Non equilibrium dynamics in *Escherichia coli*'s gene regulatory network

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie

vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Luca Galbusera

aus Italien

Basel, 2022

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Erik van Nimwegen (Fakultätsverantwortlicher)

Prof. Dr. Richard Neher (Korreferent)

Basel, den 13.10.2020

Prof. Dr. Martin Spiess
Dekan

Abstract

Gene regulation is a key process in living organisms. It defines cells identity and behavior, and allows the cells to adapt to the external environment. From a theoretical point of view, a central question is how to mathematically characterize the many players and their complex interactions to understand the gene expression output as function of the regulatory inputs. A common approach is *thermodynamic modelling*, where the transcription is assumed to be in equilibrium with the concentration of transcription factors, and any fluctuation is averaged away. However, the advent of new experimental techniques providing precise measurements of gene expression at the single-cell level is challenging the general validity of the equilibrium assumption.

In this thesis, we focus on the induction of the LexA regulon in the model organism *Escherichia coli*, which is involved in the repair of DNA damages. Tracking the single-cell expression dynamics of different genes under the exclusive control of the inhibitor LexA, we show that the induction is characterized by short bursts of production, which are incompatible with a thermodynamic model where gene transcription is in equilibrium with the concentration of LexA. On the other hand, we show that the network responds to transient fluctuations in the concentration of the regulator.

Finally, we deal with the question of how to properly analyze flow cytometry data for bacterial populations. Flow cytometry is an attractive technology to quantify single-cell gene distribution in high-throughput. However, so far no systematic investigation has been carried out to estimate the accuracy of these measurements for small bacterial cells. Here, by comparing the fluorescence distribution of the same *E. coli* strain both in flow cytometry and in a microscopy setup, we show that the fluorescent signal contains a significant amount of electronic noise and background fluorescence. We then propose a robust method to correct for these spurious components, and we show that only after correcting for electronic noise and autofluorescence, measurements from the flow cytometry agree with the ones from the microscopy setup.

Contents

1. Motivation

1.1. Why biology?

One of the most frequent questions people like to ask me when they know that I left physics for biology is *why*?

When I first embarked on the study of physics, I was fascinated by the idea of being able to understand the complexity of reality, from the tiny world of subatomic particles, up to the vastness of the Universe. I can fairly say that my journey started when as a high school student I first read an article about the standard model in a science magazine. I was stunned by the fact that the complexity of the Universe can be explained by a bunch of elementary particles and a single equation describing how they interact (gravity apart). And the very question of how complexity can emerge from simplicity is what has guided my journey until the present days. However, as an undergraduate student, I soon realized that I was not fully satisfied by the current state of affair of theoretical physics, where new theoretical ideas are exploring realms well beyond the limits of our experience, and where new advancement is justified more by abstract mathematics, than by empirical evidence. And, although I appreciate the large effort to trying to reduce the complexity of the whole Universe down to the behavior of simple entities, I realized that I personally find more intellectually satisfying the investigation of complex phenomena that can be accessed with the current technological and experimental tools. One being life on Earth. We know life is the result of the arrangement of four different bricks, the nucleic acids, that form the DNA. But, how can the variety and complexity of life emerge from the simple arrangement of these four nucleic acids? It was the desire to find an answer to this question and the need for a research field grounded on empirical evidence, that led me to study biology.

The next question people like to ask is *What can a physicist do for biology?* The answer, in my opinion, is that he or she can bring their expertise and new methodological approaches to solve a specific problem. In modern English, the expression *Renaissance man* refers to a man with several interests and broad knowledge in a variety of fields [1]. That was certainly true if we think of great Renaissance thinkers like da Vinci, Pascal, or Leibniz, to name a few. On the other side, nowadays we have accumulated such a large amount of knowledge, that a single mind cannot possibly fully master more than a specialized portion of it. However, this does not mean that science must become a compartmentalized discipline, but it is pivotal for the advancement of knowledge that scientists of different backgrounds, each with their own expertise and problem-solving approach, work side by side on a common question (as exemplified in a nice article by Lazebnik [2]. The preface Schrödinger wrote for his *What is life* [3]

is also inspiring).

The question is then, what expertise can a physicist bring to complement the expertise of a biologist? I think the answer is that physicists have a natural tendency to reductionism, that is to find patterns in apparently different complex phenomena, and to identify a set of elementary rules that can explain them all, in a unified framework. That was the case, for example, when Maxwell showed that electricity, magnetism, and light are all sides of the same force. Indeed, we now know that all the many facets of life are nothing else but different arrangements of four nucleic acids, in the same way as a feather and a stone are different arrangements of subatomic particles. And understanding the many facets of life is not so different than understanding how a complex physical phenomenon emerges from elementary rules.

1.2. Why gene regulation?

In his Confessions, Augustinus wrote [4], answering the question *what is time?*: "If no one asks me, I know; if I wish to explain to him who asks, I know not". With life, we are in a very similar situation, although everybody knows intuitively what life is, and most of the time it seems straightforward to distinguish between living and non-living matter, finding a formal definition of what life is, is a harder task [5]. Of the several characteristics that over time have been considered as essential features of life, I find adaptation one of the most fascinating, since it underlies most of the complex behavior of living organisms. In particular, I am interested in gene regulation, that is, the ability of the cells to sense the external environment and react to external changes by turning on and off specific genes.

I find gene regulation fascinating for several reasons. The rules are very simple and relatively well understood: roughly speaking, there is only one basic rule: genes have specific regions that a regulator protein can bind to inhibit or enhance its transcription. Despite the simplicity of this rule, gene regulation leads to a large variety of complex behaviors and phenotypes, given the same underlying genotype. This allows living cells to optimize their gene expression based on the external environment and therefore it has a pivotal role in adaptation. It guides the precise and complex body segmentation during the development of embryos. It is responsible for the plethora of cell types that constitute a complex multi-cellular organism. And, interestingly, the difference among organisms resides in different ways their genes are regulated, more than in having different genes [6]. Also, we are living in exciting times for research in gene regulation, with new high-throughput technologies offering us a full array of experimental techniques, unthinkable until a couple of decades ago, which are providing fundamental insights about the underlying mechanisms of gene regulation [7]. The advent of the DNA-arrays technology in the late 90's and 2000's [8] provided us with efficient ways to probe the transcriptomes of different tissues or organisms, and with the rapid advancement in sequencing, we could reach single-cell resolution. Microscopy image analysis and microfluidic devices allow for time tracking of gene expression in single cells [9–16], bacterial flow cytometry allows to investigate snapshots of protein distribution in large populations [17–27], and new

versions of ChIP allow us to explore the interactions between transcription factor and DNA [28]. And this is just to name a few. At the same time, synthetic biology allows us to not only probe existing regulatory networks but also to synthesize novel architectures de novo, therefore allowing for testing new predictions and theories.

1.3. My contribution

I focused on understanding how gene expression is regulated in *Escherichia coli* single-cells, and how the promoter architecture dictates the response of a gene to the external environment. My work combined time-lapse microscopy, high-throughput estimation of fluorescent reporters, quantitative image analysis, and mathematical modeling to foster our understanding of the mechanisms underlying gene regulation.

Although from the seminal work of Jacob and Monod [29] our understanding of gene regulation has increased enormously, there are still many open questions. Even for the simplest bacterial genomes, we are still far away from having a complete picture of the full regulatory network, with 70%-90% of the regulatory interactions in the model organism *E. coli* still missing [30]. But even if we knew all the regulatory inputs, the many molecular players and their complex web of interaction make it hard to predict the phenotype of a cell from its genotype, or how changes in regulatory sequences alter the behavior of the cells [31].

I believe that we are in a similar situation as 19^{th} century physicists dealing with the behavior of gases, and it is no wonder that many mathematical concepts are shared between the field of gene regulation and thermodynamics or statistical physics. In thermodynamics, the aim is to describe thermal processes in systems with a large number of degrees of freedom. Of course, there is no new physics in the behavior of gases: if we solved Newton equations for each particle, we would be able to explain all the observations (quantum mechanical corrections aside). However, such a model will not only be mathematically intractable, but it will also be so complicated that will give us no insight into the functioning of the system. Thermodynamics is, in this sense, a more powerful description of gases, although less precise. Instead of relying on a large number of degrees of freedom, we describe a gas with just three parameters: temperature, volume, and pressure. With these three parameters, which are readily interpretable in terms of microscopic properties of the gas, we can make predictions and understand how a gas behaves. This simplifying approach is at the base of the humorous metaphor that theoretical physicists only consider spherical cows in a vacuum. But I think this coarse-graining away of microscopic details and the simplified modelization of complex systems constitutes the real power of physical theories.

Similarly to gases, gene regulation is a highly complex process, involving many components and reactions, and if we want to understand it, we need to reduce it to a set of simple players and rules. But there are still many open questions, some of which have been tackled in this thesis. To find coarse-grain models of gene expression, it is not yet completely clear what approximations we are allowed to make. For example, in describing the Lac operon it is common practice to assume that the

system is in equilibrium with the main players involved in mRNA transcription [32]. This approach works well with the Lac operon, but there might be networks where the equilibrium hypothesis is not met [33]. Moreover, it focuses only on the average gene expression, and it is, therefore, unable to discriminate among a variety of different models giving the same average number of transcripts [34]. At the same time, it became clear that also expression noise, coarse-grained away by equilibrium models, is an important feature of regulatory networks, allowing isogenic cells to have a phenotypic variability that proves to be advantageous in adapting to varying environments [35–37]. Recently, it has been shown that a sensible part of this noise comes from the regulatory network itself [18] and this might shed light on the origin and evolution of de novo regulatory networks [27].

Specifically, I was interested in the question of how transcription factors (molecules involved in the activation or repression of genes) regulate gene expression at the single-cell level, and how fluctuations in the concentration of transcription factors propagate to the regulated genes. I examined this question not only from a theoretical point of view but also based on experimental data of the SOS response in *E. coli* [38]. The SOS response has been chosen, together with the experimental team, because one regulator, LexA, inhibits different genes, some of which are under exclusive control of LexA. The concentration of LexA is easily controllable using a common antibiotic and its binding sites are well known. I consider this network (1 regulator – many targets) as the "helium atom" of gene regulation, the Lac operon (1 regulator – 1 target) being the "hydrogen atom". Despite the simplicity of the network, the presence of feedback loops and fluctuations in LexA concentration cause a complex dynamics, with a temporal pattern characterized by waves of gene activity [33]. Different kinetic models have been proposed to explain such idiosyncratic induction pattern, and they showed the importance of single-cell studies to unmask patterns that are otherwise hidden at the population level [39–42].

In this thesis, I will discuss three projects related to the mathematical modeling of gene regulation. First, I show that the equilibrium assumption, valid for the Lac operon and the regulation of the λ phage, is not appropriate for the SOS response. I then investigate which consequences a non-equilibrium dynamics has on the response of the cells to the external environment. Second, any mathematical model is useful only when it can be applied to experimental data, and for this reason, it is important to have accurate high-throughput measurements. During my research, I carried out a rigorous comparison of flow cytometry with microscopy setups and I showed how to correctly extract quantitative information from high-throughput measurements of bacterial gene expression. Third, equilibrium models can become quite cumbersome when promoters allow the regulators to bind on different sites. For this reason, to analyze data of the SOS response previously collected in the lab, I have developed an efficient algorithm to computationally solve complex equilibrium models.

2. Introduction

2.1. Transcription initiation and regulation

2.1.1. Transcription initiation

One of the main players in gene expression is the DNA-directed RNA polymerase (RNAP), an enzyme able to synthesize RNA strains from a DNA template in a process termed transcription. In most bacteria, the RNAP is a 400 kDa enzyme with a characteristic shape of a crab claw and composed of five domains, $\alpha_2\beta\beta'\omega$ [43]. The C-terminal of the two alpha domains (αCTDs) are separated by the main body by an unstructured link which confers a degree of flexibility, in such a way that the two αCTDs can be thought of as antennae that the polymerase can use to identify specific regions of the genome [44].

The RNAP alone is not able to find a gene and to initiate the transcription process. As a result, a group of molecules, called *sigma factors*, are required to bind to the polymerase and to direct it to specific genetic loci, called *promoters*, situated upstream the transcription start site (TSS) of a specific gene. The very stable complex of the RNAP with the bound sigma factor is called *holoenzyme* and a variety of sigma factors are present inside the bacterial cell, each with different specificities for different promoters. However, all bacteria have a housekeeping sigma factor responsible for the transcription of essential genes in the growing cell. In *E. coli* the housekeeping sigma factor is the σ^{70}, so called because its molecular weight is 70 kDa. The promoter sequence of a gene regulated by σ^{70} contains 4 elements that the sigma factor recognizes and binds to: the -35 region, centered around 35 base pairs upstream the TSS, the -10 region, situated roughly 10 base pairs upstream the TSS, the extended -10 region and the discriminator. The -35 and -10 regions have the consensus sequences TTGACA and TATAAT respectively, and the optimal space in between is 17 base pairs and is called *spacer*. A further element is the UP element, found upstream the -35 region and recognized by the two αCTDs of the RNA polymerase [45].

Once the holoenzyme has recognized a promoter, a series of reactions, driven by thermal fluctuations and free energy accumulated in the holoenzyme, lead to the unwinding of the DNA and the correct positioning of the template strand in the catalytic center of RNAP, so that the transcription process can begin. As the transcription proceeds, the DNA is pulled inside the holoenzyme, which on the other hand stays firmly bound to the promoter, creating a "scrunching" of the DNA. Furthermore, σ^{70} blocks the exit channel of the newly synthesized mRNA. The scrunching of the DNA and the collision of the mRNA with the sigma factor increase the free energy that can be released in two ways. In many cases, the mRNA dissociates

from the complex, is released, probably through a secondary channel [43], and the transcription must be restarted again (abortive initiation cycles). Alternatively, the mRNA manages to displace the σ^{70}, freeing the exit channel and destabilizing the bond between the holoenzyme and the -35 region of the promoter sequence. In this way, the holoenzyme can leave the promoter (promoter escape) and can continue the transcription of the gene [46].

2.1.2. Transcription regulation

Due to the different environmental conditions that the cells may experience, the genes to be expressed and their amount of transcription must be regulated. For this reason, the activity of the holoenzyme is affected by other molecules able to sense environmental cues and orchestrate the transcriptional program. The first regulatory element ever identified was the LacI repressor described in the seminal work of Jacob and Monod in 1961 [29]. For some time the mainstream idea was that the regulatory program was carried out entirely by repressors, giving little credence to the existence of activators [44]. Today we know that the regulation of the transcription activity can be achieved both by repressors and activators, which act by interfering with the promoter accessibility or with the polymerase affinity for specific promoters.

We present here only a selection of the most common mechanisms, referring to [44] and [45] for a more comprehensive discussion.

Promoter-centred regulation

Molecules that control the expression of a gene by binding to specific sites of the promoter are called *transcription factors* (TFs) and the sites they bind to are called *operator* A repression action can be carried out simply by steric hindrance, preventing the holoenzyme to bind to the promoter. Alternatively, TFs can introduce looping at the promoter region, modifying the local DNA structure and decreasing the affinity of the promoter to the sigma factor. The TF can also interact directly with activators bound to the promoters (see below) and prevent them to recruit the holoenzyme. On the other hand, most native promoters have elements whose sequences are far away from the consensus. For these promoters, the affinity to the σ^{70} is low or, in some cases, even null. As a result, TFs acting as activators are required. One activating mechanism is to bind to elements of the promoters that are far away from the consensus sequence, acting as molecular "velcro" to increase the affinity to the RNAP [45]. Other TFs can bind between the -35 and -10 region causing a conformation change of the promoter to set an optimal spacer length. Some sigma factors, like the σ^{54} in E. Coli, require energy to unwind the DNA. Therefore, TFs called Enhancer Binding Proteins, are required to provide the necessary energy requirements from ATP hydrolysis. To notice that TFs may also be regulated, for example, they can be actively degraded or sequestrated by specific anti-repressors. This is, for example, the case of recA and lexA involved in the SOS response of *E. coli* and described in more detail in the following sections.

Finally, some promoters, like the one controlling the fim operon in *E. coli*, are in the wrong direction and must be flipped by the FimB and FimE recombinases to correctly direct the holoenzyme to the gene.

Polymerase-centred regulation

The presence of different sigma factors is an example of regulation at the level of the RNAP. While in normal conditions the most abundant factor is the housekeeping σ^{70}, in stress conditions other sigma factors can become predominant and compete with σ^{70} in the binding of RNAP and guiding it to a specific subset of promoters. The availability of alternative sigma factors is regulated by covalent modifications, subcellular localization, and different rates of synthesis and degradation. Moreover, they can be sequestrated by anti-sigma factors, which, in turn, can be sequestrated by anti-anti-sigma factors. An example is the induction of the σ^{38} factor (RpoS) in response to stress conditions or at entry into the stationary phase. Here an increase in the anti-σ^{70} Rsd causes σ^{38} to predominate in the binding of the RNAP [46]. Sigma factors can also be sequestrated, for example by up-regulation of 6S RNA in response to slow growth, which binds and sequestrates the sigma factors, allowing for coupling the growth and gene expression. Alternatively, some phages produce molecules which inhibit the bacterial RNAP in favor of the phage polymerase, or that redirect it to the viral promoters. Some of these appropriators are also auto-produced by the bacterium, e.g. E. Coli SoxS binds to the αCTDs to redirect them to promoters containing a SoxS box, which are associated to genes involved in coping with the oxidative stress.

2.2. Mathematical modeling of gene expression

Let's suppose we want to model the gene expression of a population of cells. The most precise way to proceed would be to consider all the processes involved: RNAP and sigma factor binding, open complex formation, promoter escape, ribosome binding... This way is, however, doomed to fail, because of the high number of degrees of freedom, some of the mechanisms are still not well understood, and experimental methods are not enough precise to have a complete quantitative description of gene expression[47].

To gain insights on the functioning of complex genetic networks, researchers have extensively looked at simple systems, where most of the components and interactions are known. The two archetypal examples are the *lac* operon [48] and the regulation of the lytic/lysogenic cycle of the λ phage [49] in *E. coli*. Despite the simplicity of these networks, they already entail many subtleties inherent to gene regulation and show the challenges one has to face to model more complex networks. An extensive toolbox of quantitative models, resting on different assumptions of the regulatory process, has been proposed based on the studies of these two regulatory networks.

2.2.1. Thermodynamic models

We have already noticed that the modeling of the complex and many processes that lead to gene expression is similar to when physicists describe a gas, whose behavior is determined by the complex motion of a large number of individual molecules. Since we cannot solve the equations of motions for the large number of molecules in a gas, we consider the single molecules subject to random motion, and we compute average values of the system as a whole, like pressure, temperature and volume. The framework that allows us to connect the complex microscopic dynamics to the simpler macroscopic one is *statistical physics*.

From thermodynamics, we know that at a macroscopic level, the system is completely determined by a reduced set of macroscopic quantities, for example pressure P, volume V, number of particles N, and energy E (or equivalently temperature T). Since the microscopic system is dynamic, we can expect fluctuations in the *instantaneous* values of these macroscopic quanities [50]. For example, the pressure is given by the force per unit area exerted by the molecules of gas on a wall, and since the molecules move over time, the pressure itself is a function of time. Nonetheless, because of the rapid motion of the molecules, we should also expect that changes in pressure happen at very fast time scales and therefore are somehow averaged out, so that the *measured* macroscopic pressure is an effective time-average of the instantaneous pressure [50]. The question is how to derive this time average.

The answer was first proposed by Gibbs in its 1902 seminal work "Elementary Principles in Statistical Mechanics" [51], which is considered to be the foundation of statistical mechanics. His idea is to use a mental construct called *ensemble*. An ensemble is a collection of a large number $\mathcal{N}$ of mental copies of the system, all having the same macroscopic state, but differing in the microscopic configuration. Indeed, the reduced set of macroscopic parameters is not enough to fully determine the state of the 10^{20} molecules typically contained in a gas, and we must expect that a large number of different microscopic configurations give rise to the same macroscopic state. The assumption Gibbs makes is that we can exchange the time average of the macroscopic quantities with their *ensemble* average. If we assume that over time the system under consideration explores all the microstates in the ensemble, then, in the limit $\mathcal{N} \to \infty$, the ensemble average is equal to the time average (*ergodic hypothesis*) [50]. Notice that the ensemble average must be independent on time for this assumption to hold, that is: *the system must be in equilibrium*. However, let's suppose we change some characteristics of the system, e.g. its internal energy. The system will undergo a phase of "adaptation" where the macroscopic quantities will change to reach their new values determined by the new energy. In this phase, the system is out of equilibrium and cannot be described by thermodynamics (although recently, researches have made significant progress in the field of non-equilibrium thermodynamics [52]). When this adaptation phase is over, the system has reached a new equilibrium state and can be again described by thermodynamics.

It remains to specify how the ensemble average has to be taken, that is, what is the probability distribution of the different microstates. The postulate of *equal a priori probabilities* states that the copies of the ensemble are distributed uniformly,

that is, with equal probabilities, over the possible microscopic states compatible with a given macroscopic state [50]. The consequence of the postulate of equal a priori probabilities and the ergodic hypothesis is that the real system we observe spends an equal amount of time in each of the possible microscopic states. The ergodic hypothesis and the postulate of equal a priori probabilities cannot be deduced from the laws of mechanics and are based on the agreement between the experiments and the predictions. However, using an information theoretical approach to statistical mechanics, we can derive them by requiring that the probability distribution of the microstates contains all and only the information available from the macroscopic system (maximum entropy principle) [53].

In biology, we are interested in systems characterized by a definite volume V, number of particles N and temperature T (*canonical ensemble*). Since the energy is not fixed, the copies of the ensemble can have different energetic levels E_1, E_2,... It can be shown that for the canonical ensemble the probability of having a microscopic state with energy E_j is given by the *Boltzmann distribution* [50]

$$P(E_j) = \frac{e^{-\beta E_j}}{\sum_i e^{-\beta E_i}} \tag{2.1}$$

where β is a positive constant and the index i runs over all possible energies. The denominator, called *partition function*, ensures that the probability of having any of the possible energies sums up to 1.

Let's see how the thermodynamic approach helps to describe the gene expression in cells. A cell can be seen as a system with a definite number of particles, volume, and temperature, therefore we can use the canonical ensemble approach to model the concentration of protein. We use the concentration in order to have a fixed volume despite the growth of the cell. Two assumptions are made. First, the fluctuations caused by the inherent randomness of the biochemical reactions happen at very fast time scales, such that they are averaged out and can be ignored (similar to the thermodynamic description of constant temperature, pressure, and volume). Second, although gene expression is the result of many and complex processes (loading of the polymerase, transcription initiation, translation...), we assume that the loading of the polymerase and TFs on the promoter happens at much slower scales. That is, we assume that the binding and unbinding of the regulators is the rate-limiting step and all other processes can be ignored. Therefore, we assume that each promoter state produces a definite amount of protein, and the observed gene expression is the time average of the promoter states, where each state is weighted by the amount of protein it produces [54]. Notice the similarity with the thermodynamics of gases: the molecular microstates of the gas correspond to the state of the promoter and the measured pressure, volume or temperature corresponds to the measured gene expression.

As an example, we consider the scenario where there is only one repressor, as this is relevant for the SOS network studied in the following sections. We refer to [55] for a more general description. First of all, we need to enumerate all possible states, that is, all possible positions of the repressor in the cell. Let's divide the volume V of the

cell in $N \to \infty$ small boxes such that the operator is contained in exactly one box. In order for the repressor to be unbound, it must be in one of the $N-1$ boxes away from the operator. If we call R the total number of repressors, the number of unbound states corresponds to the number of possible arrangements of the repressors in the $N-1$ boxes, that is

$$\mathcal{N}_u = \frac{(N-1)!}{R!(N-1-R)!} \tag{2.2}$$

The non-normalized probability for all the R repressors to be unbound is given by the number of states multiplied by the Boltzmann factor of Eq (2.1)

$$Z_u \propto \frac{(N-1)!}{R!(N-1-R)!} \times e^{-R\,E_u} \tag{2.3}$$

where E_u is the energy of the unbound state in units of β. On the other hand, the non-normalized probability of being bound is the probability of having $R-1$ repressors unbound and one bound

$$Z_b \propto \frac{(N-1)!}{(R-1)!(N-R)!} \times e^{-(R-1)E_u} \times e^{-E_b} \tag{2.4}$$

Putting Eq (2.3) and Eq (2.4) together we have that the probability to be unbound is given by

$$P_u = \frac{Z_u}{Z_u + Z_b} = \frac{1}{1 + [R]e^{-\Delta G}} \tag{2.5}$$

where we defined the free energy $\Delta G = E_b - E_u$ and $[R] \simeq R/N$ is the concentration of repressor [54]. Notice that the probability of being unbound decreases as the concentration of repressor increases or if the free energy decreases (that is, if the bound state has less energy than the unbound state). If we call r^H and r^L the transcription rates in the repressed (low) and unrepressed (high) states, we predict the expression to be

$$\mathcal{E} = r^H \frac{1}{1 + [R]e^{-\Delta G}} + r^L \frac{[R]e^{-\Delta G}}{1 + [R]e^{-\Delta G}} \tag{2.6}$$

Thermodynamic models are appealing because they are able to describe gene expression within a very simple framework:

1. Enumerate all the possible states of the system and their corresponding Boltzmann factor.
2. Average the gene expression output of each state, weighted by the probabilities of that state.

Nonetheless, they make two strong assumptions: gene expression is in equilibrium and gene expression output depends only on the fraction of time the regulators are bound or unbound to the promoter. This last assumption is complemented with the *occupancy hypothesis*: the binding affinities determine the number of transcription factors bound at a given concentration (occupancy); the occupancy is the only

quantity that determines gene expression and other effects, like operator positioning and specific interactions between the regulators and the polymerase, are ignored. Depending on the experiment, these hypotheses may be too restrictive. For example, if we want to predict the gene expression as a function of the position of the operator with respect to the promoter, the interaction between the regulator and the polymerase becomes important[56] and the occupancy hypothesis cannot hold. On the other hand, in some scenarios, the occupancy hypothesis is reasonable and the thermodynamic models can explain the observed behavior. Indeed, these models have been used to successfully describe gene expression for important systems like the Lac operon [57–65] or the regulation of the λ phage cycle [54, 66].

As discussed at the beginning of this section, if some properties of the system change (e.g. the concentration of repressor), there will be a phase of adaptation, where the system adapts to reach a new equilibrium state. During this phase, the expression is not in equilibrium and cannot be described by a thermodynamic model. However, if the biochemical reactions take place rapidly compared to the change in the system state, the system reaches very fast the new equilibrium, and the non-equilibrium phase can be ignored. That is, at any time point we can consider the system to be in equilibrium and it can be described by a thermodynamics framework. For example, if the concentration of repressor $[R]$ changes slowly over time and the system equilibrates very fast, at each time point we can assume that the probabilities of each promoter state are those that would be attained by the system if it had time enough to reach the equilibrium with the given concentration $[R]$. In this case, the gene expression is still described by Equation (2.6), but now with $[R]$ function of time. Therefore, thermodynamic equilibrium does not strictly mean that the expression is constant over time, and indeed it has been applied to describe time-depending gene expressions [59, 67].

2.2.2. Differential Equations

Although the success of thermodynamic models, they are only able to describe the steady-state of the system. If we want to describe the time dynamics of gene expression, or if we want to add more detailed information about the processes, we need a more sophisticated approach. In the differential equations approach, we boil down the gene expression process to a bunch of relevant processes by collecting different reactions into effective ones with specific rates. For example, for a system where a promoter can switch between a bound and an unbound state, we can consider two processes

1. The repressor binds with a rate per repressor r_b and unbinds with a rate r_u.
2. The mRNA is transcribed with a rate r^H if the repressor is unbound or r^L if the promoter is bound and decays at a rate δ per unit mRNA.

In this case, we are collecting all the transcription reactions (polymerase and sigma factor binding, switching from the open to the close complex, promoter escape...) into the effective rates r^H and r^L. Moreover, we assume that the translation is much faster

than the translation, so that we can ignore it. The equations governing the system are

$$\frac{dG_u}{dt} = r_u G_b - r_b[R]G_u \tag{2.7a}$$
$$\frac{dG_b}{dt} = -\frac{dG_u}{dt} \tag{2.7b}$$
$$\frac{dm}{dt} = r^H G_u + r^L G_b - \delta m \tag{2.7c}$$

where G_u and G_b are the fraction of promoters unbound or bound in the population and m is the number of mRNAs. The first two equations can be solved to give

$$G_u(t) = \frac{r_u}{r_u + r_b[R]}\left(1 - e^{-(r_u + r_b[R])t}\right) \tag{2.8a}$$
$$G_b(t) = 1 - G_u(t) \tag{2.8b}$$

Notice that in the limit of large times we reach the steady state

$$G_u(t) = \frac{1}{1 + [R]\frac{r_b}{r_u}} \tag{2.9}$$

At steady state the third equation for the mRNA gives

$$m = r^H \frac{1}{1 + [R]\frac{r_b}{r_u}} + r^L \frac{[R]\frac{r_b}{r_u}}{1 + [R]\frac{r_b}{r_u}} \tag{2.10}$$

If we identify $r_u = e^{-E_u}$ and $r_b = e^{-E_b}$ we see that at steady state G_u approaches the thermodynamic probability of being unbound, while m approaches the average gene expression. But, in addition to the thermodynamic model, we have here the possibility to solve for the time dynamics, as shown by Eq (2.8). On the other hand, a differential equation approach can quickly become cumbersome as more reaction steps are added.

2.2.3. Stochastic models

The differential equation approach assumes that each reaction happens at a constant rate and the amount of molecules changes continuously. However, the reactions are stochastic and the number of molecules can change only by discrete integer amounts. It is therefore more precise to look at the reaction rates as probabilities per unit time and to model the gene expression using stochastic equations. One way is to write down the master equation [68] for the probability of the gene to be bound (P_b) or unbound (P_u)

$$\frac{dP_b}{dt} = r_b[R]P_u - r_u P_b \tag{2.11a}$$
$$1 = P_b + P_u \tag{2.11b}$$

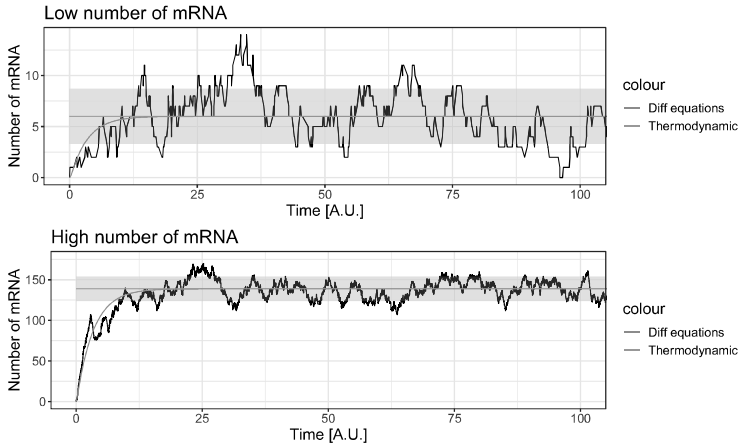

Figure 2.1.: Simulated traces of gene expression and comparison with the predictions from different models. The gray region is the steady state standard deviation predicted with the stochastic equations. As the number of mRNA decreases, the stochastic fluctuations become more important.

This gives $P_u(t) = \frac{r_u}{r_u + r_b[R]} \left(1 - e^{-(r_u + r_b[R])t}\right)$, which agrees with the differential equation model.

For the mRNA the master equation can be solved exactly if we suppose $r^L = 0$, that is, the gene is completely turned off by the repressor. In this case we have [69, 70]

$$\langle m \rangle = \tilde{P}_u \langle \tilde{m} \rangle_u \left[1 + \frac{\tau_g}{\tau_m - \tau_g} e^{-t/\tau_g} - \frac{\tau_m}{\tau_m - \tau_g} e^{-t/\tau_m}\right] \tag{2.12a}$$

$$\lim_{t \to \infty} CV^2 = \frac{1}{\tilde{m}} + \frac{1 - \tilde{P}_u}{\tilde{P}_u} \frac{\tau_g}{\tau_m + \tau_g} \tag{2.12b}$$

where $\tau_g = (r_u + r_b[R])^{-1}$ and $\tau_2 = \delta^{-1}$ are the characteristic times of the gene and mRNA dynamics. A tilde means that the value is computed at steady state, that is, $\tilde{P}_u = r_u \tau_g$ and $\langle \tilde{m} \rangle_u = r^H \tau_m$.

We can see that the fluctuations become more important, the lower the number of mRNA molecules. Since in bacterial cells the mRNA is present usually in less than 10 copy numbers [13], we can expect that the stochastic fluctuations might not be ignored.

Stochastic Simulations

Stochastic modeling is the most accurate method to infer the distribution of gene expression in the cells. Still, the method is not amenable for analytic treatments, since the stochastic equations can quickly become impossible to solve as the number of

reactions increases. For example, in the previous sections, we described the scenario where a gene can switch between a bound and an unbound state. Let's suppose we want to add more details to the model. We can expect that some processes, like fluctuations in cell growth and binomial sampling of the species at the division, can increase the CV^2, while other processes like protein translation and folding can decrease the CV^2 if they happen at slower time scales (thus averaging out the more rapid dynamic considered so far). A system of this kind cannot be solved analytically and it becomes important to have efficient ways to simulate them. The most common algorithm used for this task has been introduced by Gillespie, whose description can be found in the original paper [71].

Figure 2.1 shows a comparison of the different models of gene expression. The black trace is a realization of the stochastic simulation of a gene that can switch between an inactive and an active state (the inactive state is not expressing). The thermodynamic model (blue line) is able to describe only the mean behavior at steady-state, while the differential equations introduce more precision in describing also the initial transient dynamics. The expected standard deviation predicted by the stochastic equations is shown as a grey region and is valid only in the steady-state. When the number of mRNAs is high (bottom row), the CV^2 is low enough and a deterministic treatment can be enough; but when the number of mRNAs is low (upper row), as it is usual for bacterial genes, the stochastic fluctuations become important.

2.2.4. Hybrid models

Sometimes it is beneficial to combine different features of the previous models. For example, as explained in the section of thermodynamic models, we can add some time dynamics to the system if we suppose that the concentration of regulators changes over time according to some differential equations. If the change is slow compared to the time scale of the biochemical reactions, the system is at any given time point in equilibrium, but the equilibrium state changes over time, depending on the concentration of regulators. Using this approach, Shea and Ackers managed to explain key properties of the λ phage network, like the stability of the lysogenic state and the induction of lysis [66]. A similar approach to explain the stability of the lysogenic state was used also by Santillan and Mackey [67]. More recently, Aurell et al. [72] introduced stochasticity in thermodynamic models. At each time point, the given concentration of the two main autoregulators cro and cI is used to compute their mean expression based on a thermodynamic model. These are then varied by adding Poisson noise and new concentrations are used to find the next equilibrium state. Moreover, the cell grows and divides. Hasty and coauthors [73] included stochastic fluctuations in the λ phage system to explain the steep all-or-none response. They first used the differential equation approach to find all the possible steady states of the system. Then they added to the differential equation a random Gaussian term to explain transitions among the steady states. The idea of this approach is that the noise is small enough that it can be seen as a perturbation to the deterministic model of differential equations. It is again visible the contribution to physics, where the same approach was used by Einstein to explain the Brownian motion [68] and is the

foundation of perturbation theory in quantum mechanics [74].

2.3. SOS response in E. Coli

The SOS response is a mechanism used by *E. coli* and other bacteria to repair double-strand DNA breaks. It was first proposed in 1974 by Radman [75] to solve a conundrum, initiated by Weigle, that puzzled biologists. Weigle observed that the survival of a previously UV-irradiated bacteriophage λ was greatly improved by UV-irradiating also the host cell, and among reactivated phages, there was an overrepresentation of mutant λ phages. Moreover, while the phage λ requires normally 40 pyrimidine dimers (induced by UV) to die, reactivated phages require up to 10^3 dimers, similar to the host *E. coli*. This conundrum led Radman to propose the existence of an inducible SOS response, which can cause mutations and recombinations of the DNA during the repair process. Over the year, this process turned out to be of fundamental importance to understand the emergence of antibiotic resistance. It also proved that mutations were no more unavoidable rare stochastic accidents, but an active inducible cellular process [76].

The main players of the SOS response are the transcriptional repressor LexA and the recombinant protein RecA. The combined action of these two molecules allows the cell to sense for single-strand DNA (ssDNA) that signal the presence of a break to be resolved. The induction of the network starts when RecA polymerizes on a ssDNA. These filaments are able to increase the autocleveage rate of LexA and lead to the derepression of several other genes involved in the SOS response. These activate different pathways, like homologous recombination (HR), nucleotide excision repair (NER), and translesion synthesis (TLS) [77]. The activation of the different pathways is believed to follow a strict temporal order, where the repair mechanisms with the high fidelity (NER, HR) are induced before the most mutagenetic ones (TLS) [78].

RecA

The Recombinant Protein A (RecA) is able to recognize and bind ssDNA and it is the first actor involved in the activation of the SOS response. The binding affinity is increased by the binding of an ATP molecule between two monomers, hence the minimal unit for the filament formation is the dimer [79]. The nucleoprotein filament of RecA-ssDNA has two different functions. It can either search for a homologous dsDNA sequence to use as a template to repair the DNA damage, or it can stimulate the autocleveage of free LexA.

The inducing activity of the RecA filament is regulated through two processes. First of all, not ever ssDNA comes from a damage, for example during DNA replication some single strands are exposed. To prevent RecA to bind to these ssDNAs, single strands are quickly coated and hidden by Single Strand Binding proteins (SSB). In order for RecA to compete with SSBs, an additional set of recruiting proteins are required, namely RecFOR and RecBC [38]. Second, once RecA is loaded on the ssDNA, it can lead to HR, without requiring the activation of the SOS pathway, or

to the autocleavage of LexA. When the RecA filament is engaged in homologous recombination, it requires hydrolysis of ATP, which in turn prevents the interaction with LexA, leading to a down-regulation of the SOS response [80].

How RecA is able to find a homologous sequence among the huge searching space of the entire genome has been a riddle until recently. The ssDNA inside the filament is stretched by doubling the distance every three nucleotides [79]. A dsDNA binds the filament to a secondary binding site and it is unwound and stretched using the free energy contained in the ATP molecules attached to the RecA dimers. This stretch allows the candidate homologous filament to be tested against the ssDNA in triplets of nucleotides. If a sufficient degree of matching (more than 8bp) is found, the pairing is stabilized and the HR process is initiated [79]. The scanning of the dsDNA is performed by combining a 1D sliding along the dsDNA sequence and a 3D diffusion among different parts of the dsDNA [79, 81]. This search strategy is in common with other regulatory proteins, as outlined by Berg and von Hippel [82].

LexA

LexA, which stands for locus for X-ray sensitivity A [83], is the transcription factor responsible to repress the genes of the SOS regulatory network. It is composed of 202 amino acids which fold into two domains linked by a flexible hinge region [84] and it is predominantly found as a homodimer. Each monomer contains a catalytic site that recognizes a specific region of the same protein, inducing the auto-cleavage of LexA. In normal conditions, the protein is folded in such a way that the catalytic site and the cleavable region are separated, but upon interaction with the RecA filament, a conformational change is induced, that allows the active site to cleave the monomer. The cleaved monomer not only cannot bind the DNA efficiently, but it also exposes residues that are recognized by proteases, and it is thus quickly degraded [84].

To notice that binding with the DNA makes the structure of LexA more rigid, hindering the conformational change induced by RecA [85], and the DNA sterically precludes the interaction with the RecA filament [86]. Therefore, only free LexA can be cleaved, which means that upon induction of the SOS response, the promoter has to wait for the spontaneous fall off of LexA before being derepressed.

Finally, the dynamic of LexA degradation and recovering after the damage has been solved is very quick. At the population level, after induction with UV light, LexA level decays with a half-time of one minute and after the resolution of the break, it regains its initial level in a time that depends on the strength of the UV irradiation [87].

2.4. Outline of the thesis

In this thesis, I will discuss three projects related to gene regulation.

The first project deals with data analysis from flow cytometers. This is an established technology for the analysis of eukaryotic cells, and in recent times it has been applied also for the gene expression distribution in bacterial populations. Due to its

high-throughput, it is a highly attractive technology, but we are still lacking a systematic investigation of its accuracy for bacterial populations. In particular, I show that, contrary to eukaryotic cells, the fluorescence signal contains a substantial amount of non-biological components, like background fluorescence and electronic noise. I therefore show an accurate pipeline to extract quantitative information about bacterial gene expression. The results are collected in a paper, currently under revision at Plos One.

The second project deals with the mathematical modeling of the SOS response. Here I show that, contrary to well-established models of Lac operon and λ phage regulation, the SOS network cannot be described by an equilibrium model. On the other hand, I propose that rather than reading out an effective time average LexA concentration, the induction of the SOS promoters is a non trivial function of stochastic transient drops in LexA concentration. I also describe the consequences an equilibrium and a non-equilibrium model have on the cellular response to the external environment. The proposed dynamic model is described in a manuscript, currently under preparation for Plos Computational Biology.

Finally, I describe a side project done to understand unpublished data of flow cytometry measurements collected in the lab. The software MotEvo [88], previously developed in the lab, exploits the established techniques of positional weight matrices to infer all the possible binding sites and their energies in a promoter of interest. My goal was to use the binding sites inferred by MotEvo to compute the probability for the system to be in a transcribing state (at least one σ_{70} bound, and no repressor between the last σ_{70} and the transcription starting site). I used a thermodynamic framework and I wrote the program in a highly optimized C++ code, which is able to deal with the large combinatorial number of possible binding configurations arising from a promoter with multiple binding sites.

Notice that, for ease of reading, the papers are included as a whole, together with their supplementary and bibliography. That is, the supplementary and the bibliography of each paper are kept separated and not merged together at the end of the thesis.

3. Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria

Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria

Luca Galbusera[1], Gwendoline Bellement-Theroue[1], Arantxa Urchueguia[1], Thomas Julou[1,*], and Erik van Nimwegen[1,*]

1 Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

* E-mails: thomas.julou@unibas.ch, erik.vannimwegen@unibas.ch

Abstract

Fluorescence flow cytometry is increasingly being used to quantify single-cell expression distributions in bacteria in high-throughput. However, there has been no systematic investigation into the best practices for quantitative analysis of such data, what systematic biases exist, and what accuracy and sensitivity can be obtained. We investigate these issues by measuring the same *E. coli* strains carrying fluorescent reporters using both flow cytometry and microscopic setups and systematically comparing the resulting single-cell expression distributions. Using these results, we develop methods for rigorous quantitative inference of single-cell expression distributions from fluorescence flow cytometry data.

First, we present a Bayesian mixture model to separate debris from viable cells using all scattering signals. Second, we show that cytometry measurements of fluorescence are substantially affected by autofluorescence and shot noise, which can be mistaken for intrinsic noise in gene expression, and present methods to correct for these using calibration measurements. Finally, we show that because forward- and side-scatter signals scale non-linearly with cell size, and are also affected by a substantial shot noise component that cannot be easily calibrated unless independent measurements of cell size are available, it is not possible to accurately estimate the variability in the sizes of individual cells using flow cytometry measurements alone. To aid other researchers with quantitative analysis of flow cytometry expression data in bacteria, we distribute *E-Flow*, an open-source R package that implements our methods for filtering debris and for estimating true biological expression means and variances from the fluorescence signal. The package is available at `https://github.com/vanNimwegenLab/E-Flow`.

1. Introduction

It is has become well recognized that, due to the intrinsic stochasticity of the gene expression process, even isogenic populations of microbial cells growing in homogeneous environments exhibit significant heterogeneity in their gene expression, e.g. [1–4]. Therefore, the traditional studies at the population level, by smoothing out

this heterogeneity, tend to hide crucial information [5] [6] that is required to correctly understand and interpret the observed behavior of microbes [7].

Although most studies of single-cell gene expression in bacteria use fluorescent reporters in combination with microscopy to quantify gene expression in single cells, fluorescence flow cytometry (FCM) is also an attractive alternative methodology for single-cell gene expression studies in bacteria. In particular, given that flow cytometers can quantify the fluorescence of thousands of cells per second, flow cytometry allows for high-throughput characterization of the single-cell expression distributions of a large number of fluorescent reporters [8, 9]. Indeed, in recent years there has been a large number of studies in which standard commercially available flow cytometers were used in combination with fluorescent reporters to measure gene expression at the single-cell level in bacteria [10–31], as well as in single-celled eukaryotes [32, 33].

However, so far there has been little systematic investigation into the accuracy of flow cytometry in quantifying gene expression in single cells, or a systematic comparison with the results from microscopy measurements. Here we aim at filling this gap by systematically comparing flow cytometry measurements with measurements from a microscopy setup. In particular, there are several technical challenges in analyzing fluorescence flow cytometry data of individual bacterial cells:

1. *Differentiating cells from debris*. Bacterial cells are typically one thousandth the volume of mammalian cells, which places them near the edge of instrument detection. At this size it can be challenging to differentiate viable cells from debris of similar size [9, 34–37]. In the literature different approaches are used to separate debris from viable cells. Most of these approaches use *ad hoc* combinations of the scatter measurements to retain a fraction of the measurements.
   We here perform a careful analysis of all the scatter signals reported by the flow cytometer and propose a principled way of identifying debris from viable cells using a Bayesian mixture model that considers all the information available in the scatter signals.
2. *Distinguishing measurement noise from biological variability*. In order to quantify the amount of biological gene expression variation in a population of isogenic cells, it is important to quantify to what extent variation in measured fluorescence intensity derives from biological variation, and to what extent it derives from measurement noise.
   We show that flow cytometry measurements contain a substantial amount of shot-noise which can be easily mistaken for true biological variability, and develop a method to correct for this shot-noise using measurements of reference beads that are commonly used to calibrate flow cytometers. Using a mixture modeling approach, we develop a rigorous method for estimating the true mean and variance in expression levels of a population of cells.
3. *Accounting for autofluorescence*. Because of their small size, bacterial cells have a relatively low copy number of proteins and mRNA [38], which results in a low fluorescent emission signal, that can be close to the autofluorescence level[39]. Therefore, gene expression estimates require careful correction for autofluorescence.
   We here provide methods for correcting both the estimated mean and variance

in fluorescence levels for autofluorescence using measurements of cells that do not express GFP.

4. *Estimating the distribution of GFP concentrations.* While we provide methods for accurately estimating the distribution of total GFP levels in a population of cells from the flow cytometry measurements, microscopy measurements show that total GFP levels correlate strongly with cell size and that GFP concentrations vary significantly less across cells than total GFP.
   Estimating the distribution of GFP concentrations directly using flow cytometry requires to not only estimate the total GFP but also the volume of individual cells. Although forward- and side-scatter signals can be used to distinguish the average size of populations of cells of sufficiently different shapes and sizes [40–43], it is substantially more challenging to accurately quantify the relatively small cell-to-cell variations in cell volume for populations of isogenic bacteria growing in a homogeneous environment. In line with previous works [35, 44–46] we find that, because forward- and side-scattering measurements depend on cell volume in a complex non-linear manner and contain a substantial amount of shot noise that cannot be easily calibrated, it is impossible to accurately quantify the sizes of individual cells. Consequently, it is not possible to directly estimate the distribution of GFP concentrations from flow cytometry measurements. However, we show that because GFP concentrations and cell sizes fluctuate approximately independently, it is still possible to obtain reasonably accurate quantifications of the *relative* amounts of GFP concentration fluctuations for different genes.

Although the precise flow cytometer used will of course affect the precise values of the measurements and calibrations, the methods for separating true cells from debris, estimating and correcting for autofluorescence, and correcting for measurement shot noise, are general and should be applicable to data from most flow cytometers. Our methods have been implemented as an R package called *E-Flow*, which we make publically available and can be easily integrated in any flow cytometry data analysis pipeline.

2. Materials and methods

2.1. Strains and growth conditions

We measured the fluorescence distributions for a number of different *Escherichia coli* MG1655 strains carrying fluorescent transcriptional reporters (a GFP gene downstream of a given promoter, either on a low-copy number plasmid, or integrated into the chromosome) both using flow cytometry of batch cultures and time lapse microscopy in a microfluidic device (Mother Machine). We considered a number of

different promoters, that have different means and variances of expression levels.

In particular, we considered *E. coli* strains with a lacZ-GFP fusion integrated in the chromosome [47], and a set of *E. coli* strains that carry a transcriptional reporter expressed from a low copy number plasmid [48]. These reporters included known target promoters of the LexA transcription factor (dinB, ftsK, lexA, polB, recA, ruvA, or uvrD) [49] and two synthetic promoters that were obtained by experimental evolution and express at levels corresponding to the median and the 97^{th} percentile of all native *E. coli* promoters [23]. Throughout the paper, we refer to these two synthetic promoters as high and medium expressers.

To estimate autofluorescence in both the FCM and microfluidic experiments, we used two strains that carry plasmids where the GFP sequence is downstream of a random sequence (pUA66 and pUA139) [48] and hence do not express GFP [23].

In the microfluidic experiments, cells carrying a lacZ-GFP fusion were tracked using time-lapse microscopy while growing in a microfluidic device in M9 minimal media supplemented with 0.2% lactose (which leads to full induction of the lac operon), taking measurements every 3 minutes [47]. Detailed experimental procedures are available in the corresponding publication [47]. Microfluidic experiments with strains carrying a transcriptional reporter expressed from a plasmid were performed following the same procedure, using M9 + 0.4% glucose (supplemented with $50\mu g$ / mL of kanamycin during the overnight preculture only) and acquiring data over 4 hours.

To obtain comparable measurements with flow cytometry (FCM), the same strains were grown in the same conditions as for the microfluidic measurements. Practically, strains expressing from a plasmid were inoculated from frozen glycerol stocks and grown overnight in $200\mu L$ of M9 + 0.4% glucose supplemented with $50\mu g/$mL of kanamycin. After $100\times$ dilution in fresh medium without kanamycin, strains were grown to saturation again, and re-diluted $100\times$ to fresh medium without kanamycin. For the lacZ-GFP strain, we used $200\mu L$ of M9 + 0.2% lactose with only one overnight culture. For all strains, expression was measured in mid-exponential phase (typically after 4h), adjusting the cell concentration with PBS if necessary. All cultures used for FCM measurements were incubated in 96-well plates at 37 °C with shaking at 600-650 rpm.

To study the accuracy of the scatter signal for estimating cell size, we used the data acquired for a previous project in the lab [31]. In summary, *E. coli* strains were grown in different media characterized by different size distributions: M9 supplemented with either 0.2% glucose (w/v), 0.2% glycerol (v/v) or 0.2% lactose (w/v); a MOPS based synthetic rich media (Teknova, M2105) supplemented with 0.2% glucose. We refer to the original study for more information about the cell cultures and growth conditions [31].

2.2. Flow cytometry

The flow cytometry measurements were obtained with a BD FACSCanto II cytometer and were managed using the Diva 8 software. The excitation beam for the GFP was

set at 488 nm and the emission signal was captured with a 530/30 nm bandpass filter. The gain voltage were set by default to 625V, 420V, and 600V for FSC, SSC, and GFP acquisition respectively, and events were created for measurements where $FSC > 200$ & $SSC > 200$. For each sample, 5×10^4 events were recorded at a typical flow rate ranging from 1×10^4 to 2×10^4 per second.

2.3. Calibration beads

CS&T (Cytometer Setup and Tracking Beads) are artificial fluorescent beads that are used to calibrate fluorescence measurement values [50]. To calibrate the measurement shot noise we used beads of lot 41720 that contains beads of two different sizes, which have high, medium ($3\mu m$ in size) and low fluorescence ($2\mu m$ in size) levels.

2.4. Microscopy size estimation

To estimate cell sizes, strains containing a plasmid without promoter were selected from 4 different media with different size distributions (M9 + glucose, lactose or glycerol; MOPS + glucose. See *Strains and growth conditions*). Cells were then placed on a 1% agarose pad and phase contrast images were obtained with a Nikon Ti-E microscope using a 100 × Ph3 objective (NA 1.45) and an Hamamatsu Orca-Flash 4.0 v2 camera. Cell outlines were identified using a custom MATLAB pipeline [31].

2.5. R package E-Flow

The analysis pipeline presented in this paper has been implemented in the R package *E-Flow* available on GitHub *https://github.com/vanNimwegenLab/E-Flow*. Here the methods were tested with flow cytometers manufactured by BD and operated through the DIVA software. Nonetheless we kept the methods as general as possible, such that they should be applicable to flow cytometers of other manufacturers.

For a detailed explanation of the package, we refer to the GitHub page, including the vignette and the documentation of the individual functions. Here we list the main components of the software:

1. *Filtering*: The cells are filtered based on their scattering profile and an estimate of the mean and variance of the population is obtained. This is the most resource-intensive step and therefore can be parallelized.
2. *Mean and variance*: The mean and variance of the population of cells is computed. Measurements that are outliers in the fluorescence are accounted for using a mixture model.
3. *Autofluorescence removal*: Using the fluorescence distribution of non-expressing cells, an estimate of the autofluorescence is obtained and subtracted from the mean and variance of the population.

4. *Shot noise removal*: The shot noise introduced by the machine is removed and a corrected variance is calculated. This can be regarded as a proxy for the biological gene expression noise.

3. Results

3.1. Signals reported by the cytometer

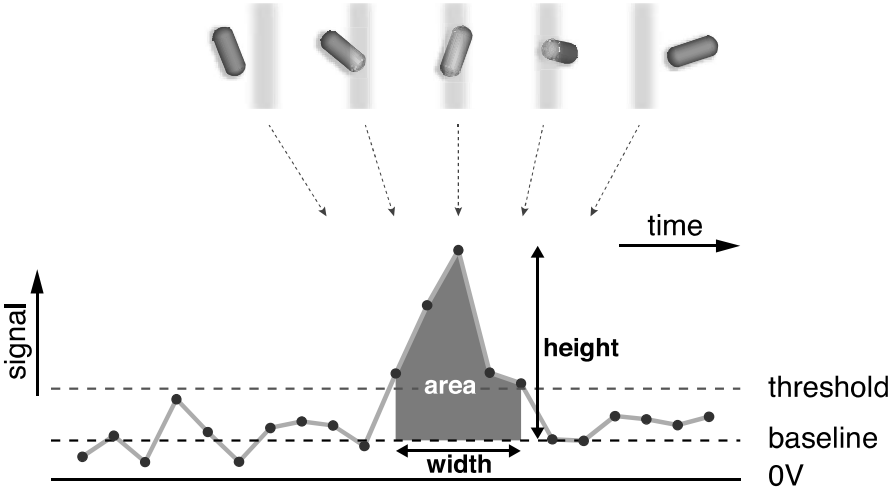

Figure 3.1.: **The signals reported by the cytometer.** As a particle enters the laser beam, an electric signal (pulse) is generated which reaches its maximum when the particle is in the middle of the beam and trails off as the particle leaves the beam. Each pulse with height over a certain threshold is recorded and three quantities are reported: height, area, and width of the pulse.

In flow cytometry, a beam of light is used to illuminate cells that flow one by one through a channel; a series of detectors is able to record the light scattered by the single cells at right angles or in the forward direction and the cell fluorescence stimulated by the incident light beam. Most flow cytometers, including the BD Canto II used here, report for each measured 'event' (typically corresponding to a single measured cell) a forward-scatter signal, a side-scatter signal, and a fluorescence signal. Each of these signals is in turn represented by 3 statistics of the electrical impulse, namely height, area, and width of the impulse (Fig 3.1). The height corresponds to

the maximal value of the impulse, the area to the area under the curve and the width is its time duration [51] (see Supplementary Material section A.1).

We noticed that these statistics are not all independent. In particular, for all three signals, the area is always directly proportional to the product of height and width (Suppl. Fig. A.1 and Supplementary Material section A.2). Moreover, while height and width vary approximately independently across events, the area correlates significantly with both (Suppl. Fig. A.2). Therefore, we only use height and width for the subsequent analysis of the forward- and side-scatter signals.

For the fluorescence signal we were unable to find any systematic dependence between the width of the fluorescence signal and any biological signal, such as cell size or total fluorescence. In addition, for the calibration beads there is clearly no information in the width of the fluorescence signal (Suppl. Fig. A.3 and Supplementary Material section A.3). Therefore, for the fluorescence signal we will only use the height statistic as a proxy for the total fluorescence of the cells. While we believe that all these considerations apply generally to flow cytometers, we also observed anomalous behavior of the signal at very low fluorescence levels that may be specific to the BD machine used here (see Suppl. materials section A.4). Due to this anomalous behavior, quantitative analysis is restricted to constructs for which the GFP fluorescence is at least as high as the autofluorescence of the cells (see Suppl. Fig. A.4).

3.2. Filtering events based on their forward- and side-scatter

In comparison to eukaryotic cells, bacterial cell produce only relatively weak scattering signals, and we used permissive settings of the device to call events. This increases the likelihood of having spurious observations that correspond to non-viable cells and other debris. Consequently, we needed a strategy for using the measured forward- and side-scatter of the events to separate viable cell measurements from debris. As explained above, the scatter of each event is characterized by 4 statistics, namely the height and width of both the forward- and side-scatter. Thus, the measured scatter of each event can be represented by a point in a 4-dimensional space, and a given dataset corresponds to a distribution of points in this 4-dimensional space. To separate viable cells from debris we fit this distribution with a mixture of a multivariate Gaussian distribution and a uniform distribution, as detailed in the Supplementary Material, section A.4. The rationale behind this mixture modeling is that most of the data represents good cells and should cluster in this 4-dimensional space, whereas the outliers are relatively rare and more widely distributed. In this model, the Gaussian part of the mixture captures the cluster of good cells, while the uniform component takes care of outliers, i.e. fragments of dead cells and other debris.

Figure 3.2 shows 2D projections of the 4D scatter of forward- and side-scatter for events taken from *E. coli* cells that carry a lacZ-GFP fusion (see [47] for a description of the strain used) while growing in M9 minimal media supplemented with lactose. Besides the scatter of measurements, Fig. 3.2 also shows the multivariate Gaussian

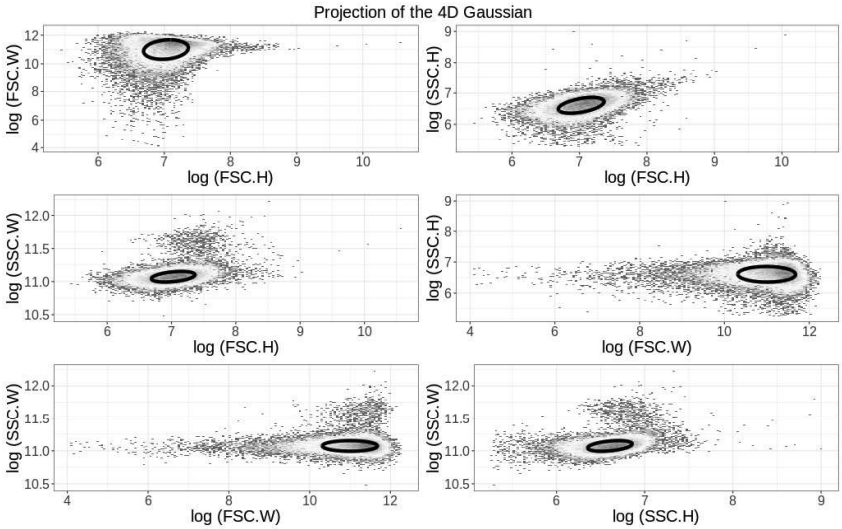

Figure 3.2.: **Mixture model fitting of the scatter signals.** The panels show different two-dimensional projections of the full 4D distribution of heights (H) and widths (W) of forward- (FSC) and side-scatter (SSC) measurements for 5×10^4 events obtained from *E. coli* cells growing in M9 minimal media with lactose. The ellipses show the contour of the fitted multivariate Gaussian distribution, one standard deviation away in each principal direction. Note that the color indicates the local density of points.

fitted to the data, showing that this Gaussian indeed captures the bulk of the measured events.

Once the mixture model has been fitted to a dataset, a posterior probability p_i is calculated for each measured event i to correspond to a viable cell, i.e. the probability that the observation derives from the multivariate Gaussian component of the mixture as opposed to deriving from the uniform distribution. By default the *E-flow* software retains all events with posterior probability $p_i \geq 0.5$ and discards as outliers events with $p_i < 0.5$, but the user can change this threshold probability if desired. Suppl. Fig. B.1 shows the same scatter of measured events as shown in Fig. 3.2, but now with selected events in red and events that were filtered out in black when using the default threshold of $p = 0.5$.

As the forward- and side-scatter should reflect the size, shape and composition of the objects measured in each event, one may wonder to what extent filtering out events based on their forward- and side-scatter may bias measurements towards cells of a certain size. Indeed, in previous work, e.g. [19], researchers have attempted to select subsets of cells with similar shapes and size by very strictly gating on forward- and side-scatter, retaining only those cells that lie near the center of the Gaussian distribution. To check the viability of such an approach, we compared the distribution of measured fluorescence levels with two extreme filtering strategies: one very lenient in which all events with $p > e^{-10}$ are retained and one very strict in which only cells with $p > 1 - e^{-10}$ are retained. As shown in Suppl. Fig. B.2, there is virtually no difference in the observed distribution of fluorescence levels between the very lenient and very strict filtering. Given that we expect total fluorescence to scale with cell size, this observation suggests that strict filtering on forward- and side-scatter is not effective for selecting out a subset of cells with similar size.

3.3. Flow cytometer measurements are affected by substantial measurement noise

When using the flow cytometer to estimate single-cell gene expression, we aim to quantify the variation in gene expression across a population of isogenic cells growing in a homogeneous environment. In such conditions, bacteria at different stages of their cell cycle vary by roughly two-fold in size, and their total fluorescence is typically proportional to cell size.

In a previous work we have established that time-lapse microscopy measurements of cells growing in microfluidic devices can measure cell size with an accuracy of around 3% error and GFP copy-number G with an error of about $\sqrt{G}$ [47]. Using such microscopy measurements on *E. coli* cells carrying a lacZ-GFP fusion gene in its native locus while growing in M9 minimal media with lactose, we find a high correlation between lacZ-GFP levels and cell size (Fig. 3.3, top panels). That is, because lacZ-GFP concentrations fluctuate only moderately from cell to cell, and both size and GFP level measurements have high accuracy, the measured cell length explains around 70% of the variance in total fluorescence.

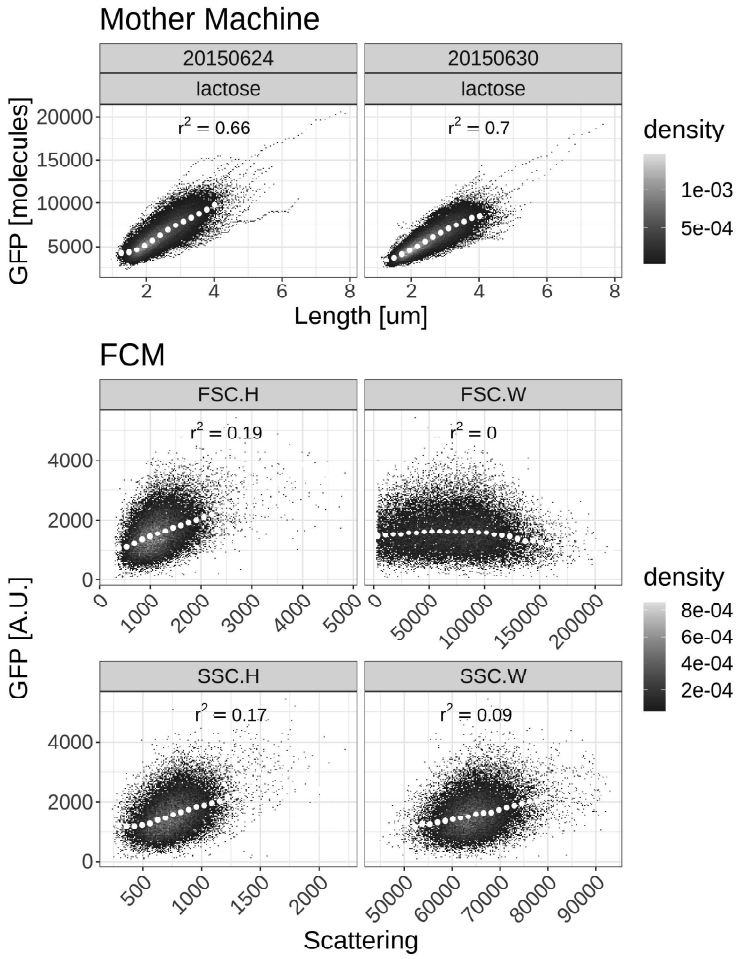

Figure 3.3.: **Correlation between cell size and fluorescence measurements for microscopy and cytometer measurements.** Each panel shows measured GFP fluorescence (vertical axis) and cell size estimates (horizontal axis) of cells growing in M9 minimal media with lactose. The top 2 panels show microscopy measurements from a microfluidic device [47]. The lower 4 panels show fluorescence measurements as a function of size estimates based on forward- (middle 2 panels) or side-scatter (bottom 2 panels) measurements in the flow cytometer (FCM). The squared Pearson correlations between fluorescence and size measurements are indicated in each panel. Note that the color indicates the density of points. The white dots show median values of equally spaced bins along the horizontal axis.

We calculated the analogous correlation between size and total fluorescence in the flow cytometer for the same strain growing in the same environment, using the scatter signals as representing the cell size. We see that, in contrast to the microscopy measurements, there is only a very weak correlation between total fluorescence and scattering measurements (Fig. 3.3, bottom 4 panels).

The lack of correlation between size and fluorescence measurements in the cytometer strongly suggests that either the fluorescence measurements, the size measurements, or both are much more heavily affected by measurement noise than in the microfluidic experiments. In the following we will look at different sources of noise and how to deal with them.

3.4. Estimating the mean and variance of the fluorescence distribution

As has been observed by others [38], we observed that for virtually all *E. coli* promoters, the distribution of fluorescence levels is fitted very well by a log-normal distribution [23], i.e. the log-fluorescence follows a Gaussian distribution. Our *E-Flow* package fits a Gaussian distribution to the measured log-fluorescence levels of single cells, estimating a mean μ and variance v for a given population of cells. However, we noticed that, even after filtering events on forward- and side-scatter as explained above, there are still clear outlying events, i.e. with fluorescence levels that lie far outside the range observed for almost all other events. To separate these outliers from valid measurements we modeled the distribution of log-fluorescence levels as a mixture of a Gaussian and a uniform distribution, fitting its parameters using expectation maximization (see section A.4 of the Supplementary Material for details). The *E-Flow* package calculates an estimated mean μ and variance v of the log-fluorescence levels of a set of measurements, together with error bars σ_μ and σ_v on these estimates. In addition, transforming from log-fluorescence back to fluorescence in linear scale, the package also calculates mean and variance of the distribution of fluorescence levels, together with error bars (Suppl. Mat. A.4).

3.5. Autofluorescence estimation

It is well known that the laser used to excite the GFP can also excite other cellular components of the cell, resulting in an "autofluorescence" signal that also occurs in cells without GFP molecules. In addition, the fluorescence signal may also contain a background fluorescence component coming from sources other than the cell's autofluorescence. In order to estimate GFP levels, we need to correct for these other sources of fluorescence and the *E-Flow* package allows for such correction by using measurements of cells that do not express GFP. Let's call I_M the measured fluorescence intensity, I_T the true intensity (deriving from GFP molecules) and A the component from other sources of fluorescence, which for simplicity we will refer to

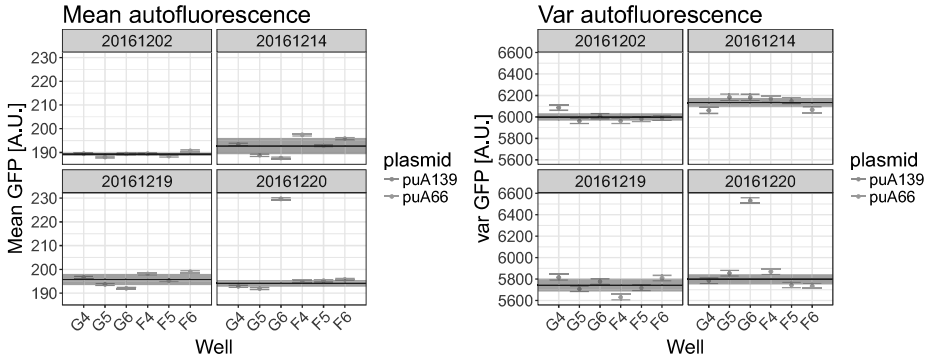

Figure 3.4.: **Autofluorescence measurements.** Each panel shows the measured mean fluorescence (left 4 panels) and variance in fluorescence (right 4 panels) on one day, with each bar indicating the measured value and error bar for one replicate. Two different strains were used (indicated in red and blue) and each was measured in triplicate on each day. The black line and grey bar indicate the estimated averages μ_d and corresponding error-bars σ_d for each day d. Note that well G6 on 20/12/2016 appears to be an outlier, possibly due to contamination of the well, which was excluded from the analysis.

as autofluorescence. We have the relation

$$I_M = I_T + A. \tag{3.1}$$

Assuming that the component A fluctuates independently from the true fluorescence I_T, we obtain

$$\langle I_T \rangle = \langle I_M \rangle - \langle A \rangle \tag{3.2a}$$
$$\mathrm{var}(I_T) = \mathrm{var}(I_M) - \mathrm{var}(A). \tag{3.2b}$$

Thus, in order to correct for autofluorescence, it suffices to estimate both its mean $\langle A \rangle$ and variance $\mathrm{var}(A)$. These can be easily estimated by performing fluorescence measurements on cells that either lack GFP, or where the GFP gene is known not to be expressed, and applying the same Bayesian mixture model described above. Once $\langle A \rangle$ and $\mathrm{var}(A)$ have been estimated in this way, the true mean and variance of GFP expression in cells carrying an active reporter can be calculated using equation (3.2).

We measured autofluorescence levels A using strains carrying two different plasmids not expressing GFP, designed as negative controls (see materials and methods) on 4 different days, measuring each strain in triplicate on each day. Figure 3.4 shows the estimated mean fluorescences (left 4 panels) and variances in fluorescences (right 4 panels) for each replicate of each strain (red an blue) on each day (one panel per day). Using a procedure described in Suppl. Mat. section A.4, we averaged over different replicates on each day to calculate a mean fluorescence μ_d for each day (black line in each panel) and an error bar on this estimate (grey region in each panel), and similarly for the variances on each day (right 4 panels). We then additionally

averaged over different days to calculate an overall average mean autofluorescence $\bar{\mu}$ and an overall average variance in autofluorescence $\bar{v}$ (see Suppl. Mat. A.4).

3.6. Mean fluorescence levels agree between microscopy and FCM across the entire range of expression levels

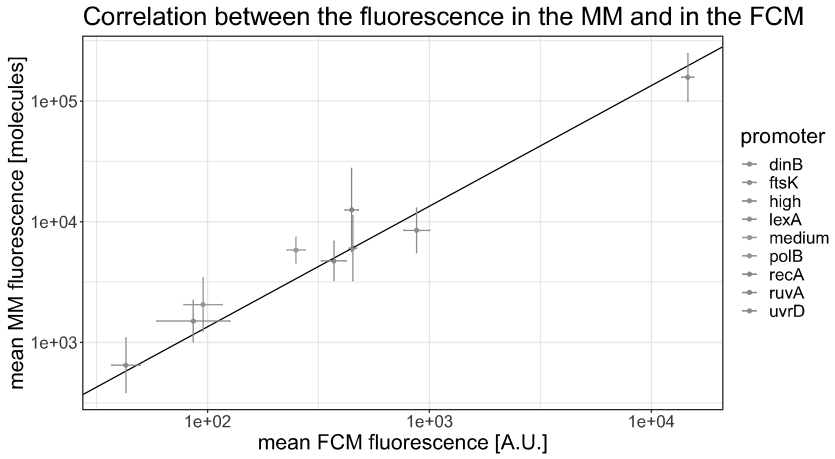

Figure 3.5.: **Estimated mean expression levels of different promoters as estimated by FCM and microscopy.** After correcting for autofluorescence, mean fluorescence levels of different promoters (colors) are perfectly linearly correlated between microscopy and FCM measurements, over the entire range of expression levels. The scales of the axes are in natural log and the error bars show the standard error of the mean. Note that the slope of the black line is 1.

Although commercial flow cytometers have been designed to ensure a linear relationship between GFP content and fluorescence measurements over a wide range and previous gene expression studies studies using FCM have operated under this assumption, we here tested this assumption by comparing estimated mean fluorescence levels of different promoters between FCM and microscopy measurements. To do so we calculated the mean fluorescence levels, corrected for autofluorescence, of promoters with a wide range of expression levels using both the FCM and microscopy measurements. As shown in Fig. 3.5, we indeed find that there is a perfectly linear relationship between the average expression levels of the different promoters as estimated by FCM and microscopy, over the entire expression range.

3.7. Cytometer fluorescence measurements exhibit significant shot noise

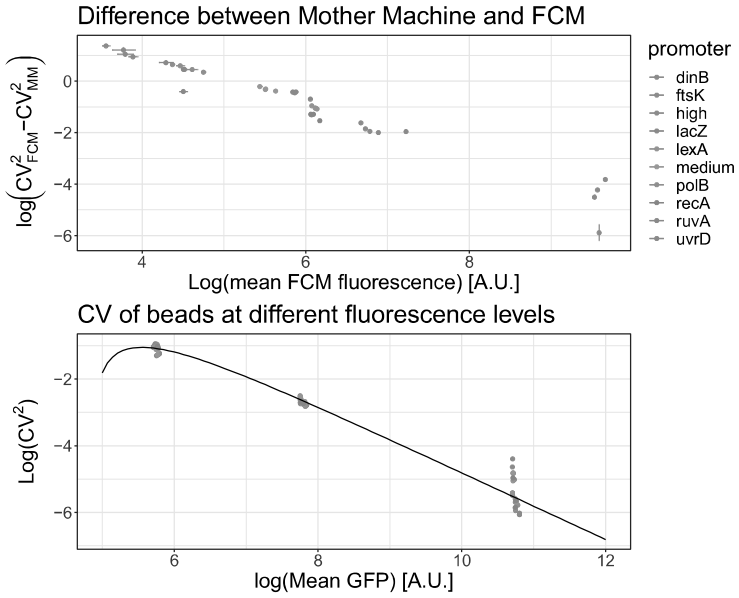

Figure 3.6.: **Difference in CV^2 between the FCM and microscopy measurements shows FCM measurements contain substantial shot noise.** *Top:* Difference between the CV^2 as measured by the FCM and the microscopy setup for different transcriptional reporters of *E. coli* promoters (colored points). Both axes are shown on a logarithmic scale. The difference in CV^2 scales inversely with mean expression. *Bottom:* The observed CV^2 of calibration beads of three different intensities also decreases as the inverse of mean intensity and this dependence can be well modeled by shot noise (black line), as given by equation (3.3).

We used equation (3.2) to remove the autofluorescence contribution from the mean expression and variance of the population for a number of different transcriptional reporters and calculated the observed squared coefficient of variation CV^2 for each promoter. Next, we took microscopy measurements from our microfluidic setup of the same *E. coli* strains growing in the same conditions and measured CV^2 for each of these promoters as well. As shown in the top panel of Fig. 3.6, we observe systematically higher CV^2 in the FCM than in the microscopy setup and the difference in the two CV^2s decreases almost exactly inversely with the mean expression level.

Since the growth conditions in the FCM and the microfluidic setup were kept as close as possible, the true CV^2 of the distribution of total GFP levels should be highly similar, so that the difference between the measured CV^2 must derive from measurement noise. Indeed, one source of noise whose contribution to CV^2 is expected to scale inversely with mean intensity is shot noise from the photomultiplier

tube, whose CV^2 scales as 1/mean [52]. Due to this noise, one generally has the following relationship between the measured fluorescence intensity I_M and the true intensity I_T:

$$I_M = I_T + \epsilon\sqrt{I_T} + O, \tag{3.3}$$

where ϵ is a Gaussian random variable with mean 0 and an (unknown) variance δ^2 which quantifies the size of the shot noise. The constant term O is an offset that is added in BD devices in order to prevent the clipping of negative values during the digital conversion, when true intensities I_T are close to zero [51].

Flow cytometers are often calibrated using synthetic fluorescent beads of known intensities and such beads can also be used to estimate the size δ of the measurement shot noise. As shown in the bottom panel of Fig. 3.6 (and Suppl. Fig. E.1) the CV^2 of the artificial beads also drops inversely with mean expression. If we assume that the true variation of the beads can be ignored, we get from equation (3.3) that the measured CV^2 is

$$CV_M^2 = \frac{\delta^2}{\langle I_M \rangle} - \frac{\delta^2 O}{\langle I_M \rangle^2} \tag{3.4}$$

If we define $Y = CV_M^2 \langle I_M \rangle$ and $X = \frac{1}{\langle I_M \rangle}$, we obtain

$$Y = \delta^2 - \delta^2 O X \tag{3.5}$$

and we can infer both the strength δ and the offset O by fitting Y as a simple linear function of X. This simple approach leads to an inferred value of $\delta = 13.4$ and $O = 128$. In the Supplementary Material section A.4 we also present a more sophisticated Bayesian mixture model approach to inferring these quantities, which does not ignore the true variability of the beads, but assumes that the CV^2 of the true intensities I_T is the same for all three types of beads. Using this more rigorous procedure, the resulting strength and offset are: $\delta = 12.7 \pm 0.6$, $O = 97 \pm 29$ (Suppl. Fig. E.1), which are close to the values we would have obtained with the more simple linear model of equation (3.4). Using this result we can now fit the observed CV^2 that we expect to see; the fit describes well the observed data, as shown in the bottom panel of Fig. 3.6 (and in the top left panel of Suppl. Fig. E.1).

Finally, section A.4 of the Supplementary Material investigates two more subtle technical points that one might think could affect the direct comparison of FCM measurements and microscopy measurement from growth in the microfluidic device. First, one could argue that the age-distributions of the population of cells in the microfluidic device and in a population that is growing exponentially (i.e. as used in the FCM) are different. That is, since in the microfluidic device some newborn daughters are constantly washed out of the growth channels, there are relatively fewer cells close to birth and more cells close to division in the microfluidic device than in a population undergoing exponential growth in bulk (Suppl. Fig. F.1). Since total fluorescence correlates with cell size, which again correlates well with time since birth, the access of 'old' cells could in principle effect the distribution of total fluorescence one observes. However, as shown in Suppl. Mat. section F.1, we derive theoretically that the effects of the altered age-distribution are small enough to be

neglected (Suppl. Fig. F.2). Second, since in the microfluidic setup we measure the fluorescence of a cell multiple times during its cell cycle, there are clearly substantial correlations between different measurements and one might wonder whether this could significantly affect the observed statistics. In Suppl. Mat. section F.2 we show that this effect is also negligible (Supplementary Fig F.2).

3.8. Correcting for autofluorescence and shot noise

After having estimated the mean and variance of the autofluorescence, and the strength of the FCM's shot noise, we can now correct the measured means and variances of transcriptional reporters for these two components. Combining the autofluorescence contribution from equation (3.1) and the shot noise component from equation (3.3), we can write the measured intensity I_M as

$$I_M = I_T + A_T + \epsilon\sqrt{I_T + A_T} + O, \tag{3.6}$$

and the measured autofluorescence as

$$A_M = A_T + \epsilon\sqrt{A_T} + O, \tag{3.7}$$

where variables with subscript T correspond to true values and variables with subscript M correspond to measured values, ϵ is again a Gaussian distributed variable with mean zero and variance δ^2 and O is a constant offset. From these equations we find for the mean and variance of the measured intensities I_M:

$$\langle I_T \rangle = \langle I_M \rangle - \langle A_M \rangle, \tag{3.8a}$$

and

$$\mathrm{var}(I_T) = \mathrm{var}(I_M) - \mathrm{var}(A_M) - \delta^2 \langle I_T \rangle. \tag{3.8b}$$

Using these expressions we calculated $\langle I_T \rangle$, $\mathrm{var}(I_T)$ and the resulting CV^2 for a set of different *E. coli* promoters and compared the results with the CV^2 measured for the same promoters in the microscopy setup. As shown in Fig. 3.7, the estimated CV^2 are much closer to the results obtained with the microscopy measurements and the difference no longer systematically depends on the mean expression level. In addition, whereas the CV^2 of the raw FCM measurements show little correlation with the CV^2 of the microscopy measurements, after correcting for autofluorescence and shot noise there is a much better agreement between the CV^2 as measured by the FCM and microscopy (Fig. 3.8).

3.9. Estimating mean and variance of GFP concentration

As shown in Fig. 3.3, microscopy measurements show a strong correlation between the size of the cells and total GFP of the cells, indicating that cell size variations

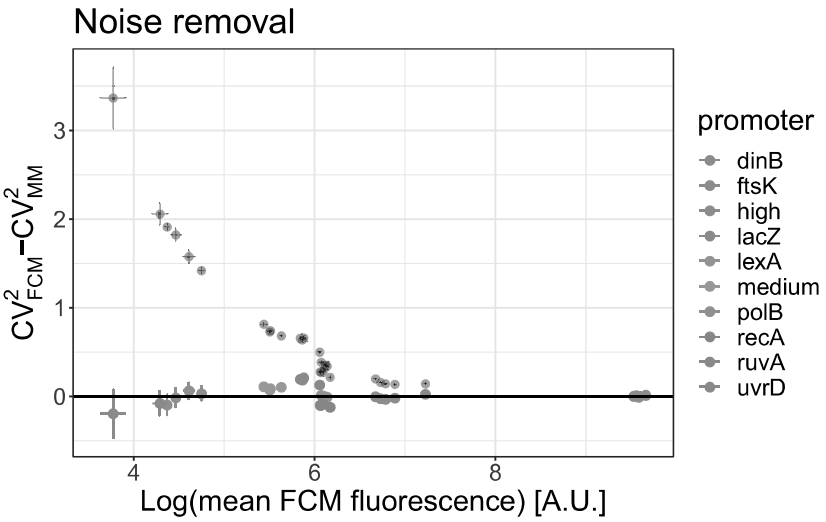

Figure 3.7.: **Comparison of CV^2 from FCM and microscope measurements after correcting for autofluorescence and shot noise.** Absolute difference of the CV^2 of different transcriptional reporters of native and synthetic *E. coli* promoters as estimated from FCM and microscope measurements. The black transparent dots use uncorrected FCM measurements and reproduce Fig. 3.6 in linear scale, while the colored dots are obtained when using the CV^2 that are corrected for the FCM shot noise.

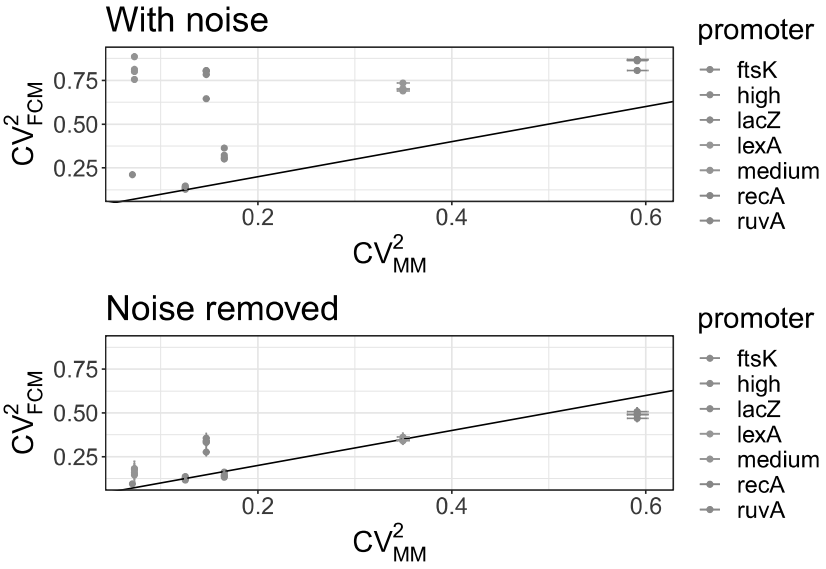

Figure 3.8.: **Correlation of CV^2 in the FCM and microscope measurements before and after correcting for autofluorescence and shot noise.** *Top*: The CV^2 of the raw FCM fluorescence measurements is consistently higher than the CV^2 of fluorescence in the microscope measurements, and there is little correlation between the two. *Bottom*: Once the FCM measurements are corrected for autofluorescence and shot noise, there is now a good agreement between the CV^2 as estimated by FCM and microscopy. Measurements for different promoters are indicated by different colors (see legend) and different points of the same color represent replicate FCM measurements. Only promoters expressing more than $\exp(4)$ above the background are shown and the black line in both plots is a line with slope 1 and intercept 0.

are responsible for a large fraction of the variation in total GFP, and that GFP *concentration* fluctuates significantly less than total GFP. It would thus be desirable to be able to estimate the mean and variance of GFP concentrations from the FCM measurements as well. However, the fact there is a much weaker correlation between raw fluorescence and scatter measurements in FCM (Fig. 3.3) suggests that it may be difficult to accurately estimate GFP concentrations for single cells. In particular, to estimate the GFP concentration of a single cell, we need to not only take the autofluorescence and shot noise of the fluorescence measurement into account, we also need to quantify how the cell's volume relates to the forward- and side-scatter measurement, which is known to be very challenging.

3.9.1. Scattering signals are non-linear functions of cell size

The extent to which forward- and side-scatter measurements of FCMs can be used to estimate the size of the measured object is a topic of considerable debate in the flow cytometry literature. It is generally assumed that forward scatter mostly reflects cell size, and that side scatter reflects surface properties such as granularity [53]. Several previous studies have established that FCM can be successfully used to distinguish bacteria of different shapes and sizes [40–43], i.e. the average scattering of a population of cells reflects the average size of the cells in the population.

To confirm that, also within our setup, the average size of a population of cells can be inferred from averages of scatter measurements, we made use of flow cytometry measurements from a recent study from our lab in which *E. coli* cells were grown in a number of different conditions and cell sizes were measured using microscopy in each condition [31]. Notably, the growth-rate of the cells varied considerably across these conditions and *E. coli* cells are known to increase size with growth-rate. For each condition, we calculated both the average cell size from the microscopy measurements as well as the average height and width of both forward- and side-scatter.

As shown in Figure 3.9, we found a very good correlation between forward-scatter and cell size in each condition, confirming results from previous studies that average scatter can indeed be used to estimate average cell size. However, it should be noted that the observed relationship between cell size and scatter is highly non-linear. That is, whereas the height of the forward-scatter grows approximately quadratically with cell area, the width of the forward-scatter grows approximately as area to the power $1/3$. Previous studies indicate that the mathematical relationship between cell size and scattering signal can be highly dependent on the specific experimental setup and is often at odds with predictions of mathematical theories of light scattering [8, 36, 37]. In [54] it is further shown that even if a particular non-linear relation between scattering and single-cell size can be established in a given setting, this relationship is not universal and it can vary even for bacteria of similar sizes and geometric properties. Thus, although we could here make use of the microscopy cell size measurements to calibrate the non-linear relationship between forward-scatter and cell size, it is highly doubtful that this relationship would apply in other settings.

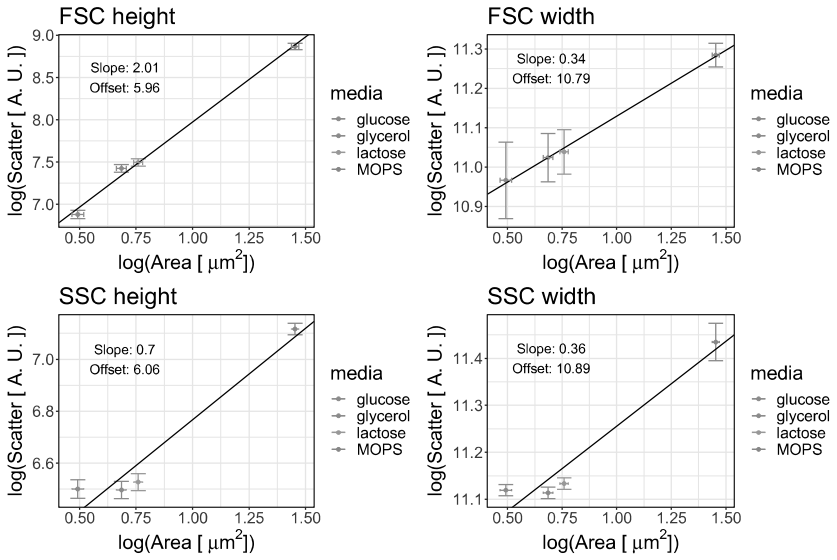

Figure 3.9.: **Average forward- and side-scatter of cells show approximate power-law dependence on average cell size.** Each panels shows the average of the logarithm of one of the four scattering signals, i.e. height or width of either forward- (FSC) or side-scatter (SSC), as a function of the average logarithm of cell area for *E. coli* cells growing in different media (M9 + glucose, glycerol or lactose; MOPS + glucose, see legend) as measured by microscopy [31]. The error bars represent the standard errors of the mean over replicate experiments.

3.9.2. Scattering signals contain a substantial shot noise component

Moreover, in order to be able to estimate GFP concentrations in individual cells, we have to go beyond relating population averages of scatter and size, and estimate sizes of individual cells from the scattering measurements. Several previous studies have reported that it is difficult to use individual scattering measurements to measure variations of the sizes of single cells in a homogeneous population [35, 44–46]. To investigate this within our setup we focused on height of the forward scattering, since based on Fig. 3.9 this signal most strongly correlates with cell size, calculated the CV^2 of the scattering as a function of the average scatter, and compared this with the CV^2 in cell area as a function of average cell area, as measured by microscope (Fig. 3.10).

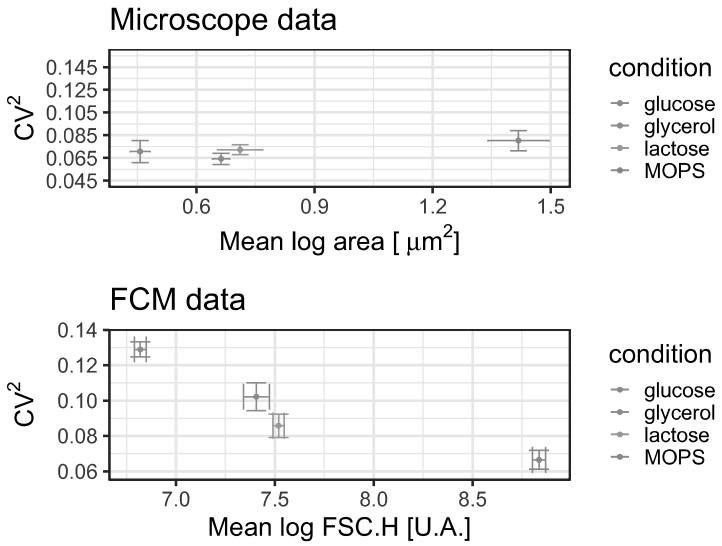

Figure 3.10.: **The CV^2 of the scattering distribution is affected by shot noise.** *Top panel*: The CV^2 of the cell areas as a function of mean cell area across growth conditions, as measured by microscopy. *Bottom panel*: The CV^2 of the height of the forward scattering signal as function of the mean height of forward scattering across growth conditions, as measured by FCM. In both panels the colors corresponds to different growth media as indicated in the legend (M9 + glucose, M9 + glycerol or M9 + lactose, and MOPS + glucose).

We see that, whereas the microscopy measurements indicate that the CV^2 in cell size is roughly equal in all conditions, the FCM measurements show a clear decrease of CV^2 with mean, similarly to what was observed for the fluorescence signal. As the scatter signal is generated by converting a light signal into an electrical impulse, it is to be expected that scattering measurements are also affected by shot noise, and the results in Fig. 3.10 confirm that this is the case. Thus, in order to estimate the

variation in cell sizes from the forward-scatter signals, we not only have to take into account the non-linear relationship between scattering and size, but also the shot noise on the scattering measurements. However, in contrast to the situation with the fluorescence measurements, where we used the calibration beads to estimate the shot noise, we cannot use these beads for estimating the shot noise on the scattering measurements since these are strongly influenced by the geometry and material of the particles. Therefore, the relationship between size and scatter will likely be very different for the beads than for living cells.

In summary, both the complex non-linear relationship between scattering measurements and size, and the absence of a general procedure for estimating the size of the shot noise in the scattering measurements, make it very difficult to estimate the true variability of cell sizes using FCM measurements only. Consequently, we currently do not see a simple way for using FCM measurements to directly measure the GFP concentrations in individual cells.

3.9.3. FCM measurements can be used to quantify the relative sizes of variation in GFP concentrations of different genes

Although we do not believe that, absent of calibration with an independent measurement technology such as microscopy, it is possible to reliably estimate the true sizes of single cells using forward- and side-scattering measurements, FCM measurements can still be used to learn a great deal about the relative noise levels of different genes. Indeed, as confirmed in Fig. 3.7, provided that autoflurescence and shot noise are taken into account, the CV^2 of total fluoresence levels of different promoters can be estimated reasonably accurately from FCM fluorescence measurements. Given that each of these fluorescent promoter reporter constructs are embedded in identical cells growing in the same environment, these cells will all exhibit the *same* variation in cell sizes, so that the differences in CV^2 in total fluorescence must reflect differences in the CV^2 of the GFP concentrations for these reporter constructs.

Without loss of generality, the total GFP intensity I of a cell can be written as the product $I = C \cdot V$ of GFP concentration C and cell volume V, and we can additionally write C as the average concentration $\langle C \rangle$ plus a deviation δ_C, and similarly for volume:

$$I = (\langle C \rangle + \delta_C)\,(\langle V \rangle + \delta_V)\,, \tag{3.9}$$

where both δ_C and δ_V have average zero.

From the microscopy measurements we know that the fluctuations in the GFP concentration C are approximately independent of fluctuations in cell volume V (Suppl. Fig. F.3). Using this, we can derive relationships between both the means and coefficients of variation of the total GFP I, and the concentration C and volume V, respectively. We find

$$\langle I \rangle = \langle C \rangle \langle V \rangle \tag{3.10a}$$

$$CV_I^2 = CV_C^2 + CV_V^2 + CV_C^2 CV_V^2. \tag{3.10b}$$

We can use this to rewrite the coefficient of variation of concentration, in terms of the coefficient of variation of total GFP (which we have shown how to estimate) and the (unknown) coefficient of variation in cell size, i.e.

$$CV_C^2 = \frac{CV_I^2}{1 + CV_V^2} - \frac{CV_V^2}{1 + CV_V^2}. \tag{3.11}$$

Thus, if the coefficient of variation of cell volume CV_V^2 in the growth condition of interest can be estimated using independent measurements, then equation (3.11) can be used to estimate the coefficient of variation of concentration in terms of the CV_I^2 for total GFP, as given by equations (3.8). Importantly, since the CV_V^2 is the same for all reporter constructs, such a measurement would only have to be done once.

Lastly, even if the CV_V^2 is not known, we note that it will be the same for each of the promoter reporter constructs. Therefore, the difference dCV_C^2 of the coefficients of variation in GFP for two promoters is directly proportional to the difference dCV_I^2 in coefficient of variation of total GFP, i.e.

$$dCV_C^2 = \frac{dCV_I^2}{1 + CV_V^2}. \tag{3.12}$$

Although this still depends on the CV_V^2, for all conditions we tested we found that $CV_V^2 \ll 1$, so that a reasonable estimate of the relative size of variation in concentrations is given by simply setting $CV_V^2 = 0$ in the above equation.

4. Discussion

Although flow cytometry is an attractive technology for single-cell analysis of gene expression in high-throughput, we have shown that for data from bacterial cells there are a number of challenges to overcome in data analysis in order to obtain accurate quantification. We here developed a number of procedures for measuring single-cell expression distributions in bacteria using FCM data and implemented them in an R package called *E-Flow*.

We first analyzed the forward- and side-scatter signals and their correlation structure. There seems to be little agreement in the literature as to when to use forward-scatter or side-scatter and whether to use height, width or area. We showed that only width and height provide independent measurements and developed a Bayesian mixture model for separating viable cell measurements from debris and other outliers using the full 4-dimensional distribution of forward- and side-scatter measurements. In general the filter we developed is much broader than the very strict gating strategies that are sometimes used and typically only a small fraction of the events are discarded.

We next developed a Bayesian mixture model to estimate the mean and variance in single-cell fluorescences of a population of cells carrying a fluorescent reporter. However, by comparing of the means and variances estimated by FCM with the means and variances estimated from microscopy measurements of the same strains growing in the same conditions, we observed systematic differences because of two effects. First, the amount of autofluorescence per cell differs systematically between FCM and microscopy and we developed methods for estimating and removing the autofluorescence from the FCM measurements. We show that, after correcting for autofluorescence, there is a perfect agreement between the means of different reporters as estimated by FCM and microscopy, over the entire range of expression levels. However, FCM measurements systematically overestimate the variation in fluorescence levels due to shot noise in the FCM measurement. We developed a method to correct for the contribution of shot noise to the estimated variation that uses calibration beads to estimate the size of the FCM shot noise. We showed that, only after correcting for shot noise do gene expression noise measurements from the FCM converge to those obtained from microscopy measurements. Although the precise size of the shot noise and autofluorescence will likely vary between different flow cytometers, the methods we presented here are general, can be applied to data from any flow cytometer, and provide a step-by-step procedure for both estimating the size of autofluorescence and shot noise, and correcting for these components.

Finally, we investigated whether FCM can be used to directly measure the distribution of GFP concentration across cells by using forward- and side-scatter measurements to estimate the volumes of individual cells. In line with previous work, we show that because scattering measurements depend on cell size in a complex non-linear manner and contain a shot noise component that is difficult to calibrate, it is not possible to accurately estimate the fluctuations in volumes of single cells from scattering measurements. However, because GFP concentration and cell size fluctuate independently across cells, we showed that the relative sizes of GFP fluctuations for different reporter constructs can still be estimated from the variation of total GFP with reasonable accuracy.

Supplementary

Supplementary Material A.

Flow cytometers signals and their statistical properties

A.1. Signal acquisition

In the flow cytometer, each cell is made to pass through a set of laser beams in order to measure its scattering profile and its fluorescence intensity at several wavelengths. The scatter has two components, forward- and side-scatter, which are usually interpreted as follows [55]:

1. Forward-scatter (FSC) generally reflects the size of the cells or particles.
2. Side-scatter (SSC) generally reflects the internal complexity or granularity of the cells or particles.

The scatter and fluorescence of each event (typically a cell) is converted by a photomultiplier tube into a pulse of electrical signal that is characterized by a height and a duration (Fig. 3.1). Roughly speaking, a pulse begins when a particle enters the laser beam and it reaches its maximum when the particle reaches the middle of the beam. In order to represent the pulse, the signal is sent to an analog-to-digital converter (ADC) that discretizes the continuous pulse into a digital one. For the machine used in this study, the ADC does this by sampling the pulse every 0.1 μsec and the height of the signal is mapped to an integer between 0 and 2^{14}, i.e. 14 bits are used to quantify the height of the signal [51].

A.2. Correlations among the signals

The cytometer gives three statistics of a measured pulse (height, width, and area), and we investigated correlations between these statistics. For the machine used in our study, we find that the area is in fact exactly proportional to the product of the height and width of the signal (Suppl. Fig. A.1).

Consequently, of the three statistics reported, only two are independent and we next measured the correlations between all pairs of statistics to determine which of these are most independent. As shown in Suppl. Fig. A.2 we can see that area correlates significantly with both height and width whereas height and width do not show significant correlation. We thus decided to characterize each signal by height and width.

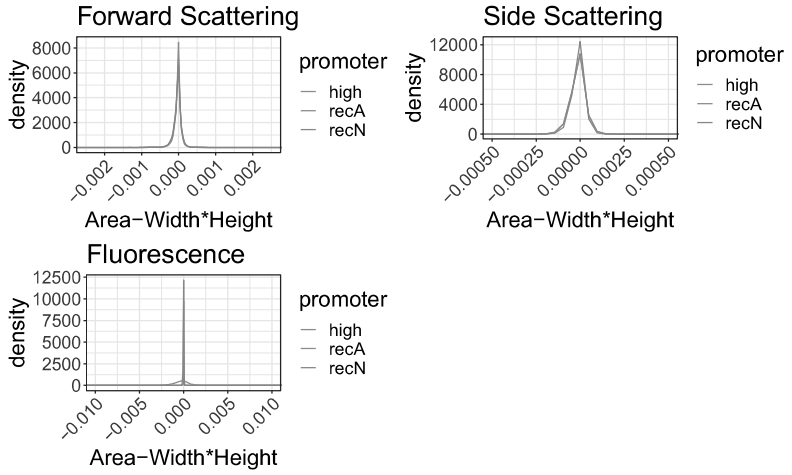

Figure A.1.: **Pulse area is proportional to the product of height and width.** Apart from very rarely occurring negative or saturated signals, the reported area of each pulse is equal to a constant C times the product of the reported height and width. The histograms show the distribution of the difference area $- C \cdot \text{width} \cdot \text{height}$, which is highly concentrated around zero. Different colors represent different promoters.

A.3. Signal width is not informative for fluorescence measurements

To assess the relative importance of the height and width statistics for fluorescence measurements we made use of the calibration beads manufactured by BD, which come in a set of three different expression levels. Supplementary Fig. A.3 shows that, whereas the heights of the signal pulses show three main peaks, corresponding to the three intensities of the beads, the distribution of measured widths is almost identical for all beads. Consequently, we chose to only use signal height to quantify fluorescence.

A.4. Anomalous behavior at low fluorescence intensities

As can already be observed from the measurements of the calibration beads (Suppl. Fig. A.3), due to shot noise in the fluorescence measurements, the coefficient of variation generally increases as the fluorescence level decreases. This general negative correlation between mean fluorescence and its coefficient of variation can not only be observed for the calibration beads, but also for a library of transcriptional reporters of native *E. coli* promoters (Suppl. Fig. A.4). However, as Suppl. Fig. A.4 also shows, when fluorescence levels become very low, the CV^2 of the measurements reaches a maximum and starts to *decrease* as the mean decreases further, and the lowest CV^2 is

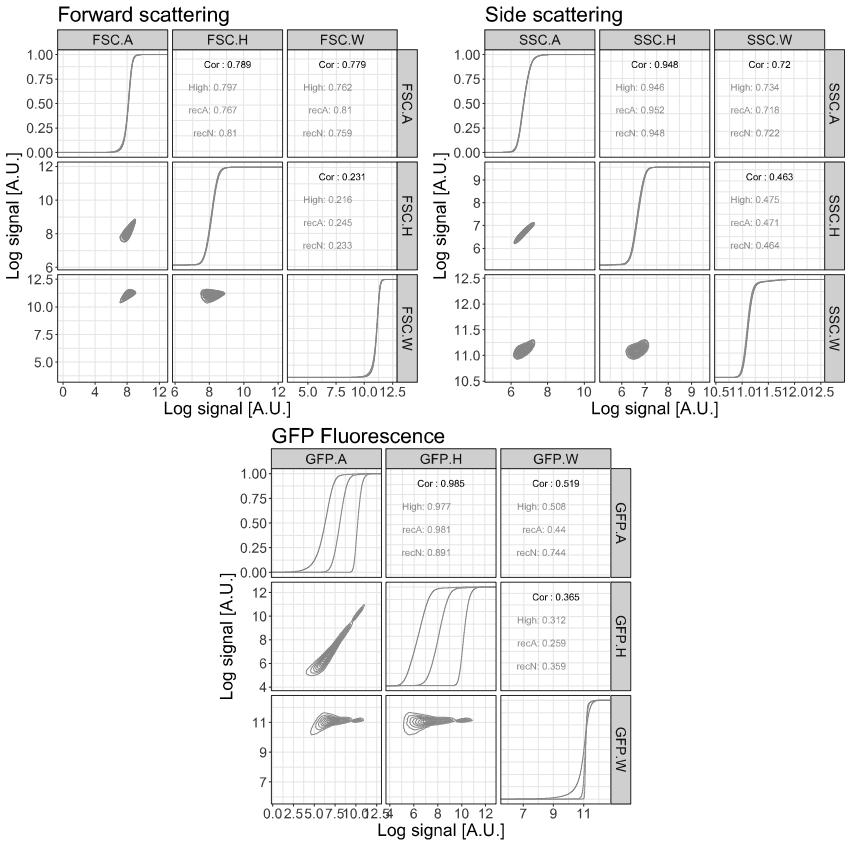

Figure A.2.: **Correlations among the different statistics reported for each signal pulse.** The plots show the correlations among the height (H), area (A) and width (W) of the signal pulses for the forward- (FSC) and side-scatter (SSC) as well as the fluorescence measurements (GFP). The colors represent 3 different promoters, two native promoters (recA and recN) and one synthetic promoters (high).

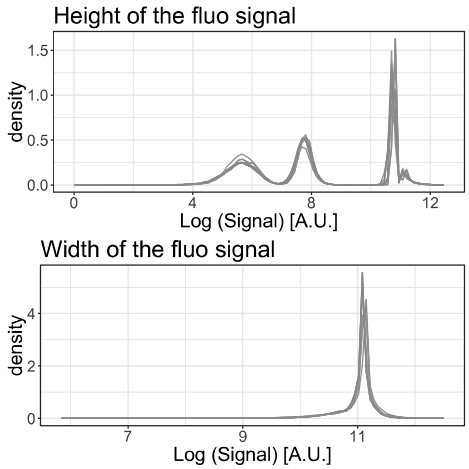

Figure A.3.: **Population separation of calibration beads based on their fluorescence signal pulses.** Distributions of the measured heights (top) and widths (bottom) of the fluorescence signal pulses for populations of artificial beads. Each color corresponds to one measurement run of a set of artificial beads consisting of a mixture of beads of three different fluorescent intensities. We see that, whereas signal height shows 3 clearly separated peaks, the distribution of signal widths is narrowly peaked at a single value. Note also the notably wider variance of the peaks in fluorescence height at low intensities than at high intensities, which results from measurement shot noise.

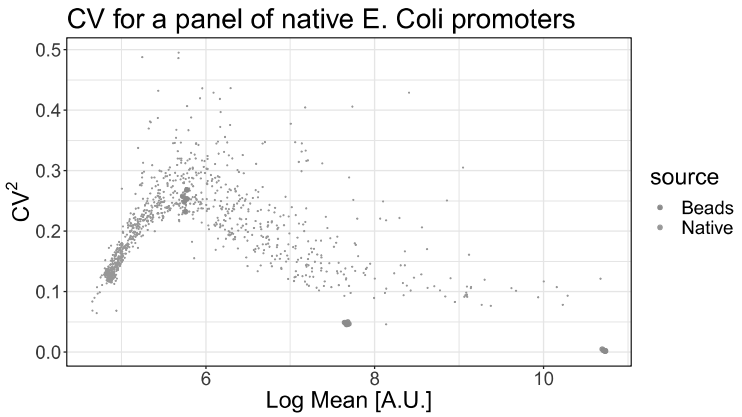

Figure A.4.: **Coefficient of variation as a function of mean fluorescence level exhibits anomalous behavior at low fluorescence.** The blue scatter shows the measured coefficient of variation squared CV^2 (vertical axis) as a function of the logarithm of mean fluorescence level (horizontal axis) for a library of transcriptional reporters of native *E. coli* promoters [48] as obtained in a recent study from our lab [31], with each blue point corresponding to one promoter of the library. The red points show the same statistics for the calibration beads at three intensities. Note that, due to shot noise in the fluorescence measurements, the CV^2 increases as the mean fluorescence decreases, including for the calibration beads (that have low intrinsic variance). However, at very low fluorescence, the CV^2 reaches a maximum and decreases for even lower fluorescence.

observed for cells without any GFP expression. We hypothesize that this decrease in CV^2 for very low fluorescence levels derives from the fact that autofluorescence levels vary less than GFP fluorescence levels, in combination with some specific signal processing that is performed by the machine at low signal intensities. Although we invested a significant amount of time in trying to reverse engineer what signal processing may cause this anomalous behavior at very low fluorescence levels, these attempts were ultimately not successful. In addition, our repeated requests to BD to bring us into contact with the relevant technicians went unanswered. Consequently, we were forced to restrict our quantitative analysis to cells whose GFP levels were at least twice the autofluorescence level (corresponding the log-mean levels roughly equal to those of the dimmest calibration beads in Suppl. Fig. A.4).

Supplementary Material B.

Mixture model of the scatter measurements

The scatter of each measured event is characterized by 4 statistics: the height and width of both the forward- and side-scatter. We fit the set of 4D scatter measurements for a given dataset by a mixture of a multivariate Gaussian and a uniform distribution. The idea is that the Gaussian distribution models the viable cells, while the uniform distribution represents outliers that may correspond to fragments of dead cells or other debris. Concretely, we assume that the probability to observe the 4D vector $\vec{y}$ for a single measured event is given by

$$P(\vec{y} \,|\, \vec{\mu}, \Sigma, \rho) = \frac{\rho}{\sqrt{(2\pi)^4 |\Sigma|}} e^{-\frac{1}{2}(\vec{y}-\vec{\mu})^T \Sigma^{-1} (\vec{y}-\vec{\mu})} + \frac{1-\rho}{\Delta}, \tag{B.1}$$

where $\vec{\mu}$ is the center of the 4D multivariate Gaussian, Σ its covariance matrix, ρ the fraction of viable cell measurements in the dataset, and Δ is the volume of the 4D hypercube spanned by all the data points. Note that $|\Sigma|$ denotes the determinant of the covariance matrix. The *E-Flow* package that we distribute contains a C++ implementation where the model (B.1) is fit to a given dataset using expectation maximization. An example of the results of this fitting to a dataset of *E. coli* cells grown in minimal media with lactose is shown in Fig. 3.2 of the main paper, showing that the observed 4D scatter can be well approximated by the mixture model.

Once the mixture model is fitted, we can calculate a posterior probability p_i for each observation i that it derives from the Gaussian mixture component, i.e. that it is a viable cell measurement. By default *E-Flow* filters out all events i for which

this posterior probability is less than 0.5 and retains all measurements with posterior probability larger than 0.5, but users can choose to alter this threshold in posterior probability. Supplementary Figure B.1 shows the same 4D scatter of measurements as shown in Fig. 3.2 of the main text, but now with all selected events in red, and all events that were filtered out in black.

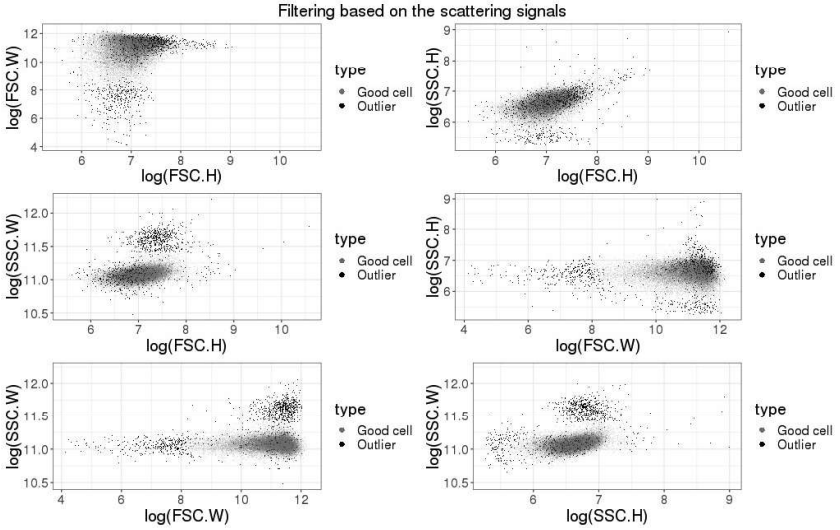

Figure B.1.: **Filtering of events based on their scattering profiles.** The panels show a scatter of the forward- and side-scatter measurements for a set of *E. coli* cells grown in minimal media with lactose. Events that have posterior probability larger than 0.5 to stem from the Gaussian component of the mixture are indicated in red, and events with posterior less than 0.5 are superimposed in black. The red events are considered viable cell measurements and the black events are discarded. Note that around 4% of the measured events are discarded for this dataset.

To investigate the effect of the filtering strategy on the observed fluorescence levels we calculated the distribution of fluorescence levels of several reporters when using either a very strict threshold, retaining only cells with posterior $p > 1 - e^{-10}$ or a very lenient threshold, retaining all cells with $p > e^{-10}$. As shown in Suppl. Fig. B.2 there is virtually no difference between the observed distribution of fluorescence with the two thresholds. This observation strongly suggests that it is not possible to effectively pick out a subpopulation of cells of similar size by strict gating on forward- and side-scatter.

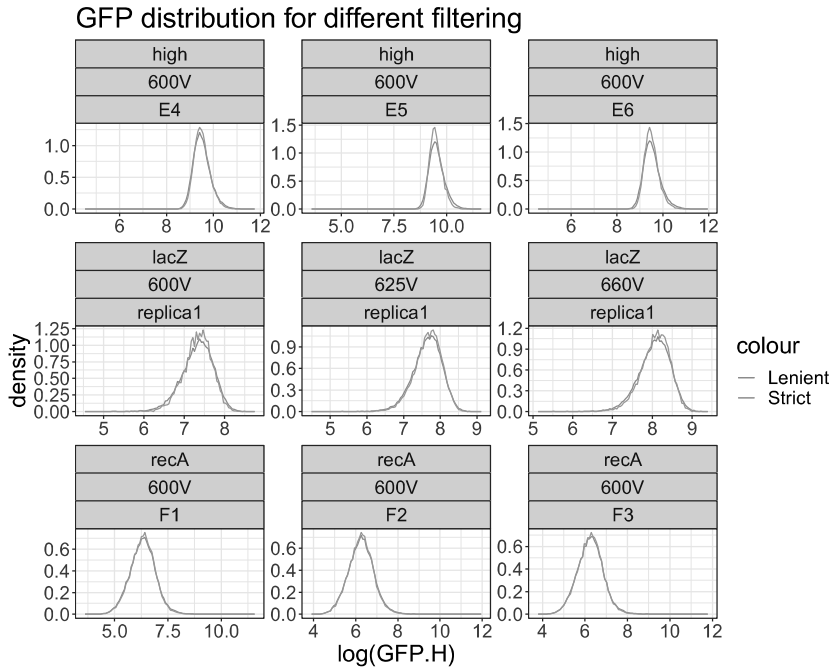

Figure B.2.: **Fluorescence distribution when using very strict or very lenient gating of forward- and side-scatter.** Each panel shows the observed distribution of fluorescence levels of a transcriptional reporter (rows) using different replicate measurements (columns) when using either a very strict threshold $p > 1 - e^{-10}$ or a very lenient threshold $p > e^{-10}$ for filtering events. Note that the name of the reporter and the voltage of the flow cytometer's laser used in a given replicate are indicated at the top of each panel.

Supplementary Material C.

Inferring the mean and variance of the fluorescence levels

After filtering events based on forward- and side-scatter, we next fit the distribution of fluorescence by a mixture of a log-normal distribution, which captures valid measurements and corresponds to the bulk of the measurements, and a uniform distribution capturing outliers. In particular, we assume that the probability for a fluorescence measurement to have log-fluorescence x is given by the mixture:

$$P(x|\rho,\mu,v) = \frac{\rho}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right] + \frac{1-\rho}{\Delta}, \tag{C.1}$$

where μ and v are the mean and variance of log-fluorescence levels, ρ is the fraction of 'valid' measurements and $\Delta = x_{\max} - x_{\min}$ is the observed range of log-fluorescence levels. The log-likelihood for a dataset of n measurements x_i is then simply given by

$$L(\rho,\mu,v) = \sum_i \log\left[P(x_i|\rho,\mu,v)\right]. \tag{C.2}$$

We maximize the log-likelihood (C.2) using expectation maximization to obtain the fitted parameters ρ_*, μ_* and v_*. To obtain error bars σ_μ and σ_v on the mean μ and variance v we obtain the Hessian matrix of the log-likelihood by expanding to second order around its maximum and we take the diagonal elements of its inverse. For any given set of fluorescence measurements, our package *E-Flow* returns both the fitted values (μ_*, v_*) and their error bars (σ_μ, σ_v).

Below we develop methods for correcting the observed fluorescence levels for both cell autofluorescence and for the shot noise in the FCM measurements. In order to do this we also need the mean and variance of fluorescence levels, rather than just the mean and variance of *log*-fluorescence levels. That is, if we write for the fluorescence of a cell $f = e^x$, we need to obtain the mean and variance of f. Using the well-known expressions of the mean and variance of a log-normal distribution we have

$$\langle f \rangle = e^{\mu_* + v_*/2}, \tag{C.3}$$

and

$$\mathrm{var}(f) = \left[e^{v_*} - 1\right] e^{2\mu_* + v_*}. \tag{C.4}$$

Further, given that the error bars on the mean and variance are generally small compared to their means, we use a linear approximation to calculate error bars on the mean and variance of f, and find

$$\sigma_{\langle f \rangle} = \langle f \rangle \sqrt{\sigma_\mu^2 + \frac{\sigma_v^2}{4}}, \tag{C.5}$$

and

$$\sigma_{\mathrm{var}(f)} = \sqrt{4\mathrm{var}(f)^2\sigma_\mu^2 + \left(2\mathrm{var}(f) + \langle f\rangle^2\right)^2 \sigma_v^2}. \tag{C.6}$$

The *E-Flow* package also returns these estimates and error bars for fluorescence levels f.

Supplementary Material D.

Averaging of replicate autofluorescence measurements

As shown in Fig. 3.4 in the main text, we measured the autofluorescence distributions of cells that do not express GFP in triplicate on 4 separate days, and for two different strains. Noticing that there appears to be a small but systematic variation in both the means and variances of the autofluorescence levels, we proceeded as follows to estimate an overall average for the mean and variance of autofluorescence, adapting a method first presented in [56].

We assume that there is an overall mean autofluorescence μ and that on any given day d, the mean autofluorescence μ_d on that day deviates from μ by some amount δ_d, i.e.

$$\mu_d = \mu + \delta_d. \tag{D.1}$$

We will assume that the deviation δ_d varies by an (unknown) amount τ across days, following a Gaussian distribution. That is, the probability of having mean autofluorescence μ_d on any given day is

$$P(\mu_d|\mu,\tau) = \frac{1}{\sqrt{2\pi}\tau}\exp\left[-\frac{(\mu_d-\mu)^2}{2\tau^2}\right]. \tag{D.2}$$

Let μ_{id} be the estimated mean autofluorescence of replicate i on day d, and let σ_{id} be the associated error-bar on this mean. Assuming that the true mean autofluorescence on day d was μ_d, the probability to obtain a measured mean autofluorescence μ_{id} is also given by a Gaussian distribution:

$$P(\mu_{id}|\sigma_{id},\mu_d) = \frac{1}{\sqrt{2\pi}\sigma_{id}}\exp\left[-\frac{(\mu_{id}-\mu_d)^2}{2\sigma_{id}^2}\right]. \tag{D.3}$$

The probability of the data D, i.e. all replicate measurements μ_{id} given an overall mean μ and variance across days τ^2 is given by taking the product over all measurements

and days, and marginalizing over all day-dependent means μ_d:

$$P(D|\mu,\tau) = \prod_d \left[\int d\mu_d P(\mu_d|\mu,\tau) \prod_i P(\mu_{id}|\sigma_{id},\mu_d) \right]. \tag{D.4}$$

These integrals can all be performed analytically and we obtain

$$P(D|\mu,\tau) = \prod_d \frac{1}{\sqrt{2\pi(\tau^2+\sigma_d^2)}} \exp\left[-\frac{(\bar{\mu}_d-\mu)^2}{2(\tau^2+\sigma_d^2)} \right], \tag{D.5}$$

where we have defined the measured mean $\bar{\mu}_d$ on day d as

$$\bar{\mu}_d = \frac{\sum_i \mu_{id}/\sigma_{id}^2}{\sum_i 1/\sigma_{id}^2}, \tag{D.6}$$

and the squared error σ_d^2 on the measured mean on day d as

$$\sigma_d^2 = \left[\sum_i \frac{1}{\sigma_{id}^2} \right]^{-1}. \tag{D.7}$$

If we define the weighted overall average

$$\bar{\mu} = \frac{\sum_d \bar{\mu}_d/(\tau^2+\sigma_d^2)}{\sum_d 1/(\tau^2+\sigma_d^2)}, \tag{D.8}$$

and the overall squared error

$$\sigma^2 = \left[\sum_d \frac{1}{\tau^2+\sigma_d^2} \right]^{-1}, \tag{D.9}$$

then we can rewrite equation (D.5) as

$$P(D|\mu,\tau) = \left[\prod_d \frac{1}{\sqrt{2\pi(\tau^2+\sigma_d^2)}} \right] \exp\left[-\frac{(\mu-\bar{\mu})^2}{2\sigma^2} + \frac{\bar{\mu}^2}{2\sigma^2} - \sum_d \frac{\bar{\mu}_d^2}{2(\tau^2+\sigma_d^2)} \right]. \tag{D.10}$$

We can then finally marginalize over μ to obtain a likelihood that depends only on τ and the measurements:

$$P(D|\tau) = \sigma \left[\prod_d \frac{1}{\sqrt{2\pi(\tau^2+\sigma_d^2)}} \right] \exp\left[\frac{\bar{\mu}^2}{2\sigma^2} - \sum_d \frac{\bar{\mu}_d^2}{2(\tau^2+\sigma_d^2)} \right]. \tag{D.11}$$

We maximize equation (D.11) with respect to τ to find the optimal value τ_*. We then substitute this value τ_* into equations (D.8) and (D.9) to obtain a final estimate $\bar{\mu}$

of the overall mean autofluorescence and an error bar σ on this estimate. The same procedure is used to estimate average variances $\bar{v}_d$ for each day and an overall average variance $\bar{v}$. The *E-Flow* package returns all these values as the overall estimates of autofluorescence averaged over multiple replicates and days.

Supplementary Material E.

Fitting the shot noise strength using calibration beads

The electronic cascade in the photomultiplier tube introduces a noise whose variance can be described as a Gaussian with a variance proportional to the square root of the incoming signal [52]. This means that, for every measurement of fluorescence, the relationship between measured intensity I_M and true intensity I_T is given by

$$I_M = I_T + O + \epsilon\sqrt{I_T}, \tag{E.1}$$

with $\epsilon \sim \mathcal{N}(0, \delta)$ and O is an offset introduced by the electronics in order to avoid negative values when I_T is very small.

To infer the strength δ of the shot noise term we used a set of calibration beads consisting of a mixture of beads of three different expression levels. The top right panel of Suppl. Fig. E.1 shows the observed distribution of expression levels for one set of beads showing three main peaks and one small additional peak at very high intensity. We believe that this highest peak is an artifact due to aggregation of multiple beads. We fit the distribution of bead intensities to a mixture of a uniform distribution (to account for outliers) plus four Gaussians. We discard all points assigned to the uniform distribution and the highest Gaussian component, and estimated the means, variances, and CV^2 for each of the remaining three Gaussians. This procedure was repeated for multiple datasets. In total we estimated means, variances, and CV^2 for 31 datasets of calibration beads. The top left panel of Suppl. Fig E.1 shows the CV^2 as a function of the mean intensities of the beads across all datasets showing that CV^2 drops approximately inversely with mean expression, as expected for shot noise.

Given one dataset of beads expressions, we can now fit a model based on Eq. (E.1). First of all it can be proved that under this model the mean and variance of the expression intensities of one single peak i are given by

$$M_M^i = M_T^i + O \tag{E.2}$$

$$S_M^i = C^2(M_T^i)^2 + \delta^2 M_T^i \tag{E.3}$$

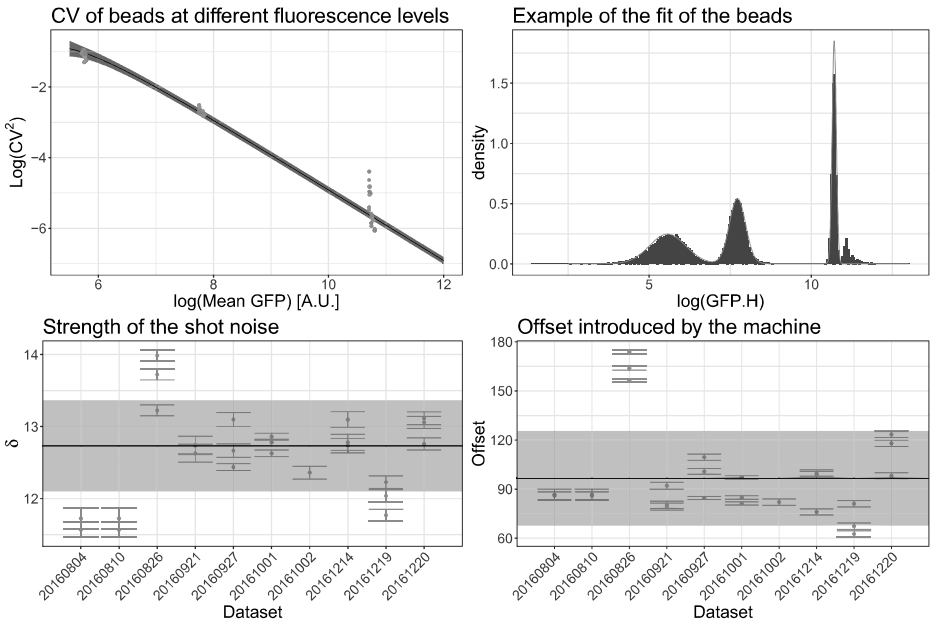

Figure E.1.: **Estimation of the shot noise strength using calibration beads.** Top left: CV^2 as a function of mean intensity for each Gaussian component of each mixture of beads analyzed. Both axes are shown on a log-scale and different colors refer to different datasets. The black line and the shaded region represent the fit obtained using a shot noise strength $\delta = 12.7 \pm 0.6$. Top right: Example of the intensity distribution of a single dataset of beads; the red line shows the fit with a mixture of three Gaussians and one uniform. Bottom: Estimates of the strength δ (left) and offset O (right) for different datasets; the estimation is done fitting a model of the form given by equations (E.6) and (E.7). The black lines correspond to the means of δ and O and the gray regions correspond to their variation among datasets.

where we have supposed that the true coefficient of variation C is the same for every peak (we assume that each of the 3 types of beads has the same manufacturing accuracy in their fluorescence intensities). From the histogram in Fig. E.1 (top right) it is clear that in logarithmic space the intensities are well described by Gaussian distributions, which means that M_M^i and S_M^i are the mean and variance of a log-normal distribution. We can work out the μ and σ of a normal distribution in log-space that would have M_M^i and S_M^i as mean and variance in real space

$$s_m^i = \log\left[1 + \left(C\frac{M_T^i}{M_T^i + O}\right)^2 + \delta^2\frac{M_T^i}{(M_T^i + O)^2}\right] \tag{E.4}$$

$$\mu_m^i = \log\left(M_T^i + O\right) - \frac{s_m^i}{2} \tag{E.5}$$

It follows that in logarithmic space the probability of observing a data point x_k (including also a uniform distribution that takes care of possible outliers) is

$$\begin{aligned} P(x_k \mid \vec{M}_T, O, C, \delta, \vec{\rho}) &= \sum_{i=1}^{3} \rho_i \mathcal{N}[x_k;\, \mu_m^i(M_T^i, O, C, \delta), s_m^i(M_T^i, O, C, \delta)] + \\ &+ \frac{1 - \sum_{i=1}^{3}\rho_i}{\Delta} \end{aligned} \tag{E.6}$$

where $\vec{\rho}$ is a vector of weights whose elements sum to one and Δ is the range of the data in logarithmic space. The log-likelihood of the data is

$$\mathcal{L}(\vec{x} \mid \vec{M}_T, O, C, \delta, \vec{\rho}) = \sum_{k=1}^{N} \log\left[P\left(x_k \mid \vec{M}_T, O, C, \delta, \vec{\rho}\right)\right] \tag{E.7}$$

This likelihood is maximized using a coordinate descent algorithm and the error bar on δ was estimated by expanding the log-likelihood to second order around its maximum, keeping all other parameters fixed. The same procedure was used to estimate an error bar on the offset O.

After obtaining estimates of δ and O and their error bars for each dataset, we averaged the δs and Os from different datasets in the same way as we did for autofluorescence measurements above. The bottom part of Suppl. Fig. E.1 shows the estimates of the shot noise and offset for different datasets and the final average of all these estimates. We find that δ has a value of 12.7 with a variation of 0.6 among datasets. Finally, the black line on the top left panel of Suppl. Fig. E.1 shows that the model correctly describes the observed decrease in CV^2 with mean.

Supplementary Material F.

Mother machine statistics

F.1. Age distribution in the Mother machine

Especially for highly expressed genes the GFP concentration varies relatively little across cells, so that the total amount of GFP in a cell correlates well with its size. Since each cell expands exponentially along its division cycle in exponential growth, cell size correlates well with the 'age' of a cell, i.e. where it is in its division cycle. For this reason we expect a correlation between the total GFP distribution and the age distribution of the population, i.e. the relative frequencies of cells in different stages of their division cycle. Thus, in order to meaningfully compare GFP distributions from the FCM and from cells growing in a microfluidic device, we must check that the age distributions of the two populations are indeed comparable.

The cell population analyzed in the FCM is expected to have an over-representation of younger cells, since for every old cell that divides two young cells are produced whereas, in the microfluidic device, we expect less over-representation of young cells due to the fact that cells continuously leave the growth channel.

To infer the population structure of a growing population of cells, we use the Leslie approach [57]. We discretize the time in steps of 3 minutes, corresponding to the acquisition time of the time lapse microscopy, and we look for the probability $P(a,t)$ of having cells of age a at time t. The Leslie theory is based on two quantities, the fecundity $f(x,t)$ which represents the expected contribution from an individual aged x at time point t to the population aged 1 at time point $t+1$, and the probability $P(x,t)$ of survival over one time point for individuals who are present in age-class x at time t. These quantities form the Leslie matrix, which is defined as

$$L(t) = \begin{pmatrix} f(1,t) & P(1,t) & 0 & \dots & 0 \\ f(2,t) & 0 & P(2,t) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(d-1,t) & 0 & 0 & \dots & P(d-1,t) \\ f(d,t) & 0 & 0 & \dots & 0 \end{pmatrix} \tag{F.1}$$

this matrix is a squared $d \times d$ matrix where d is an upper limit on the age of the cells, after which $P(x,t) = 0$. The population structure at time t is given by a vector $\vec{n}(t)$ of dimension d, where the entry i is the number of individual in the population of age i. The theory states that

$$\vec{n}(t) = \vec{n}(0) \cdot L^t \tag{F.2}$$

where we use the convention that vectors are represented by rows. If the matrix L is diagonalizable and its entries are independent of time, it can be shown that for very

large t the population reaches a stable age distribution with

$$\vec{n}(t) \sim \lambda_1^t \vec{n}(0) \cdot \vec{U}_1 \vec{U}_1^{-1} \tag{F.3}$$

$$n_{tot} \propto e^{t \log(\lambda_1)} \tag{F.4}$$

where λ_1 and $\vec{U}_1$ are the leading eigenvalues and eigenvectors of L.

From the data from the microfluidic device we can estimate the probability for a cell of age x at time t to survive one more time point by computing the probability that the division occurs at an age larger than x; the fecundity is then given as twice the probability that the cell doesn't survive. Suppl. Fig. F.1 shows the computed $P(x,t)$ by considering cells at different times t from the beginning of the experiment and it shows that they are all the same, suggesting that the Leslie matrix is indeed independent of time. We thus decided to compute a single curve for $P(x)$ by pooling all the cells together independently of the time at which they were measured. The resulting curve shows that the survival probability becomes negligible after 150 minutes, with a medium age at division of 80 minutes.

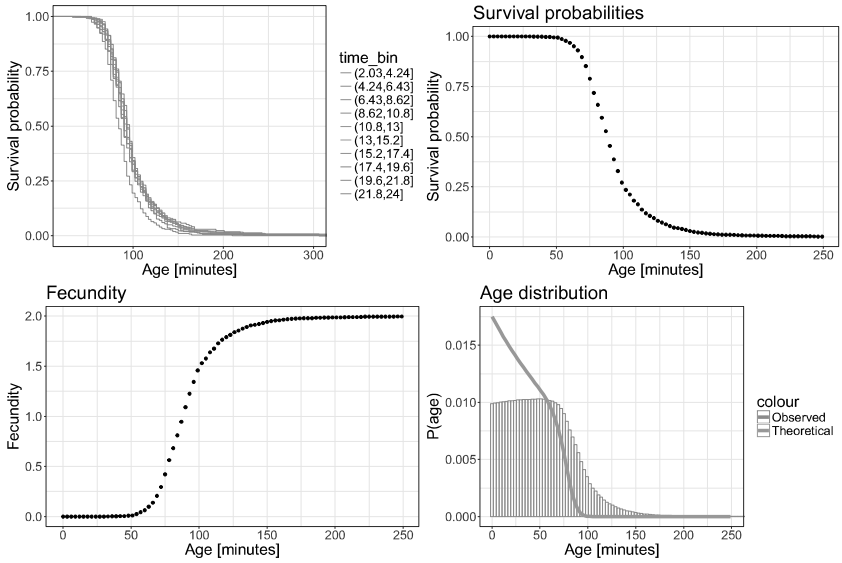

Figure F.1.: **The Leslie model.** *Top left*: The one-time point survival probabilities don't depend on the time and they show that the survival probability becomes negligible for cells older than about 150 minutes. The medium age at division is around 80 minutes. *Top right*: Since the probabilities don't depend on time, we pooled data from all times together to obtain a single survival curve. *Bottom left*: Fecundity as function of cell age. *Bottom right* Asymptotic age distribution of the population for large times as observed (red) and predicted by the theory (blue). Note that the observed distribution has an over-representation of old cells relative to the theoretically predicted distribution.

We found that the exponential growth rate of the population, Eq (F.4), is $\rho \sim$

$0.0089/min$, which is the same as the measured single cell growth rate. As expected, the population age structure differs markedly between the observed and the theoretical one, as shown on the bottom right part of Fig F.1. In particular the theoretical distribution predicts more young cells than old, while in the microfluidic device we have a nearly uniform age distribution with a right tail; hence, as expected, young cells are relatively under-represented in the population in the microfluidic device.

Assuming that the GFP concentration is constant, the total GFP must depend on the cell age through the length and in computing the mean and variances we have to weight the GFP by the ratio between the theoretical and the observed age fractions. Specifically, if $P_{obs}(\text{age})$ and $P_{th}(\text{age})$ are the observed and the theoretical age distributions, then for the $\log(GFP)$ distribution it holds

$$\bar{g} = \frac{\sum_{i=1}^{N} g_i\, \omega(\text{age}_i)}{\sum_{i=1}^{N} \omega(\text{age}_i)} \tag{F.5}$$

$$\text{Var}\, g = \frac{\sum_{i=1}^{N} (g_i - \bar{g})^2\, \omega(\text{age}_i)}{\sum_{i=1}^{N} \omega(\text{age}_i)} \tag{F.6}$$

where i runs over all cell observations, $g \equiv \log(G)$ and $\omega(\text{age})$ is a weighting factor that accounts for the over-representation of older cells

$$\omega(\text{age}) = \frac{P_{th}(\text{age})}{P_{obs}(\text{age})} \tag{F.7}$$

Going into log-space, we have estimates more robust against outliers and we can approximate the CV^2 in the real space as the variance in the log-space. Suppl. Fig. F.2 shows that the age structure correction makes the CVs only slightly lower than the one measured ignoring the age structure effects.

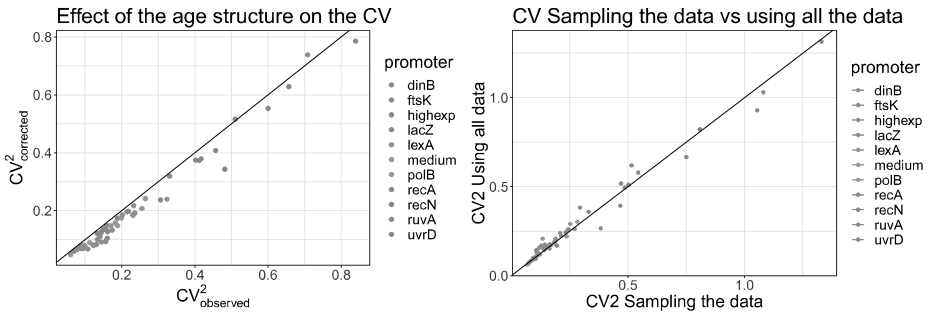

Figure F.2.: **Effect of the age structure on the measured CV^2.** *Left*: CV^2 corrected by the age structure in the microfluidic device versus the CV^2 as directly measured, ignoring the age structure effect. The age structure has the effect of making the measured CV^2 slightly higher than the CV^2 corrected by the age structure of the population. *Right*: CV^2 obtained by taking all the measures in the microfluidic device versus the CV^2 when taking only one point, chosen at random, per cell. We can see that the difference is very small and the points accumulate around the bisector.

F.2. Correlation in the measurements

As we are recording the size and GFP expression of cells every 3 minutes in the microfluidic device, observations from nearby time points are clearly correlated, and it could be suspected that this may affect the statistics we calculate. In contrast, in the FCM, we have cells measured at only one time point, so we don't have correlations coming from the fact that we have multiple data points coming from the same cell. To check whether the presence of these correlated data points can substantially bias the statistics, we measured the CV^2 for different promoters and media on different dates in two different ways:

1. Taking all the data, regardless of the fact that they may come from the same cell.
2. Taking only one randomly selected data point for each cell.

The first condition is the one that we use throughout the paper to compute the statistics, while the second one should be closer to the FCM setup. Suppl. Fig. F.2 shows that the CV^2 are almost identical whether correlated time points are included or not. In conclusion, neither the age structure nor the correlated measurements affect the comparison of CV^2 between measurements from the FCM and from the microfluidic device.

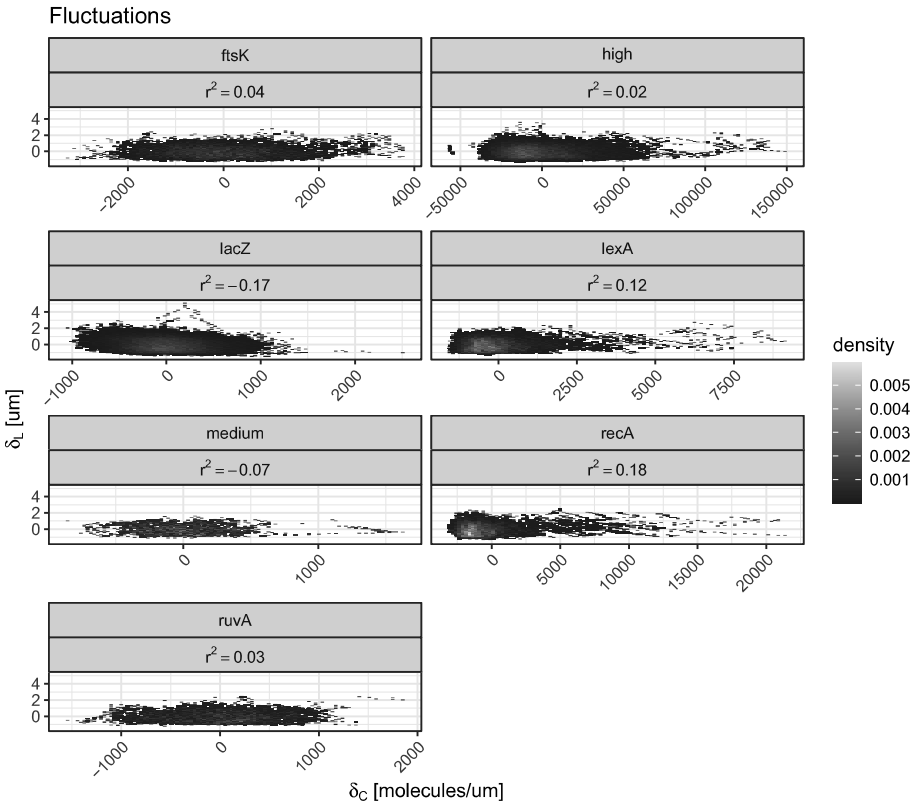

Figure F.3.: **Fluctuations in size are independent on fluctuations on concentration.** The fluctuations in length ($\delta_L = L - \langle L \rangle$) are independent of the fluctuations in GFP concentration (δ_C). Each subpanel is a different promoter measured in the mother machine setup and the squared Pearson correlation coefficient is shown under the promoter name.

Bibliography

[1] M. B. Elowitz et al. "Stochastic gene expression in a single cell." In: *Science* 297.5584 (Aug. 2002), pp. 1183–1186 (cit. on p. 1).

[2] Mukund Thattai and Alexander van Oudenaarden. "Stochastic Gene Expression in Fluctuating Environments." In: *Genetics* 167.1 (2004), pp. 523–530. ISSN: 0016-6731. DOI: 10.1534/genetics.167.1.523. eprint: https://www.genetics.org/content/167/1/523.full.pdf. URL: https://www.genetics.org/content/167/1/523 (cit. on p. 1).

[3] Nitzan Rosenfeld et al. "Gene Regulation at the Single-Cell Level." In: *Science* 307.5717 (2005), pp. 1962–1965. ISSN: 0036-8075. DOI: 10.1126/science.1106914. eprint: https://science.sciencemag.org/content/307/5717/1962.full.pdf. URL: https://science.sciencemag.org/content/307/5717/1962 (cit. on p. 1).

[4] Jonathan M. Raser and Erin K. O'Shea. "Noise in Gene Expression: Origins, Consequences, and Control." In: *Science* 309.5743 (2005), pp. 2010–2013. ISSN: 0036-8075. DOI: 10.1126/science.1105891. eprint: https://science.sciencemag.org/content/309/5743/2010.full.pdf. URL: https://science.sciencemag.org/content/309/5743/2010 (cit. on p. 1).

[5] James C. W. Locke and Michael B. Elowitz. "Using movies to analyse gene circuit dynamics in single cells." In: *Nature Reviews Microbiology* (2009). DOI: 10.1038/nrmicro2056. URL: https://doi.org/10.1038/nrmicro2056 (cit. on p. 2).

[6] Nathalie Q. Balaban et al. "Bacterial Persistence as a Phenotypic Switch." In: *Science* 305.5690 (2004), pp. 1622–1625. ISSN: 0036-8075. DOI: 10.1126/science.1099390. eprint: https://science.sciencemag.org/content/305/5690/1622.full.pdf. URL: https://science.sciencemag.org/content/305/5690/1622 (cit. on p. 2).

[7] Mary E Lidstrom and Michael C Konopka. "The role of physiological heterogeneity in microbial population behavior." In: *Nature Chemical Biology* (2010). DOI: 10.1038/nchembio.436. URL: https://doi.org/10.1038/nchembio.436 (cit. on p. 2).

[8] H M Davey and D B Kell. "Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses." In: *Microbiology and Molecular Biology Reviews* 60.4 (1996), pp. 641–696. ISSN: 0146-0749. eprint: https://mmbr.asm.org/content/60/4/641.full.pdf. URL: https://mmbr.asm.org/content/60/4/641 (cit. on pp. 2, 19).

[9] Michael K. Winson and Hazel M. Davey. "Flow Cytometric Analysis of Microorganisms." In: *Methods* 21.3 (2000), pp. 231–240. ISSN: 1046-2023. DOI: https://doi.org/10.1006/meth.2000.1003. URL: http://www.sciencedirect.com/science/article/pii/S104620230091003X (cit. on p. 2).

[10] J. D. Chung et al. "Gene expression in single cells of Bacillus subtilis: evidence that a threshold mechanism controls the initiation of sporulation." In: *Journal of bacteriology* 176 (7 1994). DOI: 10.1128/jb.176.7.1977-1984.1994 (cit. on p. 2).

[11] R. H. Valdivia and S. Falkow. "Bacterial genetics by flow cytometry: rapid isolation of Salmonella typhimurium acid-inducible promoters by differential fluorescence induction." In: *Mol. Microbiol.* 22.2 (Oct. 1996), pp. 367–378 (cit. on p. 2).

[12] R. L. Wilson et al. "Identification of Listeria monocytogenes in vivo-induced genes by fluorescence-activated cell sorting." In: *Infect. Immun.* 69.8 (Aug. 2001), pp. 5016–5024 (cit. on p. 2).

[13] E. M. Ozbudak et al. "Regulation of noise in the expression of a single gene." In: *Nat. Genet.* 31.1 (May 2002), pp. 69–73 (cit. on p. 2).

[14] K. Hakkila et al. "Monitoring promoter activity in a single bacterial cell by using green and red fluorescent proteins." In: *J. Microbiol. Methods* 54.1 (July 2003), pp. 75–79 (cit. on p. 2).

[15] Yanina Sevastsyanovich et al. "Exploitation of GFP fusion proteins and stress avoidance as a generic strategy for the production of high-quality recombinant proteins." In: *FEMS Microbiology Letters* 299.1 (2009), pp. 86–94. DOI: 10.1111/j.1574-6968.2009.01738.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-6968.2009.01738.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-6968.2009.01738.x (cit. on p. 2).

[16] H. Miao et al. "Dual fluorescence system for flow cytometric analysis of Escherichia coli transcriptional response in multi-species context." In: *J. Microbiol. Methods* 76.2 (Feb. 2009), pp. 109–119 (cit. on p. 2).

[17] J. B. Kinney et al. "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence." In: *Proc. Natl. Acad. Sci. U.S.A.* 107.20 (May 2010), pp. 9158–9163 (cit. on p. 2).

[18] R. Anand, N. Rai, and M. Thattai. "Promoter reliability in modular transcriptional networks." In: *Meth. Enzymol.* 497 (2011), pp. 31–49 (cit. on p. 2).

[19] Olin K. Silander et al. "A Genome-Wide Analysis of Promoter-Mediated Phenotypic Noise in Escherichia coli." In: *PLOS Genetics* 8.1 (Jan. 2012), pp. 1–13. DOI: 10.1371/journal.pgen.1002443. URL: https://doi.org/10.1371/journal.pgen.1002443 (cit. on pp. 2, 9).

[20] D. Madar et al. "Promoter activity dynamics in the lag phase of Escherichia coli." In: *BMC Syst Biol* 7 (Dec. 2013), p. 136 (cit. on p. 2).

[21] M. A. Sanchez-Romero and J. Casadesus. "Contribution of phenotypic heterogeneity to adaptive antibiotic resistance." In: *Proc. Natl. Acad. Sci. U.S.A.* 111.1 (Jan. 2014), pp. 355–360 (cit. on p. 2).

[22] M. Utratna et al. "Effects of growth phase and temperature on IfB activity within a Listeria monocytogenes population: evidence for RsbV-independent activation of IfB at refrigeration temperatures." In: *Biomed Res Int* 2014 (2014), p. 641647 (cit. on p. 2).

[23] Luise Wolf, Olin K Silander, and Erik van Nimwegen. "Expression noise facilitates the evolution of gene regulation." In: *eLife* 4 (June 2015). Ed. by Ido Golding, e05856. ISSN: 2050-084X. DOI: 10.7554/eLife.05856. URL: https://doi.org/10.7554/eLife.05856 (cit. on pp. 2, 4, 11).

[24] J. Baert et al. "Phenotypic variability in bioprocessing conditions can be tracked on the basis of on-line flow cytometry and fits to a scaling law." In: *Biotechnol J* 10.8 (Aug. 2015), pp. 1316–1325 (cit. on p. 2).

[25] Q. Yan and S. S. Fong. "Study of in vitro transcriptional binding effects and noise using constitutive promoters combined with UP element sequences in Escherichia coli." In: *J Biol Eng* 11 (2017), p. 33 (cit. on p. 2).

[26] N. Nordholt et al. "Effects of growth rate and promoter activity on single-cell protein expression." In: *Sci Rep* 7.1 (July 2017), p. 6299 (cit. on p. 2).

[27] J. Rohlhill, N. R. Sandoval, and E. T. Papoutsakis. "Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol." In: *ACS Synth Biol* 6.8 (Aug. 2017), pp. 1584–1595 (cit. on p. 2).

[28] M. Razo-Mejia et al. "Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction." In: *Cell Syst* 6.4 (Apr. 2018), pp. 456–469 (cit. on p. 2).

[29] N. M. Belliveau et al. "Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria." In: *Proc. Natl. Acad. Sci. U.S.A.* 115.21 (May 2018), E4796–E4805 (cit. on p. 2).

[30] M. N. M. Bahrudeen et al. "Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry." In: *J. Microbiol. Methods* 166 (Nov. 2019), p. 105745 (cit. on p. 2).

[31] Arantxa Urchueguía et al. "Noise propagation shapes condition-dependent gene expression noise in Escherichia coli." In: *BioRxiv* (2019). DOI: 10.1101/795369 (cit. on pp. 2, 4, 5, 19, 20, 30).

[32] M. Acar, J. T. Mettetal, and A. van Oudenaarden. "Stochastic switching as a survival strategy in fluctuating environments." In: *Nat. Genet.* 40.4 (Apr. 2008), pp. 471–475 (cit. on p. 2).

[33] L. B. Carey et al. "Promoter sequence determines the relationship between expression level and noise." In: *PLoS Biol.* 11.4 (2013), e1001528 (cit. on p. 2).

[34] D.A Veal et al. "Fluorescence staining and flow cytometry for monitoring microbial cells." In: *Journal of Immunological Methods* 243.1 (2000). Flow Cytometry, pp. 191–210. ISSN: 0022-1759. DOI: https://doi.org/10.1016/S0022-1759(00)00234-9. URL: http://www.sciencedirect.com/science/article/pii/S0022175900002349 (cit. on p. 2).

[35] V. Ambriz-Avina, Jorge A. Contreras-Garduno, and Mario Pedraza-Reyes. "Applications of Flow Cytometry to Characterize Bacterial Physiological Responses." In: *J BioMed Research International* (2014). DOI: 10.1155/2014/461941. URL: http://dx.doi.org/10.1155/2014/461941 (cit. on pp. 2, 3, 21).

[36] Gerhard Nebe-von-Caron. "Standardization in microbial cytometry." In: *Cytometry Part A* 75A.2 (2009), pp. 86–89. DOI: 10.1002/cyto.a.20696. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.20696. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.20696 (cit. on pp. 2, 19).

[37] Susann Müller and Gerhard Nebe-von-Caron. "Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities." In: *FEMS Microbiology Reviews* 34.4 (2010), pp. 554–587. DOI: 10.1111/j.1574-6976.2010.00214.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-6976.2010.00214.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-6976.2010.00214.x (cit. on pp. 2, 19).

[38] Yuichi Taniguchi et al. "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells." In: *Science* 329.5991 (2010), pp. 533–538. ISSN: 0036-8075. DOI: 10.1126/science.1188308. eprint: https://science.sciencemag.org/content/329/5991/533.full.pdf. URL: https://science.sciencemag.org/content/329/5991/533 (cit. on pp. 2, 11).

[39] Lingling Yang et al. "Detection and Quantification of Bacterial Autofluorescence at the Single-Cell Level by a Laboratory-Built High-Sensitivity Flow Cytometer." In: *Analytical Chemistry* 84.3 (2012). PMID: 22243282, pp. 1526–1532. DOI: 10.1021/ac2031332. eprint: https://doi.org/10.1021/ac2031332. URL: https://doi.org/10.1021/ac2031332 (cit. on p. 2).

[40] T Akerlund, K Nordström, and R. Bernander. "Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of Escherichia coli." In: *J Bacteriol.* (1995). DOI: 10.1128/jb.177.23.6791-6797.1995 (cit. on pp. 3, 19).

[41] H B Steen and E Boye. "Escherichia coli growth studied by dual-parameter flow cytophotometry." In: *Journal of Bacteriology* 145.2 (1981), pp. 1091–1094. ISSN: 0021-9193. eprint: https://jb.asm.org/content/145/2/1091.full.pdf. URL: https://jb.asm.org/content/145/2/1091 (cit. on pp. 3, 19).

[42] Ramunas Stepanauskas et al. "Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles." In: *Nature Communications* (2017). DOI: 10.1038/s41467-017-00128-z. URL: https://doi.org/10.1038/s41467-017-00128-z (cit. on pp. 3, 19).

[43] Benjamin Volkmer and Matthias Heinemann. "Condition-Dependent Cell Volume and Concentration of Escherichia coli to Facilitate Data Conversion for Systems Biology Modeling." In: *PLOS ONE* 6.7 (July 2011), pp. 1–6. DOI: 10.1371/journal.pone.0023126. URL: https://doi.org/10.1371/journal.pone.0023126 (cit. on pp. 3, 19).

[44] Henrik Christensen, Lars R. Bakken, and Rolf A. Olsen. "Soil bacterial DNA and biovolume profiles measured by flow-cytometry." In: *FEMS Microbiology Ecology* 11.3-4 (Apr. 1993), pp. 129–140. ISSN: 0168-6496. DOI: 10.1111/j.1574-6968.1993.tb05804.x. eprint: http://oup.prod.sis.lan/femsec/article-pdf/11/3-4/129/18115266/11-3-4-129.pdf. URL: https://doi.org/10.1111/j.1574-6968.1993.tb05804.x (cit. on pp. 3, 21).

[45] J. Vives-Rego, R. López-Amorós, and J. Comas. "Flow cytometric narrow-angle light scatter and cell size during starvation of Escherichia coli in artificial sea water." In: *Letters in Applied Microbiology* 19.5 (1994), pp. 374–376. DOI: 10.1111/j.1472-765X.1994.tb00479.x. eprint: https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1472-765X.1994.tb00479.x. URL: https://sfamjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1472-765X.1994.tb00479.x (cit. on pp. 3, 21).

[46] R. López-Amorós et al. "Variations in flow cytometric forward scatter signals and cell size in batch cultures of Escherichia coli." In: *FEMS Microbiology Letters* 117.2 (Apr. 1994), pp. 225–229. ISSN: 0378-1097. DOI: 10.1111/j.1574-6968.1994.tb06769.x. eprint: http://oup.prod.sis.lan/femsle/article-pdf/117/2/225/19090855/117-2-225.pdf. URL: https://doi.org/10.1111/j.1574-6968.1994.tb06769.x (cit. on pp. 3, 21).

[47] M. Kaiser et al. "Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software." In: *Nature Communications* 9 (Jan. 2018) (cit. on pp. 4, 7, 9, 10).

[48] Alon Zaslaver et al. "A comprehensive library of fluorescent transcriptional reporters for Escherichia coli." In: *Nature Methods* 3 (Aug. 2006) (cit. on pp. 4, 30).

[49] Socorro Gama-Castro et al. "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond." In: *Nucleic Acids Research* 44.D1 (Nov. 2015), pp. D133–D143. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1156. eprint: http://oup.prod.sis.lan/nar/article-pdf/44/D1/D133/16661521/gkv1156.pdf. URL: https://doi.org/10.1093/nar/gkv1156 (cit. on p. 4).

[50] BD™. *Cytometer Setup and Tracking Beads*. http://www.bdbiosciences.com/ds/is/tds/23-9141.pdf (cit. on p. 5).

[51] B. Verwer. *BD FACSDiVa Option. White Paper*. http://www.bdbiosciences.com/ds/is/others/23-6579.pdf (cit. on pp. 7, 15, 26).

[52] E. H. Eberhardt. "Noise in photmultiplier tubes." In: *IEEE* (Apr. 1967) (cit. on pp. 15, 37).

[53] H. M. Shapiro. *Practical Flow Cytometry*. 34th ed. John Wiley and Sons, 2003 (cit. on p. 19).

[54] O. Julià, J. Comas, and J. Vives-Rego. "Second-order functions are the simplest correlations between flow cytometric light scatter and bacterial diameter." In: *Journal of Microbiological Methods* 40.1 (2000), pp. 57–61. ISSN: 0167-7012. DOI: https://doi.org/10.1016/S0167-7012(99)00132-3. URL: http://www.sciencedirect.com/science/article/pii/S0167701299001323 (cit. on p. 19).

[55] BD Biosciences. *BD FACSCanto II Flow Cytometer Reference Manual*. Part No. 640806 Rev. A. May 2006 (cit. on p. 26).

[56] Piotr J. Balwierz et al. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." In: *Genome Research* 24.5 (2014), pp. 869–884. DOI: 10.1101/gr.169508.113. eprint: http://genome.cshlp.org/content/24/5/869.full.pdf+html. URL: http://genome.cshlp.org/content/24/5/869.abstract (cit. on p. 35).

[57] B. Charlesworth. *Evolution in age-structured populations*. 2nd ed. Cambridge University Press, 1994 (cit. on p. 40).

4. Non-equilibrium dynamics in the single-cell regulation of the lexA operon

Non-equilibrium dynamics in the single-cell regulation of the lexA operon

Luca Galbusera[1], Gwendoline Bellement-Theroue[1], Thomas Julou[1,*], and Erik van Nimwegen[1,*]

1 Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Basel, Switzerland

∗ E-mails: thomas.julou@unibas.ch, erik.vannimwegen@unibas.ch

Abstract

To investigate how noise propagates from transcription factors to the target promoters at the single-cell level, we focused on the LexA regulon in *Escherichia coli*, which is activated in response to DNA damage. We selected a set of promoters that are known to be only regulated by LexA and we studied the in vivo dynamics of the gene expression in a microfluidic device, using time-lapse microscopy. We found that LexA target promoters in single cells show a different behavior compared to constitutive expressed promoters, and they are characterized by short bursts in transcription which increase in frequency when LexA is induced. Depending on the promoter, the signal-to-noise ratio of the bursts can be high or low.

We show that such a behavior cannot be explained by a simple equilibrium model where fluctuations in the LexA concentration can be ignored. We hypothesize, on the other hand, that these transcriptional bursts correspond to short transient periods of low LexA concentrations that are precipitated by individual DNA strand breaks. That is, we propose that, rather than reading out an effective time average concentration of LexA, induction of different LexA target promoters is a non trivial function of the frequency and amplitude of transient drops in LexA concentration.

We then discuss how the architecture determines the ability of a promoter to detect fluctuations in the regulator levels of different durations, and to integrate the past history in its expression level.

1. Introduction

Since the discovery of the *lac* operon and gene regulation in the seminal work of Jacob and Monod in 1961 [1], we know that gene regulation plays an important role in the ability of cells to respond and adapt to environmental cues [2, 3]. Although a lot of effort has since been devoted to experimentally and theoretically dissect gene regulation, the complex web of interactions in the regulatory networks poses a great challenge for the development of quantitative models. For this reason, a full, quantitative understanding of gene regulation still remains an open question in systems biology, despite more than fifty years of research.

One open question is the functional role of gene expression noise in shaping the behavior and evolution of regulatory networks [4]. In previous works in our lab, we showed that noise is fundamental for the accurate evolution of gene regulation from a state without regulation, and that a large fraction of the noise is shaped by noise propagation from the regulator to the target promoter [5, 6]. It is, however, still not clear what the origin of the propagation is. In one regime, the fluctuations in the regulator are on a different time scale compared to the gene expression reactions, and can therefore be ignored. In this scenario, the target promoter is blind to the fluctuations of the regulator, and the noise is solely determined by the stochastic switching between a repressed and an active state, caused by the binding and unbinding of the regulator. In a dynamic regime, the fluctuations in the level of the regulator are on the same time scale as the reactions leading to mRNA production, and they explicitly contribute to the noise of the target gene.

To dissect the origin of noise, we leveraged today's technological advances in single-cell tracking and analysis, which are providing us with an unprecedented level of detail allowing for new, fundamental insights about the underlying mechanisms of gene regulation [7] and noise propagation. For example, we are now able to track real-time gene expression at the single-cell level [8–14], going beyond the analysis of the mean behavior of a population of cells. This new level of information allows us to study the time dynamics of the full distribution of molecular species and is believed to be pivotal to advance our knowledge of the regulatory schemes [15–17]. Indeed, despite the success of batch studies in describing the average protein production in a population of cells, single-cell studies are starting to reveal that average measurements might not fully characterize network responses, and the time evolution of distribution of the molecular species in single-cells also carries valuable information to dissect the mechanisms of gene regulation [18–25] and its evolution [6].

In this study, we investigated the diversity in responses of different native promoters under control of the same transcription factor at the single-cell level. In particular, we focused on the SOS pathway in *Escherichia coli* [26, 27], which is activated in response to DNA damage. This regulatory network presents several characteristics that make

it well suited to study noise propagation. First, the induction mechanism has been dissected in several studies, resulting in a good characterization of the binding sites and the transcription factors. Second, the network is under strong selection pressure, since it is crucial for the survival of the cells. This means that its noise propagation has been shaped by natural evolution. Finally, the network can be easily induced in the lab using common DNA damaging agents.

Briefly, the SOS pathway comprises genes involved in the repair of DNA double-strand breaks (DSBs) and is regulated through the interaction between the repressor LexA and the molecule RecA, both of which are also under control of the network. In the absence of DSB, LexA is bound to the promoter of the SOS genes keeping them repressed. When a DSB occurs, RecA starts to filament around the single-strand DNA and induces the autocleavage of unbound LexA, which in turn leads to an induction of the network, including LexA and RecA themselves. Once the break is repaired, the RecA filaments disassemble, LexA is not cleaved anymore and the network returns to a repressed state [26–30]. For the SOS response, it has been shown that, despite the simplicity of the network, the presence of feedback loops and fluctuations in LexA concentration cause a complex dynamics, with a gene-specific temporal pattern characterized by waves of gene activity [16, 24, 29, 31–33]. These studies also show the importance of single-cell studies to unmask patterns that are otherwise hidden at the population level.

In the present work, we constantly induced the network to a specific level, by giving the cells different concentrations of the DNA damaging agent Ciprofloxacin (Cipro) [34]. The activity of a selection of promoters, known to be only regulated by LexA, were measured by placing them upstream a GFP coding sequence in plasmids [35]. We then studied the in vivo dynamics of gene regulation in a microfluidic device using time-lapse microscopy [36] and found that LexA target promoters in single cells show short bursts in transcription which increase in frequency as Cipro increases.

We show that a simple equilibrium model, where the fluctuations in the regulator are ignored, is not able to explain the observed bursts in transcription. On the other hand, we propose that the bursts in production are caused by transient drops in LexA concentration caused by a random DNA damage and happening at the same time scale of the binding and unbinding of the regulator. In particular, our data points to a kinetic model where, instead of reading an effective time-average concentration of the regulator, the network responds to rapid and stochastic fluctuations in the LexA concentration.

We finally investigated how the promoter architecture affects the response of the promoter to fluctuations in the regulator. We show that the unbinding rate of LexA determines the level of sensitivity of the promoter to the fluctuations of the regulator, while the concentration of free LexA in the cell sets a biological limit on the highest sensitivity a promoter can achieve.

2. Materials and Methods

2.1. Strains and growth conditions

We measured fluorescence in single cells for a set of *Escherichia coli* MG1655 strains that carry a transcriptional reporter inserted upstream of a gene for the fluorescence protein GFP-mut2 and expressed from a low copy number plasmid [35]. All the sequences were verified by Sanger sequencing. These reporters are either for promoters regulated by LexA only (dinB, ftsK, lexA, polB, recA, recN, ruvA, or uvrD) [37] or for synthetic promoters obtained by experimental evolution so as to express at levels corresponding to the median or to 97.5^{th} percentile of all native *E. coli* promoters [6]. Throughout the paper, we refer to these two synthetic promoters as Synthetic High and Synthetic Medium.

2.2. Time lapse microscopy

Mother Machine experiments were performed as described in [36], using M9 + 0.4% glucose (supplemented with $50\mu g$ / mL of kanamycin during the overnight preculture only). Data were acquired every 3 minutes for 30 hours, supplementing the growing media with 3 different concentrations of the antibiotic Ciprofloxacin (Sigma Aldrich ref.: 17850, CAS Number 85721-33-1). The first six hours have no antibiotic supplemented; from 6h to 18h and from 18h to 30h we supplemented respectively 1ng/mL and 2ng/mL of antibiotic. We refer to [36] for a detailed description of the Mother Machine design that allows to dynamically change the growing media in the microfluidic device.

2.3. Analysis of the data

The GFP number of molecules correlates strongly with the cell size, as shown in the Supplementary Fig 2. Therefore, in this study we focused on the GFP concentration in order to remove variations in gene expression caused by variations in cell size rather than by induction of the LexA pathway.

Moreover, in computing the average concentration levels we need to keep in mind that the measurements coming from the same cell trace are highly correlated. Therefore, the usual computation of the standard error, based on independent samples, gives underestimated error bars and we need to use a specific algorithm to infer the corrected standard errors. The algorithm is described in the Supplementary Material section A.1.

To account for day-to-day variability, in all the experiments we measured a constitutive high expresser [6] and we used its average concentration to normalize across different days. More details are provided in the Supplementary Material section A.2.

Finally, we noticed that due to the high illumination, cells required about 2 hours before adapting to the Mother Machine environment and we therefore discarded all measurements taken earlier than the first two hours (Supplementary Material Section A.3). Moreover, we discarded all cells that approached the exit channel of the Mother Machine before the division, as their measured length, and therefore their GFP concentration, cannot be reliably estimated.

3. Results

3.1. Genes belonging to the LexA regulon show different inductions at the population level

We studied the induction of the LexA regulatory network focusing on 8 promoters that are under the exclusive control of LexA and the sigma factor, according to RegulonDB [37]. In addition, we added two constitutively expressed synthetic promoters, identified as synthetic high and synthetic medium, which exhibit a mean expression level that is respectively at the 97.5^{th} and 50^{th} percentile of all *E. coli* promoters [6]. The network is induced using the DNA damaging antibiotic Ciprofloxacin in 3 different concentrations, which are much lower than the recommended EC_{50}, and do not cause major damages to the cells. Indeed, the growth rates are not affected by the presence of antibiotic (Supplementary Fig A.2B).

Our microfluidic setup enables controlled variation of the Cipro concentration over time, allowing direct observation of gene regulatory responses to different induction levels in single cells [36]. In particular, we monitored the cells over a period of 30 hours, starting with no antibiotic in the first 6 hours, switching to 1 ng/mL in the following 12 hours and finally to 2 ng/mL in the final 12 hours. In Fig 3.1A, we show time snapshots of the fluorescence level for a growth lane expressing GFP from the recA promoter, and a growth lane expressing GFP from the synthetic high promoter. We can see that, while the synthetic promoters express similarly in all three conditions, recA promoters increase their expression as the concentration of the DNA damaging agent increases. This is shown more quantitatively in Fig 3.1B, where we show the fold change of the mean fluorescence level over time for the synthetic high promoters and the LexA targets which we found to be induced in our setup. Importantly, the figure shows that different promoters are induced differently in response to the same regulatory network. Supplementary Fig 1 shows the basal GFP concentration (in the absence of Cipro), and the fold changes for all the LexA regulated promoters. The same figure shows the agreement between the behavior of recA and polB over two different days.

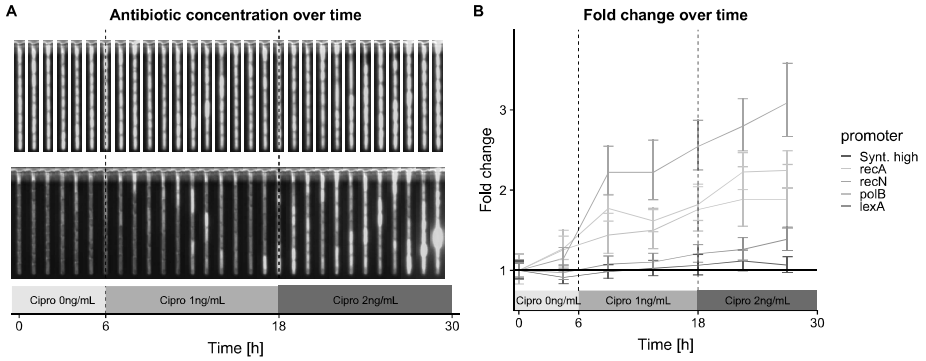

Figure 3.1.: **Induction of the genes under control of the LexA pathway.** A: Fluorescence time snapshots of two representative growth lanes of our microfluidic setup. The top lane includes cells expressing GFP from synthetic high promoter, while the bottom lane are cells expressing GFP from the recA promoter. The horizontal axis show the experiment time, with the three different Cipro concentrations. While the expression of the synthetic promoters is similar across all the Cipro concentrations, recA promoters express more as the Cipro increases. B: Fold change in the mean gene expression for a population of cells expressing 4 representative LexA target promoters, and the unregulated Synthetic promoters. It can be seen that different LexA targets are induced at different levels.

3.2. LexA regulated promoters have a single cell GFP production characterized by bursts

To better understand how different promoters respond to the induction of the same regulatory network, we turn our attention to the expression dynamics of single cells.

For each cell, the microfluidic setup enables us to track the total GFP intensity over time, as shown in the top panel of Fig 3.2A. To study how the regulatory network determines the expression of a promoter, we assume that the network affects the instantaneous production $q(t)$ of GFP, and we model the total GFP $G(t)$ at time t through the following process:

$$\frac{dG}{dt} = q(t)V(t) - \beta G(t) \tag{3.1}$$

where the term β represents the loss in fluorescence given by bleaching and dilution. Notice that the production is multiplied by the volume $V(t)$ of the cell, to factorize out any contribution to the production coming from the cell cycle (e.g. duplication of the plasmids), rather than from the regulatory network induction. Solving Eq (3.1) allows us to infer $q(t)$, as shown in the lower panel of Fig 3.2A. We refer to the Supplementary Material section A.3 for an in depth discussion of Eq (3.1) and how to solve it for $q(t)$.

Examples of time traces of GFP production over the entire experiment are shown in Fig 3.2B for three representative promoters. There is a striking difference in the time

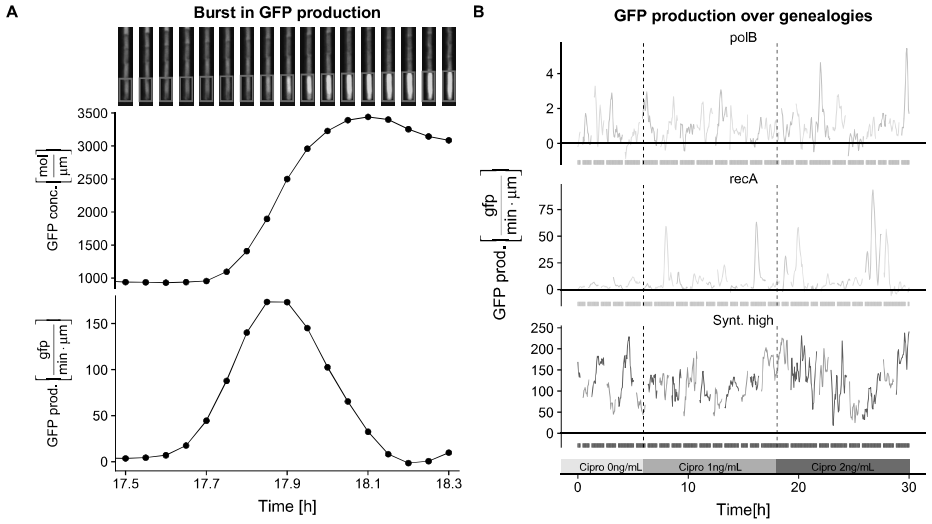

Figure 3.2.: **GFP production over time for single cells**. A: The top plot shows the GFP trace for a single cell expressing the promoter recA. The GFP production begins at time 18h and continues until about 18.3h. The bottom plot shows the corresponding GFP production, computed from Eq (3.1). The picture on top shows microscope snapshots of the tracked cell at the different acquisition times. B: Example of GFP production over the entire duration of the experiment for three promoters. For each promoter, only one representative cell genealogy is displayed, and the boundary between the cells of the genealogy are shown by the rectangles below each trace. The dotted vertical lines separate the three different Cipro concentrations.

traces for constitutively expressed and regulated promoters. The regulated promoters recA and polB are characterized by low production interspersed by periods of high production bursts, although for recA the bursts are clearer than for polB. On the other hand, the constitutive expressed synthetic promoter shows a pattern that is more similar to a random walk, with the GFP production fluctuating around a mean value.

3.3. Characterization of the bursts

To shed light on the mechanism of LexA induction, we set out to characterize the bursts in GFP production. We focused on lexA, recA, and recN, where the high signal to noise ratio allowed a peak calling algorithm (see Supplementary Material Section A.3) to reliable identify and characterize the bursts. The results are shown in Fig 3.3.

We noticed that the height of the bursts (Fig 3.3A) is characterized by an exponential distribution, whose slope depends on the specific promoter, but not on the concentration of Cipro. On the other hand, the duration, or width, of the bursts has a distribution that is not only independent on the concentration of antibiotic, but also very similar for all the promoters, with a median duration around 7.5 minutes

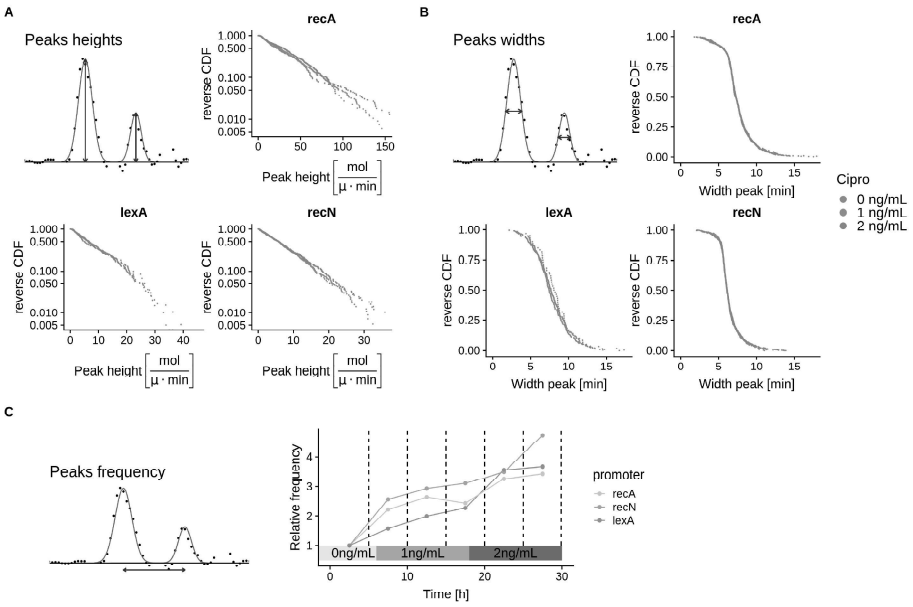

Figure 3.3.: **Characteristics of the bursts in GFP production**. A: Reverse cumulative distribution of the bursts height for different promoters and Cipro concentrations. The straight lines in a log scale shows that the distribution is exponential, with a slope determined by the promoter, but not by the antibiotic concentration. B: Reverse cumulative distribution for the bursts duration, showing that the distribution does not depend on the Cipro or on the promoter. C: Frequency of the bursts relative to the frequency without antibiotic, showing that the frequency increases with the antibiotic concentration in a similar way for all the promoters.

(Fig 3.3B). Finally, the frequency of the bursts, relative to the frequency without Cipro, increases with the concentration of antibiotic, in a similar way for the different promoters (Fig 3.3C).

All in all, Fig 3.3 shows that the induction of the regulatory network, caused by an increase in the antibiotic concentration, affects only the frequency of the bursts, while the promoter architecture determines the height of the bursts. The widths do not depend on either the Cipro concentration or the promoter architecture.

3.4. Gene expression model

In order to understand how a promoter responds to fluctuations in the regulator, we describe the gene expression using the reactions shown in Fig 3.4A.

The promoter can switch between a LexA-bound state and a LexA-free state, with the LexA unbinding rate determined by the architecture, and the binding rate given by the concentration of free LexA. When free from LexA, the promoter can be transcribed

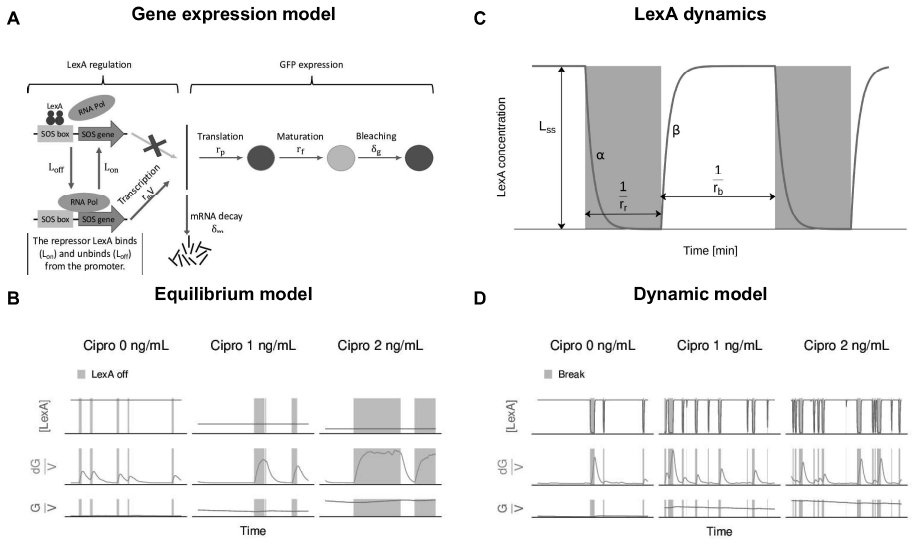

Figure 3.4.: **Theoretical models for the induction of the LexA regulatory network.** A: Model of gene expression for a promoter that can switch between a transcribing state (free from LexA) and a repressed state (bound by LexA). B: Simulation of the equilibrium model where the concentration of LexA is assumed to be constant over time and determined by the Cipro concentration (top row). The bursts correspond to periods of time when LexA is unbound from the promoter (middle row). As the concentration of LexA decreases, LexA rebinds slower, causing wider and higher bursts and leading to an increase in GFP concentration (bottom row). C: LexA dynamics in the presence of double strand breaks (DSB). Without DSB (white areas), LexA has a steady state concentration L_{ss}, which precipitates at a rate α when a DSB occurs (yellow areas). The break is repaired at a rate r_r, and the LexA concentration recovers with an exponential rate β. The DSBs occur with a frequency r_b. D: Simulation of the bursts in a non equilibrium model, where the fluctuations of LexA in panel C are explicitly taken into account. The bursts correspond to period of time when the LexA concentration is low. We refer to the Supplementary Section C.3 for a detailed explanations of the simulation algorithm.

into mRNA, which in turn can either decay or be translated into GFP. The GFP protein is initially immature and it undergoes some post-translational modification to become a mature, fluorescing protein [38]. The fluorescence of GFP decays very slowly over time, mainly due to photobleaching and dilution [36]. Notice that the transcription rate is proportional to the volume of the cell to take into account that as the cell grows the GFP carrying plasmid doubles.

All the reactions that lead to the expression of the promoter occur in a growing and dividing cell, according to an adder model [39]. When a cell has added a randomly selected volume, it divides and the daughter inherits all the molecular species according to a binomial sampling with probability 0.5.

To understand how LexA determines the induction of the gene expression, it is critical to understand how the concentration of free LexA depends on the different

concentrations of antibiotic. In the next sections, we explore two regimes. In an equilibrium model, the concentration of LexA at a given antibiotic level is assumed to be constant over time, and any fluctuation in the target expression are due to stochastic binding and unbinding of the repressor. On the other hand, in a dynamic model we explicitly take into account fluctuations in the concentration of the regulator. We will show that only a dynamic model can recapitulate the observed data.

3.5. An equilibrium model cannot explain the observed data

To explain the induction of a promoter under the model in Fig 3.4A, one of the approaches most frequently adopted is based on statistical equilibrium [40]. Starting from the seminal work of Ackers et al. [41], this strategy has been successfully applied to explain gene expression, also in presence of complex regulatory patterns [42–50]. A crucial assumption is that, at each Cipro concentration, the system is in equilibrium with the concentration of regulator. That is, we can ignore any fluctuation in the LexA level, and assume an effective constant concentration. This concentration is set by the amount of antibiotic, it is higher when there is no Cipro, a little bit lower when Cipro is present at a concentration of 1 ng/mL and even lower when the Cipro increases to 2 ng/mL (Fig 3.4B, top row) [40, 51]. The probability for the promoter to be in the active state (free from LexA) can then be computed using a Boltzmann distribution, and the total amount of protein is proportional to this probability [17, 51–53].

In this regime, the bursts in Fig 3.2B correspond to periods of time when LexA is unbound from the promoter. As the concentration of Cipro increases, the concentration of LexA decreases, lowering its binding rate and allowing the promoter to be free for longer periods of time. This causes wider and higher bursts (Fig 3.4B middle row), leading to the induction of the promoter (Fig 3.4B bottom row). That is, fluctuations in the expression of the target gene are caused by the stochastic switching of the promoter between an active and a repressed state, rather than from fluctuations in the level of LexA.

Notice that the induction is really controlled by only two parameters. The LexA unbinding rate, determined by the promoter architecture, sets how frequent a burst occurs. The LexA binding rate, determined by the concentration of free LexA, sets the duration and height of the bursts. Therefore, as the Cipro concentration increases, the model necessarily predicts longer and higher bursts, occurring at the same frequency. This is in clear contrast to the experimental data in Fig 3.3.

3.6. The SOS regulatory network responds to stochastic fluctuations in the regulator LexA

As an alternative to the previous model, we reject the equilibrium assumption, and explicitly model the fluctuations in LexA concentration. From the mechanism of

action of Cipro, we can assume that an increase in the concentration of antibiotic leads to an increase in the frequency of double-strand breaks (DSB) [54, 55], and from studies on the SOS response, we know that every time a DSB occurs there is a drop in functional LexA concentration, which is then restored when the break is repaired [26, 56] (Fig 3.4C).

Given this model, we can think of a scenario where the LexA concentration is normally so high that as soon as LexA falls off from the promoter, it immediately rebinds and the promoter cannot stay free enough time to produce a sensible amount of mRNA. However, a DSB causes a transiently low concentration of LexA, slowing down its rebinding and allowing the promoter to produce a burst of transcripts (Fig 3.4D). In this regime, the bursts correspond not to LexA unbinding events, but to transient drops in LexA concentration caused by DSBs. It is clear that the width and height of the bursts are determined by the length of time the LexA concentration is low, that is, by the repair time of the break, which is assumed to be independent of the Cipro concentration. On the other hand, the frequency of the breaks is determined by the frequency of the DSBs, which increases with the Cipro.

This model depends on several parameters (Tab 3.1), and it is important to verify that the observed behaviour can be recapitulated using values that are biologically relevant. Most of the parameters can be determined from the literature, and there are only four unknown rates, the LexA binding rate at steady state, the frequency of the breaks, the transcription rate, and the repair rate. However, a first approximation for the LexA binding rate can be estimated with the following argument. It is known that in the cells there are $1100 - 1300$ LexA dimers [28, 57], and we also know that a transcription factor needs $3 - 5$ minutes to find its binding site[58, 59]. Therefore, the LexA binding rate to the promoter can be estimated to be on the order of $400/\text{min}$.

The remaining three parameters (orange cells in Tab 3.1) need to be fitted from the data. In the Supplementary Material section C.3 we show that their values are constrained by the observed duration of the peaks and the exponential distribution of the heights. Briefly, the duration of the peaks is determined by different time scales, including the repair rate r_r of the break, the decay time δ_m of the mRNA, and the folding time r_f of the GFP. Since we know that the folding time is around 8 min [38, 60] and it is reasonable to assume that δ_m is at least $3 - 5$ minutes [11], they already make up a good fraction of the observed peak duration (Fig 3.3B). Therefore, the DSBs must be repaired quickly. Moreover, it can be shown that in the regime of quick breaks repair, the duration of the peaks are independent from the promoter (Supplementary Material E.1), and the heights are exponentially distributed (Supplementary Material section E.2), in agreement with the experimental data. The slope of the exponential distribution is set by the ratio of the transcription and repair rate, and the range of the repair rate can be bounded. Indeed, if the repair rate is faster than about $1/\text{min}$, the concentration of LexA does not have time to fall low enough to allow the promoter to be transcribed. On the other hand, if the repair rate is slower than $0.5/\text{min}$, the breaks durations become too long and we lose the exponential distribution of the heights (Supplementary Material section E.2 and E.3). We refer to the Supplementary Material section C.3 for a more detailed discussion.

Rate	Description	Value / Constraint
δ_g	GFP bleaching	1×10^{-4}/min[36]
r_g	GFP maturation time	$6 - 8$ min[38, 60]
r_p	Translation initiation rate	20/min[61–64]
α	LexA decay rate during the break	1/min[28, 65]
β	LexA decay outside breaks	1/min
L_{SS}	LexA binding outside breaks	400/min
L_{off}	LexA unbinding rate	0.1/min (consensus sequence), 3/min (recA)[31, 66, 67]
δ_m	mRNA decay rate	Between 0.52/min and 0.1/min[11]
r_m	Transcription initiation per unit volume	$4.5 - 9$/min
r_r	Repair rate of the breaks	Between 0.5/min and 1/min
r_b	Break frequency	Low enough to allow recover of LexA between two breaks

Table 3.1.: **Parameters of the gene expression model**. Green cells: we can fix a value for the parameter; orange cells: a suitable range to explain the observation data is examined in the Supplementary Material section C.3.

3.7. Response of the promoters to double strand breaks

Having determined that the bursts in gene expression are caused by fluctuations in the regulator LexA, we set out to understand how the promoter architecture determines the sensitivity of a promoter to LexA fluctuations of different time scales. All the analysis of this section are based on simulated data, using an algorithm described in the Supplementary Section C.3.

First, we want to understand whether a promoter can be tuned to respond and discriminate among fluctuations of different durations. To do so, we simulate a growing cell for 15 minutes, with one break repaired after a fixed amount of time ranging from 1 minute to 10 minutes (Fig 3.5A left). The response of the promoter is given by the total amount of mRNA produced in the 15 minutes of simulation. Notice that the break is repaired after a deterministic time, but the LexA unbinding and mRNA production are still stochastic, leading to a distribution of mRNA production over different simulations (Fig 3.5A right). In Fig 3.5A right, we show that for very strong binding sites (LexA unbinding rate of 0.1/min, red box plots), the shortest breaks are missed by the promoter, since LexA does not have time enough to unbind before the break is repaired. At higher LexA unbinding rates, the promoter is able to react to shorter breaks, and to discriminate among breaks of different durations (Fig 3.5A right, green box plots). This shows that the sensibility of a promoter to short breaks can be tuned by increasing the LexA binding site. However, there is a limit on the sensitivity. Indeed, we notice that one assumption of the non equilibrium model is that outside the breaks, LexA rebinds immediately after it unbinds, preventing a burst of mRNA production. That is, outside the breaks, the promoter is most of the time repressed, with the mRNA being produced mainly during a DSB. When the unbinding rate becomes of the same order of the steady sate binding rate (blue

box plots in Fig 3.5 right), the promoter starts to get transcribed also outside the break, leading to an mRNA production all the time. In this regime, the promoter is not able to sense any break. Altogether, to sense shorter fluctuations in the regulator level, a cell needs to increase the unbinding rate of LexA from the promoter, but the steady state LexA concentration outside a break sets a biological limit to the fastest unbinding rate.

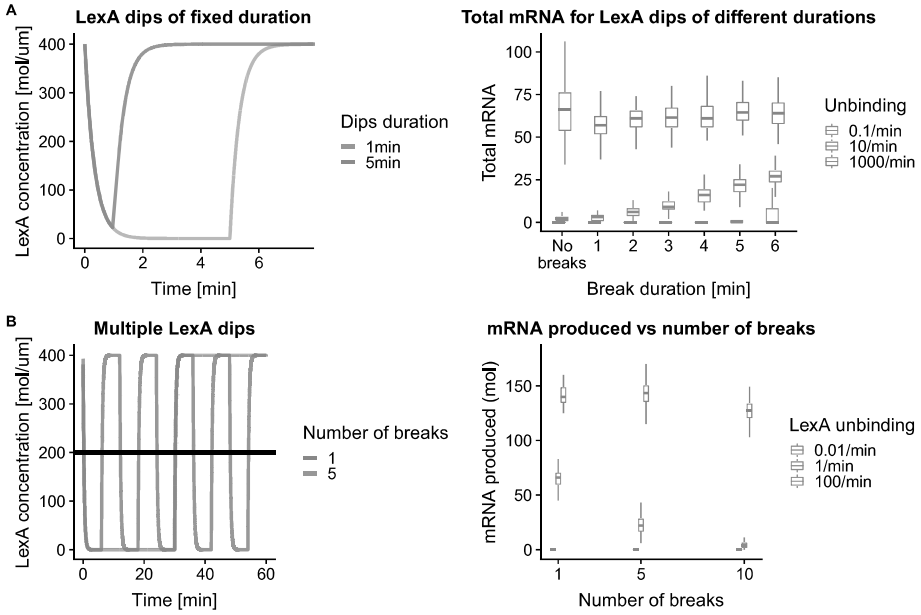

Figure 3.5.: **Response of a promoter to different kinds of breaks.**. A: mRNA production over 15 minutes of simulation with one break repaired after a fixed time, from 1min to 10min. On the left, LexA concentration during a break of duration 1min and 5min. On the right, distribution of total mRNA production for different repair times. For each repair time, the simulation has been repeated several time, leading to a distribution of mRNA represented by the box plots. The different colors correspond to promoters with different LexA unbinding rates. B: We consider the response to a different number of breaks in the previous hour. The duration of the breaks is set such that the average LexA concentration is always the same. On the left, an example of 1 long break vs 5 shorter breaks. The black horizontal line shows the average LexA concentration in both cases. On the right, distribution of total produced mRNA for different amount of breaks (x axis) and different LexA unbinding rates (different colors).

Next, we wondered how the architecture determines the ability of a promoter to integrate the past damage history in its expression level. That is, we investigated how well a promoter is able to discriminate whether in the past only one long fluctuation in LexA happened, or multiple shorter fluctuations. To this end, we simulated one hour of growing cell, subjected to one, five, or ten breaks. The duration of the breaks is set such that the average LexA level over the 1 hour simulation is always the

same (Fig 3.5B left). We can see that for very strong binding sites (e.g. 0.01/min, red distribution in Fig 3.5B right), LexA never unbinds and independently on the number of breaks the promoter never expresses. As the unbinding rate increases, the promoter is able to discriminate between the number of breaks happened in the past (green distribution in Fig 3.5B right). However, similarly as the previous scenario, if the binding site is too weak, LexA starts to unbind also outside the breaks and the promoter is always expressed, independently on the number of breaks (blue distribution in Fig 3.5B right). Therefore, also in this case the steady concentration of LexA sets a biological limit on the sensitivity of the promoter to the break history.

4. Discussion

Despite 50 years of research, a quantitative description of regulatory programs remains elusive, even for the model organism *E. coli*. In this work, we focused on the SOS response in *E. coli* to understand how fluctuations in the regulator propagate to the target genes. In particular, we induced the network by exposing the cells to three different concentrations of the DNA damaging agent Ciprofloxacin, and we followed the time dynamics of gene expression of single cells using a microfluidic device.

Looking at the time traces of single cells exposed to different levels of inducer, we found that the induction of the regulatory network is characterized by bursts of mRNA production. If the equilibrium hypothesis holds, these bursts must correspond to periods where the promoter is unbound from the repressor LexA, that is, the noise originates solely from the stochastic binding and unbinding of the regulator. In this scenario, the induction of the network with Cipro is explained by supposing a constant concentration of the repressor LexA, which decreases with the antibiotic level. As the level of Cipro increases, a lower concentration of repressor causes LexA to take more time to rebind, resulting in higher and longer bursts. On the other hand, since the occurrence of the bursts is determined by the unbinding of LexA, which depends only on the architecture of the promoter, the frequency of the bursts cannot depend on the Cipro level. In contrast, the experimental data show the opposite behavior, where the frequency of the bursts increases with Cipro, and the height and the durations are independent, with the durations being also independent on the specific promoter.

Based on the known mechanism of the SOS derepression, we hypothesized that this behavior can be explained by supposing a non-equilibrium regime, where the bursts correspond to stochastic transient low levels of LexA which are precipitated by a random DNA damage event (DSB). We imagined a scenario where, in the absence of DSB, the level of the repressor is so high that the promoters are basically always bound and high transcriptional activity is prevented. But when a DSB occurs, the level of repressor LexA falls so low that it takes a longer time for it to rebind after an unbinding event, causing a burst of transcription. In this scenario, the frequency of

the bursts is determined by the frequency of DSBs, which in turn increases with the amount of Cipro. The duration of the breaks is determined by how long it takes for the DSB to be repaired and for LexA to recover, which, for low levels of antibiotic, can be assumed to be independent on the amount of Cipro and the specific promoter.

The model requires the specification of a set of parameters and it was fundamental to verify that the regime described above can be achieved using biological relevant values. Most of the parameters have been determined independently in the literature. The only free parameters are the frequency of the breaks, the transcription rate, and the repair rate. In order to explain the short duration of the bursts, we need to assume that the DSBs are repaired quickly, on the order of minutes. In this regime, it can be shown that the heights are geometrically distributed, in agreement with the experimental data. Moreover, the slope of the geometric distribution uniquely fixes the ratio of the transcription and repair rate. The repair rate can be further bounded. Indeed, if the repair rate is faster than about $1/\text{min}$, the concentration of LexA does not have time to fall low enough to allow the promoter to be transcribed. On the other hand, if the repair rate is slower than $0.5/\text{min}$, the breaks are too long and we lose the geometric distribution of the heights of the peaks.

Finally, we discussed how the promoter architecture sets the sensitivity of the promoter to fluctuations in the regulator LexA. We have shown that higher unbinding rates allow the promoter to sense faster fluctuations, although this sensibility cannot be increased indefinitely. In fact, for LexA unbinding rates on the order of the steady state binding rate, the promoter starts to produce mRNA independently from the breaks, losing its sensitivity to the fluctuations in the regulator.

Supplementary

Supplementary Material A.

Data analysis

A.1. Correlated statistics

In the Mother Machine, the measurements coming from different time points in the cell cycle are not independent, but show a high correlation, with the effect of increasing the error of the mean by reducing the "effective" number of independent points. Therefore, the usual formula to compute the error of the mean by dividing the standard deviation of the data by the square root of the number of points is not adequate.

The problem of correctly estimating the error of the mean in the presence of correlation is recurring in the analysis of time series and one way to tackle it is to use the "blocking" method, as described in [68]. The intuition behind this method is to transform the dataset by substituting to each pair of consecutive numbers their mean. In such a way, the mean and the standard deviation of the original set of data does not change, but the correlation between two consecutive points diminishes. If this operation is repeated over and over, at a certain point the correlation disappears, and we are left with a set of uncorrelated data for which we can infer the mean and standard error using the usual formulas.

Briefly, let's suppose we want to infer the average of some correlated data $x_1, ..., x_n$. Let's call m the mean and s the variance of the estimate of m. We know that

$$s = \frac{1}{n^2}\sum_{i,j+1}^{n}\gamma_{i,j} = \frac{1}{n}\left[\gamma_0 + 2\sum_{t=1}^{n-1}\left(1-\frac{t}{n}\right)\gamma_t\right] \tag{A.1}$$

$$\gamma_{i,j} = \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle \tag{A.2}$$

$$\gamma_t = \gamma_{i,j},\ t = |i-j| \tag{A.3}$$

Let's transform the data into a half as large dataset $x'_1, ..., x'_{n'}$, where

$$x'_i = \frac{1}{2}(x_{2i-1} + x_{2i}) \tag{A.4}$$

$$n' = \frac{1}{2}n \tag{A.5}$$

since the transformation is linear, the mean and variance are invariant, that is $m' = m$ and $s' = s$.

It can be proved that γ_0 increases after each blocking transformation (A.4) and the vector $\left(\frac{\gamma_t}{n}\right)_{t=0,\dots,n-1}$ converges to $\delta_{t,0}$. At this point, the standard deviation is given, from equation (A.1), by the usual formula for uncorrelated data $\gamma_0 = \langle x^2 \rangle - \langle x \rangle^2$. The problem of computing s has been transformed to the problem of computing γ_0. For γ_0 we use the estimator

$$\hat{\gamma}_0 = \frac{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2}{n} \tag{A.6}$$

Therefore the strategy is to iteratively halve the data using the blocking transformation (A.4) and each time to estimate the standard deviation with the usual formula for uncorrelated data. After a sufficient number of blocking has been applied, the computed standard deviation reaches a plateau, which is the correct estimate for the standard deviation. Notice that if the data are uncorrelated, then $\left(\frac{\gamma_t}{n}\right)_{t=0,\dots,n-1} = \delta_{t,0}$ from the very beginning.

A.2. Day-to-day variations

In Fig A.1A we show the average GFP concentration for the Synthetic high expressers over different days, and stratified in different time bins. The standard error of the mean is computed using the blocking strategy described above.

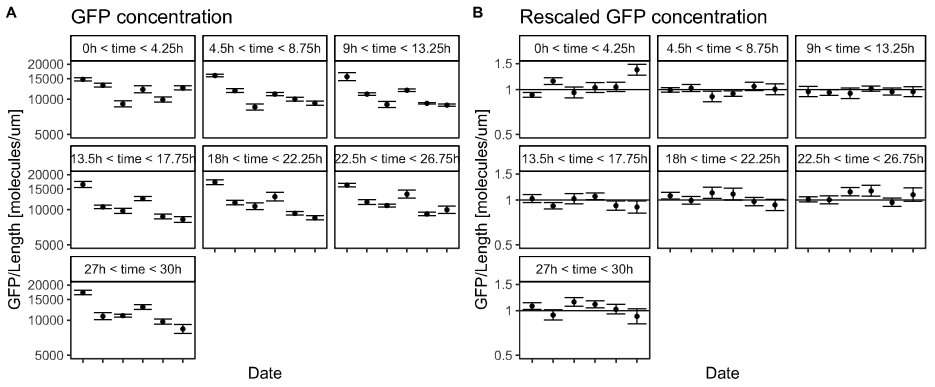

Figure A.1.: **Average GFP concentration**. *A*: Average GFP concentration of a population of Synthetic High expressers on different days, stratified by time bins of 4.5 hours. A day to day variation in the GFP levels is visible. *B*: Same as panel A, but we divided each concentration by the average concentration over the entire day. Now the concentrations look compatible among days.

From the figure, it is clear that even taking care of the correlation in the data, the error bars show a difference among different days. This difference is likely due to variations in the experimental conditions which were outside our control and we took care of them by rescaling the concentrations of all the promoters by the average

concentration of the synthetic promoters measured on that same day. The average is computed over all the experiment, without time stratification. On panel B of Fig A.1, we can see that the rescaled concentrations for the high expressers are now inside the error bars over different days and time bins. In Supplementary Figure 1, we show the average GFP concentration in glucose and the fold changes for all the promoters, and we can see that for recA and polB, the mean concentrations agree over two days.

A.3. Adaptation to the Mother Machine

Due to the strong illumination the cells are subjected to, it takes some time before the cells adapt to the experimental environment. We considered the growth rate as a proxy for the cell physiology and we looked at how they change over the time of the experiment. The growth rates have been inferred through a linear fit between the time and the log size of the cells. Figure A.2 shows that the distribution of growth rates is higher at the beginning of the experiment, but then it starts to go down to reach a new stable distribution after around 2 hours. For this reason, all measurements from cells born in the first two hours are discarded.

Supplementary Material B.

Inferring the GFP production

Let $G(t)$ be the amount of GFP inside a cell at a time t. We suppose that the GFP is created with a production rate per unit volume per unit time $q(t)$ and destroyed at a rate β. Notice that β is the sum of bleaching and dilution. In this scenario we have

$$\frac{dG}{dt} = q(t)V(t) - \beta G(t) = q(t)V_0 e^{\lambda t} - \beta G(t) \tag{B.1}$$

where we also assumed that the volume grows exponentially at a rate λ. In this last equation we only have one unknown that is $q(t)$ and this is indeed the quantity that characterizes the protein production. We can solve this equation by rewriting it as

$$\frac{d}{dt}[Ge^{\beta t}] = q(t)V_0 e^{(\lambda+\beta)t} \tag{B.2}$$

Integrating and setting the initial condition $G(t=0) = G_0$ we find

$$G(t) = G_0 e^{-\beta t} + V_0\, e^{-\beta t} \int_0^t d\tau\, q(\tau) e^{(\lambda+\beta)\tau} \tag{B.3}$$

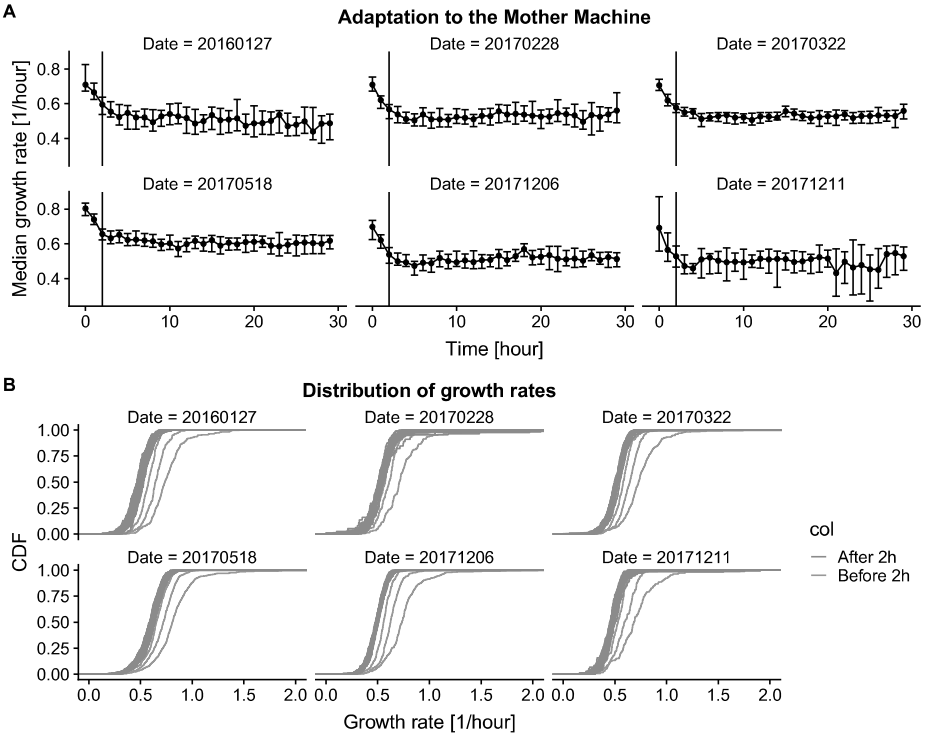

Figure A.2.: **Adaptation of the cells to the Mother Machine environment**. The growth distribution in the first two hours of the experiment are significantly higher. *A)* Median growth rate of the cells measured on different days. As the illumination is turned on (time oh), the growth rate decreases until it reaches a plateau after about 2 hours. *B)* Distribution of growth rate stratified by time bin. The blue lines show the distributions coming from bins earlier than 2 hours and they are clearly higher than the red distributions coming from bins after 2 hours.

We suppose that $q(t)$ varies slowly between one time sample and the next in the Mother Machine. This is justified by the fact that q is determined by the dynamic of the mRNA which is slower than the 3 minutes sampling of the Mother Machine. In this case, the previous equation can be solved analytically to give (we now substitute the time t with the sample i)

$$G_i = \alpha_i q + G_0 e^{-\beta t} \tag{B.4}$$

where we defined

$$\alpha_i = \frac{V_0}{\lambda + \beta}\left[e^{\lambda t} - e^{-\beta t}\right] \tag{B.5}$$

Now we can suppose that the measured G_i is the true G_i plus Gaussian noise

$$G_i^m = \alpha_i q + G_0 e^{-\beta t_i} + \mathcal{N}(0, \sigma^2) \tag{B.6}$$

Given a set of T consecutive measurements $\{G_i^m\}$ we can write the probability of the data as

$$P(\{G_i^m\} \mid q, \{\alpha_i\}, \beta, \sigma, G_0) = \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=0}^{T}\left(G_i^m - q\alpha_i - G_0 e^{-\beta t_i}\right)^2\right] \tag{B.7}$$

From which we have that the probability for q is

$$P(q \mid \{G_i^m\}, \{\alpha_i\}, \beta, \sigma, G_0) \propto \frac{1}{\sigma^T} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=0}^{T}\left(G_i^m - q\alpha_i - G_0 b_i\right)^2\right] P(\{G_i^m\}, \{\alpha_i\}, \beta, \sigma)\, P(G_0) \tag{B.8}$$

where we defined the bleaching term as $b_i = e^{-\beta t_i}$. We want to integrate out the unknown G_0 supposing a uniform prior. To do that we rewrite the exponent as

$$\begin{aligned} & -\frac{1}{2\sigma^2}\sum_{i=0}^{T}\left(G_i^m - q\alpha_i - G_0 b_i\right)^2 \\ = \; & -\frac{1}{2\sigma^2}\sum_{i=0}^{T}\left[G_0^2 b_i^2 + (G_i^m - q\alpha_i)^2 - 2G_0 b_i (G_i^m - q\alpha_i)\right] \\ = \; & -\frac{T < b^2 >}{2\sigma^2}\left[G_0^2 + \frac{< (G^m - q\alpha)^2 >}{< b^2 >} - 2G_0 \frac{< b(G^m - q\alpha) >}{< b^2 >}\right] \\ = \; & -\frac{T < b^2 >}{2\sigma^2}\left(G_0 - \frac{< b(G^m - q\alpha) >}{< b^2 >}\right)^2 \\ & -\frac{T}{2\sigma^2}\left(< (G^m - q\alpha)^2 > - \frac{< b(G^m - q\alpha) >^2}{< b^2 >}\right) \end{aligned} \tag{B.9}$$

Using this exponent we have

$$\begin{aligned} & P(q \mid \{G_i^m\}, \{\alpha_i\}, \beta, \sigma) \propto \\ & \left\{\int_{-\infty}^{\infty} dG_0 \, \frac{1}{\sigma} \exp\left[-\frac{T < b^2 >}{2\sigma^2}\left(G_0 - \frac{< b(G^m - q\alpha) >}{< b^2 >}\right)^2\right]\right\} \\ & \times \frac{1}{\sigma^{T-1}} \exp\left[-\frac{T}{2\sigma^2}\left(< (G^m - q\alpha)^2 > - \frac{< b(G^m - q\alpha) >^2}{< b^2 >}\right)\right] \\ & \times P(\{G_i^m\}, \{\alpha_i\}, \beta, \sigma) \\ = \; & \frac{1}{\sigma^{T-1}} \exp\left[-\frac{T}{2\sigma^2}\left(< (G^m - q\alpha)^2 > - \frac{< b(G^m - q\alpha) >^2}{< b^2 >}\right)\right] \\ & \times P(\{G_i^m\}, \{\alpha_i\}, \beta, \sigma) \end{aligned} \tag{B.10}$$

Now we rewrite the exponent in order to isolate q

$$\begin{aligned} & < (G^m - q\alpha)^2 > - \frac{< b(G^m - q\alpha) >^2}{< b^2 >} \\ = \; & < (G^m)^2 > + q^2 < \alpha^2 > - 2q < G^m \alpha > - \\ & \frac{< b(G^m) >^2 + q^2 < \alpha >^2 - 2q < bG^m >< b\alpha >}{< b^2 >} \\ = \; & q^2 \widetilde{Var}(\alpha) - 2q\widetilde{Cov}(G^m, \alpha) + \widetilde{Var}(G^m) \\ = \; & \widetilde{Var}(\alpha) \left[q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right]^2 - \frac{\widetilde{Cov}^2(G^m, \alpha)}{\widetilde{Var}(\alpha)} + \widetilde{Var}(G^m) \\ = \; & \widetilde{Var}(\alpha) \left[q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right]^2 + \widetilde{Var}(G^m) \left(1 - \widetilde{\rho}^2 \right) \end{aligned} \tag{B.11}$$

where $\widetilde{Var}(x)$, $\widetilde{Cov}(x, y)$ and $\widetilde{\rho}$ are the variance, the covariance and the Pearson correlation coefficient corrected by the bleaching, i.e.

$$\widetilde{Var}(x) = < x^2 > - \frac{< bx >^2}{< b^2 >} \tag{B.12}$$

$$\widetilde{Cov}(x, y) = < xy > - \frac{< bx >< by >}{< b^2 >} \tag{B.13}$$

$$\widetilde{\rho} = \frac{\widetilde{Cov}(x, y)}{\sqrt{\widetilde{Var}(x)\widetilde{Var}(y)}} \tag{B.14}$$

Using this exponent we get

$$\begin{aligned} & P(q \mid \{G_i^m\}, \{\alpha_i\}, \beta, \sigma) \propto \\ & \frac{1}{\sigma^{T-1}} \exp \left[-\frac{T}{2\sigma^2} \left(\widetilde{Var}(\alpha) \left[q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right]^2 + \widetilde{Var}(G^m) \left(1 - \widetilde{\rho}^2 \right) \right) \right] \\ & \times P(\{G_i^m\}, \{\alpha_i\}, \beta, \sigma) \end{aligned} \tag{B.15}$$

We assume a scale invariant prior $1/\sigma$ for the standard deviation and we integrate it out. First we define

$$x = \frac{T}{2} \left[\widetilde{Var}(\alpha) \left(q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right)^2 + \widetilde{Var}(G^m) \left(1 - \widetilde{\rho}^2 \right) \right] \tag{B.16}$$

such that the integral to be performed becomes

$$\int_0^\infty ds \, \frac{1}{\sigma^{T-1}} \exp \left(-\frac{x}{\sigma^2} \right) \frac{1}{\sigma} \tag{B.17}$$

We make the change of variable $y = \frac{x}{s}$ in order to remove the dependence on x from the integral. The final result is (We don't need to perform the integral, since it gives just a constant)

$$P(q \mid \{G_i^m\}, \{\alpha_i\}, \beta) \propto x^{-\frac{T-1}{2}} \propto \left[1 + \frac{\widetilde{Var}(\alpha) \left(q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right)^2}{\widetilde{Var}(G^m) \left(1 - \tilde{\rho}^2\right)} \right]^{-\frac{T-1}{2}} \tag{B.18}$$

Remembering that for large N it holds

$$(1 + x)^{-N} \to \exp(-Nx) \tag{B.19}$$

I obtain

$$P(q \mid \{G_i^m\}, \{\alpha_i\}, \beta) \propto \tag{B.20}$$
$$x^{-\frac{T-1}{2}} \propto \exp \left[-\frac{T-1}{2} \frac{\widetilde{Var}(\alpha) \left(q - \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \right)^2}{\widetilde{Var}(G^m) \left(1 - \tilde{\rho}^2\right)} \right]$$

from which we see that

$$q^* = \frac{\widetilde{Cov}(G^m, \alpha)}{\widetilde{Var}(\alpha)} \tag{B.21}$$

$$\sigma_q^2 = \frac{1 - \tilde{\rho}^2}{T - 1} \frac{\widetilde{Var}(G^m)}{\widetilde{Var}(\alpha)} \tag{B.22}$$

We can interpret these results. The best estimate for q is the usual best estimate for a linear model. The error is zero if the correlation is one, which means that we have a perfect straight line. It increases with the variance of the measured values G^m, which is determined by measurement noise; it decreases with the variance of the independent variable α, which means that if the measurements span a higher interval of the independent variable, we have a better estimate. Finally, the error depends as usual on the inverse of the number of points minus one. Notice that the tilde above the statistics show that the bleaching is taken into account.

Supplementary Material C.

Identification of the peaks in production

C.1. Defining a genealogy

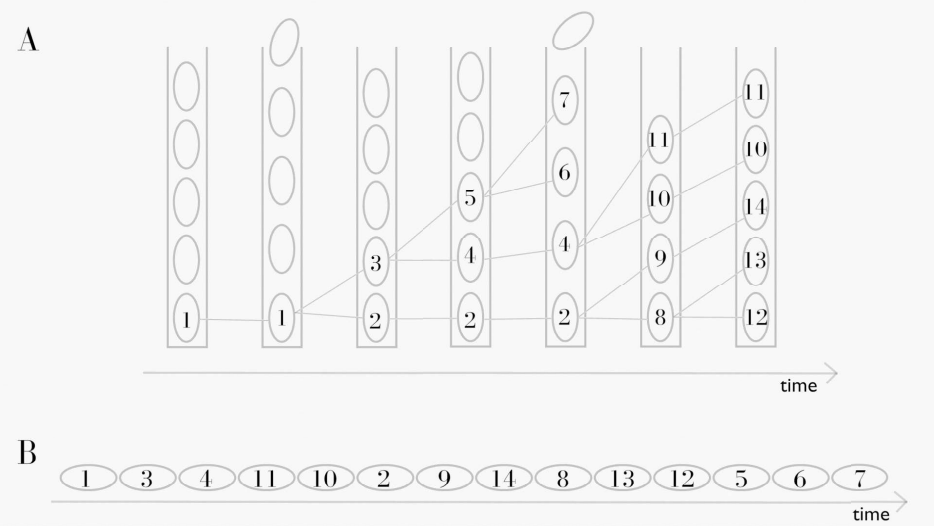

Figure C.1.: **Creating a single long time traces from all the cells in a growth lane**. *A*: Example of cells in a single growth lane. The time flows from left to right and the lines connect mothers to daughters. *B*: Final order of the cells to make up a single time trace. For graphical purpose, only the reduced set of numbered cells are considered, but in the full time trace there will be present all the cells in the growth lane.

Each experiment in the Mother Machine is characterized by cells growing in different growth lanes and we joined all the time traces of the single cells in a given growth lane into one single trace, which we call a *genealogy*. To avoid artifacts, we only retained cells for which we have data for the entire cell cycle, that is we discarded cells that at a certain point left the growth lane or that were present at the end of the experiment.

To create a genealogy, we randomly pick one of the cell in the growth lane and we trace back its genealogy until the beginning of the experiment. We then pick a second cell and we again trace back its genealogy, but this time we discard all the cells that were already present in the previous genealogy. The remaining genealogy

is placed right after the previous one. We iterate this process until all the cells have been included.

The process is illustrated with an example in Fig C.1. On the top part we show an example of a growth lane, with time flowing from left to right. As time goes on, cells grow and divide. To create a single time trace, we start by picking a cell at the end of the experiment, e.g. cell number 11 and we trace back its genealogy, that is $1-3-4-11$; this is the first part of the final time trace. We then pick e.g. cell number 10 and again trace back its genealogy $1-3-4-10$. Since cells $1-3-4$ were already in the previous genealogy, we skip them and we add only cell 10 to the final time trace. Then we pick cell number 14 and we get the genealogy $1-2-9-14$, from which we discard cell 1 which is already present in the final time trace. Iterating this process, we get the final time trace

$$1-3-4-11-10-2-9-14-8-13-12-5-6-7-...$$

C.2. Peak finding method

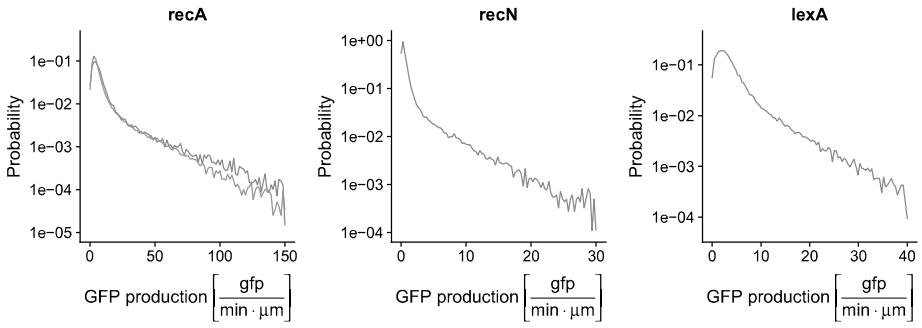

Figure C.2.: **Distribution of GFP production for the bursty promoters**. Histogram of the production rate $q = dG/(V\,dt)$ for recA (left, measured in duplicate), recN (middle) and lexA (right). The y-axis is shown on a logarithmic scale. The production rates q were pooled from the entire experiment, regardless of the Cipro concentration.

We identified peaks in the time traces of the GFP production $q = dG/V$ by finding consecutive segments in which q is over a cut-off value q_c. We used the following algorithm:

1. Find all contiguous segments in which $q > q_c$.
2. For each such segment, find all the local maxima within the segment.
3. For each local maximum, we extract a segment of data points around the maximum to fit it to a Gaussian shaped peak. To do this, the segment needs to capture the curvature around the peak. We create a segment by including contiguous data points to the left and right for as long as the their q values keep

decreasing, and as long as the q values are at least a fraction $f = 0.25$ of the maximum. For each maximum we thus include a segment of falling q values to the left and right, until q drops to a quarter of its maximum values.

4. Before we fit the q values in the segment to a Gaussian, we first perform two checks: The maximum must not lie at the edge of the segment, and it must have at least 3 data points in it. If these checks are not passed the maximum is discarded.
5. For each segment of consecutive q measurements $[q(t1), q(t2), ..., q(t_n)]$ around a maximum, we fit a Gaussian by minimizing the sum of squared deviations

$$\Delta^2 = \sum_{i=1}^{n} \left[q(t_i) - A \exp\left(-\frac{(t_i - \mu)^2}{2\sigma^2} \right) \right]^2 \tag{C.1}$$

with respect to A (height of the peak), μ (time of the peak maximum) and σ (peak width). To do this, we iterate maximizing the derivative of Δ^2 with respect to μ and σ (finding the root numerically), and A (analytically).

In this way we obtain a set of extracted peaks for each Mother Machine experiment.

C.3. The peak heights come from a distribution with exponential tail

We empirically find that, for each of the three LexA target promoters that show a clear bursty behavior (recA, lexA and recN), the distribution of production rates $q = dG/(V\,dt)$, with V the volume of the cell and G the amount of GFP, looks relatively tightly distributed around a small value, with an exponential tail, see Fig. C.2. Visual inspection of the time traces of q strongly suggests that these exponential tails correspond to peaks in the production rate.

We next set out to measure a single peak height h for each of the four experiments, assuming that the distribution of peaks heights has an exponential tail. That is, we assume that for peaks with height A larger than a cut-off A_c, the distribution is exponential, that is

$$P(A \mid A \geq A_c, h) = \lambda e^{-\lambda(A - A_c)/h} \tag{C.2}$$

Imagine that we have n peaks with $A \geq A_c$ and that $\langle A \rangle_{A_c}$ is the average heights of these n peaks. We then find for the optimal height h:

$$h(A_c) = (\langle A \rangle_{A_c} - A_c) \tag{C.3}$$

with an error bar

$$\sigma_h(A_c) = \frac{h(A_c)}{\sqrt{n}} \tag{C.4}$$

If the distribution is really exponential, we should find that the heights $h(A_c)$ are relatively insensitive to A_c in a reasonable range. Fig C.3 shows that this is indeed

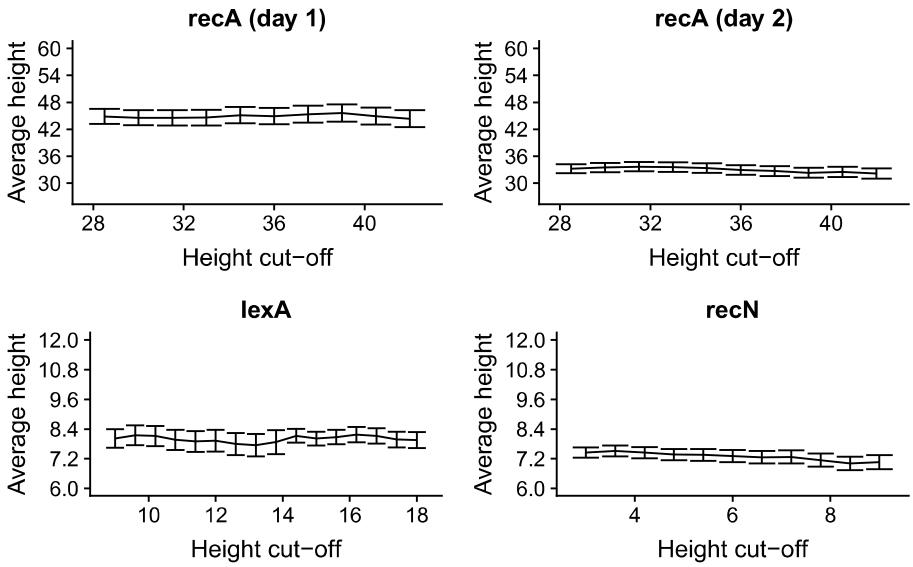

Figure C.3.: **Average peak height A as function of the cutoff A_c.** Estimated height $h(A_c)$ as a function of a cut-off on peak height A_c for each of the 4 experiments. The error bars show estimated height $h(A_c)$ plus and minus one standard-deviation $\sigma_h(A_c)$.

the case. Also, the distribution of heights above A_c, as shown in Fig 3.3 of the main text, clearly confirms that the heights are exponential.

Supplementary Material D.

Simulation of a growing cell

To simulate the GFP production in a growing and dividing cell, we wrote an algorithm in C++ that performs the following steps, depicted in Fig 3.4A in the main text:

1. We start with a cell of volume 2nm. The precise value of this initial volume is not important because, in the long run, the system loses memory of the initial conditions.
2. The cell grows exponentially with a constant growth rate of 0.0085/min (half life of 81min), until it adds a volume ΔV (adder model [39]). The added volume is drawn from a Gaussian distribution with mean 1.86 and standard deviation

0.35, computed from the experimental data.

3. After adding a volume ΔV, the cell divides into one daughter of half the volume of the mother. The daughter inherits the mRNA, the immature GFP, and the mature GFP from the mother, according to a binomial distribution of probability 0.5.
4. During the growth of the cell, the following stochastic events are simulated using a Gillespie algorithm [69]:
   - A DSB is created (rate $1/r_b$), or repaired (rate $1/r_r$).
   - LexA can bind (rate $\leq L_{on}$) or fall off (rate L_{off}) from the promoter. L_{on} is given by the concentration of free LexA, while L_{off} is determined by the architecture of the promoter.
   - If the promoter is free from LexA, an mRNA molecule can be transcribed (rate $r_m V$). The dependence of the transcription rate on the volume is because during the cell cycle, the concentration of mRNA polymerase is supposed to stay constant. Therefore, as the volume grows, also the total number of polymerases grows, increasing the transcription initiation. Hence, larger cells have a higher transcription rate.
   - A molecule of mRNA decays (rate δ_m), or is translated into immature GFP (rate r_p).
   - An immature GFP can mature into fluorescing GFP (rate r_f).
   - A mature GFP can decay because of bleaching (rate δ_g).

In an equilibrium model, the concentration of free LexA in the cell is supposed to be constant and proportional to the amount of Ciprofloxacin. In the dynamic model, the concentration is supposed to be at a constant steady-state level when no breaks are present (corresponding to a binding rate L_{SS}). During a break, the concentration decays exponentially with a rate α. When the break is repaired, the concentration recovers with a rate β, until it saturates back to the steady-state rate L_{SS}.

Supplementary Material E.

Theoretical analysis of the peaks

In this section we want to understand how the different characteristics of peaks are related to the parameters of the dynamic model (Tab 3.1 in the main text). In particular, we will show that the breaks must be repaired quickly, and that the slope of the exponential distribution sets the ratio between the transcription and the repair

rate.

E.1. The breaks must be repaired quickly

The duration of the peaks is determined by different time scales, including the repair rate r_r of the break, the decay time δ_m of the mRNA, and the folding time r_f of the GFP. Since we know that the folding time is around 8 min [38, 60] and it is reasonable to assume that δ_m is at least $3-5$ minutes [11], we see that they already make up a good fraction of the observed peak duration (gaussian shape with a standard deviation around 7 minutes). Therefore, the time to produce the mRNA must be less than $2-3$ minutes. For this reason, we consider all the mRNA to be produced in a sudden pulse of transcription, and we work out how this pulse translates into a peak in GFP production dG/V.

We assume that at time $t=0$ we are given a burst of m_0 mRNA molecules which decay at a rate δ_m and are translated into immature GFP proteins at a rate r_p. In addition, we assume unfolded GFP proteins mature into fluorescing GFP at a rate r_f. The production rate dG/V is defined as the production (per unit time) of mature GFP, per unit volume.

The time dynamic of the mRNAs is a simple exponential decay

$$m(t) = m_0 e^{-\delta_m t} \tag{E.1}$$

The differential equation for the total amount of immature protein P is given by

$$\frac{dP}{dt} = r_p m(t) - r_f P(t) \tag{E.2}$$

and the production is given by

$$\frac{dG}{V}(t) = \frac{r_f}{V} P(t) \tag{E.3}$$

Note that since the peaks are short compared to the cell growth, we will ignore the growth of the volume V. Solving for dG/V we find

$$\frac{dG}{V}(t) = \frac{r_p m_0}{V} \frac{e^{-r_f t} - e^{-\delta_m t}}{x-1} \tag{E.4}$$

where $x = \frac{\delta_m}{r_f}$ is the ratio of the mRNA decay to the protein maturation rate.

We now want to determine the values of the time t^* at which the maximum of the peak in GFP production occurs, the height of this maximum, and the time dt for which the peak is above half of its maximum. First, the maximum is reached at a time

$$t^* = \frac{1}{r_f} \frac{\log(x)}{x-1} \tag{E.5}$$

Using this result, we find for the height of the peak

$$\frac{dG}{V}^* = \frac{r_p m_0}{V} x^{-\frac{x}{x-1}} \tag{E.6}$$

Finally, there is no simple analytical expression for the time dt that $\frac{dG}{V}(t)$ is larger than half its maximum $\frac{dG}{V}^*/2$. However, we notice that it depends only on δ_m and r_f, and if we consider $r_f = 1/7\text{min}^{-1}$ and $\delta_m = 1/4\text{min}^{-1}$, we numerically obtain $dt = 13\text{min}$. This is indeed very close to the width of the peaks we observe (full width half maximum: $2\sqrt{2\log(2)}\sigma = 17\text{min}$), providing support that the peaks in $\frac{dG}{V}$ are indeed due to single transcriptional bursts, and explaining why they are the same for all the promoters.

We simulated the mRNA production for a growing cell, where the breaks are short compared to the mRNA decay and maturation time (repair rate of 0.9/min). We considered different mRNA decays and GFP maturation rates, and we fitted the peaks using the same algorithm as for the experimental data. We verified that, indeed, the mean peak duration is an increasing function of both δ_m and r_f and even when we get to very low values for these rates, the width is already close to the experimental one (Fig E.1A). If we choose an mRNA decay time of 4 min and a folding time of 7 min, we get a very nice agreement with the observed data (Fig E.1B). The details of the simulation are discussed in Section C.3 of these Supplementary Materials.

E.2. The number of mRNA transcripts produced during a break has a geometric distribution

Notice that because of equation (E.6) the distribution of the peak heights is determined by the distribution m_0 of the number of mRNA molecules produced during a short break. For this reason, given the expression model presented in Fig 3.4A in the main text, we want to determine the distribution T_n of mRNA transcripts produced during a break and how it depends on the LexA dynamics.

We describe the mRNA transcription process during a break as a dynamical system that over time visits the following states (Fig E.2):

1. $B_n(t)$: The promoter is bound by LexA and n transcripts have been produced since the beginning of the break.
2. $U_n(t)$: The promoter is unbound by LexA and n transcripts have been produced since the beginning of the break.
3. $R_n(t)$: The break is repaired after having produced n transcripts.

The transitions among the states are determined by the corresponding rates (Tab 3.1 in the main paper). Notice that only the unbound state can produce transcripts, and the repaired state is absorbing, meaning that once the system reaches an R_n state, it will stay there for all future times.

Since in the last section we showed that the breaks must have a short duration, we can assume that the volume does not change much during the time the break is repaired. Therefore, we make the simplification that the volume of the cell is just a constant and we ignore it by redefining the transcription rate $r_m \to r_m V$. A full account of the solution of this dynamical system is given in the Appendix E.3 of this Supplementary Materials. Here, we report the main results.

First of all, it turns out that the decay rate α of LexA during a break sets a time scale for the dynamics of the transcription process and can be eliminated by measuring the time in units of α.

Next, we want to understand the distribution T_n of mRNA transcripts produced by the time the break is repaired, since, for short breaks, this is related to the distribution of peak heights via Eq (E.6). If we introduce the probability $q_n(t)$ of having produced n transcripts at time t since the beginning of the break, than T_n is given by

$$T_n = \int_0^\infty dt\, q_n(t) \text{P(repaired at time t)} \tag{E.7}$$

that is, T_n is the probability of having produced n transcripts by the time t and being repaired at time t, marginalized over the repair times.

In the Appendix F.1, we show that the expression distribution $q_n(t)$ is Poisson, but with a rate which is a non trivial function of time. Nevertheless, if the break lasts long enough, the concentration of LexA falls so low that, except for the first part of the break where the concentration of LexA is still high, the promoter is basically always unbound and constitutively expressed. In this regime, the distribution $q_n(t)$ follows the Poisson distribution of a constitutively expressed promoter with a rate r_m. However, to compensate for the first part of the break, where the promoter is not constitutively expressed, we have to consider an 'effective time' $t - b_\infty$, where b_∞ is the amount of time the promoter is bound by LexA (Appendix F.3), and it is determined only by the early part of the break. That is, b_∞ is just a constant depending only on the LexA dynamics and not on the break duration. In conclusion

$$q_t(n)|_{t>t^*} = \text{Pois}\left[r_m(t - b_\infty)\right] \tag{E.8}$$

where t^* is the time we need to wait for the LexA concentration to fall so low that the promoter can be considered constitutively expressed.

As discussed in the Appendix F.4, there is no a simple closed-form formula for either t^* or b_∞, and they must be computed numerically. However, we notice that for a given LexA dynamics (fixed r_{off} and r_{on}), t^* and b_∞ are also fixed.

The probability $q_n(t)$ of having n transcripts at time t during a break is a bell-shaped function of time, peaked at $t = n/r_m + b_\infty$ (Appendix F.3 and Fig F.3). Therefore, the breaks giving a number of transcripts

$$n > n^* = r_m(t^* - b_\infty) \tag{E.9}$$

are breaks lasting more than t^*, that is breaks for which the approximation (E.8) is valid. For this reason, $q_n(t)$ simplifies to the simple Poisson distribution of a constitutive promoter either for large repair times or for large number of transcripts (Fig F.3).

Finally, we can discuss the probability T_n that n transcripts have been produced when the break is repaired. We want to show that for breaks that result in a large number of transcripts (that is, the bursts identified in the main text), T_n is geometrically distributed. Since for a large number of transcripts the condition in equation

(E.9) is valid, we expect $q_n(t)$ to be a Poisson distribution with the simplified rate of equation (E.8). The integral in (E.7) simplifies to a geometric distribution:

$$\begin{aligned} T_n|_{n>n^*} &= e^{r_r b_\infty} r_r \frac{r_m^n}{n!} \int_0^\infty dt\, (t-b_\infty)^n e^{-(t-b_\infty)(r_m+r_r)} \\ &= \lim_{n\to\infty} e^{-r_r b_\infty} \left(\frac{r_m}{r_m+r_r}\right)^n \frac{r_r}{r_m+r_r} \frac{\Gamma(n+1,b_\infty)}{n!} \\ &= e^{-r_r b_\infty} \left(\frac{r_m}{r_m+r_r}\right)^n \frac{r_r}{r_m+r_r} \end{aligned} \tag{E.10}$$

where $\Gamma(n+1,b_\infty)$ is the upper incomplete gamma function and in the last equality we used the fact that $\lim_{n\to\infty} \frac{\Gamma(n+1,b_\infty)}{\Gamma(n+1)} = 1$ [70]. Therefore, T_n is geometrically distributed for $n > (t^* - b_\infty) r_m$, but with a prefactor $e^{-r_r b_\infty}$, and in log space we expect a straight line with an offset $-r_r b_\infty$. This is shown in Fig E.3A. The diamond shows the point $n > (t^* - b_\infty) r_m$, where the geometric approximation is expected to deviate from the real solution. Both t^* and b_∞ have been computed numerically. Fig E.3B shows that the theoretical description developed so far agrees with the results of the simulation of a growing cell.

In the next section, we show that the geometric distribution of T_n results in an exponential tail for the peak heights, in agreement with the experimental data.

E.3. Constraints on the repair and transcription rates

Summarizing the theory developed so far, the distribution of the number n of transcripts follows two regimes, separated by a threshold n^*. For $n > n^*$ (i.e. the bursts identified in the main text), the transcripts are produced mostly by breaks that last long enough for the LexA concentration to fall so low that the promoter is basically constitutively expressed. This gives rise to a geometric distribution for the number of transcripts n and therefore to an exponential regime, whose slope is entirely determined by the ratio r_r/r_m of the repair rate with the transcription rate (equation (E.10) and Fig E.3). On the other hand, for $n < n^*$ the breaks are so short that the binding and unbinding of LexA affects the distribution of transcripts and the geometric distribution of equation (E.10) is not valid. The threshold n^* is given by equation (E.9) and, although it cannot be expressed as a closed-form expression, it depends only on the LexA dynamics.

In a previous section, we showed that for short breaks the peak in GFP production during a short break is proportional to the number of mRNA transcripts produced. From equation (E.6) it follows that for

$$q > q^* = \frac{r_p}{V_0} x^{-\frac{x}{x-1}} (t^* - b_\infty) r_m \tag{E.11}$$

the distribution of peak heights must follow an exponential regime, whose slope is

proportional to the slope of the mRNA transcripts

$$\text{slope} = \frac{V_0}{r_p} x^{\frac{x}{x-1}} \log\left(\frac{1}{1 + r_r/r_m}\right) \tag{E.12}$$

This agrees with the experimental data, for which we know that it exists a threshold q^*, such that the distribution for $q > q^*$ is in an exponential regime, whose slope is promoter-dependent, but the same in all Cipro concentrations (Fig 3.3A and Supplementary Fig C.3).

Notice that equation (E.12) sets the ratio r_r/r_m in order to match the observed slope of a given promoter. Equation (E.11) then sets an upper bound on r_m (or equivalently on r_r, once their ratio is fixed), because from the observed data we know the threshold q^* where the exponential regime starts

$$r_m < \frac{V_0}{r_p} x^{\frac{x}{x-1}} \frac{q^*}{t^* - b_\infty} \tag{E.13}$$

If we consider the highest expressed promoter, namely recA, we know from the literature that the unbinding rate of LexA from its promoter is of the order or 3/min [31, 66], which results in $t^* = 3.5$min and b_∞=1.83 (numerically computed in Appendix F.1 and Fig F.1). From the experimental observations, we are in exponential regime for $q > 25 - 30\,\frac{GFP}{\min \mu m}$ (Supplementary Fig C.3), with a slope of -0.02 (Fig 3.3A). The average volume of a population of growing and dividing cells is computed from the simulations to be $V_0 = 2.77$nm. From equation (E.13) we therefore obtain an upper limit on the repair rate $r_r < 0.95$/min (or equivalently $r_m < 8.55$/min).

Some examples of the distribution of peak heights obtained through simulations of a growing cell are shown in Fig E.4. Notice how the peak height q^* where the exponential regime starts increases with the repair rate.

Fig E.4 also shows that the repair rate cannot be made arbitrarily slow. Indeed, as the repair becomes slow, the distribution starts to bend (as it is visible at a repair rate of 0.2/min). This is due to the fact that the exponential distribution is expected as long as equation (E.6) is valid, that is, when we are in the short break approximation (Fig E.5). Therefore, for the heights of the peaks in GFP production to be exponentially distributed, the breaks must be long enough for the LexA concentration to fall, but also short enough for equation (E.6) to hold.

Supplementary Material F.

Appendix

F.1. The probability $q_n(t)$ of having n transcripts at time t is Poisson

The mRNA transcription process during a break is described as a system that can transit through the following states (see also Fig E.2):

1. $B_n(t)$: The promoter is bound and n transcripts have already been done.
2. $U_n(t)$: The promoter is unbound and n transcripts have already been done.
3. $R_n(t)$: The break is repaired. This is an absorbing state.

We also make the further assumption that the volume does not change much during the time the break is repaired and we can ignore it by redefining the transcription rate: $r_m \to r_m V$.

The dynamical system is described by the following differential equations

$$\partial_t B_n(t) = r_{on} U_n(t) - (r_{off} + r_r) B_n(t) \tag{F.1a}$$
$$\partial_t U_n(t) = r_{off} B_n(t) + r_m U_{n-1}(t) - (r_m + r_{on} + r_r) U_n(t) \tag{F.1b}$$
$$\partial_t R_n(t) = r_r [B_n(t) + U_n(t)] \tag{F.1c}$$

By marginalizing the previous equations, we can compute the probability of the system to be in the bound or unbound state, regardless of the number of transcripts:

$$\partial_t B(t) = r_{on} U(t) - (r_{off} + r_r) B(t) \tag{F.2}$$
$$\partial_t U(t) = r_{off} B(t) - (r_{on} + r_r) U(t) \tag{F.3}$$

Let's introduce the probability $x(t)$ that at time t the break is not yet repaired:

$$x(t) = B(t) + U(t) = P(\text{repair ¿ t}) \tag{F.4}$$

x satisfies $\partial_t x(t) = -r_r x$ and therefore $x(t) = e^{-r_r t}$. This just says that the repair happens exponentially with a rate r_r.

The probability to be in the bound and unbound states can be rewritten introducing the probability $f(t)$ of being unbound at time t

$$U(t) = f(t) x(t) \tag{F.5}$$
$$B(t) = [1 - f(t)] x(t) \tag{F.6}$$

that is, $U(t)$ is the probability of being unbound at time t and not being repaired at time t. f satisfies the following equation:

$$\partial_t f(t) = r_{off} - (r_{on} e^{-\delta_L t} + r_{off}) f(t) \tag{F.7}$$

where δ_L is the LexA decay rate during a break.

We can introduce the probability $P_n(t) = U_n(t) + B_n(t)$ that at time t n transcripts have been produced and the break has not yet been repaired:

$$\partial_t P_n(t) = -r_r P_n + r_m f P_{n-1} - r_m f P_n \tag{F.8}$$

But P_n can be rewritten as $P_n(t) = x(t) q_n(t)$, where $q_n(t)$ is the probability of having n transcripts at time t. Rewriting the previous equation in terms of q and x we obtain:

$$-r_r x(t) q_n(t) + r_m f(t) x(t) q_{n-1}(t) - r_m f(t) x(t) q_n(t) = -r_r x(t) q_n(t) + x(t) \partial_t q_n(t) \tag{F.9}$$

and therefore $\partial_t q_n(t) = r_m f(t) q_{n-1}(t) - r_m f(t) q_n(t)$, which is the master equation of a Poisson process:

$$q_n(t) = \frac{\left(r_m \int_0^t d\tau\, f(\tau) d\tau\right)^n}{n!} e^{-r_m \int_0^t d\tau\, f(\tau) d\tau} \tag{F.10}$$

We notice that

$$\frac{1}{t} \int_0^t d\tau\, f(\tau) d\tau = \langle f \rangle_t \tag{F.11}$$

is simply the average time the promoter is unbound. Therefore,

$$q_n(t) \simeq \text{Pois}(r_m \langle f \rangle_t t) \tag{F.12}$$

that is, the number of transcripts produced in a time t is Poisson with rate equal to the transcription rate r_m, rescaled by the fraction of time the promoter is unbound.

Finally, the probability of having n transcripts when the break is repaired is

$$T_n = P(\text{n transcripts at repair}) = \int_0^\infty dt\, q_n(t) r_r e^{-r_r t} \tag{F.13}$$

F.2. Probability $f(t)$ that the promoter is unbound at time t

To understand the behavior of $q_n(t)$, we need first to study in more detail the probability $f(t)$ that the promoter is unbound at time t. We start from the differential equation (F.7) for $f(t)$

$$f'(t) = r_{off} - (r_{on} e^{-\delta_L t} + r_{off}) f(t) \tag{F.14}$$

First of all notice that if we redefine $t \to \delta_L t$, $r_{on} \to r_{on}/\delta_L$ and $r_{off} \to r_{off}/\delta_L$, then the previous equation becomes

$$f'(t) = r_{off} - (r_{on}e^{-t} + r_{off})f(t) \tag{F.15}$$

that is, the decay of LexA just sets a global time scale and we can ignore it by measuring the time in units of LexA decay.

From the theory of differential equations, a solution to an equation of the form

$$f'(t) = a(t)f(t) + b(t) \tag{F.16}$$

is

$$f(t) = e^{A(t)}\left(C + \int b(t)e^{-A(t)}dt\right) \tag{F.17}$$

where $A(t) = \int a(t)dt$ is an antiderivative of $a(t)$ and C is a constant of integration. In our case,

$$a(t) = -(r_{on}e^{-\delta_L t} + r_{off}) \tag{F.18a}$$
$$b(t) = r_o ff \tag{F.18b}$$

$A(t)$ is easily found to be $A(t) = r_{on}e^{-t} - r_{off}t$

Let's now solve $\int e^{-A(t)}dt = \int dt \exp(r_{on}e^{-t} - r_{off}t)$. This can be solved with the substitution $u = r_{on}e^{-\tau}$:

$$\int e^{-A(t)}dt = -\int \frac{du}{u}e^{-u}\left(\frac{u}{r_{on}}\right)^{-r_{off}} = -r_{on}^{r_{off}}\Gamma\left(-r_{off}, r_{on}e^{-t}\right) \tag{F.19}$$

where $\Gamma(z, x)$ is the incomplete upper gamma function. A general solution for f is therefore given by

$$f(t) = e^{r_{on}e^{-t} - r_{off}t}\left[C - r_{on}^{r_{off}}\Gamma\left(-r_{off}, r_{on}e^{-t}\right)\right] \tag{F.20}$$

The expression can be simplified by replacing the incomplete gamma function with the generalized exponential integral [70] $E_n(x) = \int_1^\infty dt\, \frac{e^{-xt}}{t^n} = x^{n-1}\Gamma(1-n, x)$

$$f(t) = r_{off}e^{r_{on}e^{-t}}\left[E_{1+r_{off}}\left(r_{on}e^{-t}\right) - e^{-r_{off}t}C\right] \tag{F.21}$$

The constant of integration C is determined by imposing that at the beginning of the break $f(0) = \frac{r_{off}}{r_{off}+r_{on}}$.

We notice that

1. For $t \to \infty$ we have $f(t) \to 1$, since $E_n(0) = \frac{1}{n-1}$ for $n > 1$ [70]. This is expected, because for large times the LexA concentration has fallen to 0 and once LexA falls off, it cannot rebind.
2. If $r_{on} = 0$, $f(t) = 1 - e^{-r_{off}t}$. That is, the probability of being unbound at time t is the probability that the unbind happens at time lower than t.

F.3. Asymptotic analysis of $q_n(t)$

In this section we study the large time behavior of $q_n(t)$. As shown in section F.1 of this Supplementary Material, $q_t(n)$ is given by

$$q_n(t) = \frac{\left(r_m \int_0^t d\tau\, f(\tau) d\tau\right)^n}{n!} e^{-r_m \int_0^t d\tau\, f(\tau) d\tau} \tag{F.22}$$

Due to the complex nature of $f(t)$, the above equation shows that $q_n(t)$ is a Poisson distribution whose rate is a complicated function of time.

We can rewrite $f(t)$ as $f(t) = 1 - b(t)$, where $b(t)$ is the probability of being bound at time t. The integral over τ in equation (F.22) becomes:

$$\int_0^t d\tau\, f(\tau) = t - \int_0^t d\tau\, b(\tau) = t - b_t \tag{F.23}$$

where b_t is the probability the promoter is bound in the time interval t. But since we have shown that $f(t) \to 1$ for large t, $\lim_{t\to\infty} b(t) = 0$ and $\int_0^t d\tau\, b(\tau)$ converges to an unknown constant b_∞ that it can only depend on r_{on} and r_{off}, as shown in Fig F.1. By numerically solving $f(t) = 0.98$, we see that when $r_{on} = 400$ and $r_{off} = 3$, $f(t)$ reaches 0.98 at $t = 3.5$ (see also Fig F.2), which is compatible with the convergence of $\int_0^t d\tau\, b(\tau)$, as shown by the vertical lines in Fig F.1.

All in all, we can conclude that for large times

$$\lim_{t\to\infty} q_t(n) \simeq \text{Pois}\left[r_m\,(t - b_\infty)\right] \tag{F.24}$$

This equation tells us that, although the mRNA production $q_n(t)$ is a Poisson distribution whose rate is a non trivial function of time, for large times $q_n(t)$ is a Poisson process with constant rate r_m, but with the time shifted by the (constant) fraction of time the promoter is bound.

For the purposes of the main text, what we need is not $q_n(t)$ at large times, but $q_n(t)$ for large n. However, for large times we have seen that $q_n(t)$ is approximated by

$$\lim_{t\to\infty} q_n(t) = \frac{[r_m(t - b_\infty)]^n}{n!} e^{-r_m(t-b_\infty)} \tag{F.25}$$

As function of time, this is a bell shaped function with a peak at $t^M = \frac{n}{r_m} + b_\infty$ and inflection points at $t = t^M \pm \frac{\sqrt{n}}{r_m}$. But this also means that for large n, $q_n(t)$ is non-zero only when the time is large. Therefore, for large n we can use the simplified rate $r_m(t - b_\infty)$, because for small times, where the approximation does not hold, $q_n(t)$ is in any case null. Fig F.3 shows $q_t(n)$ as function of time for different values of n and r_{off}. As discussed previously, the approximation (F.25) is expected to be good when $\int dt\, P_b(t)$ has converged to a constant b_∞ and from Fig F.1 this happens at $t = 3.5$ and $t = 14.8$ for respectively $r_{off} = 3$ and $r_{off} = 0.3$. In conclusion, we can write

$$q_t(n)|_{n>(t^*-b_\infty)r_m} \simeq \text{Pois}\left(r_m\left[t - b_\infty(r_{on},\, r_{off})\right]\right) \tag{F.26}$$

F.4. Time to reach a probability 0.5 of being unbound

To have an indication of the order of magnitude of t^*, we want to investigate at what time the promoter is unbound from LexA with a probability of 0.5. Intuitively, we expect this to happen when the binding and unbinding rate becomes equal, i.e. when $r_{on}e^{-t} = r_{off}$. Moreover, we notice that the first term in bracket in equation (F.21) is the integral of a curve increasing exponentially with time, while the second term decreases exponentially. Hence, for time large enough we can assume that the first term is dominating over the second. Using this assumption and imposing the condition $r_{on}e^{-t} = r_{off}$, the probability $f(t)$ of being unbound at time $t_{1/2} = \log\left(\frac{r_{on}}{r_{off}}\right)$ becomes

$$f\left(t = \frac{r_{on}}{r_{off}}\right) = r_{off}e^{r_{off}}r_{off}^{r_{off}}\Gamma\left(-r_{off}, r_{off}\right) \tag{F.27}$$

where we have rewritten the exponential integral in terms of the upper incomplete gamma function [70]. The upper gamma function can be expanded as [71]:

$$\Gamma(-r_{off}, r_{off}) \simeq r_{off}^{-r_{off}}e^{-r_{off}}\left[\frac{1}{2r_{off}} + \mathcal{O}\left(r_{off}^{-3}\right)\right] \tag{F.28}$$

Therefore equation (F.27) becomes

$$\lim_{r_{off}\to\infty} f\left(t = \frac{r_{on}}{r_{off}}\right) = \frac{1}{2} + \mathcal{O}\left(r_{off}^{-2}\right) \tag{F.29}$$

This shows that when the unbinding is fast (the timescale is set by the LexA decay rate), the probability of being unbound is 0.5 when the unbinding and binding rates are equal.

On the other hand, for very small unbinding rates, $f(t)$ can be expanded as

$$f(t) = r_{off}r_{on}^{r_{off}}e^{-r_{off}t}\Gamma\left(-r_{off}\right) + 1 \tag{F.30}$$

If we equate it to 0.5 we get that the time $t_{1/2}$ to be unbound with probability 50% is:

$$\lim_{r_{off}\to 0} t_{1/2} = \log(r_{on}) + \frac{\log(r_{off})}{r_{off}} + \log\left[-2\Gamma(-r_{off})\right]\frac{1}{r_{off}} \tag{F.31}$$

The time to reach $f(t) = 0.5$ with the fit of the fast and slow unbinding rates are show in Fig F.4.

More in general, if we are interested in knowing when $f(t) = p$, with a generic p, we can rewrite $f(t)$ in terms of a new variable $e^{\beta} = r_{on}e^{-}t$. Ignoring again the exponential decaying term containing the constant of integration, we get

$$f(\beta) = e^{\beta}r_{off}E_{1+r_{off}}(\beta) = p \tag{F.32}$$

where p is the target probability of being unbound that we want to reach. This equation cannot be solved analytically, but it shows that the time to reach a probability p of being unbound increases with the logarithm of r_{on}. In fact, let's call $\tilde{\beta}$ the value for which $f(\tilde{\beta}) = p$, then we have that $\tilde{\beta}$ is a decreasing function of only r_{off} and a probability of being unbound of p is reached at the time

$$\tilde{t} = \log r_{on} - \tilde{\beta}(r_{off}, p) \tag{F.33}$$

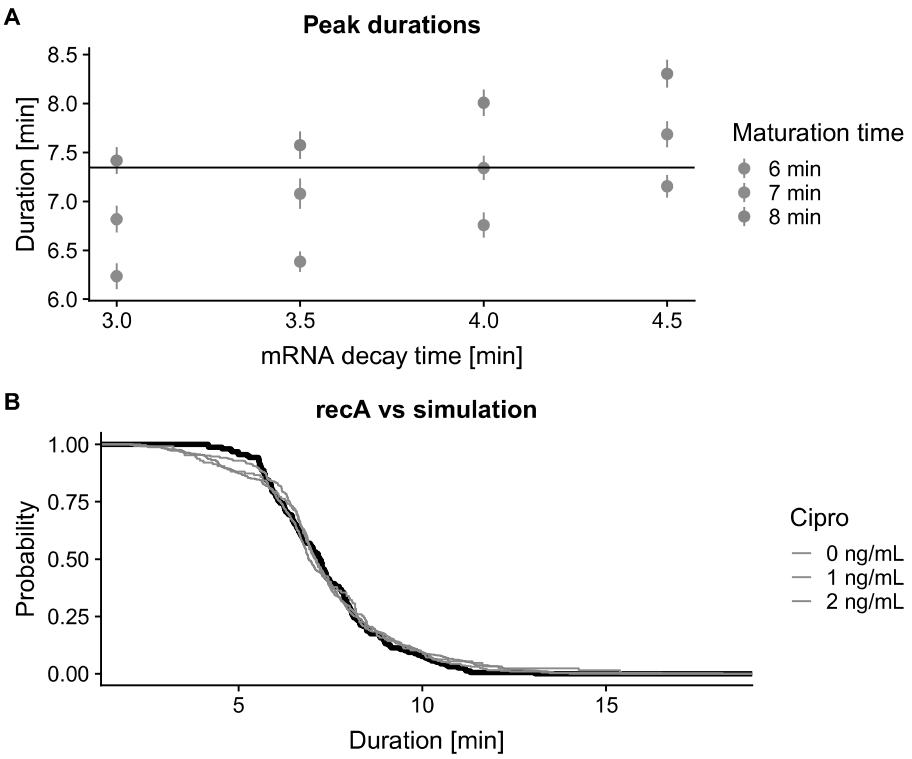

Figure E.1.: **Duration of peaks in GFP production from pulses of mRNA production.** *A*: We simulated a growing cell producing mRNA from very short breaks (repair rate = 0.9/min). We inferred the duration as the standard deviation of a Gaussian fitted to the peak, following the same procedure as for the experimental data. The horizontal line is the typical duration of the observed peaks in recA. *B*: Comparison of the widths distribution for the simulation (black) and for the measured recA in different conditions (colored lines), for mRNA decay rate $1/(4min)$ and GFP maturation rate $1/(7min)$.

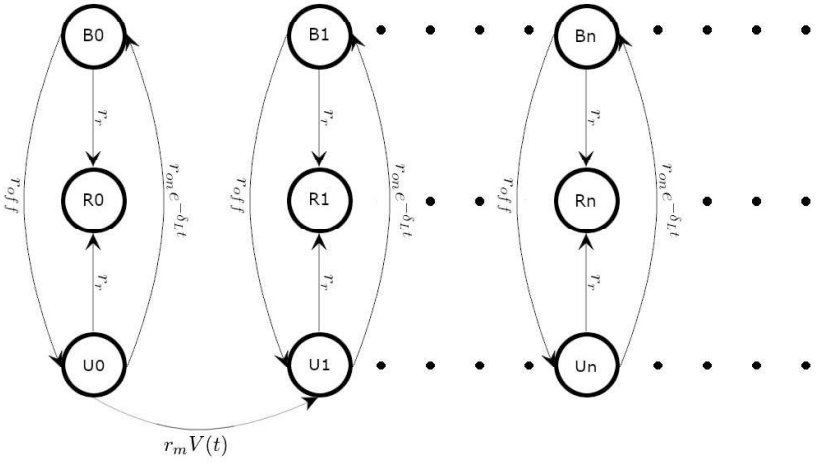

Figure E.2.: **Markov diagram of the mRNA transcription process**. The mRNA transcription process during a break is modeled as a chain of states. *Bn* is a state where the promoter is bound by LexA and n transcripts have been produced; *Un* is a state with the promoter free from LexA and n transcripts have been produced; *Rn* is a state where the break has been repaired, having produced n transcripts. Only the unbound state can produce mRNA and the repaired state is an absorbing state.

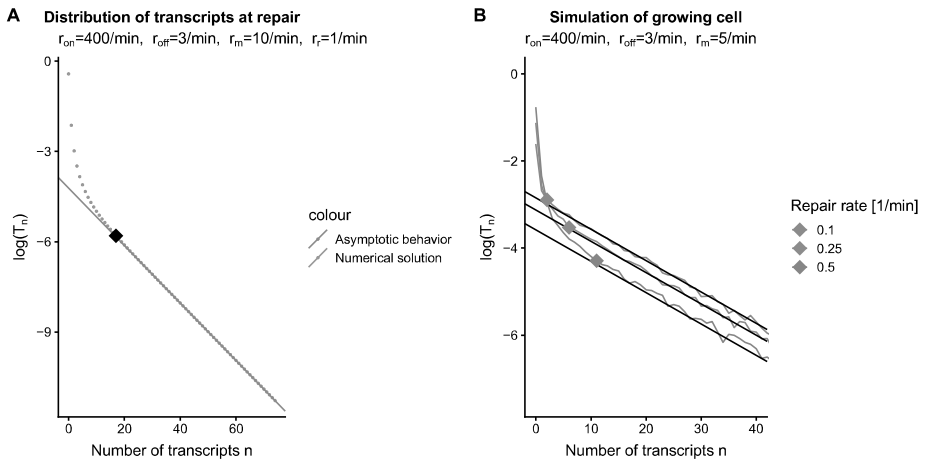

Figure E.3.: **Probability of n transcripts at repair**. *A*: In blue is the solution from the numeric integration of the dynamical system, in red is the geometric approximation valid for large n. The diamond shows the point n^* in equation (E.9) where the geometric approximation is predicted to deviate from the real solution. $t^* = 3.5$ min has been numerically computed (Appendix F.3). *B*: Same as for plot A, but from simulations of a growing cell, with breaks arrival at exponentially distributed times. Different colors show different repair rates, while the transcription rate is set such that the slope of the geometric approximation is constant.

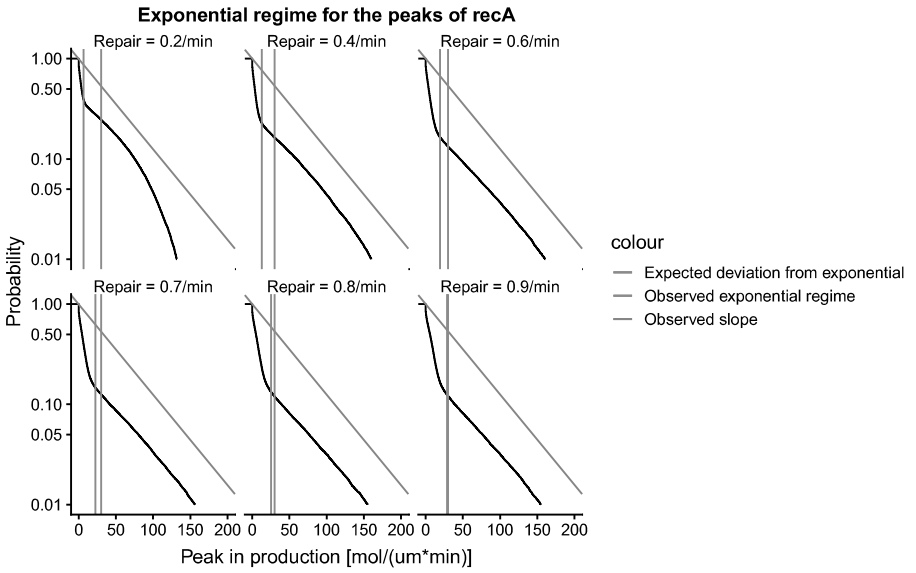

Figure E.4.: **Simulation of the recA promoter**. We show the distribution of peak heights obtained simulating a growing cell expressing from the recA promoter with different repair rates (sub-panels). The transcription rate has been set to match the slope observed experimentally (black line). The blue line shows the number of transcripts at which experimentally we are in the exponential regime, while the red line shows, for each repair rate, the point where we expect to deviate from the exponential regime.

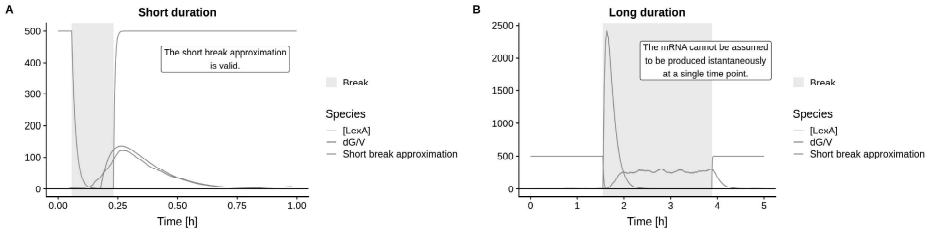

Figure E.5.: **Deviation from the short break approximation**. Example of mRNA and GFP production during breaks from the simulation of a growing cell. We show, the time of the break (shaded gray area), the LexA concentration (red curve), the observed GFP production (green) and the GFP production predicted if the mRNA was produced instantaneously (Supplementary equation (E.4), blue line). *A)*: For short breaks, the GFP predicted from the short break approximation agrees with the observed GFP production. *B)*: If the break lasts too long compared to the GFP production dynamics, the peak in GFP production deviates from the small break approximation. Notice the difference time scale in each plot.

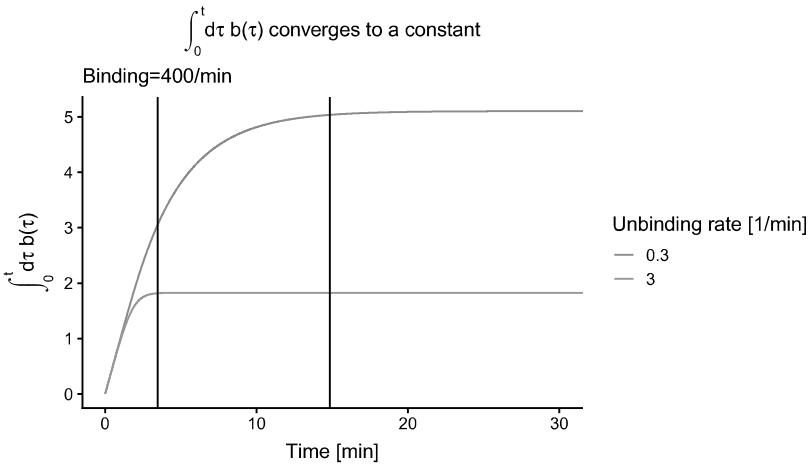

Figure F.1.: **Convergence of the probability $b(t)$ to be unbound**. Numerical computation of the integral $\int_0^t d\tau\, b(\tau)$ shows that it converges to a constant 1.83 min if $r_{off} = 3/\text{min}$ and 5.11 min if $r_{off} = 0.3/\text{min}$. The vertical lines represent the time when $f(t) = 0.98$ (3.5 min and 14.8 min) and it shows that the convergence of the integral is compatible with the convergence of $f(t)$.

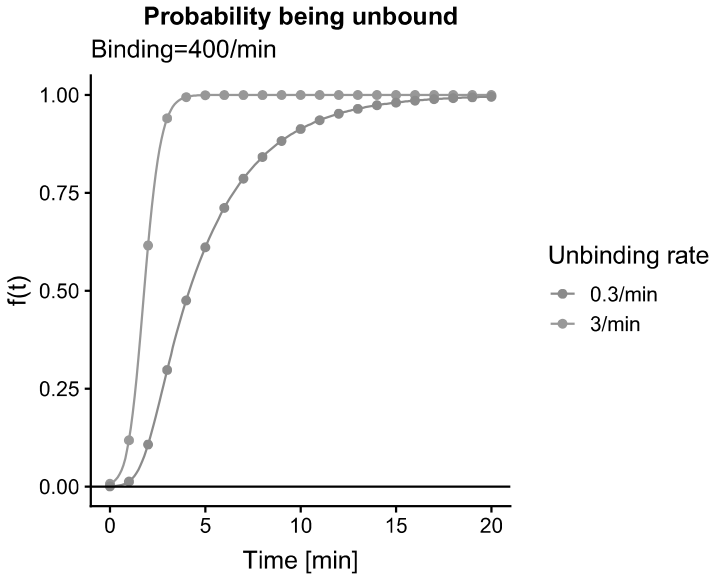

Figure F.2.: **Probability for the promoter to be unbound**. Example of the probability $f(t)$ to be unbound for two different unbinding rates. The solid curve shows the analytical solution equation (F.21), while the dots have been obtained by numerical integration of the dynamical system.

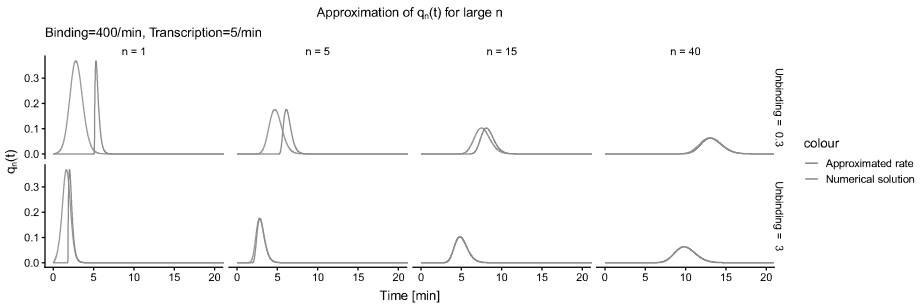

Figure F.3.: $q_n(t)$ **as function of** n. $q_n(t)$ for different values of n (columns) and r_{off} (rows). As n increases, $q_n(t)$ is concentrated at higher times, therefore it is well represented by a Poisson distribution with the approximated rate $r_m(t - b_\infty)$ (red curve). The blue curve shows the numerical solution using the full rate $r_m \int dt\, f(t)$.

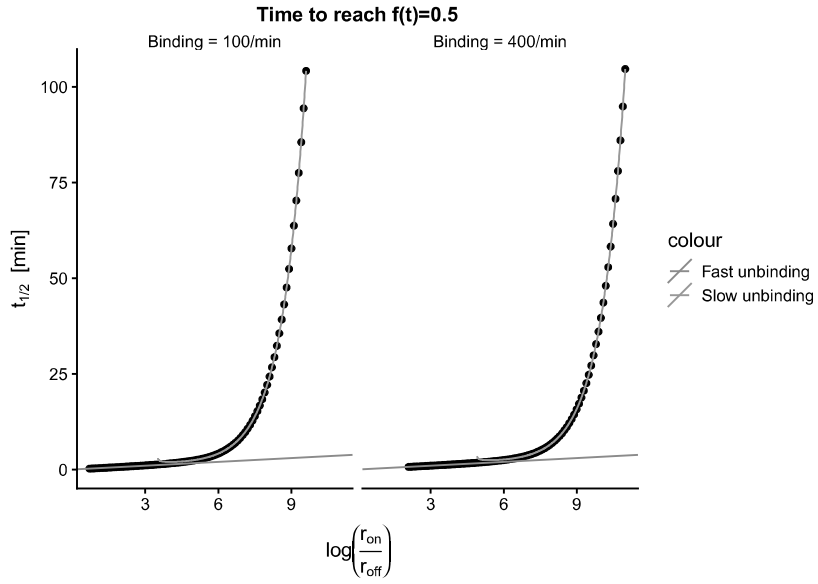

Figure F.4.: **Time to reach** $f(t) = p$. Numerical solution of the time to reach a probability 0.5 for the promoter of being unbound for two different unbinding rates. In red is the fit using the fast unbinding rate approximation; in this case $f(t) = 0.5$ when $r_{on} = r_{off}$. In blue is the fitting of slow unbinding rates.

Bibliography

[1] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins." In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356. ISSN: 0022-2836. DOI: 10.1016/S0022-2836(61)80072-7. URL: http://www.sciencedirect.com/science/article/pii/S0022283661800727 (cit. on p. 2).

[2] Mark Ptashne and Alexander Gann. *Genes & Signals*. CSHL Press, 2002 (cit. on p. 2).

[3] Albert Courey. *Mechanisms in Transcriptional Regulation*. Wiley-Blackwell, 2008 (cit. on p. 2).

[4] Avigdor Eldar and Michael B. Elowitz. "Functional roles for noise in genetic circuits." In: *Nature* 467.7312 (2010), pp. 167–173. DOI: 10.1038/nature09326. URL: https://doi.org/10.1038/nature09326 (cit. on p. 2).

[5] Arantxa Urchueguía et al. "Genome-wide gene expression noise in Escherichia coli is condition-dependent and determined by propagation of noise through the regulatory network." In: *PLOS Biology* 19 (Dec. 2021), pp. 1–22. DOI: 10.1371/journal.pbio.3001491. URL: https://doi.org/10.1371/journal.pbio.3001491 (cit. on p. 2).

[6] Luise Wolf, Olin K Silander, and Erik van Nimwegen. "Expression noise facilitates the evolution of gene regulation." In: *eLife* 4 (June 2015). Ed. by Ido Golding, e05856. ISSN: 2050-084X. DOI: 10.7554/eLife.05856. URL: https://doi.org/10.7554/eLife.05856 (cit. on pp. 2, 4, 5).

[7] Hernan G. Garcia et al. "Transcription by the numbers redux: experiments and calculations that surprise." In: *Trends in Cell Biology* 20 (6 Dec. 2010). DOI: 10.1016/j.tcb.2010.07.002. URL: https://doi.org/10.1016/j.tcb.2010.07.002 (cit. on p. 2).

[8] James C. W. Locke and Michael B. Elowitz. "Using movies to analyse gene circuit dynamics in single cells." In: *Nature Reviews Microbiology* 383.7 (May 2009). DOI: 10.1038/nrmicro2056. URL: https://doi.org/10.1038/nrmicro2056 (cit. on p. 2).

[9] Daniel R. Larson, Robert H. Singer, and Daniel Zenklusen. "A single molecule view of gene expression." In: *Trends in Cell Biology* 19 (11 Nov. 2009). DOI: 10.1016/j.tcb.2009.08.008. URL: https://doi.org/10.1016/j.tcb.2009.08.008 (cit. on p. 2).

[10] Nitzan Rosenfeld et al. "Gene Regulation at the Single-Cell Level." In: *Science* 307.5717 (2005), pp. 1962–1965. ISSN: 0036-8075. DOI: 10.1126/science.1106914. eprint: https://science.sciencemag.org/content/307/5717/1962.full.pdf. URL: https://science.sciencemag.org/content/307/5717/1962 (cit. on p. 2).

[11] Yuichi Taniguchi et al. "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells." In: *Science* 329.5991 (2010), pp. 533–538. ISSN: 0036-8075. DOI: 10.1126/science.1188308. eprint: https://science.sciencemag.org/content/329/5991/533.full.pdf. URL: https://science.sciencemag.org/content/329/5991/533 (cit. on pp. 2, 11, 12, 29).

[12] Ido Golding et al. "Real-Time Kinetics of Gene Activity in Individual Bacteria." In: *Cell* 123 (Dec. 2005). DOI: 10.1016/j.cell.2005.09.031. URL: https://doi.org/10.1016/j.cell.2005.09.031 (cit. on p. 2).

[13] Edouard Bertrand et al. "Localization of ASH1 mRNA Particles in Living Yeast." In: *Molecular Cell* 2 (4 Oct. 1998). DOI: 10.1016/S1097-2765(00)80143-4. URL: https://doi.org/10.1016/S1097-2765(00)80143-4 (cit. on p. 2).

[14] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. "Single-RNA counting reveals alternative modes of gene expression in yeast." In: *Nature Structural & Molecular Biology* 15 (12 Dec. 2008). DOI: 10.1038/nsmb.1514. URL: https://doi.org/10.1038/nsmb.1514 (cit. on p. 2).

[15] Lok-hang So et al. "General properties of transcriptional time series in Escherichia coli." In: *Nature Genetics* 43 (6 2011). DOI: 10.1038/ng.821. URL: https://doi.org/10.1038/ng.821 (cit. on p. 2).

[16] Nir Friedman et al. "Precise Temporal Modulation in the Response of the SOS DNA Repair Network in Individual Bacteria." In: *PLOS Biology* 3.7 (June 2005). DOI: 10.1371/journal.pbio.0030238. URL: https://doi.org/10.1371/journal.pbio.0030238 (cit. on pp. 2, 3).

[17] Muir Morrison, Manuel Razo-Mejia, and Rob Phillips. "Reconciling Kinetic and Equilibrium Models of Bacterial Transcription." In: *bioRxiv* (2020). DOI: 10.1101/2020.06.13.150292. eprint: https://www.biorxiv.org/content/early/2020/06/14/2020.06.13.150292.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/14/2020.06.13.150292 (cit. on pp. 2, 10).

[18] Juan M. Pedraza and Alexander van Oudenaarden. "Noise Propagation in Gene Networks." In: *Science* 307.5717 (2005), pp. 1965–1969. ISSN: 0036-8075. DOI: 10.1126/science.1109090. eprint: https://science.sciencemag.org/content/307/5717/1965.full.pdf. URL: https://science.sciencemag.org/content/307/5717/1965 (cit. on p. 2).

[19] Jeff Hasty et al. "Noise-based switches and amplifiers for gene expression." In: *Proceedings of the National Academy of Sciences* 97.5 (2000), pp. 2075–2080. ISSN: 0027-8424. DOI: 10.1073/pnas.040411297. eprint: https://www.pnas.org/content/97/5/2075.full.pdf. URL: https://www.pnas.org/content/97/5/2075 (cit. on p. 2).

[20] Erik Aurell et al. "Stability puzzles in phage λ." In: *Phys. Rev. E* 65 (5 May 2002), p. 051914. DOI: 10.1103/PhysRevE.65.051914. URL: https://link.aps.org/doi/10.1103/PhysRevE.65.051914 (cit. on p. 2).

[21] José M.G. Vilar, Călin C. Guet, and Stanislas Leibler. "Modeling network dynamics : the lac operon, a case study." In: *Journal of Cell Biology* 161.3 (May 2003), pp. 471–476. ISSN: 0021-9525. DOI: 10.1083/jcb.200301125. eprint: https://rupress.org/jcb/article-pdf/161/3/471/911832/jcb1613471.pdf. URL: https://doi.org/10.1083/jcb.200301125 (cit. on p. 2).

[22] Jerome T. Mettetal et al. "Predicting stochastic gene expression dynamics in single cells." In: *Proceedings of the National Academy of Sciences* 103.19 (2006), pp. 7304–7309. DOI: 10.1073/pnas.0509874103. eprint: https://www.pnas.org/content/103/19/7304.full.pdf. URL: https://www.pnas.org/content/103/19/7304 (cit. on p. 2).

[23] Michail Stamatakis and Nikos V. Mantzaris. "Comparison of Deterministic and Stochastic Models of the lac Operon Genetic Network." In: *Biophysical Journal* 96.3 (2009), pp. 887–906. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2008.10.028. URL: http://www.sciencedirect.com/science/article/pii/S0006349508001008 (cit. on p. 2).

[24] Lennart Hilbert, David Albrecht, and Michael C. Mackey. "Small delay, big waves: a minimal delayed negative feedback model captures Escherichia coli single cell SOS kinetics." In: *Mol. BioSyst.* 7 (9 2011), pp. 2599–2607. DOI: 10.1039/C1MB05122A. URL: http://dx.doi.org/10.1039/C1MB05122A (cit. on pp. 2, 3).

[25] Thomas Julou et al. "Subpopulations of sensorless bacteria drive fitness in fluctuating environments." In: *bioRxiv* (2020). DOI: 10.1101/2020.01.04.894766. eprint: https://www.biorxiv.org/content/early/2020/01/06/2020.01.04.894766.full.pdf. URL: https://www.biorxiv.org/content/early/2020/01/06/2020.01.04.894766 (cit. on p. 2).

[26] Zeynep Baharoglu and Didier Mazel. "SOS, the formidable strategy of bacteria against aggressions." In: *FEMS Microbiology Reviews* 38.6 (Nov. 2014), pp. 1126–1145. ISSN: 0168-6445. DOI: 10.1111/1574-6976.12077. eprint: https://academic.oup.com/femsre/article-pdf/38/6/1126/18142660/38-6-1126.pdf. URL: https://doi.org/10.1111/1574-6976.12077 (cit. on pp. 2, 3, 11).

[27] Bénédicte Michel. "After 30 Years of Study, the Bacterial SOS Response Still Surprises Us." In: *PLOS Biology* 3.7 (July 2005). DOI: 10.1371/journal.pbio.0030255. URL: https://doi.org/10.1371/journal.pbio.0030255 (cit. on pp. 2, 3).

[28] Mandana Sassanfar and Jeffrey W. Roberts. "Nature of the SOS-inducing signal in Escherichia coli: The involvement of DNA replication." In: *Journal of Molecular Biology* 212.1 (1990), pp. 79–96. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(90)90306-7. URL: http://www.sciencedirect.com/science/article/pii/0022283690903067 (cit. on pp. 3, 11, 12).

[29] Justin Courcelle et al. "Comparative Gene Expression Profiles Following UV Exposure in Wild-Type and SOS-Deficient Escherichia coli." In: *Genetics* 158.1 (2001), pp. 41–64. ISSN: 0016-6731. eprint: https://www.genetics.org/content/158/1/41.full.pdf. URL: https://www.genetics.org/content/158/1/41 (cit. on p. 3).

[30] Michal Ronen et al. "Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics." In: *Proceedings of the National Academy of Sciences* 99.16 (2002), pp. 10555–10560. ISSN: 0027-8424. DOI: 10.1073/pnas.152046799. eprint: https://www.pnas.org/content/99/16/10555.full.pdf. URL: https://www.pnas.org/content/99/16/10555 (cit. on p. 3).

[31] Matthew J. Culyba et al. "Non-equilibrium repressor binding kinetics link DNA damage dose to transcriptional timing within the SOS gene network." In: *PLOS Genetics* 14.6 (June 2018), pp. 1–29. DOI: 10.1371/journal.pgen.1007405. URL: https://doi.org/10.1371/journal.pgen.1007405 (cit. on pp. 3, 12, 33).

[32] Sandeep Krishna, Sergei Maslov, and Kim Sneppen. "UV-Induced Mutagenesis in Escherichia coli SOS Response: A Quantitative Model." In: *PLOS Computational Biology* 3.3 (Mar. 2007), pp. 1–12. DOI: 10.1371/journal.pcbi.0030041. URL: https://doi.org/10.1371/journal.pcbi.0030041 (cit. on p. 3).

[33] Yishai Shimoni et al. "Stochastic Analysis of the SOS Response in Escherichia coli." In: *PLOS ONE* 4.5 (May 2009), pp. 1–7. DOI: 10.1371/journal.pone.0005363. URL: https://doi.org/10.1371/journal.pone.0005363 (cit. on p. 3).

[34] K Drlica and X Zhao. "DNA gyrase, topoisomerase IV, and the 4-quinolones." In: *Microbiology and Molecular Biology Reviews* 61.3 (1997), pp. 377–392. ISSN: 1092-2172. eprint: https://mmbr.asm.org/content/61/3/377.full.pdf. URL: https://mmbr.asm.org/content/61/3/377 (cit. on p. 3).

[35] Alon Zaslaver et al. "A comprehensive library of fluorescent transcriptional reporters for Escherichia coli." eng. In: *Nature Methods* 3.8 (Aug. 2006), pp. 623–628. ISSN: 1548-7091. DOI: 10.1038/nmeth895 (cit. on pp. 3, 4).

[36] M. Kaiser et al. "Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software." In: *Nature Communications* 9 (Jan. 2018) (cit. on pp. 3–5, 9, 12).

[37] Socorro Gama-Castro et al. "RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond." eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D133–143. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1156 (cit. on pp. 4, 5).

[38] Enrique Balleza, J Mark Kim, and Philippe Cluzel. "Systematic characterization of maturation time of fluorescent proteins in living cells." In: *Nature Methods* (2018). DOI: 10.1038/nmeth.4509. URL: https://doi.org/10.1038/nmeth.4509 (cit. on pp. 9, 11, 12, 29).

[39] Mats Wallden et al. "The Synchronization of Replication and Division Cycles in Individual E.coli Cells." In: *Cell* 166.3 (2016), pp. 729–739. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2016.06.052. URL: http://www.sciencedirect.com/science/article/pii/S0092867416308601 (cit. on pp. 9, 27).

[40] Hernan G. Garcia et al. *A First Exposure to Statistical Mechanics for Life Scientists*. 2007. arXiv: 0708.1899 [q-bio.QM] (cit. on p. 10).

[41] Madeline A. Shea and Gary K. Ackers. "The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation." In: *Journal of Molecular Biology* 181.2 (1985), pp. 211–230. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(85)90086-5. URL: http://www.sciencedirect.com/science/article/pii/0022283685900865 (cit. on p. 10).

[42] G K Ackers, A D Johnson, and M A Shea. "Quantitative model for gene regulation by lambda phage repressor." In: *Proceedings of the National Academy of Sciences* 79.4 (1982), pp. 1129–1133. ISSN: 0027-8424. DOI: 10.1073/pnas.79.4.1129. eprint: https://www.pnas.org/content/79/4/1129.full.pdf. URL: https://www.pnas.org/content/79/4/1129 (cit. on p. 10).

[43] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. "On schemes of combinatorial transcription logic." In: *Proceedings of the National Academy of Sciences* 100.9 (2003), pp. 5136–5141. ISSN: 0027-8424. DOI: 10.1073/pnas.0930314100. eprint: https://www.pnas.org/content/100/9/5136.full.pdf. URL: https://www.pnas.org/content/100/9/5136 (cit. on p. 10).

[44] José M.G. Vilar and Stanislas Leibler. "DNA Looping and Physical Constraints on Transcription Regulation." In: *Journal of Molecular Biology* 331.5 (2003), pp. 981–989. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(03)00764-2. URL: http://www.sciencedirect.com/science/article/pii/S0022283603007642 (cit. on p. 10).

[45] Thomas Kuhlman et al. "Combinatorial transcriptional control of the lactose operon of Escherichia coli." In: *Proceedings of the National Academy of Sciences* 104.14 (2007), pp. 6043–6048. ISSN: 0027-8424. DOI: 10.1073/pnas.0606717104. eprint: https://www.pnas.org/content/104/14/6043.full.pdf. URL: https://www.pnas.org/content/104/14/6043 (cit. on p. 10).

[46] Robert Daber, Matthew A. Sochor, and Mitchell Lewis. "Thermodynamic Analysis of Mutant lac Repressors." In: *Journal of Molecular Biology* 409.1 (2011). The Operon Model and its Impact on Modern Molecular Biology, pp. 76–87. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2011.03.057. URL: http://www.sciencedirect.com/science/article/pii/S0022283611003469 (cit. on p. 10).

[47] Hernan G. Garcia and Rob Phillips. "Quantitative dissection of the simple repression input–output function." In: *Proceedings of the National Academy of Sciences* 108.29 (2011), pp. 12173–12178. ISSN: 0027-8424. DOI: 10.1073/pnas.1015616108. eprint: https://www.pnas.org/content/108/29/12173.full.pdf. URL: https://www.pnas.org/content/108/29/12173 (cit. on p. 10).

[48] Robert C. Brewster et al. "The Transcription Factor Titration Effect Dictates Level of Gene Expression." In: *Cell* 156.6 (2014), pp. 1312–1323. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2014.02.022. URL: http://www.sciencedirect.com/science/article/pii/S0092867414002219 (cit. on p. 10).

[49] Manuel Razo-Mejia et al. "Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction." In: *Cell Systems* 6 (4 2018). DOI: doi:10.1016/j.cels.2018.02.004. URL: https://doi.org/10.1016/j.cels.2018.02.004 (cit. on p. 10).

[50] Justin B. Kinney et al. "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence." In: *Proceedings of the National Academy of Sciences* 107.20 (2010), pp. 9158–9163. ISSN: 0027-8424. DOI: 10.1073/pnas.1004290107. eprint: https://www.pnas.org/content/107/20/9158.full.pdf. URL: https://www.pnas.org/content/107/20/9158 (cit. on p. 10).

[51] Rob Phillips. "Napoleon Is in Equilibrium." In: *Annual Review of Condensed Matter Physics* 6.1 (2015), pp. 85–111. DOI: 10.1146/annurev-conmatphys-031214-014558. eprint: https://doi.org/10.1146/annurev-conmatphys-031214-014558. URL: https://doi.org/10.1146/annurev-conmatphys-031214-014558 (cit. on p. 10).

[52] Lacramioara Bintu et al. "Transcriptional regulation by the numbers: models." In: *Current Opinion in Genetics & Development* 15.2 (2005). Chromosomes and expression mechanisms, pp. 116–124. ISSN: 0959-437X. DOI: https://doi.org/10.1016/j.gde.2005.02.007. URL: http://www.sciencedirect.com/science/article/pii/S0959437X05000304 (cit. on p. 10).

[53] Lacramioara Bintu et al. "Transcriptional regulation by the numbers: applications." In: *Current Opinion in Genetics & Development* 15.2 (2005). Chromosomes and expression mechanisms, pp. 125–135. ISSN: 0959-437X. DOI: https://doi.org/10.1016/j.gde.2005.02.006. URL: http://www.sciencedirect.com/science/article/pii/S0959437X05000298 (cit. on p. 10).

[54] María Tamayo et al. "Rapid assessment of the effect of ciprofloxacin on chromosomal DNA from Escherichia coli using an in situ DNA fragmentation assay." In: *BMC Microbiology* 9 (Apr. 2009). DOI: 10.1186/1471-2180-9-69. URL: https://doi.org/10.1186/1471-2180-9-69 (cit. on p. 11).

[55] Katie J. Aldred, Robert J. Kerns, and Neil Osheroff. "Mechanism of Quinolone Action and Resistance." In: *Biochemistry* 53.10 (2014). PMID: 24576155, pp. 1565–1574. DOI: 10.1021/bi5000564. eprint: https://doi.org/10.1021/bi5000564. URL: https://doi.org/10.1021/bi5000564 (cit. on p. 11).

[56] M. Butala et al. "The bacterial LexA transcriptional repressor." In: *Cellular and Molecular Life Sciences* (2008). DOI: 10.1007/s00018-008-8378-6. URL: https://doi.org/10.1007/s00018-008-8378-6 (cit. on p. 11).

[57] Patrice L. Moreau. "Effects of overproduction of single-stranded DNA-binding protein on RecA protein-dependent processes in Escherichia coli." In: *Journal of Molecular Biology* 194.4 (1987), pp. 621–634. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(87)90239-7. URL: https://www.sciencedirect.com/science/article/pii/0022283687902397 (cit. on p. 11).

[58] Gene-Wei Li, Otto G. Berg, and Johan Elf. "Effects of macromolecular crowding and DNA looping on gene regulation kinetics." In: *Nature Physics* 5 (2009), pp. 294–297. DOI: 10.1038/nphys1222 (cit. on p. 11).

[59] Petter Hammar et al. "The lac Repressor Displays Facilitated Diffusion in Living Cells." In: *Science* 336.6088 (2012), pp. 1595–1598. ISSN: 0036-8075. DOI: 10.1126/science.1221648. eprint: https://science.sciencemag.org/content/336/6088/1595.full.pdf. URL: https://science.sciencemag.org/content/336/6088/1595 (cit. on p. 11).

[60] Judith A. Megerle et al. "Timing and Dynamics of Single Cell Gene Expression in the Arabinose Utilization System." In: *Biophysical Journal* 95.4 (2008), pp. 2103–2115. ISSN: 0006-3495. DOI: https://doi.org/10.1529/biophysj.107.127191. URL: http://www.sciencedirect.com/science/article/pii/S0006349508701681 (cit. on pp. 11, 12, 29).

[61] Marlena Siwiak and Piotr Zielenkiewicz. "Transimulation - Protein Biosynthesis Web Service." In: *PLOS ONE* 8.9 (Sept. 2013), pp. 1–8. DOI: 10.1371/journal.pone.0073943. URL: https://doi.org/10.1371/journal.pone.0073943 (cit. on p. 12).

[62] David Kennell and Howard Riezman. "Transcription and translation initiation frequencies of the Escherichia coli lac operon." In: *Journal of Molecular Biology* 114.1 (1977), pp. 1–21. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(77)90279-0. URL: http://www.sciencedirect.com/science/article/pii/0022283677902790 (cit. on p. 12).

[63] Doron Levin and Tamir Tuller. "Genome-Scale Analysis of Perturbations in Translation Elongation Based on a Computational Model." In: *Scientific Reports* (2018). DOI: https://doi.org/10.1038/s41598-018-34496-3. URL: 10.1038/s41598-018-34496-3 (cit. on p. 12).

[64] Namiko Mitarai, Kim Sneppen, and Steen Pedersen. "Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization." In: *Journal of Molecular Biology* 382.1 (2008), pp. 236–245. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2008.06.068. URL: http://www.sciencedirect.com/science/article/pii/S002228360800795X (cit. on p. 12).

[65] R. Khanin, V. Vinciotti, and E. Wit. "Reconstructing repressor protein levels from expression of gene targets in Escherichia coli." In: *Proceedings of the National Academy of Sciences* 103.49 (2006), pp. 18592–18596. ISSN: 0027-8424. DOI: 10.1073/pnas.0603390103. eprint: https://www.pnas.org/content/103/49/18592.full.pdf. URL: https://www.pnas.org/content/103/49/18592 (cit. on p. 12).

[66] Yi Luo et al. "Nucleosomes accelerate transcription factor dissociation." In: *Nucleic Acids Research* 42.5 (Dec. 2013), pp. 3017–3027. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1319. eprint: https://academic.oup.com/nar/article-pdf/42/5/3017/28910228/gkt1319.pdf. URL: https://doi.org/10.1093/nar/gkt1319 (cit. on pp. 12, 33).

[67] F. Kuhner et al. "LexA-DNA Bond Strength by Single Molecule Force Spectroscopy." In: *Biophysical Journal* 87.4 (2004), pp. 2683–2690. ISSN: 0006-3495. DOI: https://doi.org/10.1529/biophysj.104.048868. URL: http://www.sciencedirect.com/science/article/pii/S000634950473739X (cit. on p. 12).

[68] Henrik Flyvbjerg and H.G. Petersen. "Error Estimates on Averages of Correlated Data." In: *The Journal of Chemical Physics* 91 (July 1989). DOI: 10.1063/1.457480 (cit. on p. 17).

[69] Daniel T. Gillespie. "Exact stochastic simulation of coupled chemical reactions." In: *The Journal of Physical Chemistry* 81.25 (1977), pp. 2340–2361. DOI: 10.1021/j100540a008. eprint: https://doi.org/10.1021/j100540a008. URL: https://doi.org/10.1021/j100540a008 (cit. on p. 28).

[70] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. ninth Dover printing, tenth GPO printing. New York: Dover, 1964 (cit. on pp. 32, 36, 38).

[71] Gergő Nemes. "The Resurgence Properties of the Incomplete Gamma Function II." In: *Studies in Applied Mathematics* 135.1 (2015), pp. 86–116. DOI: 10.1111/sapm.12077. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sapm.12077. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sapm.12077 (cit. on p. 38).

5. Thermodynamic model in the presence of multiple regulatory sites

MotEvo is a software that infers the position and energy of binding sites for a specific transcription factor, given a promoter sequence ([88]). Using the thermodynamic approach described in Sec 2.2.1, we extended the functionality of MotEvo to predict the probability for a promoter to be in the active state in the presence of multiple σ_{70} and repressor binding sites. The idea is that the probability for the gene to be active is the fraction of times at least one polymerase is bound downstream all repressors. We assume that this scenario applies for the SOS pathway, where we have to consider the binding sites for the repressor LexA and binding sites for the polymerase and σ_{70} holoenzyme.

First of all, we start from the output of MotEvo to get the number, positions and energies of the binding sites. For illustrative purposes, let's consider a simple promoter as in Fig 5.1, which has 3 σ_{70} and 2 LexA binding sites of different energies and sizes. The strongest LexA is on the negative strand. The solid ellipses and rectangles show the presence of the σ_{70} or LexA on a specific binding site. To apply the thermodynamic model, we need to enumerate all possible configurations, keeping in mind that two transcription factors cannot occupy two overlapping binding sites at the same time. Then, we have to check which configurations are active, that is which configurations have at least one σ_{70} bound upstream all the LexA. Finally, we need to multiply the Boltzmann factors of the active configurations and divide by the Boltzmann factors of all the configurations.

Let's call $\{\alpha\beta\}$ a specific active configuration with α σ_{70} and β LexA bound to the promoter. The Boltzmann factor associated with $\{\alpha\beta\}$ is

$$W^A_{\{\alpha\beta\}} = \exp\left[-\sum_{i\in\{\alpha\}}\Delta G_i - \sum_{j\in\{\beta\}}\Delta G_j\right] \tag{5.1}$$

Each configuration with α σ_{70} and β LexA bound is multiplied by $[\sigma_{70}]^\alpha[L]^\beta$, where $[\sigma_{70}]$ and $[L]$ are the concentration of polymerase and LexA respectively. Summing over all possible active configurations, we have that the unnormalized probability to be active is

$$Z_A = \sum_{\alpha,\beta}[\sigma_{70}]^\alpha[L]^\beta W^A_{\alpha,\beta} \tag{5.2}$$

where $W^A_{\alpha,\beta} \equiv \sum_{\{\alpha\beta\}} W^A_{\{\alpha\beta\}}$. It follows that the normalized probability to be in the active state is

$$P = \frac{Z_A}{Z_A + Z_I} \tag{5.3}$$

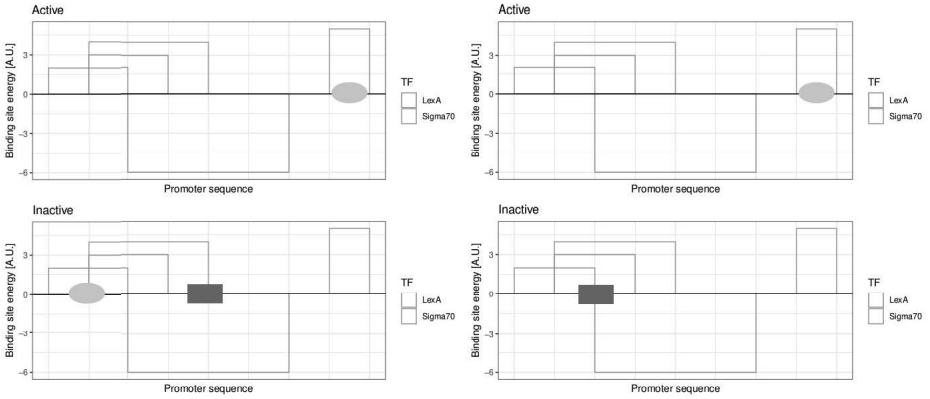

Figure 5.1.: Example of a simple promoter architecture. The wired rectangles show the different binding sites, with blue representing the polymerase holoenzyme and red the repressor LexA. The height of the rectangles corresponds to the binding energy. Negative energies show that the binding site is on the reverse strand. The solid ellipses and rectangles show where the σ_{70} and LexA are binding. The picture shows 4 possible configurations, but only the ones on the top row are active, i.e. they have a σ_{70} not blocked by LexA.

where Z_I is the analogous of Z_A for the inactive configurations.

Let's compute the probability to be active using the simple architecture in Fig 5.1. Let's suppose that the energies of the binding sites from left to right are $\log(2)$, $\log(3)$, $\log(4)$ for the polymerase and $\log(6)$, $\log(5)$ for LexA. Then the matrix $W^A_{\alpha\beta}$ is

$$W^A_{\alpha\beta} = \begin{pmatrix} 0 & 0 & 0 \\ 9 & 44 & 0 \\ 20 & 40 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.4}$$

For example, the 0^{th} row is the sum of the Boltzmann factors of the active configurations with 0 polymerases, but without polymerases there are no active configurations. Therefore, the first row is always 0. $W_{2,0} = 20$, this is because we have two configurations with two polymerases and no LexA, i.e. the polymerase bound to the first and third site, with Boltzmann factor $2 \cdot 4$, and the polymerase on the second and third sites, with Boltzmann factor $3 \cdot 4$. Notice that the configuration with the polymerase on the first and second sites is not physically possible because the two sites overlap. The total Boltzmann factor is then $2 \cdot 4 + 3 \cdot 4 = 20$. In the same way, the matrix $W^I_{\alpha\beta}$ for the inactive configurations is

$$W^I_{\alpha\beta} = \begin{pmatrix} 1 & 11 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.5}$$

The probability to be in an active configuration if $[\sigma_{70}] = 1.5$ and $[LexA] = 0.5$ is, from equation (5.3),

$$P = \frac{(1.5^0, 1.5^1, 1.5^2, 1.5^3)\, W^A\, (0.5^0, 0.5^1, 0.5^2)^T}{(1.5^0, 1.5^1, 1.5^2, 1.5^3)\, (W^A + W^I)\, (0.5^0, 0.5^1, 0.5^2)^T} \tag{5.6}$$

Inserting the values of the matrices, we finally get $P = 0.9$. In the Appendix we show how the implementation of the model can be optimized using low level operations in C++, resulting in a fast algorithm.

6. Summary and future perspectives

6.1. Summary

Gene regulation is a key mechanism for the survival and evolution of organisms. It is also of fundamental importance in clinical research, since it can improve our ability to use bacteria to synthesize therapeutic proteins, increase the efficacy of current therapies, and improve our understanding of diseases. Although the mechanisms have been studied since the 60s, there are still many open questions. In particular, the many players and the complex regulatory network topologies make it hard to come up with a mathematical model complex enough to give reliable predictions about the amount of protein produced as a function of the input signal, and simple enough to provide useful insights on the complex mechanisms of regulation. In this work, I contributed to improving our understanding of gene regulation, both from data analysis and theoretical modeling point of view.

Relatively recent advancements in experimental techniques are allowing us to collect detailed description of gene expression in high throughput. One attracting technology is flow cytometry, originally developed for large eukaryotic cells, and now routinely applied to report single cell distribution of gene expression. However, data analysis must be done carefully, using proper statistical analysis. Comparing measurements of *E. coli* strains carrying fluorescent reporters in both flow cytometry and microscopic setups, we proposed a principled and systematic way of carrying out analysis of flow cytometry data. In particular, we showed that the expression of proteins at low copy numbers, typical of bacteria, results in a large component of the fluorescence signal to come from autofluorescence and electronic shot noise. The R package *E-Flow* enables us to remove these spurious components and to extract biologically relevant results. Also, while for eukaryotic cells is an established practice to use forward and side scatter to identify cells of different sizes and granularity, there seems to be little agreement in the literature on how to use scatter for discerning small bacterial cells. We showed that there is little correlation between scatter signal and viability of the cells, and we showed that a broad filter that discards only a small fraction of the cells is to prefer to stricter filterings sometimes used in the literature. Finally, although the flow cytometer measures total fluorescence, it is preferable to measure fluorescence concentration, as this does not strongly depend on the cell size. We could prove that, although scatter signals carry information about the average cell size of a population, extracting the size of the single cells is much trickier. Nevertheless, if an independent estimation of the distribution of sizes in the population of cells is available, then measurements of total fluorescence from the flow cytometer can be used to get an estimate of the concentration distribution.

Having a lot of data is not enough to extract meaningful insights. A fundamental step is to describe the data with a mathematical model able to highlight the relevant relationships between the properties of the system and its behavior. For gene regulation, thermodynamic models provide a powerful tool that enabled scientific community to gain meaningful insights on the mechanisms of LacZ expression [57–65] and λ phage induction [54, 66]. However, for a general network, it is still not clear which simplifying assumptions we are allowed to make, and how to mathematically characterize the many interacting players involved in gene regulation. We showed that for the well-known SOS response in *E. coli* a simple equilibrium model is not enough to describe the behavior of the network, and fluctuations in the regulator LexA must be explicitly taken into account in the mathematical model. We proposed a non-equilibrium model and we showed that it is able to recapitulate the experimental data, using biological relevant values for the different parameters. We also investigated how the cells can exploit the non-equilibrium dynamics to tune the promoter architecture in order to either change the responsivness to the external environment, or the amount of noise propagated from the regulator to the target gene.

6.2. Future perspectives

The work on the SOS regulatory network showed the importance of having time resolution data of single-cell gene expression data to understand the mechanisms of gene regulation. Building upon what has been done on this thesis, we see four possible follow-up studies.

First, it would be interesting to simultaneously measure the expression dynamics of two LexA target promoters. I am convinced that the correlation structure among the production peaks of different promoters will provide important insights into the gene regulation mechanism. Moreover, it has been proposed that a regulatory strategy used by the cell is to regulate the activation time of different genes in the network, and some works have shown that the SOS response is indeed characterized by "early" activated genes and "late" genes[38]. Having the possibility to study the GFP production of two promoters at the same time, can shed light on the mechanisms that lead to different activation times. The technology that allows for such an experiment can be the dual-color Mother Machine, but although already developed, some challenges still have to be overcome. From a theoretical point of view, the most compelling one is that the breaks have been shown to be of very short duration, and the largest part of the peak in production is determined by GFP folding time, more than by the mRNA transcription. In our study, we used a fast-folding mutant of GFP, but for tracking the expression of a second gene we need a fluorescent report expressing in a channel far away from the one of GFP. However, it is difficult to find such reporters with a fast-folding time, and a longer folding time will smooth out the mRNA production signal, losing in resolution. Therefore, the comparison between the peaks in production of this slow folding reporter with the peaks in GFP must be done in a careful way.

Another follow-up project is to use a different induction strategy. In this thesis, we constantly induced the network by giving a constant amount of antibiotic

Ciprofloxacin. However, it would be also interesting to expose the cells to a high UV radiation at the very beginning of the experiment and observe how the network induction evolves over time once the UV source is turned off. This has been shown to have a very interesting dynamic behavior [33], although the study was based on time snapshots of gene expression distribution in a population of cells, rather than on time tracking of single cells. Preliminary experiments in our lab, have shown that with our Mother Machine setup it should be possible to study the network behavior under UV irradiation.

From a more general theoretical point of view, it is extremely interesting to ask why the cells have evolved a non-equilibrium dynamics. It is worth to explore what the advantages of having a non-equilibrium dynamics are and how they depend on the external environment.

Finally, although we measured gene expression using fluorescent reporters, it would be highly informative to directly look at single mRNA molecules. A technique called single molecule Fluorescence In Situ Hybridization (smFISH) [89] allows us to directly visualize single molecules of mRNA, and it has been recently applied by the lab of Golding [89] to study the distribution of transcripts in *E. coli*. Building on some trials already made in our lab, it would be beneficial to further develop this technology for bacteria and to be able to integrate it with the Mother Machine.

Considering projects outside the SOS regulatory network, an interesting theoretical question is how gene regulation is optimized by the cells. During the transition to another food source, bacteria need a random amount of time to adapt to the new environment, resulting in a lag phase of no growth. In a study from our lab, it has been shown how single-cell lag time affects the fitness of the entire population, and under which circumstances it is beneficial for the population to reduce the lag phase. Building upon this, it would be interesting to investigate how the cells can optimize the concentration of a regulator. Let's suppose an inhibitor is bound to DNA with energy E. This means that we need to wait on average a time e^E before it falls off. For example, for the Lac operon it means that when lactose first enters the cell, the machinery is not immediately induced, but there will be an exponentially distributed waiting time. This results in a lag time in the single-cell growth, which in turn affects the population growth. The cell can reduce the lag time by decreasing the binding E, but this will require producing more repressor LacI in order to have the same repression power when no lactose is present. This in turn reduces the fitness (growth rate) of the single cell. On the other hand, at the population level, a higher LacI binding energy means that the growth of the cell when it wakes up it's going to be fast since the cells spend less energy in producing TFs. Therefore, high binding energies means low levels of inhibitor and fast growth, but also large lag times for the single cells. On the contrary, low binding energies means high levels of inhibitor and slower growth, but faster wake up times. If the population is very large, it seems reasonable to imagine that the first strategy is favorable, because the waiting times are exponentially distributed and in a very large population we can expect to find a cell that wakes up immediately and starts to reproduce exponentially. But in a small population, if the mean waiting time is large, it is less likely to find a cell that wakes up early. It is therefore interesting to explore what the trade-off between high and

low TF levels is, and what it depends on (e.g. population size or external conditions).

Appendix

Appendix A.

Implementation of the combinatorial thermodynamic model

Due to the large amount of computations required (the number of configurations to consider grows exponentially with the number n of binding sites as 2^n) the code to compute the probability of being in a transcriptionally active state has been written in C++ using low-level bit-wise operations to optimise the combinatorial analysis. Using the simplified architecture of Fig 5.1 and focusing only on σ_{70}, we can write the state of the promoter with only one polymerase bound as a string of bits, where 0 means that the specific base is unbound and 1 means bound

$$\begin{array}{c} 11000000 \\ 01110000 \\ 00000001 \end{array} \tag{A.1}$$

To each string is associated a weight W_{config} which is the Boltzmann factor of the specific configuration.

Since we have three binding sites, we have $2^3 = 8$ possible combinations of bound σ_{70}, which can be written as a string of three bits

$$\begin{array}{c} 000 \\ 001 \\ 010 \\ 011 \\ \vdots \\ 111 \end{array}$$

For example, the string 011 means that the first site is unbound, while the second and the third ones are bound. Notice that the number of possible configurations is simply 2^n, which is how many binary numbers can be represented with n bits.

Let's consider the binding configuration 011. This means that the polymerase is bound on the second and third site, therefore the promoter state is given by 01110001. How do we build this string?

1. We can start from the null sequence of bits $c = 00000000$, with a Boltzmann factor $B = 1$ (nothing is bound to the promoter).
2. We loop through all the bites of the binding configuration 011.
3. For the i-th bite b_i we construct the new string $c' = c \,|\, (b_i \,\&\, c[i])$. Where $c[i]$ is the i-th string in (A.1), describing the promoter with the i-th binding site occupied. We multiply the current Boltzmann factor by $B' = B \times B_i^{b_i}$, where B_i is the Boltzmann factor of the i-th string in (A.1).

One caveat is that we need to consider the binding configuration only if it doesn't have any molecules bound to overlapping sites. So, for example, 110 is not valid. To check for this, every time we loop on the bites we check $c \,\&\, (b_i \,\&\, c[i])$. If this is different from zero, then there is an overlap and we discard the binding configuration. Notice that if b_i is zero we are not adding any polymerase and the Boltzmann factor doesn't increase.

The previous algorithm gives the set of all valid promoter configurations with some σ_{70} bound. Let's generalise it to the case where there is also a repressor. The first step is to compute the possible valid binding configurations, separately for the σ_{70} and the inhibitor. Once we have the possible configurations for the inhibitor and the polymerase and the corresponding Boltzmann factors, we can build the matrices $W^A_{\alpha\beta}$ and $W^I_{\alpha\beta}$.

But now we have to check which configurations of σ_{70} and inhibitor are compatible among each other, i.e. they don't give a repressor overlapping a σ_{70}. Let's call c_s and c_L the promoter states with respectively the polymerases and the inhibitors bound. To check that there are no overlapping molecules, we use the same algorithm used above to check the overlapping of the different σ_{70}s. Once we know that the combination of bound LexA and σ_{70} is compatible we need to check whether there is at least a polymerase at the right of all the LexA. We can do that by considering the position of the right-most set bit in c_L and setting do 0 all the bits to the left of this position in c_s. If the resulting c_s is different from 0, then the combination of c_i and c_s leads to an active promoter. This gives all the compatible configurations of the active promoters, together with their Boltzmann factors and allows to build the matrices $W^A_{\alpha\beta}$ and $W^I_{\alpha\beta}$.

All the above operations can be achieved by bit-wise operations, making the algorithm particularly optimised.

Bibliography

[1] *Oxford English Dictionary*. Oxford University Press, 2020. URL: https://oed.com/view/Entry/162352?redirectedFrom=Renaissance+man#eid26055692 (cit. on p. 1).

[2] Yuri Lazebnik. "Can a biologist fix a radio?—Or, what I learned while studying apoptosis." In: *Cancer Cell* 2.3 (2002), pp. 179–182. ISSN: 1535-6108. DOI: https://doi.org/10.1016/S1535-6108(02)00133-2. URL: http://www.sciencedirect.com/science/article/pii/S1535610802001332 (cit. on p. 1).

[3] Erwin Schrödinger. *What is Life?* Cambridge University Press, 1944 (cit. on p. 1).

[4] Augustinus von Hippo. *Confessions*. Vol. XI. XIV. 397AD-401AD (cit. on p. 2).

[5] Daniel E. Koshland. "The Seven Pillars of Life." In: *Science* 295.5563 (2002), pp. 2215–2216. ISSN: 0036-8075. DOI: 10.1126/science.1068489. eprint: https://science.sciencemag.org/content/295/5563/2215.full.pdf. URL: https://science.sciencemag.org/content/295/5563/2215 (cit. on p. 2).

[6] Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. "Distribution of Initiation Times Reveals Mechanisms of Transcriptional Regulation in Single Cells." In: *Biophysical Journal* 114.9 (2018), pp. 2072–2082. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2018.03.031. URL: http://www.sciencedirect.com/science/article/pii/S0006349518304077 (cit. on p. 2).

[7] Hernan G. Garcia et al. "Transcription by the numbers redux: experiments and calculations that surprise." In: *Trends in Cell Biology* 20 (6 Dec. 2010). DOI: 10.1016/j.tcb.2010.07.002. URL: https://doi.org/10.1016/j.tcb.2010.07.002 (cit. on p. 2).

[8] Roger Bumgarner. "Overview of DNA Microarrays: Types, Applications, and Their Future." In: *Current Protocols in Molecular Biology* 101.1 (2013), pp. 22.1.1–22.1.11. DOI: 10.1002/0471142727.mb2201s101. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb2201s101. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb2201s101 (cit. on p. 2).

[9] M. Kaiser et al. "Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software." In: *Nature Communications* 9 (Jan. 2018) (cit. on p. 2).

[10] James C. W. Locke and Michael B. Elowitz. "Using movies to analyse gene circuit dynamics in single cells." In: *Nature Reviews Microbiology* 383.7 (May 2009). DOI: 10.1038/nrmicro2056. URL: https://doi.org/10.1038/nrmicro2056 (cit. on p. 2).

[11] Daniel R. Larson, Robert H. Singer, and Daniel Zenklusen. "A single molecule view of gene expression." In: *Trends in Cell Biology* 19 (11 Nov. 2009). DOI: 10.1016/j.tcb.2009.08.008. URL: https://doi.org/10.1016/j.tcb.2009.08.008 (cit. on p. 2).

[12] Nitzan Rosenfeld et al. "Gene Regulation at the Single-Cell Level." In: *Science* 307.5717 (2005), pp. 1962–1965. ISSN: 0036-8075. DOI: 10.1126/science.1106914. eprint: https://science.sciencemag.org/content/307/5717/1962.full.pdf. URL: https://science.sciencemag.org/content/307/5717/1962 (cit. on p. 2).

[13] Yuichi Taniguchi et al. "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells." In: *Science* 329.5991 (2010), pp. 533–538. ISSN: 0036-8075. DOI: 10.1126/science.1188308. eprint: https://science.sciencemag.org/content/329/5991/533.full.pdf. URL: https://science.sciencemag.org/content/329/5991/533 (cit. on pp. 2, 13).

[14] Ido Golding et al. "Real-Time Kinetics of Gene Activity in Individual Bacteria." In: *Cell* 123 (Dec. 2005). DOI: 10.1016/j.cell.2005.09.031. URL: https://doi.org/10.1016/j.cell.2005.09.031 (cit. on p. 2).

[15] Edouard Bertrand et al. "Localization of ASH1 mRNA Particles in Living Yeast." In: *Molecular Cell* 2 (4 Oct. 1998). DOI: 10.1016/S1097-2765(00)80143-4. URL: https://doi.org/10.1016/S1097-2765(00)80143-4 (cit. on p. 2).

[16] Daniel Zenklusen, Daniel R Larson, and Robert H Singer. "Single-RNA counting reveals alternative modes of gene expression in yeast." In: *Nature Structural & Molecular Biology* 15 (12 Dec. 2008). DOI: 10.1038/nsmb.1514. URL: https://doi.org/10.1038/nsmb.1514 (cit. on p. 2).

[17] Luca Galbusera et al. "Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria." In: *bioRxiv* (2019). DOI: 10.1101/793976. eprint: https://www.biorxiv.org/content/early/2019/10/05/793976.full.pdf. URL: https://www.biorxiv.org/content/early/2019/10/05/793976 (cit. on p. 2).

[18] Arantxa Urchueguia et al. "Noise propagation shapes condition-dependent gene expression noise in Escherichia coli." In: *bioRxiv* (2019). DOI: 10.1101/795369. eprint: https://www.biorxiv.org/content/early/2019/10/07/795369.full.pdf. URL: https://www.biorxiv.org/content/early/2019/10/07/795369 (cit. on pp. 2, 4).

[19] J. D. Chung et al. "Gene expression in single cells of Bacillus subtilis: evidence that a threshold mechanism controls the initiation of sporulation." In: *Journal of bacteriology* 176 (7 1994). DOI: 10.1128/jb.176.7.1977-1984.1994 (cit. on p. 2).

[20] R. H. Valdivia and S. Falkow. "Bacterial genetics by flow cytometry: rapid isolation of Salmonella typhimurium acid-inducible promoters by differential fluorescence induction." In: *Mol. Microbiol.* 22.2 (Oct. 1996), pp. 367–378 (cit. on p. 2).

[21] R. L. Wilson et al. "Identification of Listeria monocytogenes in vivo-induced genes by fluorescence-activated cell sorting." In: *Infect. Immun.* 69.8 (Aug. 2001), pp. 5016–5024 (cit. on p. 2).

[22] E. M. Ozbudak et al. "Regulation of noise in the expression of a single gene." In: *Nat. Genet.* 31.1 (May 2002), pp. 69–73 (cit. on p. 2).

[23] K. Hakkila et al. "Monitoring promoter activity in a single bacterial cell by using green and red fluorescent proteins." In: *J. Microbiol. Methods* 54.1 (July 2003), pp. 75–79 (cit. on p. 2).

[24] Yanina Sevastsyanovich et al. "Exploitation of GFP fusion proteins and stress avoidance as a generic strategy for the production of high-quality recombinant proteins." In: *FEMS Microbiology Letters* 299.1 (2009), pp. 86–94. DOI: 10.1111/j.1574-6968.2009.01738.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1574-6968.2009.01738.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-6968.2009.01738.x (cit. on p. 2).

[25] J. B. Kinney et al. "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence." In: *Proc. Natl. Acad. Sci. U.S.A.* 107.20 (May 2010), pp. 9158–9163 (cit. on p. 2).

[26] Olin K. Silander et al. "A Genome-Wide Analysis of Promoter-Mediated Phenotypic Noise in Escherichia coli." In: *PLOS Genetics* 8.1 (Jan. 2012), pp. 1–13. DOI: 10.1371/journal.pgen.1002443. URL: https://doi.org/10.1371/journal.pgen.1002443 (cit. on p. 2).

[27] Luise Wolf, Olin K Silander, and Erik van Nimwegen. "Expression noise facilitates the evolution of gene regulation." In: *eLife* 4 (June 2015). Ed. by Ido Golding, e05856. ISSN: 2050-084X. DOI: 10.7554/eLife.05856. URL: https://doi.org/10.7554/eLife.05856 (cit. on pp. 2, 4).

[28] Anil Ozdemir and Angelike Stathopoulos. "Exciting times: bountiful data to facilitate studies of cis-regulatory control." In: *Nature Methods* 8.12 (Dec. 2011), pp. 1016–1017. DOI: 10.1038/nmeth.1795. URL: https://doi.org/10.1038/nmeth.1795 (cit. on p. 3).

[29] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins." In: *Journal of Molecular Biology* 3.3 (1961), pp. 318–356. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(61)80072-7. URL: http://www.sciencedirect.com/science/article/pii/S0022283661800727 (cit. on pp. 3, 6).

[30] Alberto Santos-Zavaleta et al. "RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12." In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp. D212–D220. ISSN: 0305-1048. DOI: 10.1093/nar/gky1077. eprint: https://academic.oup.com/nar/article-pdf/47/D1/D212/27437535/gky1077.pdf. URL: https://doi.org/10.1093/nar/gky1077 (cit. on p. 3).

[31] Gabriela G. Loots. "Chapter 10 Genomic Identification of Regulatory Elements by Evolutionary Sequence Comparison and Functional Analysis." In: *Long-Range Control of Gene Expression*. Vol. 61. Advances in Genetics. Academic Press, 2008, pp. 269–293. DOI: https://doi.org/10.1016/S0065-2660(07)00010-7. URL: http://www.sciencedirect.com/science/article/pii/S0065266007000107 (cit. on p. 3).

[32] Rob Phillips. "Napoleon Is in Equilibrium." In: *Annual Review of Condensed Matter Physics* 6.1 (2015), pp. 85–111. DOI: 10.1146/annurev-conmatphys-031214-014558. eprint: https://doi.org/10.1146/annurev-conmatphys-031214-014558. URL: https://doi.org/10.1146/annurev-conmatphys-031214-014558 (cit. on p. 4).

[33] Nir Friedman et al. "Precise Temporal Modulation in the Response of the SOS DNA Repair Network in Individual Bacteria." In: *PLOS Biology* 3.7 (June 2005). DOI: 10.1371/journal.pbio.0030238. URL: https://doi.org/10.1371/journal.pbio.0030238 (cit. on pp. 4, 128).

[34] Muir Morrison, Manuel Razo-Mejia, and Rob Phillips. "Reconciling Kinetic and Equilibrium Models of Bacterial Transcription." In: *bioRxiv* (2020). DOI: 10.1101/2020.06.13.150292. eprint: https://www.biorxiv.org/content/early/2020/06/14/2020.06.13.150292.full.pdf. URL: https://www.biorxiv.org/content/early/2020/06/14/2020.06.13.150292 (cit. on p. 4).

[35] Martin Ackermann. "A functional perspective on phenotypic heterogeneity in microorganisms." In: *Nature Reviews Microbiology* 13 (Aug. 2015). DOI: 10.1038/nrmicro3491. URL: https://doi.org/10.1038/nrmicro3491 (cit. on p. 4).

[36] Edo Kussell and Stanislas Leibler. "Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments." In: *Science* 309.5743 (2005), pp. 2075–2078. ISSN: 0036-8075. DOI: 10.1126/science.1114383. eprint: https://science.sciencemag.org/content/309/5743/2075.full.pdf. URL: https://science.sciencemag.org/content/309/5743/2075 (cit. on p. 4).

[37] Avigdor Eldar and Michael B. Elowitz. "Functional roles for noise in genetic circuits." In: *Nature* 467 (Sept. 2010). DOI: 10.1038/nature09326. URL: https://doi.org/10.1038/nature09326 (cit. on p. 4).

[38] Bénédicte Michel. "After 30 Years of Study, the Bacterial SOS Response Still Surprises Us." In: *PLOS Biology* 3.7 (July 2005). DOI: 10.1371/journal.pbio.0030255. URL: https://doi.org/10.1371/journal.pbio.0030255 (cit. on pp. 4, 15, 127).

[39] Matthew J. Culyba et al. "Non-equilibrium repressor binding kinetics link DNA damage dose to transcriptional timing within the SOS gene network." In: *PLOS Genetics* 14.6 (June 2018), pp. 1–29. DOI: 10.1371/journal.pgen.1007405. URL: https://doi.org/10.1371/journal.pgen.1007405 (cit. on p. 4).

[40] Sandeep Krishna, Sergei Maslov, and Kim Sneppen. "UV-Induced Mutagenesis in Escherichia coli SOS Response: A Quantitative Model." In: *PLOS Computational Biology* 3.3 (Mar. 2007), pp. 1–12. DOI: 10.1371/journal.pcbi.0030041. URL: https://doi.org/10.1371/journal.pcbi.0030041 (cit. on p. 4).

[41] Yishai Shimoni et al. "Stochastic Analysis of the SOS Response in Escherichia coli." In: *PLOS ONE* 4.5 (May 2009), pp. 1–7. DOI: 10.1371/journal.pone.0005363. URL: https://doi.org/10.1371/journal.pone.0005363 (cit. on p. 4).

[42] Lennart Hilbert, David Albrecht, and Michael C. Mackey. "Small delay, big waves: a minimal delayed negative feedback model captures Escherichia coli single cell SOS kinetics." In: *Mol. BioSyst.* 7 (9 2011), pp. 2599–2607. DOI: 10.1039/C1MB05122A. URL: http://dx.doi.org/10.1039/C1MB05122A (cit. on p. 4).

[43] Katsuhiko S Murakami and Seth A Darst. "Bacterial RNA polymerases: the wholo story." In: *Current Opinion in Structural Biology* 13.1 (2003), pp. 31–39. ISSN: 0959-440X. DOI: https://doi.org/10.1016/S0959-440X(02)00005-2. URL: http://www.sciencedirect.com/science/article/pii/S0959440X02000052 (cit. on pp. 5, 6).

[44] David J. Lee, Stephen D. Minchin, and Stephen J.W. Busby. "Activating Transcription in Bacteria." In: *Annual Review of Microbiology* 66.1 (2012). PMID: 22726217, pp. 125–152. DOI: 10.1146/annurev-micro-092611-150012. eprint: https://doi.org/10.1146/annurev-micro-092611-150012. URL: https://doi.org/10.1146/annurev-micro-092611-150012 (cit. on pp. 5, 6).

[45] Douglas F. Browning and Stephen J. W. Busby. "Local and global regulation of transcription initiation in bacteria." In: *Nature Reviews Microbiology* 14 (2016). DOI: 10.1038/nrmicro.2016.103. URL: https://doi.org/10.1038/nrmicro.2016.103 (cit. on pp. 5, 6).

[46] Sergei Borukhov and Evgeny Nudler. "RNA polymerase: the vehicle of transcription." In: *Trends in Microbiology* 16.3 (2008), pp. 126–134. ISSN: 0966-842X. DOI: https://doi.org/10.1016/j.tim.2007.12.006. URL: http://www.sciencedirect.com/science/article/pii/S0966842X08000334 (cit. on pp. 6, 7).

[47] Ahmet Ay and David N. Arnosti. "Mathematical modeling of gene expression: a guide for the perplexed biologist." In: *Critical Reviews in Biochemistry and Molecular Biology* 46.2 (2011), pp. 137–151. DOI: 10.3109/10409238.2011.556597. eprint: https://doi.org/10.3109/10409238.2011.556597. URL: https://doi.org/10.3109/10409238.2011.556597 (cit. on p. 7).

[48] Moises Santillan and Michael C Mackey. "Quantitative approaches to the study of bistability in the lac operon of Escherichia coli." In: *Journal of The Royal Society Interface* 5.suppl_1 (2008), S29–S39. DOI: 10.1098/rsif.2008.0086.focus. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2008.0086.focus. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2008.0086.focus (cit. on p. 7).

[49] Amos B. Oppenheim et al. "Switches in Bacteriophage Lambda Development." In: *Annual Review of Genetics* 39.1 (2005). PMID: 16285866, pp. 409–429. DOI: 10.1146/annurev.genet.39.073003.113656. eprint: https://doi.org/10.1146/annurev.genet.39.073003.113656. URL: https://doi.org/10.1146/annurev.genet.39.073003.113656 (cit. on p. 7).

[50] Terrell L. Hill. *An introduction to statistical thermodynamics*. Dover Publications, 1986 (cit. on pp. 8, 9).

[51] Josiah Willard Gibbs. *Elementary Principles in Statistical Mechanics*. Charles Scribner's Sons, 1902 (cit. on p. 8).

[52] Georgy Lebon and David Jou. *Understanding Non-equilibrium Thermodynamics*. Springer-Verlag Berlin Heidelberg, 2008 (cit. on p. 8).

[53] E. T. Jaynes. "Information Theory and Statistical Mechanics." In: *The Physical Review* 106.4 (1957) (cit. on p. 9).

[54] G K Ackers, A D Johnson, and M A Shea. "Quantitative model for gene regulation by lambda phage repressor." In: *Proceedings of the National Academy of Sciences* 79.4 (1982), pp. 1129–1133. ISSN: 0027-8424. DOI: 10.1073/pnas.79.4.1129. eprint: https://www.pnas.org/content/79/4/1129.full.pdf. URL: https://www.pnas.org/content/79/4/1129 (cit. on pp. 9–11, 127).

[55] Lacramioara Bintu et al. "Transcriptional regulation by the numbers: models." In: *Current Opinion in Genetics & Development* 15.2 (2005). Chromosomes and expression mechanisms, pp. 116–124. ISSN: 0959-437X. DOI: https://doi.org/10.1016/j.gde.2005.02.007. URL: http://www.sciencedirect.com/science/article/pii/S0959437X05000304 (cit. on p. 9).

[56] Hernan G. Garcia et al. "Operator Sequence Alters Gene Expression Independently of Transcription Factor Occupancy in Bacteria." In: *Cell Reports* 2.1 (2012), pp. 150–161. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2012.06.004. URL: http://www.sciencedirect.com/science/article/pii/S2211124712001647 (cit. on p. 11).

[57] Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa. "On schemes of combinatorial transcription logic." In: *Proceedings of the National Academy of Sciences* 100.9 (2003), pp. 5136–5141. ISSN: 0027-8424. DOI: 10.1073/pnas.0930314100. eprint: https://www.pnas.org/content/100/9/5136.full.pdf. URL: https://www.pnas.org/content/100/9/5136 (cit. on pp. 11, 127).

[58] José M.G. Vilar and Stanislas Leibler. "DNA Looping and Physical Constraints on Transcription Regulation." In: *Journal of Molecular Biology* 331.5 (2003), pp. 981–989. ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(03)00764-2. URL: http://www.sciencedirect.com/science/article/pii/S0022283603007642 (cit. on pp. 11, 127).

[59] Thomas Kuhlman et al. "Combinatorial transcriptional control of the lactose operon of Escherichia coli." In: *Proceedings of the National Academy of Sciences* 104.14 (2007), pp. 6043–6048. ISSN: 0027-8424. DOI: 10.1073/pnas.0606717104. eprint: https://www.pnas.org/content/104/14/6043.full.pdf. URL: https://www.pnas.org/content/104/14/6043 (cit. on pp. 11, 127).

[60] Robert Daber, Matthew A. Sochor, and Mitchell Lewis. "Thermodynamic Analysis of Mutant lac Repressors." In: *Journal of Molecular Biology* 409.1 (2011). The Operon Model and its Impact on Modern Molecular Biology, pp. 76–87. ISSN: 0022-2836. DOI: https://doi.org/10.1016/j.jmb.2011.03.057. URL: http://www.sciencedirect.com/science/article/pii/S0022283611003469 (cit. on pp. 11, 127).

[61] Hernan G. Garcia and Rob Phillips. "Quantitative dissection of the simple repression input–output function." In: *Proceedings of the National Academy of Sciences* 108.29 (2011), pp. 12173–12178. ISSN: 0027-8424. DOI: 10.1073/pnas.1015616108. eprint: https://www.pnas.org/content/108/29/12173.full.pdf. URL: https://www.pnas.org/content/108/29/12173 (cit. on pp. 11, 127).

[62] Robert C. Brewster et al. "The Transcription Factor Titration Effect Dictates Level of Gene Expression." In: *Cell* 156.6 (2014), pp. 1312–1323. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2014.02.022. URL: http://www.sciencedirect.com/science/article/pii/S0092867414002219 (cit. on pp. 11, 127).

[63] Manuel Razo-Mejia et al. "Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction." In: *Cell Systems* 6 (4 2018). DOI: doi:10.1016/j.cels.2018.02.004. URL: https://doi.org/10.1016/j.cels.2018.02.004 (cit. on pp. 11, 127).

[64] Justin B. Kinney et al. "Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence." In: *Proceedings of the National Academy of Sciences* 107.20 (2010), pp. 9158–9163. ISSN: 0027-8424. DOI: 10.1073/pnas.1004290107. eprint: https://www.pnas.org/content/107/20/9158.full.pdf. URL: https://www.pnas.org/content/107/20/9158 (cit. on pp. 11, 127).

[65] Robert C. Brewster, Daniel L. Jones, and Rob Phillips. "Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli." In: *PLOS Computational Biology* 8.12 (Dec. 2012), pp. 1–10. DOI: 10.1371/journal.pcbi.1002811. URL: https://doi.org/10.1371/journal.pcbi.1002811 (cit. on pp. 11, 127).

[66] Madeline A. Shea and Gary K. Ackers. "The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation." In: *Journal of Molecular Biology* 181.2 (1985), pp. 211–230. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(85)90086-5. URL: http://www.sciencedirect.com/science/article/pii/0022283685900865 (cit. on pp. 11, 14, 127).

[67] Moisés Santillán and Michael C. Mackey. "Why the Lysogenic State of Phage λ Is So Stable: A Mathematical Modeling Approach." In: *Biophysical Journal* 86.1 (2004), pp. 75–84. ISSN: 0006-3495. DOI: https://doi.org/10.1016/S0006-3495(04)74085-0. URL: http://www.sciencedirect.com/science/article/pii/S0006349504740850 (cit. on pp. 11, 14).

[68] C. Gardiner. *Stochastic Methods*. Springer-Verlag, 2009 (cit. on pp. 12, 14).

[69] J. Peccoud and B. Ycart. "Markovian Modeling of Gene-Product Synthesis." In: *Theoretical Population Biology* 48.2 (1995), pp. 222–234. ISSN: 0040-5809. DOI: https://doi.org/10.1006/tpbi.1995.1027. URL: http://www.sciencedirect.com/science/article/pii/S0040580985710271 (cit. on p. 13).

[70] Johan Paulsson. "Models of stochastic gene expression." In: *Physics of Life Reviews* 2.2 (2005), pp. 157–175. ISSN: 1571-0645. DOI: https://doi.org/10.1016/j.plrev.2005.03.003. URL: http://www.sciencedirect.com/science/article/pii/S1571064505000138 (cit. on p. 13).

[71] Daniel T. Gillespie. "Exact stochastic simulation of coupled chemical reactions." In: *The Journal of Physical Chemistry* 81.25 (1977), pp. 2340–2361. DOI: 10.1021/j100540a008. eprint: https://doi.org/10.1021/j100540a008. URL: https://doi.org/10.1021/j100540a008 (cit. on p. 14).

[72] Erik Aurell et al. "Stability puzzles in phage λ." In: *Phys. Rev. E* 65 (5 May 2002), p. 051914. DOI: 10.1103/PhysRevE.65.051914. URL: https://link.aps.org/doi/10.1103/PhysRevE.65.051914 (cit. on p. 14).

[73] Jeff Hasty et al. "Noise-based switches and amplifiers for gene expression." In: *Proceedings of the National Academy of Sciences* 97.5 (2000), pp. 2075–2080. ISSN: 0027-8424. DOI: 10.1073/pnas.040411297. eprint: https://www.pnas.org/content/97/5/2075.full.pdf. URL: https://www.pnas.org/content/97/5/2075 (cit. on p. 14).

[74] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. 2nd ed. Cambridge University Press, 2017. DOI: 10.1017/9781108499996 (cit. on p. 15).

[75] Miroslav Radman. "SOS Repair Hypothesis: Phenomenology of an Inducible DNA Repair Which is Accompanied by Mutagenesis." In: *Molecular Mechanisms for Repair of DNA: Part A*. Ed. by Philip C. Hanawalt and Richard B. Setlow. Boston, MA: Springer US, 1975, pp. 355–367. ISBN: 978-1-4684-2895-7. DOI: 10.1007/978-1-4684-2895-7_48. URL: https://doi.org/10.1007/978-1-4684-2895-7_48 (cit. on p. 15).

[76] Miroslav Radman. "SOS Hypothesis and the Emergence of Integrative Biology." In: *Supramolecular Structure and Function 9*. Ed. by Greta Pifat-Mrzljak. Dordrecht: Springer Netherlands, 2007, pp. 307–313. ISBN: 978-1-4020-6466-1. DOI: 10.1007/978-1-4020-6466-1_16. URL: https://doi.org/10.1007/978-1-4020-6466-1_16 (cit. on p. 15).

[77] Zeynep Baharoglu and Didier Mazel. "SOS, the formidable strategy of bacteria against aggressions." In: *FEMS Microbiology Reviews* 38.6 (2014), pp. 1126–1145. DOI: 10.1111/1574-6976.12077. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1574-6976.12077. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1574-6976.12077 (cit. on p. 15).

[78] Justin Courcelle et al. "Comparative Gene Expression Profiles Following UV Exposure in Wild-Type and SOS-Deficient Escherichia coli." In: *Genetics* 158.1 (2001), pp. 41–64. ISSN: 0016-6731. eprint: https://www.genetics.org/content/158/1/41.full.pdf. URL: https://www.genetics.org/content/158/1/41 (cit. on p. 15).

[79] Jason C. Bell and Stephen C. Kowalczykowski. "RecA: Regulation and Mechanism of a Molecular Search Engine." In: *Trends in Biochemical Sciences* 41.6 (2016), pp. 491–507. ISSN: 0968-0004. DOI: https://doi.org/10.1016/j.tibs.2016.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0968000416300056 (cit. on pp. 15, 16).

[80] Katarzyna H. Maslowska, Karolina Makiela-Dzbenska, and Iwona J. Fijalkowska. "The SOS system: A complex and tightly regulated response to DNA damage." In: *Environmental and Molecular Mutagenesis* 60.4 (2019), pp. 368–384. DOI: 10.1002/em.22267. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/em.22267. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/em.22267 (cit. on p. 16).

[81] Kevin Hiom. "Homologous Recombination: How RecA Finds the Perfect Partner." In: *Current Biology* 22.8 (2012), R275–R278. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2012.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0960982212002588 (cit. on p. 16).

[82] P H von Hippel and O G Berg. "Facilitated target location in biological systems." In: *Journal of Biological Chemistry* 264.2 (1989), pp. 675–678. eprint: http://www.jbc.org/content/264/2/675.full.pdf+html. URL: http://www.jbc.org/content/264/2/675.abstract (cit. on p. 16).

[83] P. Howard-Flanders, R. P. Boyce, and L Theriot. "Three Loci in ESCHERICHIA COLI K-12 That Control the Excision of Pyrimidine Dimers and Certain Other Mutagen Products from DNA." In: *Genetics* 53.6 (June 1966) (cit. on p. 16).

[84] M. Butala, D. Žgur-Bertok, and S. J. W. Busby. "The bacterial LexA transcriptional repressor." In: *Cellular and Molecular Life Sciences* 66.1 (Aug. 2008), p. 82. ISSN: 1420-9071. DOI: 10.1007/s00018-008-8378-6. URL: https://doi.org/10.1007/s00018-008-8378-6 (cit. on p. 16).

[85] Matej Butala et al. "Interconversion between bound and free conformations of LexA orchestrates the bacterial SOS response." In: *Nucleic Acids Research* 39.15 (May 2011), pp. 6546–6557. ISSN: 0305-1048. DOI: 10.1093/nar/gkr265. eprint: http://oup.prod.sis.lan/nar/article-pdf/39/15/6546/16775565/gkr265.pdf. URL: https://doi.org/10.1093/nar/gkr265 (cit. on p. 16).

[86] Lidija Kovačič et al. "Structural insight into LexA–RecA* interaction." In: *Nucleic Acids Research* 41.21 (Aug. 2013), pp. 9901–9910. ISSN: 0305-1048. DOI: 10.1093/nar/gkt744. eprint: http://oup.prod.sis.lan/nar/article-pdf/41/21/9901/16805251/gkt744.pdf. URL: https://doi.org/10.1093/nar/gkt744 (cit. on p. 16).

[87] Mandana Sassanfar and Jeffrey W. Roberts. "Nature of the SOS-inducing signal in Escherichia coli: The involvement of DNA replication." In: *Journal of Molecular Biology* 212.1 (1990), pp. 79–96. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(90)90306-7. URL: http://www.sciencedirect.com/science/article/pii/0022283690903067 (cit. on p. 16).

[88] Phil Arnold et al. "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences." In: *Bioinformatics* 28.4 (Dec. 2011), pp. 487–494. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr695. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/28/4/487/16910366/btr695.pdf. URL: https://doi.org/10.1093/bioinformatics/btr695 (cit. on pp. 17, 123).

[89] Samuel O Skinner et al. "Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization." In: *Nature Protocols* 8 (June 2013). DOI: 10.1038/nprot.2013.066. URL: https://doi.org/10.1038/nprot.2013.066 (cit. on p. 128).

Acknowledgments

During these years in Basel, I was lucky enough to met many interesting people, both inside and outside the University, and all of them have been foundamental for the success of my PhD.

I am grateful to my supervisor, Erik van Nimwegen, who gave me the opportunity to join his lab and initiated me to the world of computational biology. I admired his way of doing science, and his passion spanning different fields of knowledge. I enjoyed any single discussion about physics, biology, phylosophy, and many other fields. Doing a PhD in his lab was a great opportunity to learn how science is done, and helped me shaping my analytical and critical mind.

I would also like to thank Thomas Julou, who introduced me to the experimental side of biology. I always appreciated your ability to master both experimental work and theoretical analysis. I am also grateful for your invaluable insights on data visualization and analysis and the support you gave me during my projects.

A special thanks to Arantxa, I could not thank you enough for your support, not only inside the lab, but also outside. Thanks also to Athos, Dorde and Gwendoline, I consider you not only as nice colleagues, but also as good friends. And a big thanks to all the people that I met in van Nimwegen's lab. I have always had a great time with all of you guys.

Finally, I would also like to thank Richard and Marco for having accepted to be part of my Ph.D. committee.

Curriculum Vitae

Luca Galbusera

Physicist, computational biologist

Computational biologists with a **physics** background. I enjoy diversity in **cross-functional teams** that are not afraid of tacking on **challenging** tasks and **responsibilities**. I like to develop **software solutions** to advance projects with a **great** degree of **purpose**.

Personal Data

Phone:	+41 764001621
E-mail:	lucagalbu@gmail.com
Nationality:	Italy
Residence:	Switzerland
Work permit:	B permit (CH)

Internet

- LinkedIn
https://www.linkedin.com/in/luca-galbusera/
- GitHub
https://github.com/lucagalbu

Key Skills

- *Data science*: Computational biology, statistics, mathematics, physics, machine learning, NLP.
- *Programming*: C++, Python, R, CUDA, Boost, STL, Bash, git, Object oriented.
- *Soft skills*: Reliable, collaboration, communication, driven to learn, written and verbal English, proactivity, problem-solving, prioritization.

Languages

Italian	☒☒☒☒	Native
English	☒☒☒☒	C1/C2
German	☒☒☒☐	C1
French	☒☒☒☐	C1
Russian	☒☐☐☐	A1/A2

Interests

Cosmology, Dancing, Acting, Languages, 3D Art.

Work Experience

2019 – 2020 **Data scientist** at **Nestle**, Lausanne (CH)

DATA-DRIVEN SOLUTIONS FOR THE R&D AND BUSINESS

- I leveraged data science to drive the development of high priority Nestle projects.
- I deployed Shiny apps to support R&D analyses.
- I planned and supported food safety studies to ensure the compliance with company standards.

Research Experience

2015 – 2020 **PhD student in Bioinformatics** at **University of Basel**, Basel (CH)

MATHEMATICAL MODELING OF GENE REGULATION

- I applied theoretical modeling to understand the complexity of bacterial gene regulation.
- I wrote packages in R and C++ to appropriately interpret high-throughput biological data.
- Improve existing software features of SWISS-MODEL.

2013 – 2014 **Data Scientist at Institute for Oncology Research**, Bellinzona (CH)

IDENTIFICATION OF MOLECULAR BIOMARKERS

- I implemented a meta-analytic approach to identify a robust set of molecular biomarkers for lymphoma.
- I identified a set of molecular biomarkers and I used them to build a classifier for different lymphoma subgroups.
- I supported biologists and medical doctors to analyze omics data.

Academic degrees

2020 (November)

PhD in bioinformatics - University of Basel (CH)

2011-2014

Master in physics - University of Milan/Bicocca (IT)

Personal development

- **Communication** trainings in Italian, French, German.
- **Teaching** experience for high school and bachelor students.
- **Project management and marketing** for cultural associations in Switzerland and Italy.