

Fig. 6: *Estimation of the time required for the open complex formation by the abortive initiation assay on the bacteriophage T7 D promoter. The product in this experiment is pGpUpU. The intermediate steps in transcription initiation delay the formation of the product, resulting in a lag time to reach steady-state. Image taken from [2].*

These experiments were carried out in an *in vitro* environment. Measurements for the rate constants *in vivo* are expected to differ widely (e.g. the media viscosity differs).

Later on, we describe advances in fluorescent live cell imaging and in computational image processing tools that have opened new possibilities for the characterization of this process *in vivo*.

2.3.2 Transcription and Translation Elongation

Transcription elongation corresponds to the process of transcript synthesis. It initiates when the RNAp clears the promoter region and proceeds with the addition of the required nucleotides to form the RNA transcript.

The transcription elongation complex (TEC) can be seen as an integrated macromolecular machine that performs a group of specific activities: first, it works as a helicase to open the DNA genome, exposing the DNA template strand; second, it carries out RNA synthesis; and, third, it acts as a regulator of its own stability and rate, mainly through the binding of transcription factors [35].

In each step of RNA synthesis, the TEC can enter alternative reaction pathways that can lead pausing, arrests, misincorporation and editing, pyrophosphorolysis, and premature termination. The probability of TEC entering any of these pathways is closely related with the interactions between the complex, the template DNA, the nascent RNA and regulatory transcription factors [35]. These events heavily affect the times for completion of

RNA molecules. For example, the duration of the pausing events have been shown to vary from less than a second to a few minutes [36], [37], thus influencing not only the mean transcription rate but also the level of noise in this process [38].

The TEC usually 'protects' ~35 bp of the double stranded DNA. Within this region, the transcription bubble is formed and occupies 12-14bp. The 3' end of the nascent RNA is where RNA synthesis takes place, while the 5' end of the nascent RNA is free to form secondary structures or interact with other components (Fig. 7) [35].

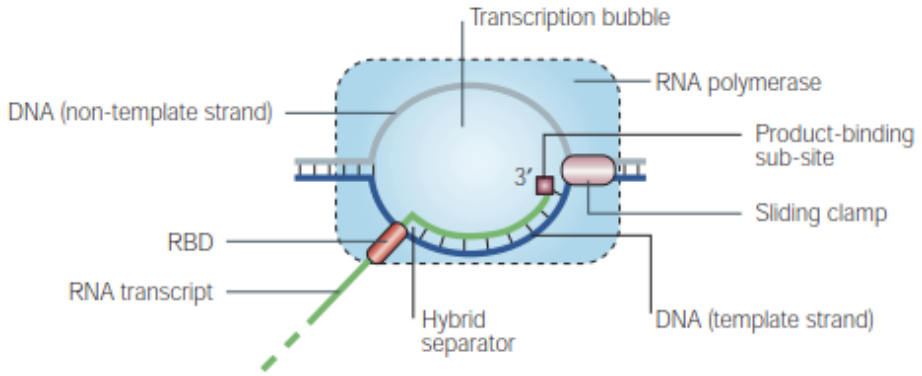

***Fig. 7**: Typical template positon of the Transcription Elongation Complex. The elongation complex protects ~35 bp of the DNA. The transcription bubble formed covers between 12-14bp. RNA synthesis takes place at the 3' end, forming a complementary duplex RNA-DNA whose length is around 8-9bp. The 5' region of the RNA transcript lies within the 'RNA-exit channel'. Image taken from* [35].

If none of the alternative reaction pathways is followed by TEC, the transcript continues to be elongated, until the complex finds a termination site on the DNA template. At this stage, the RNA transcript is released from the template, forming an mRNA.

In 2008, Wen *et al.* revealed that translation elongation occurs as a series of translocation-pause cycles [39]. Each translocation consists of the ribosome moving three bases (one codon) at a time, creating a peptide-bond formation between amino-acids. The pauses time-lengths are the main responsible for the overall rate of translation and depend on the secondary structure of the mRNA.

Translation rates have also been shown to be codon-specific [40]. The fact that different codons can correspond to the same amino-acid but different amino-acids cannot correspond to the same codon, allows an additional level of regulation for translation, because different codons can happen to not change the coded protein but, instead, affect the translation elongation rate.

In prokaryotes, transcription and translation are dynamically coupled in the cytoplasm and translation initiates as soon as the RBS emerges from the TEC [16].

2.3.3 Regulation of Transcription

The control of the number of RNA and proteins inside a cell seems to happen due to the cell's ability to control the frequency and timing with which each gene is expressed [41]. This capability of self-regulation is essential for the cell survival, when facing changes in the internal or external environment.

In order to initiate transcription, the RNA polymerase and the promoter region of the DNA have to undergo highly specific interactions with each other, turning transcription initiation one of the key points of regulation in gene expression [41], [42].This regulation can be carried out by different mechanisms, such as the promoter sequence itself, σ factors, small ligands and transcription factors.

There are more than 2000 promoter sequences in *E. coli* [43], with each promoter sequence presenting a specific affinity for RNAp binding. Thus, the promoter sequence plays a role in the rate of the closed complex formation. However, since these sequences cannot be tuned, they only provide static regulation [26].

The fifth subunit (σ) of the RNA polymerase holoenzyme, also known as σ factor, is essential for specific promoter recognition [44]. There are seven types of σ factors, one main σ factor is σ^{70}, as it allows to recognize most promoters. The other six σ factors present in the *E. coli* accumulate in numbers in response to specific stress conditions [45]. This ability of expressing different σ factors in specific conditions, allows initiating transcription of different sets of genes which causes global changes in the dynamics of the gene regulatory network of *E. coli*.

Small ligands are another means by which extrinsic regulation of transcription is achieved in *E. coli*. An example of a small ligand is ppGpp. ppGpp is able to regulate the synthesis of the machinery for translation (e.g. ribosome synthesis), by destabilizing the open complex formation of *rrn* promoters [26]. Recent studies have also demonstrated that ppGpp can also up-regulate genes for amino acid biosynthesis [46].

Although promoter sequences, σ factors and small ligands contribute to the regulation of transcription, the main regulation occurs by the binding of gene-specific transcriptional factors to the promoter region. In *E. coli*, more than 300 genes encode for transcriptional factors, which can either activate or repress transcription initiation depending on the mode of regulation [26]. Some transcription factors can act either as an activator or as a repressor, depending on the target promoter.

Activator molecules upregulate transcription by several mechanisms. The activation mechanisms can be divided in three different classes: In Class I and Class II, the activator

molecules interact directly with the RNAp, while in Class III the DNA conformation is altered by the binding of activators, resulting in the augmenting of the RNAp binding affinity (e.g. araBAD promoter) [47].

Repressor molecules downregulate transcription by inhibiting transcription initiation. The exact mechanism of repression varies between promoters. There are at least three ways by which repression can act: the repressor can compete directly with the RNAp in binding to the promoter [48], the repressor can inhibit open complex formation and, finally, the repressor can inhibit promoter escape [49].

Experimentally, the use of inducer molecules is the most common way of controlling the binding of transcription factors to DNA [50]. In this work, we make use of the IPTG and L-arabinose inducers in order to regulate the activity of the promoters $P_{lacO3O1}$ and P_{BAD}, respectively.

2.3.4 Noise in Gene Expression

A population of genetically identical cells in the same environment can exhibit different phenotypes. This heterogeneity is known to have several sources, one of which being the stochasticity present in the process of gene expression (intrinsic noise) [51], [52], and another being differences in the numbers of molecules regulating gene expression (extrinsic noise).

The total cell-to-cell variability in RNA and protein numbers present in a population of cells has two noise sources: extrinsic and intrinsic noise. Differences between cells in the numbers of molecules that regulate transcription, such as RNA polymerases and transcription factors, will cause cell-to-cell variability in the output of a gene. These are known as sources of extrinsic noise. Meanwhile, the inherent stochasticity of gene expression due to the small number of molecules involved that also affects the rate at which a certain gene is expressed is considered an intrinsic source of noise [52].

In 2002, Elowitz *et al.* were able to measure the levels of extrinsic and intrinsic noise in *E. coli*. In order to do that, two strains of *E. coli* incorporating CFP and YFP fluorescent proteins controlled by identical promoters were built. To measure the intrinsic noise, the fluorescent levels of each protein were measured in each cell, whereas extrinsic noise was measure by the correlation between the levels of both proteins in each cell (Fig. 8) [52].

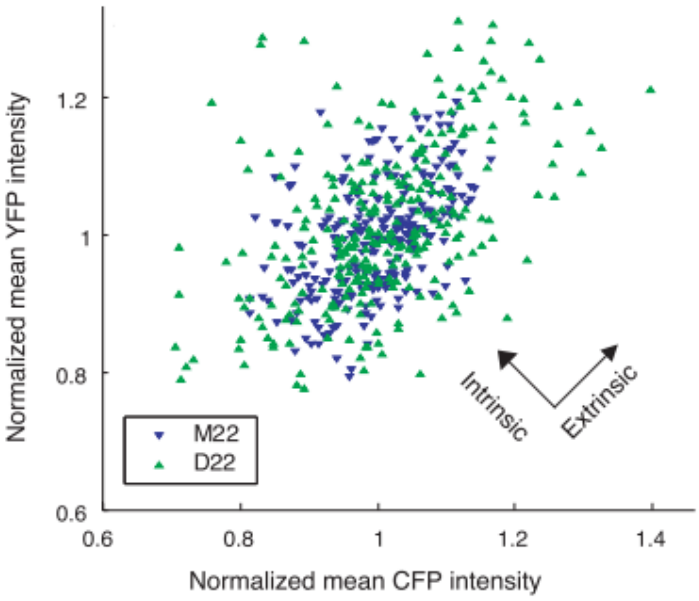

Fig. 8: *Quantification of noise. Plot of the fluorescence in two strains: one noisy and another quiet, D22 and M22, respectively. Each triangle represents a cell for the respective strain. Spread of points perpendicular to the diagonal line correspond to intrinsic noise, whereas spread of the points parallel to the line correspond to extrinsic noise. Image taken from* [52].

Cell-to-cell variability in the mRNAs and protein levels has been continuously studied. The development of new methodologies have allowed to measure with single cell sensitivity both mRNAs and proteins in single cells [53]–[55].

In 2005, Golding and co-workers were able to study single-cell transcription in *E. coli* by detecting individual mRNA molecules in individual living cells. After demonstrating that their method was reliable, they characterized the transcription kinetics in individual cells, and suggested that transcription occurs in quantal bursts, even in fully induced cells [54]. The frequency and size of these bursts affect the levels of mRNA and proteins within a cell, contributing for the noise in gene expression [56].

Real-time monitoring of protein production was possible in 2006 [55]. In order to do that, a variant of the YFP protein was used as the reporter. From the analysis of the time-traces of the fluorescent protein molecules Yu and co-workers suggested that protein molecules are also generated in bursts, and that the number of proteins produced in each burst varies. Since the distribution of the numbers of gene expression bursts per cell cycle, for all cells, fits a Poisson distribution, they presented the suggestion that gene expression bursts occur randomly and uncorrelated in time [55].

The sources of variability in RNA and protein numbers are still not completely clarified. Different studies suggest that other mechanisms, not directed related with gene expression, can also contribute to the observable variability, such as DNA supercoiling and random segregation during cell division [56]–[58].

3. MODELS OF GENE EXPRESSION DYNAMICS IN *E. COLI*

This chapter provides an overview of the concepts regarding the modeling of gene expression that were used in this thesis to characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation.

When modeling gene expression, most authors try to go after simple formulation of the process. To achieve that, only the steps that affect the overall behavior should be included in the model [59]. Usually, the models are represented as chemical reactions as the one below:

$$A + B \xrightarrow{k} AB \tag{3.1}$$

Here it is assumed that one molecule of A reacts with one molecule of B, forming a molecule AB, at a constant rate k.

In order to understand the behaviour of gene regulatory networks (GRNs), different models have been developed [59]. Models with deterministic kinetics are not able to predict the behaviour of GRNs since gene expression has been shown to be a stochastic process [55].

In vitro studies done by several researchers (see [3] for a review) focused on the regulation of transcription initiation steps, in order to establish models of transcription:

$$R+P \underset{K_B}{\rightleftharpoons} RP_C \xrightarrow{k_f} RP_O \rightarrow\rightarrow\rightarrow ... \rightarrow RNA \tag{3.2}$$

In (3.2), R stands for RNAp, P stands for a free Promoter, RP_C stands for a closed complex and RP_O for an open complex.

This scheme involves the binding of the RNAp to the promoter with a rate binding constant, K_B, to form a closed complex. Next, the closed complex isomerizes with a rate constant k_f and forms the open complex (RP_O). After, the RNAp is released from the promoter, and the elongation phase begins. This scheme was first proposed by Zillig et al. [60].

In 2006, Roussel and Zhu proposed a model gene expression along with a simulation strategy that allowed for delayed events [61]. Following this, in [59] the authors have shown that a simpler model, using the same simulation algorithm (delayed SSA), can be used to reproduce the known empirical data on the stochastic kinetics of gene expression. The model is a reaction-based model, which consist only of the following reactions:

$$\mathrm{RNAP}(t) + \mathrm{Pro}(t) \xrightarrow{k_1} \mathrm{Pro}(t+\tau_1) + \mathrm{RNAP}(t+\tau_2) + \mathrm{R}(t+\tau_2) \tag{3.3}$$

$$\mathrm{Ribosome}(t) + \mathrm{RBS}(t) \xrightarrow{k_2} \mathrm{RBS}(t+\tau_3) + \mathrm{Ribosome}(t+\tau_4) + \mathrm{P}(t+\tau_5) \tag{3.4}$$

$$\mathrm{RBS} \xrightarrow{k_s} \mathrm{decay} \tag{3.6}$$

$$\mathrm{Pro} + \mathrm{Rep} \xrightarrow{k_{rep}} \mathrm{ProRep} \tag{3.7}$$

$$\mathrm{ProRep} \xrightarrow{k_{unrep}} \mathrm{Pro+Rep} \tag{3.8}$$

Reaction (3.3) models the first step of transcription initiation with a probability rate constant k_1. The reaction between RNAP and Pro is the input of the transcription process. RNAP finds the TSS and forms the closed complex (at rate k_1). On the product side, the output events happen τ_1 and τ_2 time units later, with $\tau_2 > \tau_1$. First, the promoter is cleared and, at the same time, the region of the transcript containing the ribosome binding site (RBS) is produced. Then the RNAP is released from the DNA, and a primary mRNA transcript (R) is produced.

In Reaction (3.4), the ribosome finds the RBS (ribosome binding site region of the RNA) with a probability rate constant k_2. While the ribosome is bound to the RBS, no other ribosome can bind (τ_3). The degradation of the RBS is modelled by reaction (3.6). Reaction (3.7) represents the blocking of the promoter by a repressor molecule (Rep) and reaction (3.8) models the unbinding of the repressor.

In parallel with the validation of this delayed stochastic model, a simulator was developed that was used of for subsequent studies of more complex models, including of small genetic circuits [62].

Models of single gene expression have been continuously further developed since then. In [63], [64] a detailed model including alternative pathways to elongation was proposed and confronted with a single-step multi-delayed stochastic model. As shown, for low expression rates both models seem accurate, however for higher rates the two models differ.

In [65], the authors investigated the effect of different codons in the rate at which they are translated, thus improving the model of transcription elongation.

A recent study [6], based on the time-intervals between RNA production at single molecule-level for different RNAp concentrations, was capable of accurate characterization of the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the $P_{lac/ara-1}$ promoter. *In vivo* durations for the open and closed complex were estimated for this promoter. In order to do this, first it had to be verified that it is possible to change the RNAp concentration with different media richness, and that it is possible to infer about the relative free RNAp concentration from the total RNAp concentration, since it is this one that affects the transcription kinetics. Therefore, a plot of the reciprocal of the RNA production rate against the relative RNAp concentrations was made (Fig. 9). The result was a linear relationship that shows that, in fact, the freely diffusing RNAp concentrations can be assessed from the total RNAp concentrations and that, besides the RNAp concentration, there is no other variable affecting the target promoter kinetics.

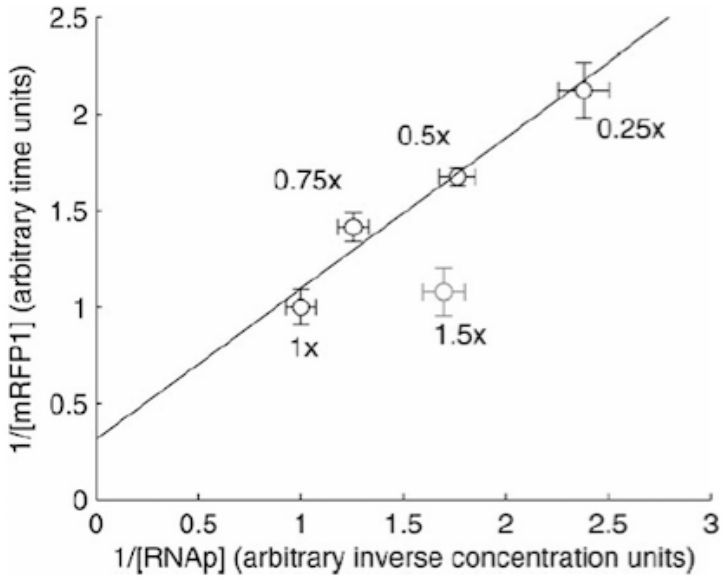

Fig. 9*: Lineweaver-Burk plot of the inverse of the production rate of mRFP1 from the* $P_{lac/ara-1}$ *against the inverse of the total RNAp concentrations for the same growth conditions. Figure taken from [6].*

Given these results, it was possible to develop the model-fitting procedure. The mean-interval distributions between RNA productions were calculated and a 'tau plot' was made from them. Using this *in vivo* technique, it is possible to extract more information besides the mean duration of the open complex-formation. In [6] information about the

mean duration and variance of the closed complex formation was extracted from the slope of the plot: $788 \pm 59 R.s$ (R is the polymerase concentration).

The model fitted to the experimental results was:

$$\mathrm{R} + \mathrm{P}_{ON} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP}_C \xrightarrow{k_2} \mathrm{RP}_O \xrightarrow{k_s} \mathrm{R} + \mathrm{P}_{ON} + \mathrm{RNA} \tag{3.9}$$

$$\mathrm{P}_{ON} \underset{k_{ON}}{\overset{k_{OFF}}{\rightleftarrows}} \mathrm{P}_{OFF} \tag{3.10}$$

In (3.9) the RNA polymerase (R) binds to the free promoter (P_{ON}) forming the closed complex (RP_C). Once the start site is reached, the RNAp must open the DNA double helix, resulting in the open complex (RP_O). Next, the polymerase begins elongation, thus clearing the promoter.

In (3.10) it is represented the transition between the active (P_{ON}) and inactive (P_{OFF}) state of the promoter. These states can occur, for example, by the binding and unbinding of the repressors and activators.

The authors make a note saying that reaction (3.9) should not be seen as elementary transitions but rather as effective rates of the rate-limiting steps in the process.

Besides this model, three simplified models were derived from this and all of them were fit to the observed dynamics of $\mathrm{P}_{lac/ara-1}$. To compare the goodness of the fits, the BIC (Bayesian Information Criterion) was used and the model that best fit the data was the one where $k_{-1} \gg k_{-2}$. From the measurements, the authors were capable to determine the time spent in each OFF state (~ 87s), the mean time taken by the initial binding of RNAp $\left(788 \pm 59\,\mathrm{R.s}\right)$, and the mean time since the polymerase has committed to transcription until it releases the promoter (193 ± 49s).

The authors were also able to identify the repressor LacI as the responsible for this ON/OFF dynamics.

Inspired by these studies, in this thesis we propose a new method that allows to estimate the time-length spent by a promoter in the OFF state, using empirical data at the single molecule level for different induction levels.

For this, a model of transcription was assumed (see chapter 4.5.1), which, when confronted with the time-intervals between RNA production events, allows extracting information regarding the promoter OFF state.

4. MATERIALS AND METHODS

Materials and Methods are presented in the conference paper [66] and are further explained here to discuss how and why were they used in the development of this work. These methods comprise: single-molecule approaches of fluorescent tagging, microscopy techniques, single-RNA detection methods, RNA quantification and methods for independent validation of the main measurement techniques.

4.1 Single-Cell, Single-RNA, time-lapse quantification methodology

In order to understand the mechanism of the system used one should first know how the lac operon and araBAD operon operates:

4.1.1 Lac operon

Thanks to the pioneering research done by Francois Jacob and Jacques Monod in the 1960s, the gene regulation mechanisms were primarily understood through the study of the lac operon, which became, by that time, one of the best understood and explained models for the control of protein production [17], [67].

The *lac* operon includes three genes: *lac*Z, *lac*Y and *lac*A. All these genes are transcribed as a single polycistronic mRNA. The gene *lac*Z is responsible for the production of $\beta-$galactosidase that catalyzes lactose molecules [68]. *lac*Y encodes for lactose permease, which facilitates the uptake of lactose into the cell trough active transportation that uses the energy of the electromagnetic proton gradient [69]. It is known also that *lac*A encodes the enzyme thiogalactoside transacetylase, although its physiological function remains unclear.

Expression of the lac operon is negatively controlled by three *lac* operators O1, O2 and O3. The *lacI* gene, which codes for LacI molecules (*lac* repressor), and its promoter lie upstream of the *lac* promoter [70]. A schema of the *lac* operon is shown in Fig. 10.A.

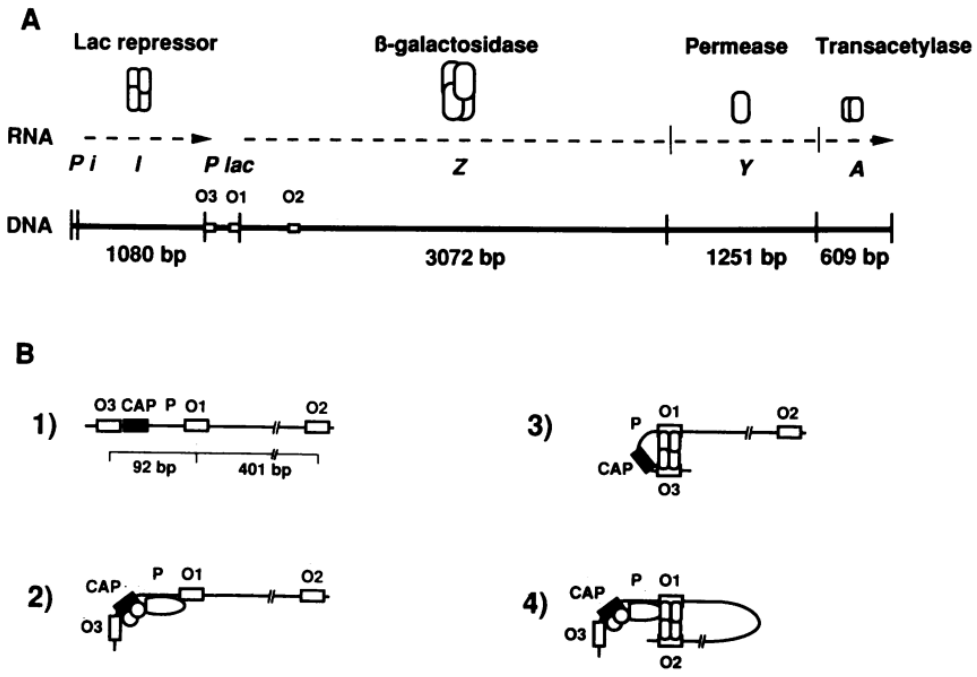

Fig. 10: *Diagram for the structure and regulation of E. coli lac operon. (A) lacZ, lacY and lacA genes are transcribed from the lac promoter (P lac). Pi is the promoter for the lacI gene that codes for the repressor LacI. The length of its gene is given in bp. (B) the control elements of the lac promoter. In between O3 and O1 there is a CAP binding site, which aids the binding of RNAp. O2 lies in the region of lacZ gene (1). Lac repressors form loops in the DNA in the absence of inducer (2), It can bind to O1 and O3 (3) or to O1 and O2 (4). Reused from* [70].

The LacI repressor molecules interact with the DNA through the two N-terminals at each end of the molecule, changing its binding affinity for the RNAp [71]. In order to have full repression, LacI molecules bind to two *lac* operators, forming a DNA loop (Fig. 10.B). This can occur by the molecules binding to O1 and O3 or to O1 and O2 [70].

On the opposite side, allolactose is the natural inducer for the *lac* operon, and results from the cleavage and isomerization of lactose and, thus, when this compound is present in the cell the transcription levels of lac operon increase.

The molecular reagent IPTG is a mimic of allolactose and it is commonly used to regulate the *lac* operon in laboratory conditions. When IPTG is added, it binds to the LacI molecules inducing a conformational change in the protein structure that turns the binding to the operator site no longer possible.

4.1.2 *araBAD* operon

The genes of the *araBAD* operon are expressed under the control of the promoter $P_{BAD.}$ Regulation of P_{BAD} is done by the transcription factor AraC. AraC expression is controlled by the promoter $P_{C,}$ which is divergently oriented from P_{BAD} (Fig. 11). In between PC and P_{BAD} there is the binding site for CAP. AraC protein acts either positively, stimulating transcription or, in the absence of arabinose, negatively by repressing transcription initiation [72].

In the absence of L-arabinose, AraC is bound simultaneously to two different DNA sites (*araI*$_1$ and *araO*$_2$), causing DNA looping that prevents the binding of RNAp to the promoter. When L-arabinose is in the system, the products from the genes that code for the arabinose transporters (araE and araFGH) take up the L-arabinose from the growth medium. L-arabinose binds to AraC, breaking the DNA loop, thus promoting the occurrence of transcription [72], [73].

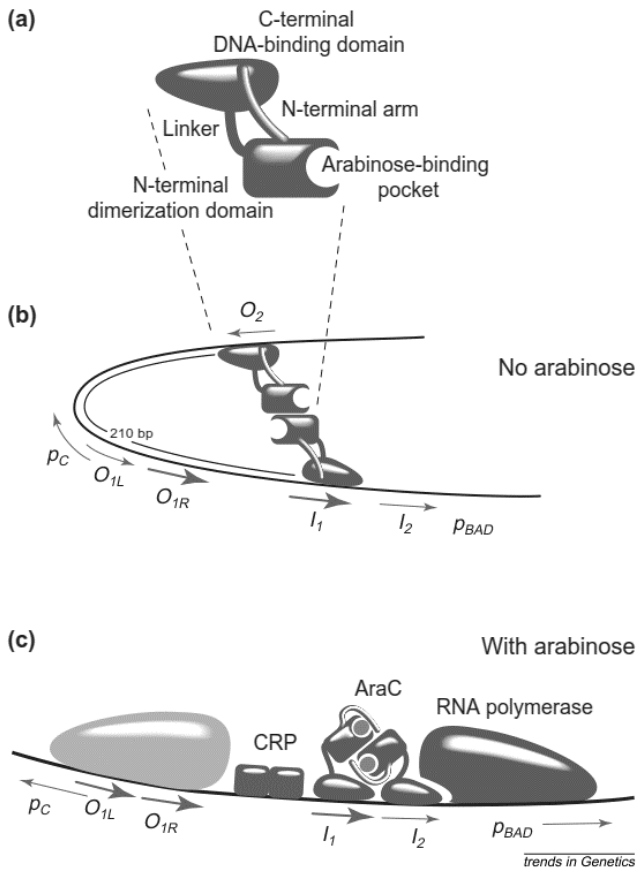

Fig. 11: *araBAD repression and activation mechanism. (a) Structure domain of AraC protein. (b) In the absence of arabinose the RNAp is prevented from binding to the P_{BAD} and P_C promoters. (c) When arabinose is present the DNA loop is broke and the RNAp has free access to the promoters. Adapted from* [72].

4.1.3 Strains

For our study we made use of two different *E. coli* strains: BW25113 and JW0336.

The *E. coli* K-12 BW25113 is the common strain background used for the generation of the Keio collection mutants. The Keio collection is a comprised of 3985 single-gene deletions of *E. coli* K-12 BW25113 [74]. Derivation of this strain can be seen in Fig. 12.

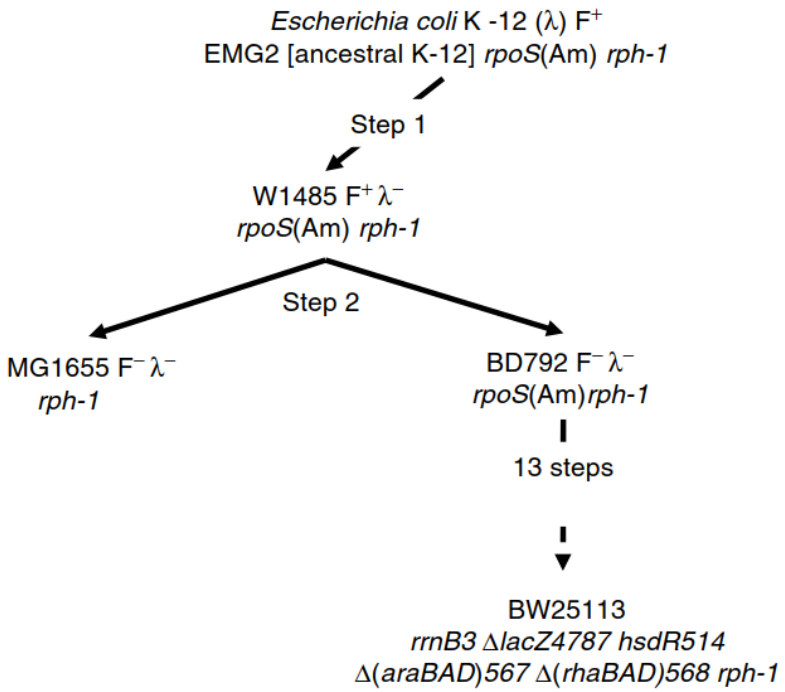

Fig. 12: *Derivation of E .coli K-12 BW25113 strain. BW25113 is a descendent of BD792, lacking the bacteriophage lambda and the F plasmid, which is a two-step descendent ancestral of E. coli K-12. BW25113 was derived from BD792 in a series of 13 steps involving transduction and allele replacements. Reused from* [74].

The other strain used, JWO336, is a deletion mutant of the Keio collection. This strain lacks the ability to express LacI repressor molecules. BW25113 contains the promoters P_{lacI+} and P_{araC} that are responsible for the expression of LacI and AraC repressors, respectively.

In both strains, a single-copy plasmid carrying a $P_{lacO3O1}$ promoter and a multi-copy plasmid with the gene P_{BAD}-MS2-GFP were introduced.

From the expression of $P_{lacO3O1}$, a target RNA containing 48 binding sites for the MS2-GFP proteins is produced. Compared with the standard *lac* operon, the sequence responsible for the expression of the 48 binding sites is placed in the *lacZ* region, which falls in the O2 region. In the system, both the repressor LacI and IPTG regulate the activity of

$P_{lacO3O1}$ as described in chapter 4.1.1. Data was collected from this system for different IPTG concentrations (0, 5, 25, 50, 100, 250, 500 and 1000 µM).

The reporter gene P_{BAD}-MS2-GFP is responsible for the production of the MS2d-GFP proteins, which bind to the target RNA, making it appear as bright spots under the HILO microscope (Fig. 15). Regulation of PBAD is done by AraC and L-arabinose as described in chapter 4.1.2. In order for L-arabinose to be present in the system, to activate the activity of the P_{BAD} promoter, 0.4% of L-arabinose was added to the culture. More details about the growth conditions can be read in [66].

4.2 RNA Detection

Over the years, different studies contributed to the understanding of transcription. These studies made use of techniques such as X-ray crystallography [75], FRET [76], footprinting based on gel electrophoresis [77] and FISH [78]. However, all of these provide a static picture of a dynamic process. Real-time *in vivo* single-molecule studies are required to understand the mechanisms of transcription [79]. By using *in vivo* single-molecule methods, we are able to create a detailed picture of the kinetics of every step in the process.

4.2.1 Fluorescent proteins

In this thesis, to study the *in vivo* kinetics of transcription in individual cells, we make use of fluorescent probing. In 1961, during the study of the jellyfish *Aequorea* (Fig. 13.A), Osamu Shimomura and colleagues discovered the luminescent substance aequorin. Aequorin is capable of storing a high amount of energy, which is released when calcium is present, generating a bright blue light. Due to its properties, it is widely used as a calcium probe. Furthermore, during the purification of aequorin, another protein with bright green fluorescence was extracted. This protein was renamed as GFP (Fig. 13. B), and the structure of its chromophore was elucidated later [80], [81].

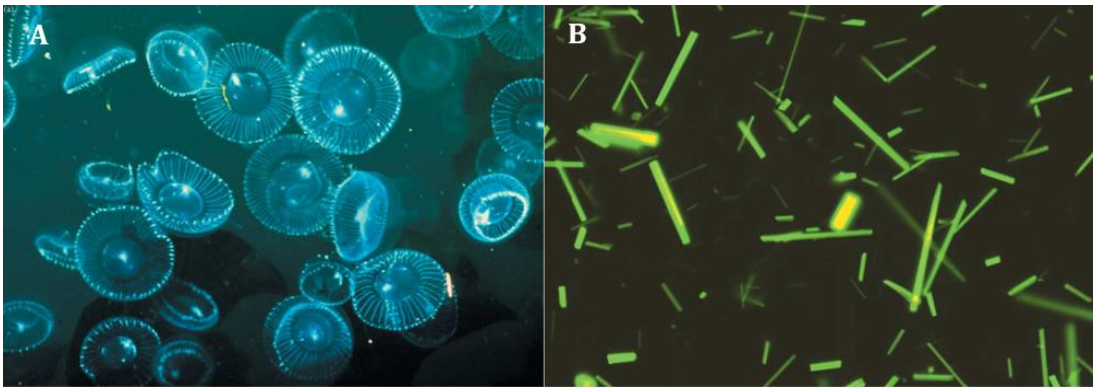

Fig. 13 : *(**A**) Jellyfish Aequorea (**B**) Crystallized GFP. Reused from* [80].

The potential of GFP was understood later, when GFP became a marker protein for gene expression. Since then, through protein engineering methods, fluorescent proteins development evolved, covering now almost the full visible spectrum of light [82].

In order for fluorescent probing to be an effective method, the fusion of the fluorescent proteins cannot impair the function of the targeting molecules. Although fluorescent proteins became a powerful tool to understand the dynamics at a spatial and temporal level simultaneously, they still require some improvements in, e.g.: maturation times, photobleaching and blinking. Namely, if maturation times were shorter, allowing for the detection of the targeting molecules as soon as they are produced, if the fluctuations in the fluorescence intensities (noise) did not exist and the molecules were not subject to photobleaching, this detection system would be more effective.

To detect fluorescent proteins precisely, the emitted fluorescent signal needs to be higher than the background fluorescence (auto-fluorescence), so fluorescent proteins need to be selected based on the conducted study, in order to not have the same wavelength excitation as the background. In our case study, a 48 tandem repeats binding sites were added to the target RNA molecule, making the fluorescent signal much higher than the background when all the binding sites are bound to the fusion protein MS2-GFP.

4.2.2 MS2-GFP Tagging Method

To study the dynamic nature of transcription, methods capable of following gene expression, in individual living cells, needed to be developed. The understanding of the potential of fluorescent proteins as sensors of this process led to a rapid development of the methods for imaging *in vivo* biological processes.

A method allowing the visualization of native RNA in living cells did not exist before Singer and colleagues, in 1998, developed a novel approach that allowed for the *in vivo* real-time visualization of mRNA molecules in eukaryotic cells [83]. An adaptation of this method, in 2004,

allowed the tracking of individual mRNA molecules for many hours in *E. coli* [84]. Since then, the study of single molecules in single cells has been possible, allowing for the quantification of gene expression dynamics *in vivo.*

In this thesis, as in [54], [84], a two plasmid system was used. On one plasmid, the GFP sequence is fused to a tandem dimer of the RNA bacteriophage MS2 coat protein, under the control of the P_{BAD} promoter. On the second plasmid, 48 tandemly repeated MS2-binding sites were inserted into a reporter mRNA, each one of them consisting in a stem-

loop structure of viral RNA with 19 nucleotides, under the control of the PlacO3O1. A schematic description of the constructs used in this study is shown in Fig. 14.

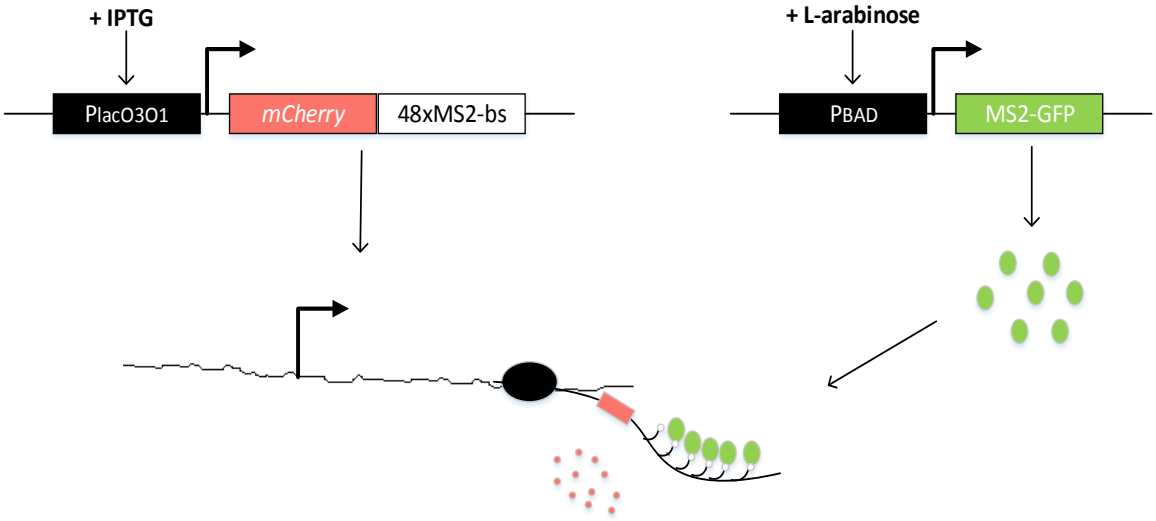

Fig. 14: *Schematic image of the constructs used for MS2-GFP tagging of RNA molecules. MS2-GFP proteins are expressed in the presence of L-arabinose. LacO3O1 promoter controls the expression of the target RNA (mCherry followed by 48 binding sites for MS2-GFP). The MS2-GFP accumulates in the cytoplasm, once a target mRNA is produced, they bind to it. The mCherry region is translated into proteins with red fluorescence.*

After induction of the reporter plasmids, many copies of MS2-GFP proteins will be in the cytoplasm and the cells cytoplasm will be bright-green. The MS2-GFP proteins will tag each target mRNA as soon as it is transcribed, creating a bright spot that can be visually separated from the background fluorescence.

Interestingly, due to the properties of the viral MS2 coat protein, when RNA molecules are bound to MS2-GFP, they become ‘immortalized’, in that they do not degrade during the course of the measurements. This allows an accurate quantitative study of RNA numbers over time without contamination by RNA degradation [54].

4.3 Microscopy

Microscopy images were acquired in the laboratory by the molecular biologists of the group. The information from each condition was obtained from two different channels: phase contrast and fluorescent microscopy. Phase contrast images were acquired with the purpose of cell segmentation, tracking and counting, while fluorescent images allowed for the visualization of the RNA expression levels.

In fluorescence microscopy, the illumination scheme commonly used is wide-field epi-illumination. In epi-illumination, the entire depth of the sample is excited making the out-of-focus molecules to contribute also for the background fluorescence.

In order to reduce the out-of-focus molecules, different microscopy methods were developed such as: confocal microscopy [85], total internal reflection florescence (TIRF) microscopy [86] and highly inclined and laminated optical sheet (HILO) [87], although TIRF is more commonly used to study processes *in vitro*.

4.3.1 HILO Time-lapse microscopy

To perform this project, HILO method was used to obtain time-lapse microscopy images. The HILO method was developed in order to restrict the illumination volume of the sample, since previous methods excited the entire depth of the sample making out-fluorescence molecules, which contributed to the background fluorescence (out-of-focus fluorescence). In HILO microscopy, the light is refracted into the sample with a high inclination angle, illuminating only an angled layer within the sample, thus reducing the out-of-focus fluorescence.

In HILO time-lapse microscopy, hundreds of images may need to be taken, with each image containing dozens of cells. For our study, in each IPTG condition, an HILO image was taken every minute for a period of two hours (Fig. 15). For excitation of the molecules, a 488 nm wavelength was applied during 100ms, while the radiation emission was in between 515 and 530 nm.

RNA molecules tagged with MS2-GFP are detected through the green channel of the microscope with single-molecule sensitivity. mCherry proteins can also be visualized through the red channel, although since they diffuse in the cytoplasm and have much weaker intensity levels, only total fluorescence intensity in the cell can be measured. Meanwhile, RNA spots are seen to move freely in the cytoplasm with tendency to accumulate in the cell´s poles, due to a nucleoid-exclusion phenomenon [88].

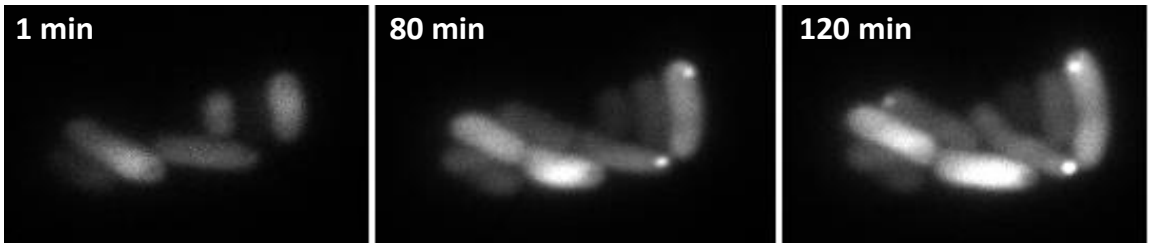

Fig. 15: *An example image of HILO time-lapse images for the 25µM IPTG condition 1 minute, 80 minutes and 2 hours after the start of the time-series.*

4.3.2 Phase-contrast microscopy

Phase-contrast microscopy was first described by the Dutch physicist Zernike [89]. This method made possible to observe high-contrast images of transparent samples without the need to colour them. This optical technique translates phase shifts in amplitude differences, thus enhancing the image contrast.

When the light rays hit a single cell in the sample, under the microscope, they propagate with less speed and therefore the scattered rays will be retarded in phase when compared with the background light by nearly -90˚. This gives rise to a slightly defocusing and not very detailed image. However, in the phase contrast method, the background light is also phase-shifted by a phase-shift ring. In our lab, we make use of negative phase contrast (darker foreground and lighter background), so that the phase-shift ring shifts the background light by +90˚, thus destructive interference happens when background light and scattered light rays meet, making the cells appear darker than the background (Fig. 16).

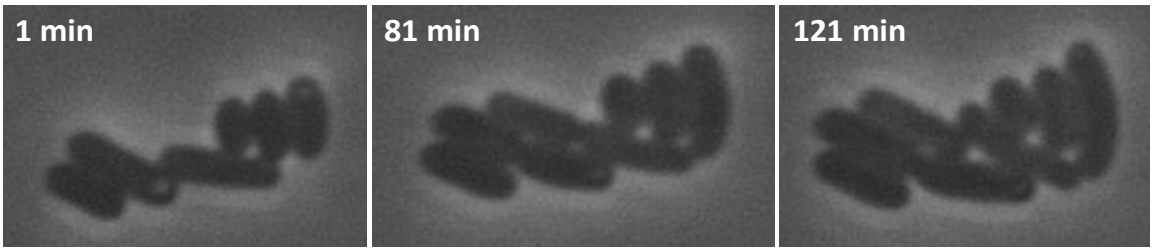

Fig. 16: *An example image of Phase-contrast time-lapse images for the 25µM IPTG condition 1 minute, 81 minutes and 121 minutes after the start of the time-series.*

For time-lapse measurements, phase-contrast images were acquired each 5 minutes during a period of two hours.

4.4 Image Analysis and Data Extraction

To conduct this project, detailed image analysis of time-lapse microscopy images was performed. To process the images, first the cells need to be segmented. The segmentation is done by an automatic method [90] able to find the approximate dimensions and orientation of the cells, after which manual correction is done. Next, automatic segmentation of the RNA spots is done by kernel density estimation [90].

The computational methods used here [90] also establish a temporal relationship between the cells, which allows creating a spatial and temporal distribution of fluorescence intensity.

From the total spot intensity in each cell (following the background correction), we can determine the number of RNAs. Since the tagged RNA is immortalized, the intensity level should increase monotonically. Jumps in intensity correspond to new transcription events that produced a new target RNA in the cell. Distributions of the time intervals between consecutive transcription events in individual cells can be obtained, and from them we can infer the number and duration of the intermediate rate-limiting steps in transcription initiation. The theory behind these steps is further explained in the following subchapters.

4.4.1 Cells and Spots Segmentation

Once the fluorescent and phase-contrast images of the cells are acquired we perform a temporally alignment using cross-correlation. This step is essential as it allows to remove drifts of the cells in the image plan, since the movement of the cells over time can make extremely difficult the process of cell tracking.

The next step in the analysis is to segment the cells. Cells were detected from the phase-contrast images using a tool developed in the lab for automatic segmentation and cell tracking [90]. The algorithm is able to identify the cell region and draw a mask over it, which, afterwards, needs to be subject to visual inspection and go through manual correction if needed. From the masks region, the location, orientation and dimensions of the cells are obtained by principal component analysis (PCA). Cells crossing the borders of the frames were not masked.

After cell segmentation, an automatic alignment of the segmented phase-contrast images and the fluorescence images was performed, followed by manual correction. An example of the results of the alignment step can be seen in Fig. 17.

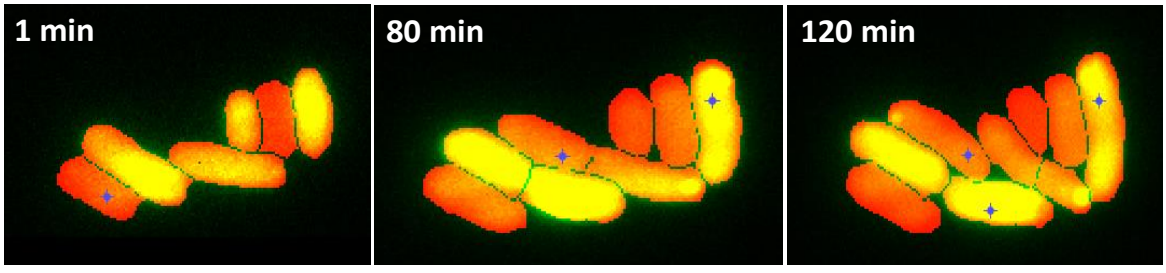

Fig. 17: *An example image of phase-contrast and fluorescence time-lapse images alignment for the 25μM IPTG condition at minutes 1, 80 and 120 of the measurements. The blue dots correspond to the points created in order to drag and anchor the phase-contrast image to overlap the HILO image in the respective place.*

4.4.2 RNA Quantification

To quantify the RNAs in each cell there is a need to segment the MS2-GFP RNA spots so as to measure their respective intensities. The spots are automatically segmented using the Kernel Density Estimation method [91]. Briefly, this method estimates the probability density function of the pixel intensities of each spot, finding a cut-off point which corresponds to the first local minimum of the KDE. Next, each pixel is analyzed and segmented only if its value is bigger than the cut-off [92].

The unbound MS2-GFP molecules cause a background fluorescence in each cell, which needs to be taken in account when estimating the total spot intensity of the cell. In order to perform this correction, the mean background intensity of the cell is multiplied by the area of the spot and then subtracted from the total intensity of the spot. Following this correction, one can then estimate the number of RNA molecules in each spot, by normalizing the spot-intensity histogram by the intensity of the first peak, which corresponds to the intensity of a single RNA [54].

4.4.3 Measurement of time intervals of RNA production events

From cell population data, it is possible to extract the mean RNA production rate during a certain time. Meanwhile, with time-series data it is possible to extract more detailed information about the dynamics of RNA production.

The MS2-GFP tagged RNA molecules do not degrade over time, thus when monitoring the total spots intensity level of each cell, with the respective background correction, we expect it to increase over time. The production of a new RNA molecule will cause a discrete jump in the total spots intensity of the cell.

Using the automated method described in [93], where a monotone piecewise-constant function by least squares fits the total spot intensity of each cell, over time, we are capable of extracting the time-intervals between consecutive RNA production events. The order of the model is selected by the use of the F-test (p-value 0.01), where for a higher order curve to be selected it has to fit the data significantly better [93]. From the result of this fitting, distributions of the intervals between consecutive RNA productions were extracted for each condition studied (5, 10, 25 and 50 μM IPTG). In Fig. 18 we show an example of the results of the jump detection method for a specific cell subject to 50 μM IPTG induction.

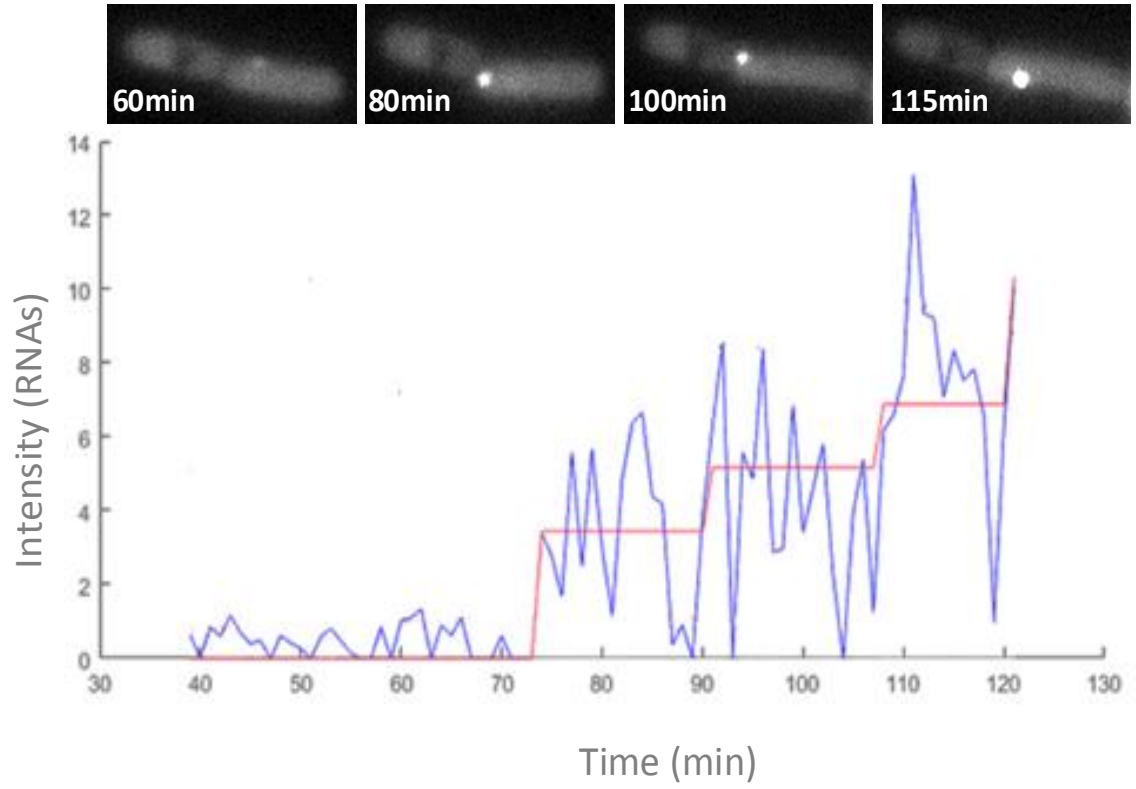

Fig. 18: *An example of the detection of RNA production events from time-lapse microscopy. This example shows the intensity series and the fit curve (red) for a cell under 50 µM IPTG. To show the correlation with the visual inspection of the cell's spots, it is shown in the top row the respective florescence microscope images at 60, 80, 100 and 115 minutes.*

4.4.4 Censored Data and uncertainties

Trough likelihood-based statistical methods we are capable of inferring the model parameters that best fit our censored data. Likelihood is used to make inferences from reliable data. Models are fit to the data for which the probability of the data is large. Model parameters are then estimated based on the combination that maximizes the likelihood function. Likelihood functions are equal to the probability of the data, assuming there were n measured intervals the likelihood function for the data set is [94]:

$$L(\theta) = \prod_{i=1}^{n} L_i(\theta; data_i) \tag{4.1}$$

Where L_i is the probability of observation i, assuming a model with the parameter θ. Since in our data there is interval and right-censored data, both will have specific likelihoods. Since in right-censoring data it is only known that the interval value is greater than some value, the likelihood for this type of data is written as follow:

$$L(\theta)_i = \int_{t_1}^{x} f(t)dt = 1 - F(t_1) \tag{4.2}$$

Here, *f(t)* represents the probability density function and *F(t)* the CDF of *f(t)*. In the case of interval-censoring data, the measured time is known to have occurred between the times t_1 and t_2, the probability of this event can then be written as:

$$L(\theta)_i = \int_{t_1}^{t_2} f(t)dt = F(t_2) - F(t_1) \tag{4.3}$$

Given this, the full likelihood function can be written as:

$$L(\theta) = \prod_{i=1}^{n} [F(t_2) - F(t_1)].[1 - F(t_1)] \tag{4.4}$$

From equation (4.4) the maximum likelihood estimator θ that maximizes the likelihood function can be calculated. In order to find this maximum there is the need to differentiate *L(θ)*. For that, it is more convenient instead to differentiate and set to zero the log-likelihood function, since the value θ that maximizes the log-likelihood is the same that maximizes the likelihood. In order to verify that indeed we obtained a maximum, we can show that the second derivative of the log-likelihood is negative. Once θ is found, we know the parameter of the model that best fits our data. This analysis was made for every inducer concentration condition and the best exponential distribution fit can be seen in Fig. 21.

For computing the standard uncertainty of the maximum likelihood estimation function, the Delta Method was used [95]. The Delta Method creates a linear approximation of the function that best fits the data, through the use of a Taylor Series expansion, and then computes its variance, otherwise it would be tremendously complex to compute the variance of the function fitting the data.

4.5 Modeling Gene Expression

4.5.1 Model of Transcription Initiation

Knowing that transcription is a sequential process composed by two or more rate-limiting steps which cannot be discarded, here we considered a model that accounts for these main features known to occur during transcription initiation.

In our model, different states of the promoter are represented allowing for the model dynamics to range from sub- to super-Poissonian. We expect this model to be applicable to a different set of promoters, although some slight changes will be needed to match specific dynamics or regulatory mechanisms.

The following set of reactions below represent the model of transcription initiation considered:

$$\mathrm{P}_{OFF} \underset{k_{OFF}*[\mathrm{Rep}]}{\overset{k_{ON}}{\rightleftarrows}} \mathrm{P}_{ON} \tag{4.6}$$

$$\mathrm{P}_{ON} \underset{k_{-1}}{\overset{k_1*[\mathrm{RNAP}]}{\rightleftarrows}} \mathrm{P}_{CC} \xrightarrow{k_2} \mathrm{P}_{OC}{}^{i} \xrightarrow{k_3} \mathrm{P}_{OC}{}^{f} \xrightarrow{k_4} \mathrm{P}_{ON} + RNA \tag{4.7}$$

RNA production from active promoter starts with the closed complex formation, which includes multiple RNA polymerase (RNAP) binding and unbinding events, as well as possible binding and unbinding of repressor molecules. When the RNAP is bound to the promoter, the open complex formation process competes with the unbinding of the RNAP. After the open complex is formed, it is followed by the promoter escape and RNA production. We represent this in reactions (4.6) and (4.7).

Reaction (4.6) shows the reversibility between the ON and OFF state of the promoter (P). Mechanisms responsible for this step were experimental identified as being, e.g., the binding and unbinding of repressor and activator molecules [96].

Reaction (4.7) represents the transcription initiation steps of an active promoter. Once the RNAP finds the TSS, the promoter proceeds to form the closed complex. This stage is dependent on the concentration of RNAP and on the rate constant to which the RNAP binds to the TSS, represented by k_1. Once in this stage, the promoter can return to the previous state. Notice that $\mathrm{P_{CC}}$ state complies all the substates until the occurrence of the first irreversible reaction. Following this, k_{-1} and k_2 are then the multiplication result of the rate constants for the elementary reactions that lead to the previous or next state. Once the promoter is in the initial stage of the open complex formation (P_{OC}^{i}) the process becomes almost irreversible and, in the presence of Mg^{2+}, the isomerization process is complete and the fully formed state of the open complex is reached (P_{OC}^{f}) [97]. Finally, the RNAP escapes the promoter and elongation begins creating a new RNA molecule. In our model, elongation is not explicitly represented since it is a fast process when compared with the binding, isomerization and promoter clearance steps.

4.5.2 τ Plots

McClure in 1985 [2] was able to demonstrate that the abortive initiation reaction method of RNA polymerase promoter bound can be used to access the time-length of the rate-limiting steps of the transcription initiation mechanism.

During these studies it was observed that a lag time was present before reaching a steady-state rate of abortive initiation, after the mixing of the promoter with the RNAp. This lag time was interpreted as the time needed for the binding of the enzyme, and subsequent isomerization, to form an open complex.

In order to do a quantitative separation of the binding and isomerization steps, McClure described the binary complexes as reversible processes and modelled the transcription initiation as a two-step process (4.8) [2].

$$\mathrm{R} + \mathrm{P} \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \mathrm{RP}_C \underset{k_{-2}}{\overset{k_2}{\rightleftarrows}} \mathrm{RP}_O \tag{4.8}$$

Where R is free RNAp, P is free promoter, RP_C and RP_O are the closed and open complex, respectively. From the derivation of this model kinetics and using the same approach as in [98], it is possible to quantify the rate-limiting steps of transcription initiation.

From the application of the steady-state condition to RP_C (see [2] and [98] for a review) and considering that the equilibrium lies far to the right:

$$k_{obs} = \frac{k_1[\mathrm{R}]k_2}{k_1[\mathrm{R}] + k_{-1} + k_2} \tag{4.9}$$

k_{obs} being then, the formation rate of RP_O which can be experimentally measured. Thus the inverse of k_{obs} corresponds to the average time to fully form the open complex:

$$\tau_{obs} = \frac{1}{k_2} + \frac{k_{-1} + k_2}{k_1[R]k_2} \tag{4.10}$$

From equation (4.10) that describes the average time spent for the formation of the open complex it is now possible to do a “t plot” between τ_{obs} and the reciprocal of the RNAp concentrations (Fig. 19).

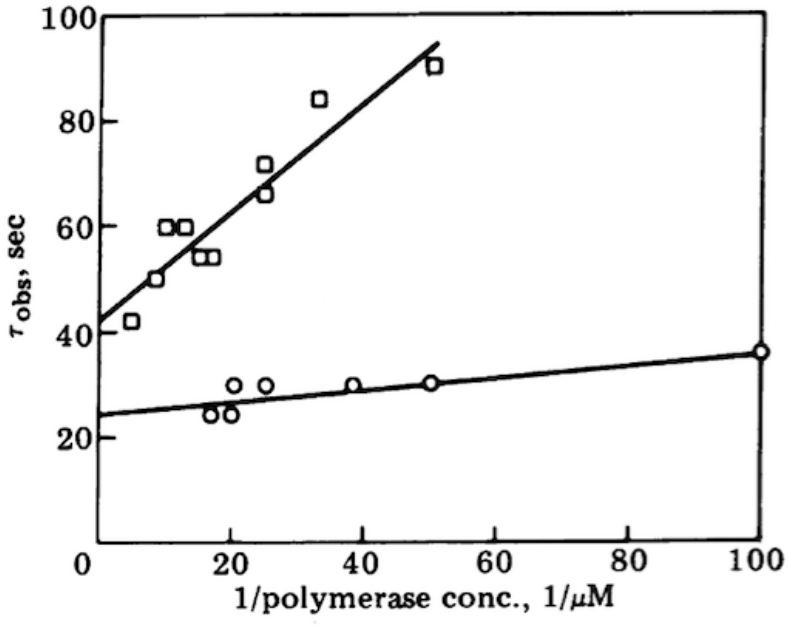

Fig. 19: *tau plot for the bacteriophage T7 D and A2 promoters. The lag times observed (τ_{obs}) for pGpUpu synthesis from the D promoter (squares) and pGpC synthesis from the A2 promoter (circles) are plotted versus the reciprocal of the RNAp concentrations. Image taken from [2].*

As the rate of the RP_C is proportional to the concentration of RNAp, it is expected a linear relationship. The slope in the plot is the mean time for the closed complex formation and the intercept with the y-axes, $\frac{1}{k_2}$ is the mean time for the open complex formation. McClure and colleagues used this method to dissect the *in vitro* the kinetics of transcription initiation.

More recently, in [6] it was shown that, given distributions of *in vivo* mean intervals between RNA production events in sets of individual cells with different RNAp concentrations (within a certain range), a similar method could be employed.

Using the method in [6], the time-length between RNA productions for an infinite amount of RNAp was inferred. For an infinite amount of RNAp, it is considered that the rate for the closed complex formation is infinitely fast, thus the y-axes value corresponds only to the rate of the open complex formation.

In our present study, a similar strategy is proposed, but instead of measuring the time-intervals between RNA production events for cells under different RNAp concentrations, we measured for cells under different inducer, namely IPTG, concentrations. Assuming our model and following the same approach as in [2] and [98], the model kinetics was derived and a τ plot was constructed allowing for the extrapolation of the time spent between RNA productions under infinite intracellular IPTG concentrations.

4.5.3 Fitting line procedure and uncertainty

In chapter 4.4.4 we described the method of parameter estimation by maximum likelihood. In order to do the τ-plot shown in Fig. 22 we fitted the data to a straight line. Each data point $\left(x_{i,}\ y_{io}\right)$ has its own standard uncertainty $\left(\sigma_i\right)$ calculated with the Delta Method as described in chapter 4.4.4 and the values of x_i are known exactly. We want our inferred straight line to have small residual values (4.1.1). A residual value is known to be the difference between the observed value $\left(y_{io}\right)$ and the value calculated by the straight line equation (y_{ic}) (4.12).

$$y_{ic} = mx_i + b \tag{4.11}$$

$$\Delta y = y_{io} - y_{ic} \tag{4.12}$$

Given this, the best fitting straight line can be calculated through the minimization of the sum of the squares of the residuals. Knowing that each data point has an uncertainty of σ_i we define:

$$\chi^2 = \sum_{i=1}^{i=N}\left[\frac{\left(y_{io} - y_{ic}\right)}{\sigma_i}\right]^2 = \sum_{i=1}^{i=N}\left[\frac{(y_{io} - (mx_i + b))}{\sigma_i}\right]^2 \tag{4.13}$$

Equation (4.13) is known has the chi-squared function. The sum considers that we have N data points to do the fitting. Minimizing the chi-square function gives us the maximum likelihood estimate of the line parameters [99].

After inferring the best fitting line, we calculated the uncertainties in the estimates of m and b, since the uncertainty associated with each measurement should contribute for the uncertainty in the estimation of these parameters. These uncertainties were calculated using the propagation of errors formula (also known as Delta Method) [99]:

$$\sigma_f = \sqrt{\sum_{i=1}^{N}\sigma_i^2\left(\frac{\partial f}{\partial y_{ic}}\right)^2} \tag{4.14}$$

In our specific case, the squared partial derivatives correspond to the derivatives of m and b with respect to y_{ic}. After computing the results, we are able to obtain the standard uncertainty associated with the slope (m) and y-axes interception (b) of the best fitting line. Results can be seen in Fig. 22.

5. RESULTS AND DISCUSSION

This chapter contains the results obtained in this study.

As described above, the kinetics of transcription is limited by the rate-limiting steps of transcription initiation which can differ between promoters. In the case of the LacO3O1 promoter, IPTG is known to affect one of these rate-limiting steps, namely, the closed complex formation, by 'inactivating' the repressor molecules, LacI [100].

Here, we propose a new strategy to dissect the rate constants of the transcription repression mechanism of the LacO3O1 promoter from live single cell, single microscopy data. The method relies on measurements of the *in vivo* kinetics of RNA production under the control of the LacO3O1 promoter subject to different levels of induction. To validate our results, we compared the results above with direct measurements of the transcription kinetics of this promoter when integrated into a mutant strain incapable of producing the repressor molecules.

5.1 Parameter estimation

Assuming the model described by equations 4.6 and 4.7, we derived the equations that support our method using the same approach as in [2]. This model can be re-written as:

$$\mathrm{P}_{OFF} \underset{k_{OFF}*[\mathrm{Rep}]}{\overset{k_{ON}}{\rightleftarrows}} \mathrm{P}_{ON} \underset{k_{-1}}{\overset{k_1*[\mathrm{RNAP}]}{\rightleftarrows}} \mathrm{P}_{CC} \xrightarrow{k_2} \mathrm{P}_{OC} \xrightarrow{k_3} \mathrm{P}_{ON} + RNA \tag{5.1}$$

In this model, P_{OC}^{i} and P_{OC}^{f} (from model 4.7) were converted in only one state P_{OC}. This assumes that the new rate constant k_2 of the model (5.1) includes all the elementary reactions necessary to form the fully formed open complex (P_{OC}) (i.e. it represents the product of k_2 and k_3 of the previous model (4.7).

According with model (5.1) the formation rate of one RNA is given by (5.2):

$$\frac{dRNA}{dt} = k_3[\mathrm{P}_{OC}] \tag{5.2}$$

Thus, the time between consecutive transcription initiation events is given by:

$$\Delta t = \frac{1}{k_3[\mathrm{P}_{OC}]} \tag{5.3}$$

However, this result does not satisfy, since we aim for a more detailed decomposition of Δt. In order to accomplish that, we now write the reaction rates associated to each state of model (2):

$$\frac{dP_{OC}}{dt} = k_2[\mathrm{P}_{CC}] - k_3[\mathrm{P}_{OC}] \tag{5.4}$$

$$\frac{dP_{CC}}{dt} = (-k_2 - k_{-1})[\mathrm{P}_{CC}] + k_1[RNAP][\mathrm{P}_{ON}] \tag{5.5}$$

$$\frac{dP_{OFF}}{dt} = -k_{ON}[\mathrm{P}_{OFF}] + k_{OFF}[\mathrm{Rep}][\mathrm{P}_{ON}] \tag{5.6}$$

The steady-state approach makes use of the assumption that the rate of production of an intermediate is equal to the rate of its consumption, one can write the above equations as follows:

$$\frac{dP_{OC}}{dt} = 0 \Leftrightarrow [\mathrm{P}_{CC}] = \frac{[\mathrm{P}_{OC}]k_3}{k_2} \tag{5.7}$$

$$\frac{dP_{CC}}{dt} = 0 \Leftrightarrow [\mathrm{P}_{ON}] = \frac{(k_2 + k_{-1})k_3[\mathrm{P}_{OC}]}{[RNAP]k_1k_2} \tag{5.8}$$

$$\frac{dP_{ON}}{dt} = 0 \Leftrightarrow [\mathrm{P}_{OFF}] = \frac{k_{OFF}[\mathrm{Rep}][\mathrm{P}_{ON}]}{k_{ON}} \tag{5.9}$$

Knowing that:

$$[\mathrm{P}_{OFF}] = 1 - [\mathrm{P}_{ON}] - [\mathrm{P}_{CC}] - [\mathrm{P}_{OC}] \tag{5.10}$$

Combining both equations (5.10) and (5.9) one can write:

$$1 - [\mathrm{P}_{ON}] - [\mathrm{P}_{CC}] - [\mathrm{P}_{OC}] = \frac{k_{OFF}[\mathrm{Rep}][\mathrm{P}_{ON}]}{k_{ON}} \tag{5.11}$$

Changing $[\mathrm{P}_{ON}]$ and $[\mathrm{P}_{CC}]$ for the respective equations (5.8) and (5.7) one gets:

$$1=\frac{(k_2+k_1)k_3[\mathrm{P}_{OC}]}{[RNAP]k_1k_2}\times\left(1+\frac{k_{OFF}[\mathrm{Rep}]}{k_{ON}}\right)+\frac{k_3}{k_2}[\mathrm{P}_{CC}]+[\mathrm{P}_{OC}] \tag{5.12}$$

From here, we can obtain Δt, based on equation (5.3), by moving the terms $k_3[\mathrm{RP}_O]$ to the right side of the equation:

$$\Delta t=\frac{(k_2+k_{-1})k_{OFF}[\mathrm{Rep}]}{[RNAP]k_1k_2k_{ON}}+\frac{(k_2+k_{-1})}{[RNAP]k_1k_2}+\frac{1}{k_2}+\frac{1}{k_3} \tag{5.13}$$

From this, note that it is possible to conceive measurement conditions such that all terms can be kept constant, while varying only [Rep] (e.g., by having conditions differing in the concentration of inducers that block the repressors' activity). As such, we expect Δt to vary linearly with [Rep].

Next, we consider the model of transcription again (reactions 5.1), and write the mean time interval between consecutive transcription events as a function of the time spent in the various stages:

$$\Delta t=\tau_{OFF}+\tau_{CC}+\tau_{OC} \tag{5.14}$$

Here, τ_{OFF} is the mean time spent by the promoter in the OFF state between consecutive RNA production events. Depending on the promoter, the value of τ_{OFF} depends both on how long the promoter remains in that state once it gets there, as well as on the number of times it goes into that state until the successful initiation of an open complex. Next, τ_{CC} is the time that it takes to successfully form a closed complex, which in this dissection of Δt corresponds to the time that the promoter spends in the ON state (P_{ON}) as well as in forming a closed complex. Finally τ_{OC} represents the time it takes to complete the formation of the open complex, once initiated. As such, it includes various isomerization events, until the open complex is completed. The open complex formation is nearly irreversible as it only happens once between two consecutive RNA production events.

From equation (5.14) that describes the average time spent for the formation of the open complex, it is now possible to do a "t plot" of Δt and the reciprocal of the IPTG concentrations (Fig. 22).

5.2 Induction curve

As discussed, repressor molecules downregulate transcription by inhibiting transcription initiation. The mechanism of repression differs between promoters. In the case of the LacO3O1 promoter, LacI repressor molecules interact with the DNA, changing the RNAp binding affinity to the promoter [69]. When inducer molecules are introduced, they will bind to the LacI molecules, inducing a conformational change in the protein structure that causes the binding to the operator site to no longer be possible, and thus, when this compound is present in the cell, the transcription levels of LacO3O1 increase.

We first study how the transcription kinetics is altered with the introduction of inducers in the media. Since IPTG is not metabolized by the cell, once reaching equilibrium with the media, the intracellular concentration is maintained constant during the measurements, which allow us to study the transcription kinetics of promoters in *E .coli* for constant levels of ITPG

Based on this, we first measure the required changes in the inducer concentrations for which transcription rates differ widely and the concentration beyond which maximum RNA production is reached. For this, for the different levels of induction of LacO3O1, we measured the mean RNA numbers per cell 2 hours after the activation of the target gene. Results are shown in Fig. 21.

When measuring the mean RNA numbers per cell from microscopy images in a given time moment following induction, one needs to have in account the effects of cell division, which acts as a 'RNA dilution' mechanism. Meanwhile, here there is no need to have into account RNA degradation effects, since the MS2 tagged RNA molecules are 'immortalized' once bound, as described previously. In [101] the dilution rate was estimated based on the optical density (OD) measurements of the cultures over time (5.15).

$$k_d = \frac{\ln(2)}{Div} = \frac{(\ln OD_2 - \ln OD_1)}{(t_2 - t_1)} \tag{5.15}$$

Here t_1 and t_2 correspond to the corresponding time points of the measurements of OD_1 and OD_2, respectively. Once the dilution rate is known, it is possible to estimate the mean rate of RNA production (λ_{RNA}):

$$\lambda_{RNA} = \frac{k_d \times RNA(t)}{1 - e^{-k_d \times t}} \tag{5.16}$$

Where RNA(t) is the number of RNAs per cell at a given time (t). Results of this calculations can be seen in Table 1 and Table 2.

Table 1. *Mean RNA production from LacO3O1 promoter at 37˚C under different levels of induction. Mean RNA numbers were extracted from single time point images captured after 2 hours following the activation of the target gene. (Methods). Standard deviations (σ) and standard errors of the mean are also presented.*

Condition [IPTG] (μM)	No. of Cells	Mean RNA	σ	SEM
0	516	0.03	0.21	0.01
5	452	0.11	0.44	0.02
50	443	0.35	0.99	0.05
100	383	0.37	0.94	0.05
250	489	0.39	0.96	0.04
500	551	0.41	1.05	0.04
1000	377	0.37	0.84	0.04

Table 2. *RNA dilution rate (5.15), and RNA production rate (5.16). Production ratio relative to the 0 μM condition is calculated. Standard error of the mean is also presented.*

Condition [IPTG] (μM)	Dilution Rate (h^{-1})	Production Rate (h^{-1})	Production Rate SEM (h-1)	Production Ratio (to 0μM)	Production Ratio SEM (to 0μM)
0	0.34	0.01	0.91	1	0.01
5	0.34	0.01	1.25	3.23	0.05
50	0.34	0.03	1.92	10.77	0.36
100	0.34	0.03	1.64	11.35	0.38
250	0.34	0.03	1.62	11.87	0.36
500	0.34	0.03	1.85	12.41	0.39
1000	0.34	0.03	1.33	11.29	0.35

In order to calculate the RNA dilution rate, the optical density of the samples was measured in the beginning (t_1=0) and after the first hour (t_2=1). For all conditions, OD_1= 0,45 and OD_2=0,99. RNA production rate was calculated with the mean RNA values presented in Table 1. To calculate the production ratio the 0 µM IPTG condition was used as reference. In Fig. 20 results of the RNA production relative to the control condition are presented.

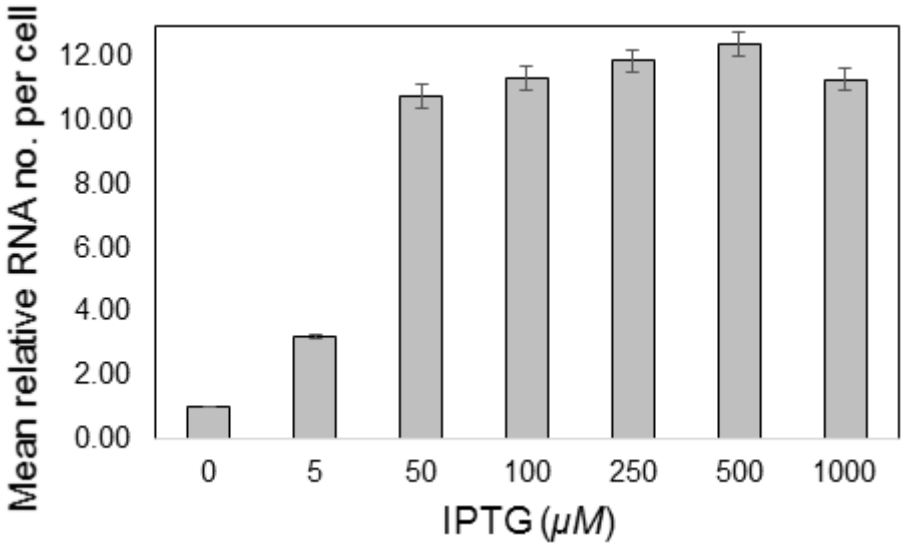

Fig. 20. *Mean relative RNA produced in individual cells. Images were taken 2 hours after the activation of the target gene. Error bars represent the standard uncertainty of the mean.*

From Fig. 20 it is noticeable that full induction occurs for an IPTG concentration of 50 µM and above. RNA production rates are most sensitive to changes in IPTG concentrations in the interval [0, 50] µM. Following this result, time-lapse microscopy images of cells under a concentration of 5, 10, 25 and 50 µM IPTG was performed (details of how the images were obtained can be read in Methods).

5.3 Interval distributions

From the induction curve, we know that maximum induction occurs for a concentration of 50 µM IPTG onwards. Since our goal is to dissect the rate constants associated with the repression mechanism, we thus limit our study to the range of induction levels for which the transcription rate is most sensitive.

For this, we performed microscopy time-series of cells under 5, 10, 25 and 50 µM [IPTG]. We also imaged a deletion mutant, without the ability to express repressor molecules, to validate our results. Details of the image analysis procedure can be read in Methods.

Based on likelihood statistical methods, we inferred the model parameter values that best fit our censored and non-censored data. Meanwhile, from the time-lapse microscopy images, we extracted the number of cells as well as the number and duration of the intervals between RNA production events in single cells (Methods).

The shapes of the distributions for the observed intervals, and the PDFs of the estimated models both for non-censored and censored data are shown in Fig. 21 on the right panel. On the left panel, CDFs for the respective models are also shown. The statistics of the measured intervals are shown in Table 3. This table shows the number of cells analyzed for each condition, as well as the total number of censored intervals between RNA production events along with the mean and respective standard uncertainty. We also estimated the CV^2 of the interval distribution. The CV^2 is used here to quantify the variability in the interval distributions and thus, is a measure of noise in transcription. Since CV^2 is slightly larger than 1 for all conditions, we conclude that RNA production is a super-Poissonian process in these conditions. Given our model of transcription, this indicates that the promoter must spend some time in the OFF state between transcription events (without this, the process would either be Poissonian or sub-Poissonian).

From Fig. 21 it is possible to see that in the absence of censored data, there is a significant underestimation of the mean interval between consecutive RNA production events, as intervals outside the measurement window are not taken into account when estimating the mean.

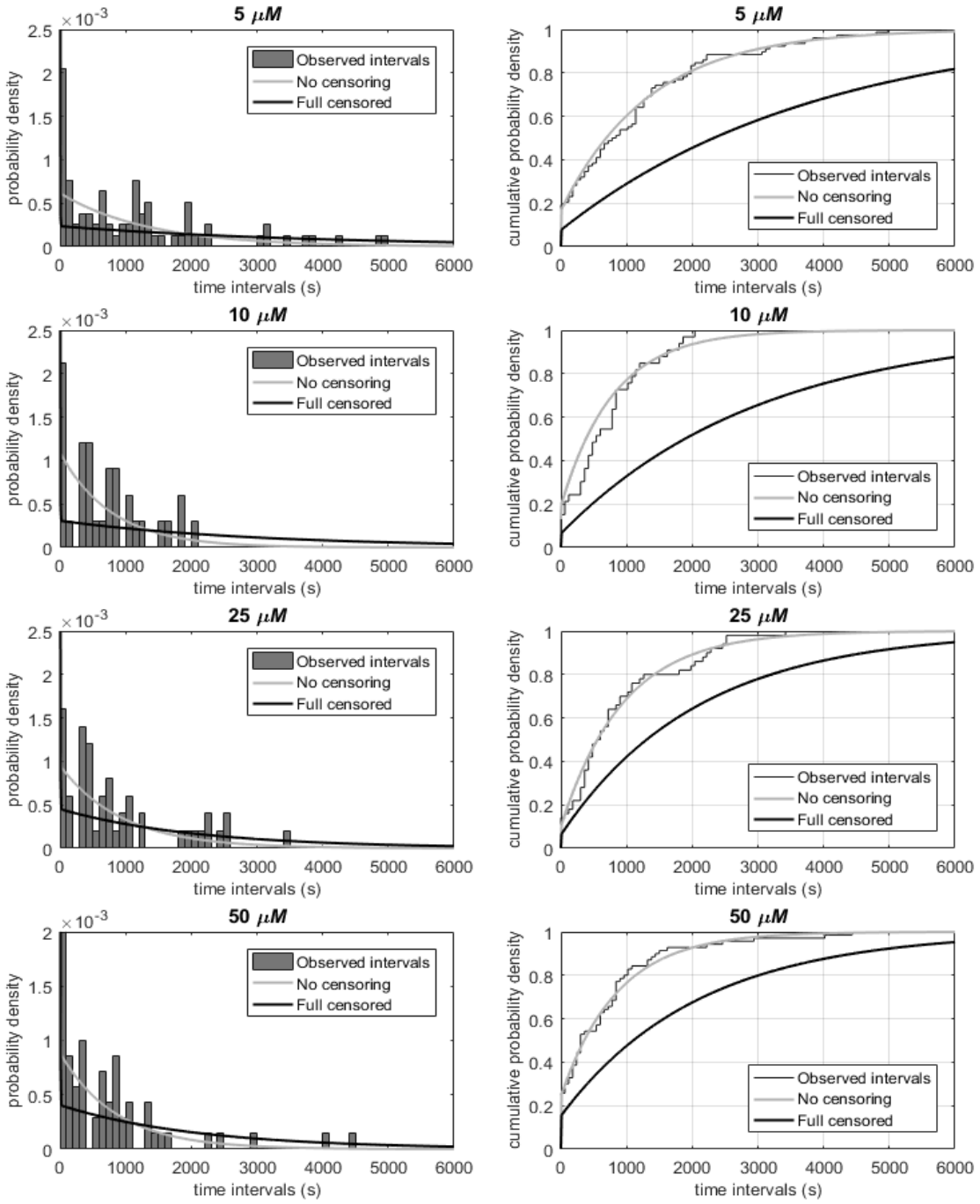

Fig. 21: *Transcription intervals for the LacO3O1 promoter. The left panels show the histograms of the observed intervals for each condition(5, 10, 25, and 50* μM*) together with the PDF's for no censored intervals and censored intervals. As expected neglecting the unobserved intervals leads to and underestimation of the mean intervals. The left panels show the corresponding CDFs.*

Table 3. *Mean and uncertainty of the interval between transcription events in individual cells for the LacO3O1 promoter. Amount of empirical data and CV^2 are also shown.*

Condition	No. of Cells	No. of intervals	Mean inferred interval and uncertainty (s)	CV^2
5 μM	360	186	3632 ± 500	1.161
10μM	246	87	2852 ± 553	1.191
25 μM	144	90	1948 ± 345	1.131
50 μM	173	119	1771 ± 254	1.359
Deletion Mutant	44	24	1607 ± 621	2.029

From Table 3 and Fig. 21, the mean interval decreases significantly with increasing IPTG concentration, as expected, since with the increase of inducer concentration it is presumed that more repressor molecules will be inactive (bound to the inducer), allowing for transcription initiation to occur. One should notice that, by altering IPTG concentrations, we should be changing k_off*[REP] (see chapter 4.5.1 for modeling details), i.e., we should be changing the number of times the promoter goes into P_{OFF} state and not how long it remains in that state.

5.4 Dissection of the *in vivo* kinetics – τ plot

Here we decompose the *in vivo* transcription kinetics of the LacO3O1 promoter of *E. coli* when integrated into the plasmid, as a function of the inducer concentration, in order to get τ_{OFF}.

From the decomposition of Δt into τ_{OFF}, τ_{CC} and τ_{OC} described in the subchapter 5.1, it is expected that altering the concentration of IPTG only affects τ_{OFF}. Meanwhile, the closed complex formation should only be affected by changes in the free RNAp concentration, while the open complex formation depends only on the promoter (and bound inducers if existing, which is not the case here). Given this, we can re-write equation 5.14 as:

$$\Delta t = \tau_{OFF} + \Delta t_{ind} \tag{5.17}$$

where Δt_{ind} represents the mean time between transcription events, excluding the time that the promoter spends in the OFF state (t_{OFF}). Next, considering the range where RNA production rate is mostly sensitive to different IPTG levels, assuming a new condition differing in [IPTG], one can write:

$$\Delta t^{new} = \tau_{OFF}{}^{new} + \Delta t_{ind} \tag{5.18}$$

Using the same approach as in [6], where a rate estimation for RNA production was made for an infinite RNAp concentration, we plotted the results from a set of measurements of the mean interval between RNA production events for different levels of IPTG concentration.

Next, a linear fit (see Methods) was made to estimate the time-length between RNA productions for infinite IPTG concentration (Fig. 22).

Note that this linear fit, and the value obtained from it of the height at which it crosses the Y axis, is only a rough estimation of the duration of Δt_{ind}. In fact, the cell has a finite number of repressors and therefore, beyond a certain concentration of inducers we do not expect further increases in RNA production rate (which is visible in Fig 20). However, the very large number of repressors in the cell allows this estimation to be realistic, as shown below.

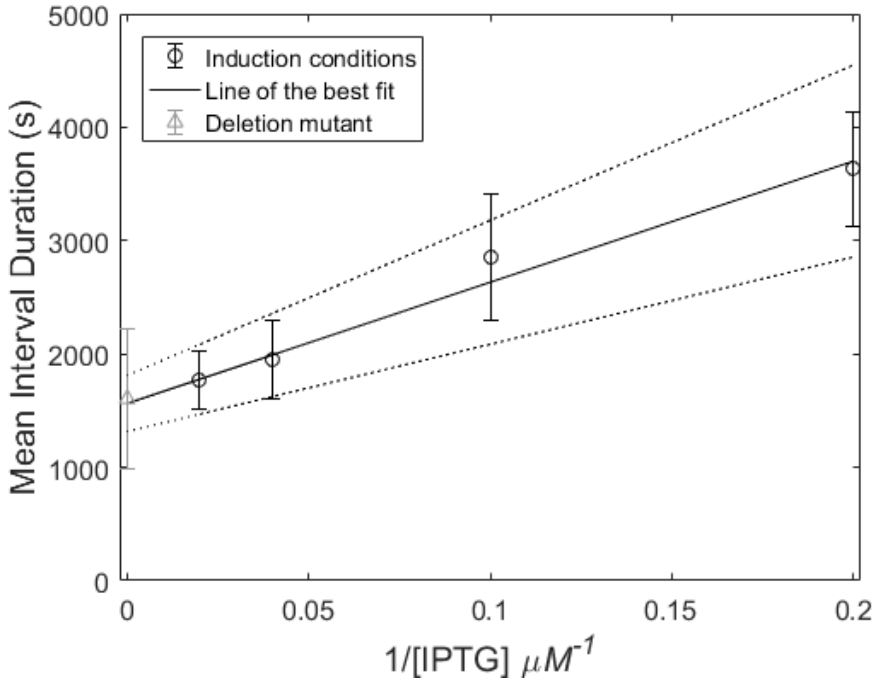

Fig. 22: *τ -plot as a function of inducer concentration for the LacO3O1 promoter. For different levels of IPTG (5, 10, 25 and 50 μM) Δt is shown (circles) along with their standard uncertainty. Also shown is the best-fit line, estimated by the chi-square merit function. Dotted lines represent the uncertainty of the best-fit line calculated by propagation of errors. Further the figure shows the data from the mutant strain lacking repressor molecules (triangle, not used for the estimation of the best-fitting line).*

In Fig. 22, it is represented the τ-plot as a function of inducer concentration for the LacO3O1 promoter. The X-axis corresponds to the inverse of [IPTG] and the Y-axis represents the mean duration of the transcription intervals (Table 3). Each circle in the picture represent a measured mean interval between RNA production events for a specific IPTG concentration along with the respective standard uncertainty (values shown in Table 3).

The line corresponds to the best fit line by minimization of the chi-squared function (Methods). The height at which the line of the fit crosses the y-axis corresponds to the Δt assuming an infinite amount of IPTG.

Physically, infinite IPTG implies that each repressor molecule is bound to an IPTG molecule, thus no repressor molecules are active in the system. Therefore, to validate this result, we measured the transcription kinetics of a mutant strain lacking the ability to express LacI molecules (triangle in Fig. 22). It is expected that, since there are no repressors in these cells, the time between RNA production events should match our extrapolation of the line for infinity [IPTG] (Δt^{inf}).

From the extrapolation of the line in Fig. 22, $\Delta t^{inf} = 1562$s, whereas the mean time between two consecutive RNA productions of the mutant strain is 1607s, which cannot be distinguished from Δt^{inf}, showing that the estimation is reliable.

Thus, this method is able to dissect the time that the promoter spends in the OFF state for each one of the conditions under study. Given the values of Δt, one can also obtain the fraction of time that the promoter spend in the OFF state Results can be seen in Table 4.

Table 4. *Results from the best-fitting line. The value of the best fitting line is shown for each condition, along with the absolute and the fraction of time that the promoter spend in the OFF state during two consecutive RNA productions.*

Condition [IPTG] (μM)	Line of best fit (s)	Time OFF (s)	%Time OFF
5	3702	2140	57.8
10	2632	1070	40.7
25	1990	428	21.5
50	1776	214	12
Deletion Mutant	1562	0	0

From Table 4, we find that the mean percentage of time spent by the promoter in OFF state in between transcription events changes with the induction level, as expected from the known effects of LacI on the dynamics of LacO3O1.

6. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a new methodology to extract the mean time spent by a promoter in the OFF state between consecutive RNA production events in live cells. To the extent of our knowledge, our study is the first to perform such dissection of these events in transcription from single-cell, single-RNA microscopy data.

When executing this method, prior knowledge in the repression mechanism is necessary. Repressors are known to downregulate transcription but far more than one mechanism exists, depending on the promoter. We expect the method to not be applicable to several cases (e.g. if the rate of transcription does not change linearly with the inducer or repressor concentration). Our measurements show that, in our case study, the time between consecutive RNA productions varies linear with the inverse of the inducer level, within the range studied. Given this linear relationship, an extrapolation of the duration of the transcription intervals time-length was possible for infinite levels of inducer concentration, allowing for the differentiation between the time that a promoter spends in the OFF state and the time it takes to form the closed and open complex.

Another factor that made this extraction possible was that the induction mechanism under study acts by releasing the promoter from an OFF state and does not interfere with the subsequent steps in transcription initiation, otherwise the time for closed and open complex formation would not remain constant for different levels of induction, which is not contemplated by the method presented here.

The methodology executed in this project was applied to the LacO3O1 promoter and its well studied repression system. This system acts by the bind of the repressor molecules to the operator site and, when inducers are present in the cell, they bind to the repressor molecules, inducing conformational changes that decrease their affinity with the operator site, freeing the promoter for transcription initiation. In this study, we used the molecular reagent IPTG as the inducer.

To validate our results, we collected and analyzed the data from a deletion mutant incapable of producing repressor molecules, which should exhibit the same behavior as the ‘infinitely’ induced system. We found that the mean duration of the intervals between RNA productions of this strain to be in clear agreement with our method’s expectation, showing that, in the absence of repressor molecules, the promoter exhibits the same kinetics as what is predicted for cells with an infinity concentration of inducers.

We note that there are mechanisms of repression that could cause OFF states in a promoter other than the one studied here. For example, the accumulation/release of local positive DNA super-coiling in chromosomal integrated genes, generated by transcription events, can cause OFF periods. While more detailed analysis of each case is needed for definitive answer, we expect our method to be applicable to several mechanisms.

Therefore, in the future, we expect our method to be applicable to a wide number of promoters in *E. coli*. Studies to compare, e.g., the efficiency of repressor binding for different inducers should be of particular interest, and may lead to interesting clues on how to tune promoter efficiency.

REFERENCES

[1] L. López-Maury, S. Marguerat, and J. Bähler, "Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation," *Nat. Rev. Genet.*, vol. 9, no. 8, pp. 583–593, Aug. 2008.

[2] W. R. McClure, "Rate-limiting steps in RNA chain initiation 1.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 77, no. 10, pp. 5634–5638, 1980.

[3] W. R. McClure, "Mechanism and Control of Transcription Initiation in Prokaryotes," *Annu. Rev. Biochem.*, vol. 54, no. 1, pp. 171–204, 1985.

[4] P. Xie, "A dynamic model for transcription elongation and sequence-dependent short pauses by RNA polymerase," *BioSystems*, vol. 93, no. 3, pp. 199–210, 2008.

[5] H. Maamar, A. Raj, and D. Dubnau, "Noise in Gene Expression Determines Cell Fate in Bacillus subtilis," *Science (80-.).*, vol. 317, no. 5837, pp. 526–529, Jul. 2007.

[6] J. Lloyd-Price, S. Startceva, V. Kandavalli, *et al.*, "Dissecting the stochastic transcription initiation process in live Escherichia coli," *DNA Res.*, vol. 23, no. 3, pp. 203–214, 2016.

[7] L. Pauling and R. B. Corey, "A Proposed Structure For The Nucleic Acids.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 39, no. 2, pp. 84–97, Feb. 1953.

[8] J. Watson and F. Crick, "Molecular Structure of Nucleic Acids." Nature, pp. 737–738, 1953.

[9] P. A. Levene, "The structure of yeast nucleic acids: IV. Ammonia hudrolysis," *J. Biol. Chem.*, vol. 40, pp. 415–424, 1919.

[10] E. Chargaff, R. Lipshitz, C. Green, *et al.*, "The composition of the deoxyribonucleic acid of salmon sperm.," *J. Biol. Chem.*, vol. 192, no. 1, pp. 223–230, Sep. 1951.

[11] L. Pray, "Discovery of DNA Structure and Function: Watson and Crick," *Nature Education*, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397. [Accessed: 10-Feb-2017].

[12] F. R. Blattner, G. Plunkett, C. A. Bloch, *et al.*, "The complete genome sequence of Escherichia coli K-12.," *Science*, vol. 277, no. 5331, pp. 1453–62, Sep. 1997.

[13] G. Cooper and R. Hausman, "The Cell: A Molecular Approach," *ASM Press*, 2009.

[14] F. H. C. Crick, "Central Dogma of Molecular Biology," *Nature*, vol. 227, no. 5258. pp. 561–563, 1970.

[15] A. Eliasson, R. Bernander, S. Dasgupta, *et al.*, “Direct visualization of plasmid DNA in bacterial cells.,” *Mol. Microbiol.*, vol. 6, no. 2, pp. 165–170, 1992.

[16] O. L. Miller, B. A. Hamkalo, and C. A. Thomas, “Visualization of bacterial genes in action.,” *Science*, vol. 169, no. 943, pp. 392–395, 1970.

[17] F. Jacob, D. Perrin, C. Sanchez, *et al.*, “Operon: a group of genes with the expression coordinated by an operator.,” *C. R. Hebd. Seances Acad. Sci.*, vol. 250, pp. 1727–9, Feb. 1960.

[18] A. E. Osbourn and B. Field, “Operons,” *Cell. Mol. Life Sci.*, vol. 66, no. 23, pp. 3755–3775, 2009.

[19] B. Alberts, A. Johnson, J. Lewis, *et al.*, *Molecular biology of the cell*. Garland Science, 2002.

[20] M. E. Karpen and P. L. deHaseth, “Base flipping in open complex formation at bacterial promoters.,” *Biomolecules*, vol. 5, no. 2, pp. 668–78, Apr. 2015.

[21] C. B. Harley and R. P. Reynolds, “Analysis of E. coli promoter sequences.,” *Nucleic Acids Res.*, vol. 15, no. 5, pp. 2343–2361, 1987.

[22] L. Bai, A. Shundrovsky, and M. D. Wang, “Sequence-dependent kinetic model for transcription elongation by RNA polymerase,” *J. Mol. Biol.*, vol. 344, no. 2, pp. 335–349, 2004.

[23] V. Ramakrishnan, “Ribosome structure and the mechanism of translation,” *Cell*, vol. 108, no. 4, pp. 557–572, 2002.

[24] R. M. Saecker, M. T. Record, and P. L. deHaseth, “Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis,” *J. Mol. Biol.*, vol. 412, no. 5, pp. 754–771, Oct. 2011.

[25] S. Borukhov and K. Severinov, “Role of the RNA polymerase sigma subunit in transcription initiation.,” *Res. Microbiol.*, vol. 153, no. 9, pp. 557–62, Nov. 2002.

[26] D. F. Browning and S. J. W. Busby, “The regulation of bacterial transcription initiation,” *Nat. Rev. Microbiol.*, vol. 2, no. 1, pp. 57–65, Jan. 2004.

[27] S. Borukhov and E. Nudler, “RNA polymerase: the vehicle of transcription,” *Trends Microbiol.*, vol. 16, no. 3, pp. 126–134, Mar. 2008.

[28] S. R. Goldman, R. H. Ebright, and B. E. Nickels, “Direct Detection of Abortive RNA Transcripts in Vivo,” *Science (80-.).*, vol. 324, no. 5929, pp. 927–928, May 2009.

[29] L. M. Hsu, “Promoter clearance and escape in prokaryotes.,” *Biochim. Biophys. Acta*, vol. 1577, no. 2, pp. 191–207, Sep. 2002.

[30] G. Bar-Nahum and E. Nudler, “Isolation and characterization of sigma(70)-retaining transcription elongation complexes from Escherichia coli.,” *Cell*, vol.

106, no. 4, pp. 443–51, Aug. 2001.

[31] A. N. Kapanidis, E. Margeat, T. A. Laurence, *et al.*, "Retention of Transcription Initiation Factor σ70 in Transcription Elongation: Single-Molecule Analysis," *Mol. Cell*, vol. 20, no. 3, pp. 347–356, Nov. 2005.

[32] M. Raffaelle, E. I. Kanin, J. Vogt, *et al.*, "Holoenzyme Switching and Stochastic Release of Sigma Factors from RNA Polymerase In Vivo," *Mol. Cell*, vol. 20, no. 3, pp. 357–366, Nov. 2005.

[33] T. T. Harden, C. D. Wells, L. J. Friedman, *et al.*, "Bacterial RNA polymerase can retain σ 70 throughout transcription," *Proc. Natl. Acad. Sci.*, 2016.

[34] W. R. McClure, C. L. Cech, and D. E. Johnston, "A steady state assay for the RNA polymerase initiation reaction," *J. Biol. Chem.*, vol. 253, no. 24, pp. 8941–8948, 1978.

[35] S. J. Greive and P. H. von Hippel, "Thinking quantitatively about transcriptional regulation," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 3, pp. 221–232, Mar. 2005.

[36] K. M. Herbert, A. La Porta, B. J. Wong, *et al.*, "Sequence-resolved detection of pausing by single RNA polymerase molecules.," *Cell*, vol. 125, no. 6, pp. 1083–94, Jun. 2006.

[37] K. M. Herbert, J. Zhou, R. A. Mooney, *et al.*, "E. coli NusG Inhibits Backtracking and Accelerates Pause-Free Transcription by Promoting Forward Translocation of RNA Polymerase," *J. Mol. Biol.*, vol. 399, no. 1, pp. 17–30, May 2010.

[38] T. Rajala, A. Häkkinen, S. Healy, *et al.*, "Effects of transcriptional pausing on gene expression dynamics," *PLoS Comput. Biol.*, vol. 6, no. 3, pp. 29–30, 2010.

[39] J.-D. Wen, L. Lancaster, C. Hodges, *et al.*, "Following translation by single ribosomes one codon at a time," *Nature*, vol. 452, no. 7187, pp. 598–603, Apr. 2008.

[40] M. A. Sørensen, C. G. Kurland, and S. Pedersen, "Codon usage determines translation rate in Escherichia coli.," *J. Mol. Biol.*, vol. 207, no. 2, pp. 365–77, May 1989.

[41] J. A. Bernstein, A. B. Khodursky, P.-H. Lin, *et al.*, "Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays," *Proc. Natl. Acad. Sci.*, vol. 99, no. 15, pp. 9697–9702, Jul. 2002.

[42] E. Bertrand-Burggraf, J. F. Lefèvre, and M. Daune, "A new experimental approach for studying the association between RNA polymerase and the tet promoter of pBR322.," *Nucleic Acids Res.*, vol. 12, no. 3, pp. 1697–706, Feb. 1984.

[43] S. Gama-Castro, H. Salgado, M. Peralta-Gil, *et al.*, "RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units).," *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D98-105, Jan. 2011.

[44] V. K. Kandavalli, H. Tran, and A. S. Ribeiro, “Effects of σ factor competition are promoter initiation kinetics dependent,” *Biochim. Biophys. Acta - Gene Regul. Mech.*, vol. 1859, no. 10, pp. 1281–1288, 2016.

[45] A. Ishihama, “Functional Modulation of *Escherichia Coli* RNA Polymerase,” *Annu. Rev. Microbiol.*, vol. 54, no. 1, pp. 499–518, Oct. 2000.

[46] W. Ross, C. E. Vrentas, P. Sanchez-Vazquez, *et al.*, “The Magic Spot: A ppGpp Binding Site on E. coli RNA Polymerase Responsible for Regulation of Transcription Initiation,” *Mol. Cell*, vol. 50, no. 3, pp. 420–429, May 2013.

[47] R. H. Ebright, “Transcription activation at Class I CAP-dependent promoters.,” *Mol. Microbiol.*, vol. 8, no. 5, pp. 797–802, May 1993.

[48] P. J. Schlax, M. W. Capp, and M. T. Record, “Inhibition of transcription initiation by lac repressor.,” *J. Mol. Biol.*, vol. 245, no. 4, pp. 331–50, Jan. 1995.

[49] A. Sanchez, M. L. Osborne, L. J. Friedman, *et al.*, “Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules,” *EMBO J.*, vol. 30, no. 19, pp. 3940–3946, Oct. 2011.

[50] H. G. Garcia, A. Sanchez, T. Kuhlman, *et al.*, “Transcription by the numbers redux: experiments and calculations that surprise.,” *Trends Cell Biol.*, vol. 20, no. 12, pp. 723–33, Dec. 2010.

[51] H. H. McAdams and A. Arkin, “Stochastic mechanisms in gene expression.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 3, pp. 814–9, Feb. 1997.

[52] M. B. Elowitz, A. J. Levine, E. D. Siggia, *et al.*, “Stochastic Gene Expression in a Single Cell,” *Science (80-.).*, vol. 297, no. 5584, pp. 1183–1186, Aug. 2002.

[53] Y. Taniguchi, P. J. Choi, G. W. Li, *et al.*, “Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells,” *Sci. (New York, NY)*, vol. 329, no. 5991, pp. 533–538, 2010.

[54] I. Golding, J. Paulsson, S. M. Zawilski, *et al.*, “Real-time kinetics of gene activity in individual bacteria,” *Cell*, vol. 123, no. 6, pp. 1025–1036, 2005.

[55] J. Yu, J. Xiao, X. Ren, *et al.*, “Probing Gene Expression in Live Single Escherichia coli Cells – One Molecule at a Time,” *Science (80-.).*, vol. 311, no. 5767, pp. 1600–3, 2006.

[56] A. Sanchez and I. Golding, “Genetic Determinants and Cellular Constraints in Noisy Gene Expression,” *Science (80-.).*, vol. 342, no. 6163, pp. 1188–1193, Dec. 2013.

[57] S. Chong, C. Chen, H. Ge, *et al.*, “Mechanism of Transcriptional Bursting in Bacteria,” *Cell*, vol. 158, no. 2, pp. 314–326, Jul. 2014.

[58] D. Huh and J. Paulsson, “Non-genetic heterogeneity from stochastic partitioning at cell division.,” *Nat. Genet.*, vol. 43, no. 2, pp. 95–100, Feb. 2011.

[59] A. S. Ribeiro, R. Zhu, and S. A. Kauffman, “A general modeling strategy for gene regulatory networks with stochastic dynamics.,” *J. Comput. Biol.*, vol. 13, no. 9, pp. 1630–1639, 2006.

[60] G. Walter, W. Zillig, P. Palm, *et al.*, “Initiation of DNA-dependent RNA synthesis and the effect of heparin on RNA polymerase.,” *Eur. J. Biochem.*, vol. 3, no. 2, pp. 194–201, Dec. 1967.

[61] M. R. Roussel and R. Zhu, “Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression,” *Phys. Biol.*, vol. 3, no. 4, pp. 274–284, Dec. 2006.

[62] A. S. Ribeiro and S. A. Kauffman, “Noisy attractors and ergodic sets in models of gene regulatory networks,” *J. Theor. Biol.*, vol. 247, no. 4, pp. 743–755, 2007.

[63] A. S. Ribeiro, O.-P. P. Smolander, T. Rajala, *et al.*, “Delayed stochastic model of transcription at the single nucleotide level. TL - 16,” *J. Comput. Biol.*, vol. 16 VN-r, no. 4, pp. 539–553, 2009.

[64] J. Mäkelä, J. Lloyd-Price, O. Yli-Harja, *et al.*, “Stochastic sequence-level model of coupled transcription and translation in prokaryotes.,” *BMC Bioinformatics*, vol. 12, no. 1, p. 121, 2011.

[65] I. Potapov, J. Mäkelä, O. Yli-Harja, *et al.*, “Effects of codon sequence on the dynamics of genetic networks,” *J. Theor. Biol.*, vol. 315, pp. 17–25, 2012.

[66] C. S. D. Palma, S. Startceva, R. Neeli-Venkata, *et al.*, “A strategy for dissecting the kinetics of transcription repression mechanisms,” *In Proceedings of the. European. Medical and. Biolgical Engineering Conference (EMBEC 2017)*, June 11-15, Tampere, Finland.

[67] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins,” *J. Mol. Biol.*, vol. 3, no. 3, pp. 318–356, Jun. 1961.

[68] D. H. Juers, B. W. Matthews, and R. E. Huber, “*LacZ* β-galactosidase: Structure and function of an enzyme of historical and molecular biological importance,” *Protein Sci.*, vol. 21, no. 12, pp. 1792–1807, Dec. 2012.

[69] H. F. Lodish, *Molecular cell biology*. W.H. Freeman, 1999.

[70] S. Oehler, M. Amouyal, P. Kolkhof, *et al.*, “Quality and position of the three lac operators of E. coli define efficiency of repression.,” *EMBO J.*, vol. 13, no. 14, pp. 3348–55, Jul. 1994.

[71] J. Chen, S. Alberti, and K. S. Matthews, “Wild-type operator binding and altered cooperativity for inducer binding of lac repressor dimer mutant R3.,” *J. Biol. Chem.*, vol. 269, no. 17, pp. 12482–7, Apr. 1994.

[72] R. Schleif, “Regulation of the L-arabinose operon of Escherichia coli.,” *Trends Genet.*, vol. 16, no. 12, pp. 559–65, Dec. 2000.

[73] A. Khlebnikov, O. Risa, T. Skaug, *et al.*, “Regulatable arabinose-inducible gene

expression system with consistent control in all cells of a culture.," *J. Bacteriol.*, vol. 182, no. 24, pp. 7029–34, Dec. 2000.

[74] T. Baba, T. Ara, M. Hasegawa, *et al.*, "Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection," *Mol. Syst. Biol.*, vol. 2, p. 2006.0008, Feb. 2006.

[75] K. S. Murakami, S. Masuda, E. A. Campbell, *et al.*, "Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex," *Science (80-.).*, vol. 296, no. 5571, pp. 1285–1290, May 2002.

[76] V. Mekler, E. Kortkhonjia, J. Mukhopadhyay, *et al.*, "Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex.," *Cell*, vol. 108, no. 5, pp. 599–614, Mar. 2002.

[77] M. L. Craig, W. C. Suh, and M. T. Record, "HO. and DNase I probing of E sigma 70 RNA polymerase--lambda PR promoter open complexes: Mg2+ binding and its structural consequences at the transcription start site.," *Biochemistry*, vol. 34, no. 48, pp. 15624–32, Dec. 1995.

[78] S. O. Skinner, L. A. Sepúlveda, H. Xu, *et al.*, "Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization," *Nat. Protoc.*, vol. 8, no. 6, pp. 1100–1113, May 2013.

[79] X. S. Xie, P. J. Choi, G.-W. Li, *et al.*, "Single-Molecule Approach to Molecular Biology in Living Bacterial Cells," *Annu. Rev. Biophys.*, vol. 37, no. 1, pp. 417–444, Jun. 2008.

[80] O. Shimomura, "The discovery of aequorin and green fluorescent protein," *J. Microsc.*, vol. 217, no. 1, pp. 3–15, Jan. 2005.

[81] O. Shimomura, "Structure of the chromophore of *Aequorea* green fluorescent protein," *FEBS Lett.*, vol. 104, no. 2, pp. 220–222, Aug. 1979.

[82] M. W. Davidson and R. E. Campbell, "Engineered fluorescent proteins: innovations and applications," *Nat. Methods*, vol. 6, no. 10, pp. 713–717, Oct. 2009.

[83] E. Bertrand, P. Chartrand, M. Schaefer, *et al.*, "Localization of ASH1 mRNA Particles in Living Yeast," *Mol. Cell*, vol. 2, no. 4, pp. 437–445, 1998.

[84] I. Golding and E. C. Cox, "RNA dynamics in live Escherichia coli cells.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 31, pp. 11310–5, Aug. 2004.

[85] J. B. Pawley, *Handbook of biological confocal microscopy*. Springer, 2006.

[86] D. Axelrod, "Cell-substrate contacts illuminated by total internal reflection fluorescence.," *J. Cell Biol.*, vol. 89, no. 1, pp. 141–5, Apr. 1981.

[87] C. A. Konopka and S. Y. Bednarek, "Variable-angle epifluorescence microscopy: a new way to look at protein dynamics in the plant cell cortex," *Plant J.*, vol. 53, no. 1, pp. 186–196, Jan. 2008.

[88] A. Gupta, J. Lloyd-Price, R. Neeli-Venkata, *et al.*, "In vivo kinetics of segregation and polar retention of MS2-GFP-RNA complexes in Escherichia coli.," *Biophys. J.*, vol. 106, no. 9, pp. 1928–37, May 2014.

[89] F. Zernike, "How I discovered phase contrast.," *Science*, vol. 121, no. 3141, pp. 345–9, Mar. 1955.

[90] A. Hakkinen, A.-B. Muthukrishnan, A. Mora, *et al.*, "CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli," *Bioinformatics*, vol. 29, no. 13, pp. 1708–1709, Jul. 2013.

[91] P. Ruusuvuori, T. Aijo, S. Chowdhury, *et al.*, "Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images," *BMC Bioinformatics*, vol. 11, no. 1, p. 248, May 2010.

[92] T.-B. Chen, H. H.-S. Lu, Y.-S. Lee, *et al.*, "Segmentation of cDNA microarray images by kernel density estimation," *J. Biomed. Inform.*, vol. 41, no. 6, pp. 1021–1027, Dec. 2008.

[93] M. Kandhavelu, A. Häkkinen, O. Yli-Harja, *et al.*, "Single-molecule dynamics of transcription of the lar promoter," *Phys. Biol.*, vol. 9, no. 2, p. 26004, Apr. 2012.

[94] W. Q. Meeker and L. A. Escobar, *Statistical methods for reliability data*. Wiley, 1998.

[95] G. Casella and R. L. Berger, *Statistical inference*. Thomson Learning, 2002.

[96] R. Lutz, T. Lozinski, T. Ellinger, *et al.*, "Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator.," *Nucleic Acids Res.*, vol. 29, no. 18, pp. 3873–81, Sep. 2001.

[97] P. L. deHaseth, M. L. Zupancic, and M. T. Record, "RNA polymerase-promoter interactions: the comings and goings of RNA polymerase.," *J. Bacteriol.*, vol. 180, no. 12, pp. 3019–25, Jun. 1998.

[98] S. Strickland, G. Palmer, and V. Massey, "Determination of dissociation constants and specific rate constants of enzyme-substrate (or protein-ligand) interactions from rapid reaction kinetic data.," *J. Biol. Chem.*, vol. 250, no. 11, pp. 4048–52, Jun. 1975.

[99] W. H. Press, *Numerical recipes : the art of scientific computing*. Cambridge University Press, 2007.

[100] H. Tran, S. M. D. Oliveira, N. Goncalves, *et al.*, "Kinetics of the cellular intake of a gene expression inducer at high concentrations.," *Mol. Biosyst.*, vol. 11, no. 9, pp. 2579–87, Sep. 2015.

[101] F. Widdel, "Theory and measurement of bacterial growth," *Di dalam Grundpraktikum Mikrobiol.*, pp. 1–11, 2007.