Stochastic gene expression and lag time in bacteria

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Athos Fiori

Basel, 2021

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Erik Van Nimwegen
Prof. Dr. Richard Neher
Prof. Dr. Zoltán Kutalik

Basel, den 17.11.2020

Prof. Dr. Martin Spiess
Dekan

Abstract

The survival of organisms in randomly fluctuating environments not only depends on their ability to grow in different conditions but also on the time needed to adapt to each new habitat. Recent works had shown that, like many other physiological quantities, the adaptation time fluctuates in a stochastic manner across single cells and that the underlying distribution can dramatically change across genotypes. To understand how natural selection may have acted on the distribution of single-cell lags we develop a mathematical theory of how the single-cell lag distribution determines the reproductive success at the population level. We show that lags at the population level are exponentially dominated by the shortest lags at the individual cell level. Consequently, analogous to the selection shadow theory of aging, there is virtually no selection against subsets of cells with very long lags, suggesting that persister-like phenotypes may very generally be expected to occur in microbial population. In addition, we show that the relationship between single-cell and population lags depends on the typical population size and that, while noisy single-cell lag distributions might be beneficial, they are only effective at large population sizes. This result suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environmental changes, small populations will require regulated sense-and-response strategies in order to ensure short population lags. Experimental validation of these results can be done trough dedicated microfluidic devices combined with time lapse microscopy images. Unfortunately, these methods often lack the direct observation of important gene expression variables as the mRNA or the ribosome levels. We developed a dedicated biophysical model of gene expression which, together with a specific Bayesian inference scheme, allows to predict the dynamics of these latent variables. We first tested this method on time series data of single cell growth. The results show that cells growing in different media have similar cell-cycle and longer scales dynamics.

Table of contents

Chapter 1

Introduction

Living systems are complex machines which build and regulate themselves with high precision. Indeed, they are able to grow and multiply, process nutrients and communicate among themselves efficiently. In order to describe the mechanisms behind and build reliable quantitative models, many studies have been conducted on bacteria, which are among the simplest living systems: they are unicellular organisms which lack membrane-bound organelles such as nucleus or mitochondria. Bacteria form one of the three domains of life, the two others being archaea and eukaryotes. This classification is based on the sequencing of a piece of ribosomal RNA known as 16S RNA. Bacterial size spans a large spectrum going from the 10^{-2} $[\mu m^3]$ to 10^{8} $[\mu m^3]$ whereas the genome size is in the order of a few million base pairs (Mb) and typically contains a few thousands of coding genes[1]. Let's now briefly discuss some aspects of bacterial growth and regulation which will be useful for the understanding of the thesis. The reader should have in mind that this introduction does not cover all the aspects of growth and regulation in bacteria as its goal is to provide the reader with a basic understanding of the most important concepts.

1.1 The Bacterial growth

One of the most striking properties of bacteria colonies is the speed at which they grow. *E. coli* for example can divide with a rate of one division every 15 minutes [44] giving rise to millions of off-springs in just a few hours. This implies that, due to resource limitation, bacterial growth can not be constantly exponential. Indeed, an *E. coli* cell of 10^{-12} $[g]$ dividing every 15 minutes will generate 6×10^{45} $[g]$ of biomass in 2 days which is more than the mass of the Earth! Other phases are part of the bacterial growth and, in ideal experimental

[1] Rule of thumb for the bacterial genome is 1 protein-coding gene per Kb

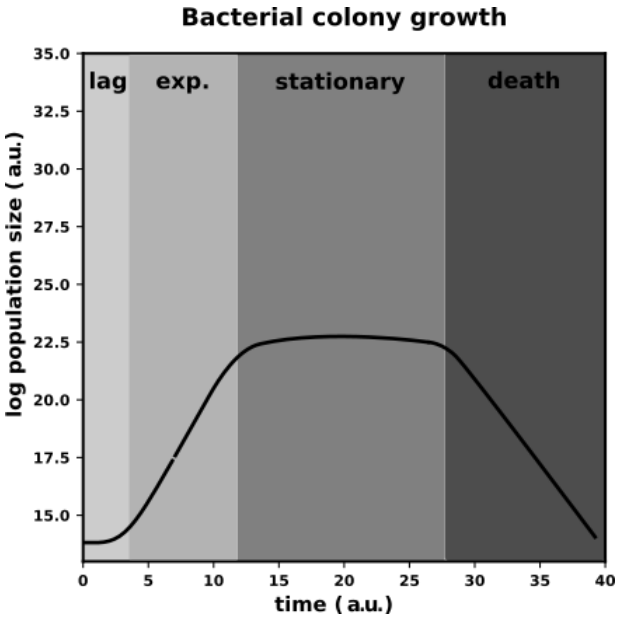

Fig. 1.1 The solid line represents a typical bacterial growth curve in batch conditions where the four phases are colored with a different shade of gray tonality.

conditions, bacterial growth is characterized by four phases, i.e. the lag, the exponential, the stationary and the death phase, as shown in figure 1.1 and discussed in detail below.

Exponential phase This phase is characterized by an exponential growth of the bacterial colony, due to cell division. Indeed, let's consider the case of binary fission, as for *E.coli*, where we assume every cell to divide every τ_d minutes. After t minutes the number of cells generated from this single bacteria will be

$$n(t) = 2^{\frac{t}{\tau_d}} = e^{rt} \tag{1.1}$$

which makes clear why this is called the exponential phase. The quantity $r = \frac{\log 2}{\tau_d}$ is called the colony growth rate and is often used when working with the natural logarithm. Note that the doubling time τ_d ranges from minutes to hours for *E. coli* and depends on several factors like the strain, the type of nutrient, the temperature and other environmental conditions. Bacterial growth may be more complicated than simple binary fission. For example *C. crescentus* divides into two morphologically different daughter cells, one motile and the other adherent and *B. subtilis* divides in a process of sporulation, but we will however ignore these particular cases in the following discussion.

Stationary phase and death phase As said, exponential growth can not continue indefinitely due to the limitation of resources. As a colony starts to run out of resources, the growth rate r decreases until growth eventually stops. The colony enter the stationary phase as depicted in figure 1.1. This phase is considered an active phase in the sense that cells are not dead: if inoculated into fresh media growth resumes [29]. However, if a colony spends too much time in an exhausted media, cells start to lyse leading to a decrease in the colony size (death phase).

Lag phase When a colony is inoculated into fresh media, growth is usually not resumed immediately. The time needed from inoculation to full speed growth is called the population (or bulk) lag time. This time delay is in part due to the lack of the correct cellular machineries needed to metabolize the nutrients [36]. More detail about this particular phase will be given in the next chapter.

1.2 Mechanisms of gene regulation

In order to grow, replicate, move or simply respond to external stimuli, cells have to build and maintain several micro and macro molecules like peptides, proteins, ribosomes, etc. The synthesis of such molecules very often involves the expression of some or several genes, and we will, in the following paragraphs, explain the mechanisms behind gene expression and gene regulation.

Gene expression First, remember that a gene is defined as a sequence of DNA that encodes a functional molecular product (e.g. proteins). The process to read out the molecular product from the gene is called gene expression and is done in two separate steps known as transcription and translation (figure 1.2). Transcription is the process of copying a section of the DNA into mRNA while translation allows the synthesis of proteins from the genetic information contained in the mRNA. Transcription starts by unwinding the DNA double helix into two single strands. This is done through the DNA helicase enzyme which breaks the hydrogen bonds between the strands. Once the DNA is unwound, one of the two strands is used as a template by the RNA polymerase enzyme (RNAP) which synthesizes RNA following the template. In order to synthesize RNA starting from the DNA template, the RNAP needs first to bind to the DNA. Unfortunately, RNAP can not directly bind to the DNA but it first has to bind to a sigma factor protein[2] and the complex formed is then able

[2] In *E. coli* 7 different sigma factor proteins exists allowing the regulation of different sets of genes. The most common sigma factor found in *E. coli* is σ^{70}.

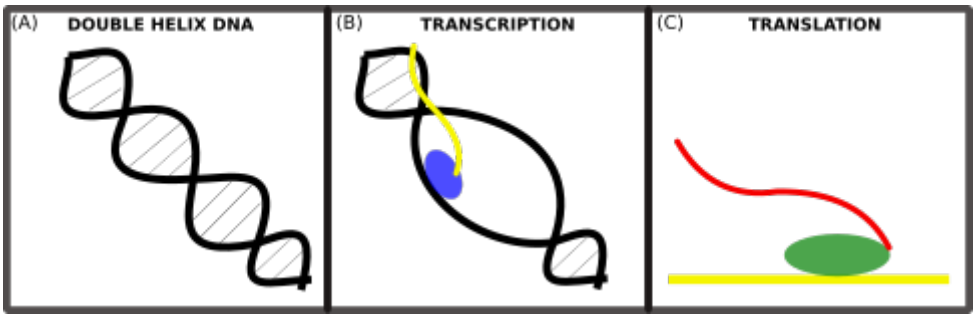

Fig. 1.2 **(a)** The double helix DNA **(b)** Transcription: the DNA is unwound, the complex RNAP-σ factor (blue) binds to the template and travel along it synthesizing the mRNA (yellow) **(c)** Translation: proteins (red) are synthesized from the released mRNA (yellow) trough ribosomes (green).

to "recognize" and bind to the DNA template. The region on the DNA where the RNAP binds is called the promoter region and is usually few nucleotides prior the transcription start site. Once bound, the RNAP travels along the DNA strand and synthesizes the nucleotide sequence into the so-called messenger RNA (mRNA). Transcription ends when the RNAP recognizes a specific termination sequence on the template, it detaches itself from the DNA strand and releases the mRNA.

As already mentioned, translation is the process to synthesize proteins from mRNA. Large macromolecular complexes called ribosomes bind to the released mRNA to start protein synthesize. Once bound, the ribosome travels along the mRNA (elongation phase) reading its genetic code and forming the corresponding amino acid chain. Finally, the ribosome unbinds from the mRNA upon recognition of a specific termination sequence.
This description of transcription and translation is both simplistic and idealistic. In real biological systems, these processes can be way more complex. For example it is known that supercoiled DNA may stops transcription [35], RNAP has difficulties to overcome DNA damages or tightly bound proteins [49] or that RNAP forms "traffic jams" on the DNA and has consequences in transcription [27][28]. However, in order to keep this introduction simple, we will not discuss any of these details.
Depending on the external conditions or its life stage, a cell may need certain molecular products instead of others. Gene regulation is, therefore, a really important process in the cell life since it controls the levels of proteins within the cell. We will show how proteins levels are regulated through one of the most well understood gene regulatory system, namely the *lac* operon in *E. coli*.

Gene regulation Gene regulation includes all the mechanisms that cells use in order to increase or decrease the levels of specific gene products. This can be achieved either by

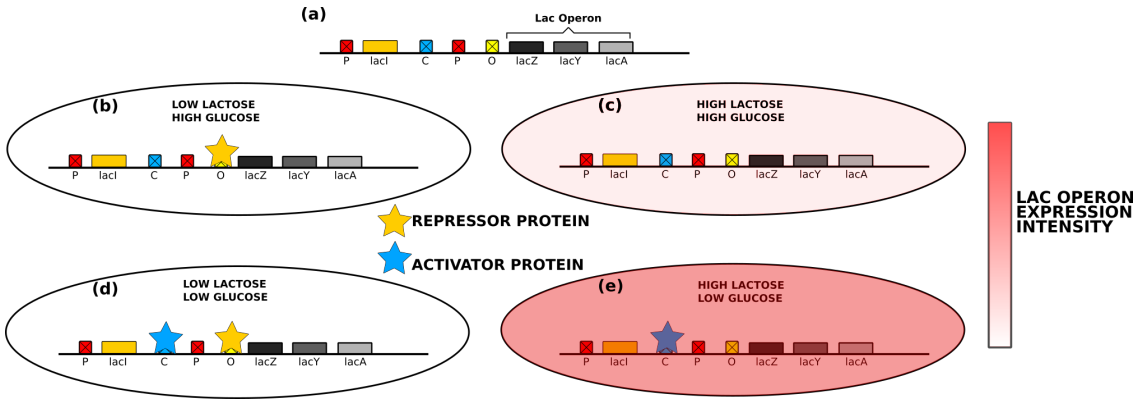

Fig. 1.3 **(a)** The *lac* operon is made of three genes *lacZ*, *lacY*, *lacA*, which molecular products are needed to metabolize lactose. Prior to the operon we find the promoter region (red crossed box), the binding site for the activator protein (blue crossed box) and the binding site for the repressor protein (yellow crossed box). The constitutively expressed *lacI* gene is also present in the *E.coli* chromosome and its molecular product is the repressor protein (yellow star).**(b)** Low lactose and high glucose. When there is no lactose, the repressor protein (yellow star) binds to DNA preventing the operon transcription. **(c)** High lactose and high glucose. When lactose molecules are presents, they bind to the repressor proteins and the complex lactose-repressor is unable to bind to DNA. This leaves the operon free to be transcribed. However, since the activator protein (blue star) levels and the glucose levels are inversely proportional, very few activators are present in the cell making transcription only moderate. **(d)** Low lactose and low glucose. Low glucose levels means high activator levels but the operon is repressed due to the absence of lactose. **(e)** High lactose and low glucose. The operon is not repressed, and the high amount of activator proteins makes expression strong.

regulating the amount of mRNA produced (transcription regulation), by regulating the amount of protein produced from the mRNA (translation regulation) or by controlling the levels of active proteins once the proteins are already formed (post-translation regulation). For transcription regulation, one of the most famous examples is the one of the *lac* operon in *E. coli* which discovery in 1961 was worth the Nobel prize to Francois Jacob and Jacques Monod [17]. Monod and Jacob observed that *E.coli* growing in a mixture of glucose and lactose do not metabolize the two sugars simultaneously but they rather consume them sequentially [41]. This observation is at the base of the theory on the *lac* operon regulation which we will now briefly summarize.

The *lac* operon (figure (1.3a)) consists of three genes *lacZ*, *lacY*, *lacA*, which molecular products are needed to metabolize lactose. The operon is "controlled" by two transcription factor proteins[3] which presence will increase or decrease the expression of the operon. These two proteins are of opposite nature. One is a repressor protein which, once bounds to the DNA, prevents the transcription of the operon; whereas the other is an activator protein which, once bound to the DNA, increases the transcription of the operon. Note that the repressor protein (named LacI) is the molecular product of the *lacI* gene contained in the *E.coli* chromosome. The *lacI* gene is continuously expressed[4] which means that no transcription factor proteins regulate its transcription activity.

When no lactose is present, the LacI repressor protein binds near the promoter region, preventing the RNAP to initiate transcription (figure (1.3b)). However, if lactose is present, the LacI repressor protein binds to the lactose molecule and the complex lactose-LacI is unable to bind to DNA. In this condition there is nothing that prevents the RNAP to initiate transcription therefore the operon is expressed (figure (1.3c)). This explain how *E. coli* can turn "on/off" the *lac* operon depending on the lactose presence but it does not explain why the two sugars are metabolized sequentially. In order to fully explain Jacob and Monod observation, we also have to consider that the operon responds to the presence of glucose by increasing/decreasing the transcription rate. Indeed, it has been discovered that the amount of activator proteins is inversely proportional to the levels of glucose [37]. Therefore, when the glucose levels are low, the amount of activator proteins is high thus, if the operon is not repressed, transcription activity is high (figure (1.3e)). However, if the glucose levels are high, the amount of activator proteins is low making the operon expression moderate even if it is not repressed (1.3c)). Obviously, in both scenario of low/high glucose levels, there is no expression of the *lac* operon if lactose is not present (1.3b,d).

The deterministic model of gene regulation presented so far clearly does not take into account

[3]Transcription factors are proteins which control the transcription of DNA by binding to specific DNA sequences.

[4]A gene which is continuously expressed is called constitutive.

the stochastic nature of the underlying phenomena. Transcription initiation, binding/unbinding of transcription factors and many other processes, happen with a certain probability but with no certainty. This means that even in unfavorable conditions like high glucose and low lactose the operon may be expressed giving rise at what we call "noise in gene expression" which is the topic of the next section.

1.3 Noise in gene expression

Due to the stochastic nature of gene expression, cells sharing the same DNA and living in similar conditions do not necessarily express the same genes at the same levels (figure 1.4). Gene expression noise is defined to be the cell to cell variation on the protein levels associated with a gene. It is usually quantified as the coefficient of variation (standard deviation divided by the mean) of the protein levels distribution. It has been experimentally [10] and theoretically [52] shown that two independent noise source (the intrinsic and extrinsic noise) contribute to the final observed variability in the protein levels. The first, the intrinsic noise, is due to the stochastic nature of the protein production and degradation. Indeed, even in the ideal case where gene expression takes place in the exactly same conditions, due to the stochastic nature of the process (e.g. binding/unbinding of RNAP and ribosomes, etc.), the final amount of protein molecules produced is not deterministic. The intrinsic noise term is modeled trough a Poisson process. Indeed, if $p(n,t)$ is the probability to have n proteins at time t and

$$p(n+1,t+\Delta t|n,t) = k\Delta t + \mathscr{O}(\Delta t^2) \tag{1.2}$$

$$p(n-1,t+\Delta t|n,t) = \gamma\Delta t + \mathscr{O}(\Delta t^2) \tag{1.3}$$

are the probability to produce/degrade one protein during the time interval Δt, the master equation governing this process reads

$$\begin{aligned} p\,(n,t+\Delta t) =& p\,(n,t)\,(1-k\Delta t-n\gamma\Delta t) \\ &+ p\,(n-1,t)\,k\Delta t + p\,(n+1,t)\,(n+1)\gamma\Delta t + \mathscr{O}(\Delta t^2) \end{aligned} \tag{1.4}$$

The steady state solution of this equation is the Poisson distribution

$$p(n) = \frac{\langle n\rangle^n e^{-\langle n\rangle}}{n!} \tag{1.5}$$

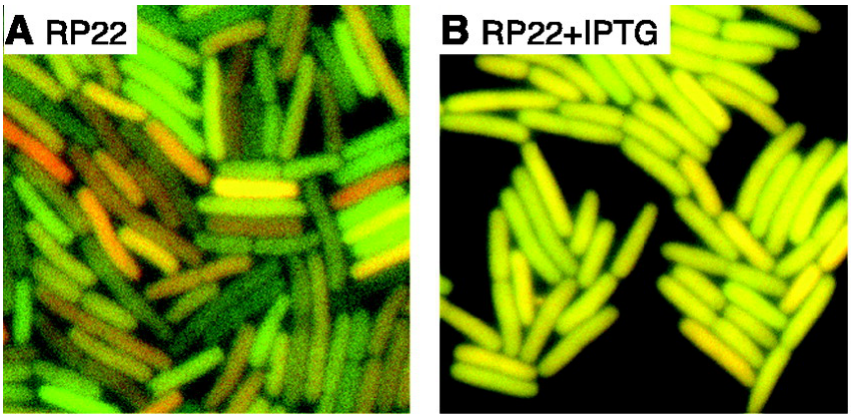

Fig. 1.4 Isogenic strains of *E.coli* incorporating the distinguishable cyan and yellow alleles of green fluorescent protein in the chromosome. In each strain, the two reporter genes were controlled by identical promoters. **(A)** In strain RP22, with promoters repressed by the wild-type lacI gene, red and green indicate significant amounts of noise. **(B)** RP22 grown in the presence of lac inducer, 2 mM IPTG. Both fluorescent proteins are expressed at higher levels and the cells exhibit less noise. Figures taken from [10].

with $\langle n \rangle = \frac{k}{\gamma}$ the expected number of proteins. Experimental evidence [54] shows that this model well explains the noise pattern observed for low expressed genes (< 10 protein molecules per cell) but it is not capable to explain the noise observed for highly expressed genes. This suggests that an additional noise source, the extrinsic noise, takes part in the gene expression. With the term extrinsic noise we denote all sources of noise which are global to a single cell but vary from one cell to the other. Concentrations, states and locations of molecules such as regulatory proteins and polymerases, variations in the levels or activity of these molecules cause fluctuations in the expression of the gene which are global to the single cell but vary from cell to cell. This will affect one cell differently from another and add an extra layer of noise on top of the intrinsic noise.

It is important to observe that transcription noise, i.e. the cell to cell variability on transcript levels, is encoded into the promoter sequence [16],[45] and therefore is under natural selection. Whereas some studies argue that natural selection acts to minimize expression noise [3],[45], others show that gene expression noise can be a beneficial trait [5],[30],[60]. Our work, presented in the following chapter, will provide an additional example of the expression noise effects on evolution.

1.4 Outline of the thesis

Some studies had shown that different genes present different levels of expression noise and this difference is, to some extend, encoded into the promoter sequence [45][16]. This implies that transcriptional noise is an evolvable trait subject to natural selection. With this, we mean that mutations[5] of the promoter sequence may lead to changes in the promoter noise levels and this might affect the organism chances to survive (fitness). For a long time, noise in gene expression had been seen as an undesirable but unavoidable trait of gene expression. It was thought that for every condition, there exists an optimal expression level and deviations from it are detrimental to the organism's fitness. In this interpretation, natural selection acts to select promoters with a low noise level [45][5]. Theoretical [30][10][9] and experimental [6][48] evidences however show that expression noise generates phenotypic diversity among isogenic cells and *Kussel et al.* [30] demonstrated that, for bacterial colonies living in fluctuating environments, phenotypic diversity (bet-hedging) is a particularly effective survival strategy. Moreover, it has been shown [60] that, in some circumstances, evolution must have acted in order to increase the noise levels of certain promoters. Part of the work presented in this thesis is based on the simple observation [22] that *E.coli* undergoing carbon source switching, resume growth with a large growth lag variability (noise). Indeed, when inoculated from glucose to lactose, the 27% of *E.coli* cells start growing within the first 45 $[min]$ whereas the 5% do not resume growth during the entire experiment duration of 240 $[min]$. Although a recent study shows the mechanisms behind this observation [20], we here present a general mathematical theory on why noisy growth lag distribution are expected in clonal populations.

In Chapter 2, we indeed show that in bacterial colonies the first bacteria resuming growth generates exponentially more offspring and so contribute more to the final fitness. This observation let us hypothesize that natural selection is strong for the first growth resuming bacteria but weak for others. With this we mean that mutations affecting the first growth resuming bacteria are strongly selected/counter-selected whereas mutations affecting the late resuming growth bacteria are weakly selected/counter-selected. This mechanism allows detrimental mutations affecting late regrowing cells to accumulate and explains the observed heterogeneity in the growth lag distribution. Therefore, the noise in the growth lag distribution is not only a beneficial trait as some studies proposed [12][43] but it is an unavoidable trait in bacterial populations. In addition, we show that this result depends on the typical population size and, while lag distributions with a large variance are expected in large

[5] DNA mutations can be beneficial, deleterious or neutral depending if they increase/decrease or unalter the organism fitness.

populations, this is not true for small colonies. This suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environmental changes, small populations will require regulated sense-and-response strategies.

In chapter 3: Tracking cell growth and gene expression at the single cell level is now possible through microfluidic devices combined with time lapse microscopy [22][59]. However, even if dedicated software are able to precisely estimate the cell size and the amount of target proteins [22], these measurements are not free from measurements errors. In this chapter we develop a dedicated biophysical model for cell growth and gene expression which, combined with a regression technique known as kriging, not only allows us to reduce the measurement errors but also to disentangle promoter specific fluctuations from other noise sources. We then apply this technique to the case of cell growth time series data (chapter 4) and reveal some new features of the cell growth dynamic.

Chapters 2 and 4 are presented as individual stand-alone publications.

Chapter 2

The benefits of a noisy lag distribution in bacterial populations

Athos Fiori[1], Erik van Nimwegen[1,*]

[1] Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland.
[*] to whom correspondence should be addressed: erik.vannimwegen@unibas.ch

Abstract

The survival of organisms in randomly fluctuating environments not only depends on their ability to grow in different conditions but also on the time needed to adapt to each new habitat. Recent works had shown that, like many other physiological quantities, the adaptation time fluctuates in a stochastic manner across single cells and that the underlying distribution can dramatically change across genotypes. To understand how natural selection may have acted on the distribution of single-cell lags we develop a mathematical theory of how the single-cell lag distribution determines the reproductive success at the population level. We show that lags at the population level are exponentially dominated by the shortest lags at the individual cell level. Consequently, analogous to the selection shadow theory of aging, there is virtually no selection against subsets of cells with very long lags, suggesting that persister-like phenotypes may very generally be expected to occur in microbial population. In addition, we show that the relationship between single-cell and population lags depends on the typical population size and that, while heterogeneous single-cell lag distributions can be beneficial, they are only effective at large population sizes. This result suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environmental changes, small populations will require regulated sense-and-response strategies in order to ensure short population lags.

2.1 Introduction

Bacterial colonies are composed of phenotypically different individuals that compete with each other giving rise to potentially complex dynamics (figure 2.1a). Predicting these collective dynamics base on the knowledge of the single-cell dynamics remains challenging. Since the bacterial colony growth underlies the organism fitness, a mathematical description of the colony dynamic based on the single cells dynamic is important to understand the genotype fitness. *Hashimoto et al.* [14] showed that growth noise causes clonal populations of *E.coli* to double faster than the mean doubling time of their constituent single cells and so growth noise is a way to increase cell proliferation. This work instead, focuses on the

consequences of the lag noise on the genotype fitness. Although *Sean et al.*[51] showed that, under favorable conditions, *E.coli* strains with short lags have an evolutionary advantage; wild type *E. Coli* has been shown to have a non negligible lag noise [20][32][50][46]. It has been suggested that lag noise [12][43] is a bet-hedging strategy where, through phenotype randomization, the bacterial colony is prepared for different kinds of conditions. Indeed, keeping a fraction of the colony in a non growing state, may make cells more resilient to stresses like heat shock or antibiotics [1][21] thus increasing the survival chances.
In this paper we show, theoretically, why a high lag noise should in general be expected in isogenic bacterial populations even without advocating bet-hedging. Indeed, due to the exponential growth of bacterial populations, the first regrowing cells largely determine the bacterial growth curve therefore the single cell lag time distribution (or just lag distribution) tail has a low impact on the genotype fitness. In analogy with the theory of senescence [39],[58], noisy lag distributions should be expected since selection strongly acts on the first regrowing cells but is weak on the lag distribution tail. Through a theoretical model and computer simulations, we investigate the consequences of the lag noise on the genotype fitness. We show the lag distribution and the bulk lag time T strongly depend on the inoculum size (the number of bacteria presents when the new environment first comes), and we discover that lag noise is expected only when the inoculum size is large. This suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environments, small populations will require regulated sense-and-response strategies in order to optimize the genotype fitness.

2.2 Why noisy lag distribution are expected in large populations

We first revisit [2],[31] the relation between the single cell lag distribution (LD) and the population lag (or bulk lag) and then focus on the consequences of the lag distribution noise on bacteria proliferation.

Let's consider a single cell inoculated into fresh media at time $t_0 = 0$. If we wait long enough this cell will generate a bacterial growth curve similar to the one in figure 2.1a. Therefore, for this specific cell i, the population size at any time t in the exponential phase $(t > \tilde{\tau}_i)$

$$N_i(t) = e^{\tilde{r}(t-\tilde{\tau}_i)} \tag{2.1}$$

where $\tilde{\tau}_i$ is the lag time and $\tilde{r}$ the growth rate of the bacterial growth curve generated by the cell i. If we inoculate N_0 cells at t_0 instead of just one, and assume they all grow with the

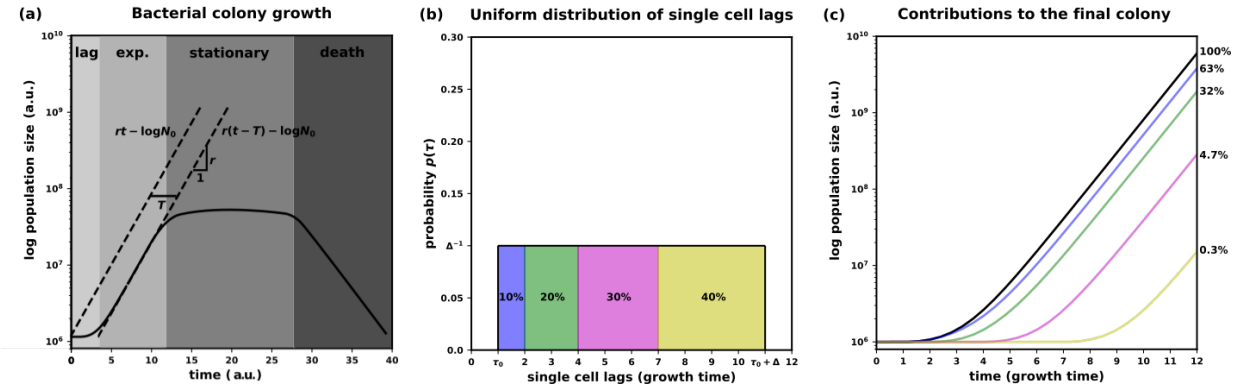

Fig. 2.1 **(a)** The simple case of bacterial growth in batch conditions. The solid line represents a typical bacterial growth curve where each phase (lag ,exponential, stationary and death phase) is colored with a different shade of gray tonality. The right dotted line: $r(t-T)-\log N_0$ describes growth in the exponential phase whereas the left dotted line represents an ideal population without the lag phase (T), but with the same growth rate (r). Both of these lines have a slope r equal to the population growth rate and their time translation T represents the lag time. The quantity $N_0 = N(t=0)$ represents the number of inoculated bacteria at time $t=0$. **(b)** The black line represents the continuous uniform lag time distribution (LD) as described in equation (2.7) with $\tau_0 = 1$ and $\Delta = 10$. This distribution has been divided in four areas, from cells with a very short lag time (blue) to the one with a very long one (yellow), and we noted their relative size (%) compared to the total area. **(c)** In black the total population growth $N(t)$ given the LD depicted in (a). The colored bacterial curves represent the bacterial growth curves $N_\star(t)$ coming from the four different regimes depicted in (a). On the right we noted the c_f i.e. the relative contributions to the total population coming from these regimes.

same population growth rate $\tilde{r}$, then the population size at time t is given by

$$N(t) = \sum_{i=0}^{N_0} N_i(t) = \sum_{i=0}^{N_0} e^{\tilde{r}(t-\tilde{\tau}_i)} \tag{2.2}$$

or, by re-scaling all the time variables by $\tilde{r}^{-1}$ i.e. working in growth time units

$$N(t) = \sum_{i=0}^{N_0} e^{(t-\tau_i)} \tag{2.3}$$

The single cells lag times τ_i can be experimentally determined through methods like the one proposed by *Kaiser et al.* [22]. The relation between the lag time distribution $p(\tau)$ and the bulk lag (T) has already been shown by *Baranyi* [2] and we here simply revisit it. Note that $p(\tau)d\tau$ represents the probability that a bacteria will generate a population growth curve with lag time τ. The equation describing the bacterial growth curve in the exponential phase is known since more than fifty years [42] and reads $(t > T)$

$$N(t) = N_0 e^{r(t-T)} \tag{2.4}$$

where r is the bulk growth rate, T the bulk lag time and N_0 the inoculum size.

First, for ease, we work out the relation between the bulk lag time and the lag time distribution in the limit $N_0 \to \infty$. Note that, if not explicitly mentioned, the results are presented in growth time units through the entire article.

In the limit $N_0 \to \infty$ the sum in (2.3) can be approximate by its expected value

$$N(t) = e^t \sum_{i=0}^{N_0} e^{-\tau_i} \approx N_0 e^t \left\langle e^{-\tau} \right\rangle \tag{2.5}$$

and, using (2.4) , the relation between the bulk lag and the lag distribution reads

$$T_\infty = -\log \left\langle e^{-\tau} \right\rangle \tag{2.6}$$

where T_∞ is the bulk lag time for the case $N_0 \to \infty$ and $\langle e^{-\tau} \rangle = \int d\tau e^{-\tau} p(\tau)$. This shows that the expected lag time is not simply the expected value of the lag distribution $\langle \tau \rangle$ but the log transform of its exponentially weighted average. To examine the consequences of this

result, let's consider, as an example, a uniform lag distribution $p(\tau)$ defined as

$$p(\tau) = \begin{cases} \Delta^{-1} & \text{if } \tau \in [\tau_0, \tau_0 + \Delta] \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

as depicted in figure 2.1b. In this case, the expected single-cell lag time

$$\langle \tau \rangle = \tau_0 + \frac{\Delta}{2} \tag{2.8}$$

and the population lag time (2.6)

$$T_\infty = -\log\left(\frac{e^{-\tau_0}}{\Delta}\left(1 - e^{-\Delta}\right)\right) \underset{\Delta \gg 1}{\approx} \tau_0 + \log\Delta \tag{2.9}$$

can be easily computed. We realize that, apart from the offset τ_0, the population lag T_∞ is exponentially shorter than the expected single cell lag $\langle \tau \rangle$. This is due to the fact that growth in bacterial populations is exponential, so the descendants of the first regrowing cells will soon dominate the entire bacterial growth. To better show this concept, let $N_\star(t)$ be the number of cells in the exponential phase coming from bacteria with lag $\tau \in [\tau^\star, \tau^\star + \Delta^\star]$ (colored areas and lines in figures 2.1b,c) and $N(t)$ be the total population size at time t. Then the fraction

$$c_f = \frac{N_\star(t)}{N(t)} = \frac{1 - e^{-\Delta^\star}}{1 - e^{-\Delta}} e^{-(\tau^\star - \tau_0)} \tag{2.10}$$

represents the contribution to the final population given by cells resuming growth within $[\tau^\star, \tau^\star + \Delta^\star]$. As shown in figure 2.1c the contributions to the final population given by the first regrowing cells (blue) equal to 63% even if they just represents the 10% of the initial population (figure 2.1b). On the contrary the last regrowing cells (yellow), which represent the 40% of the initial inoculum, contributes the 0.3% to the final colony. This example makes clear that a bacteria with a short lag will generate exponentially more descendant than a bacteria with a long lag. This observation makes us conjecture that noisy lag distribution should, in general terms, be expected in bacterial populations since selection is strong only for the first regrowing bacteria. Indeed, we consider a mutation which can increase/decrease the heterogeneity of the lag distribution by an amount $\delta > 0$ without changing its mean $\langle \tau \rangle$. The two mutant lag distribution are for example given by

$$p_\pm(\tau) = \begin{cases} \frac{1}{\Delta \pm 2\delta} & \text{if } \tau \in [\tau_0 \mp \delta, \tau_0 + \Delta \pm \delta] \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

where $+$ indicates an increase in lag noise and $-$ a decrease in lag noise. In the case where this mutation increases the lag noise ($p_+(\tau)$), the mutation has a beneficial effect on the head of the lag distribution, since it allows cells to start growing a bit earlier ($\tau_0 - \delta$), but has a deleterious one on the tail of the LD since it also let cells to start growing later ($\tau_0 + \Delta + \delta$). In the case in which this mutation decreases the LD heterogeneity $p_-(\tau)$, the picture is exactly the opposite. With this in mind, let's compute the ratio between the mutants populations sizes $N_\pm(t)$ and the wild type $N(t)$, to compare the number of descendants generated by the mutants and the wild type cells

$$\frac{N_\pm(t)}{N(t)} = e^{\pm\delta}\frac{1-e^{-\Delta\pm\delta}}{1-e^{-\Delta}}\frac{\Delta}{\Delta\pm 2\delta} \underset{\substack{\delta \ll 1 \\ \Delta \gg 1}}{\approx} 1 \pm \delta \tag{2.12}$$

This shows that the more noisy lag distribution will generate more descendants than the wild type version $\frac{N_+(t)}{N(t)} > 1$, hence this mutation has an high chance to be fixed into the population. On the contrary the less noisy lag distribution will generate less descendant than the wild type $\frac{N_-(t)}{N(t)} < 1$, and therefore is less likely to be fixed into the population. This shows that beneficial/deleterious mutations acting on the head of the lag distribution are strongly selected independently on the effects they have on the tail of the LD. This hypothesis, similar to the antagonistic pleiotropy hypothesis [39],[58], explains why deleterious mutations appearing on the LD tail may accumulate therefore why long tailed lag distributions are expected in the wild.

All these arguments are general and independent of the specific single cell lag distribution $p(\tau)$. However, all these results have been computed in the limit $N_0 \to \infty$. When the population size N_0 is small, the bulk lag time T is on average longer than T_∞. Indeed, in the extreme case $N_0 = 1$, the expected population lag time

$$\langle T \rangle_{N_0=1} = \left\langle -\log\left[\exp^{-\tau}\right] \right\rangle = \langle \tau \rangle \tag{2.13}$$

is given by the expected single cell lag time $\langle \tau \rangle$ which is exponentially longer than T_∞. Let's work out, in the following sections, the impact and the consequences of a finite inoculum N_0.

2.3 The bulk lag time distribution $p(T)$

Single cell divisions, lags and lysis are stochastic processes [22][14] which collective result determines the duration of the observed bulk lag time (T). Therefore, the bulk lag time is

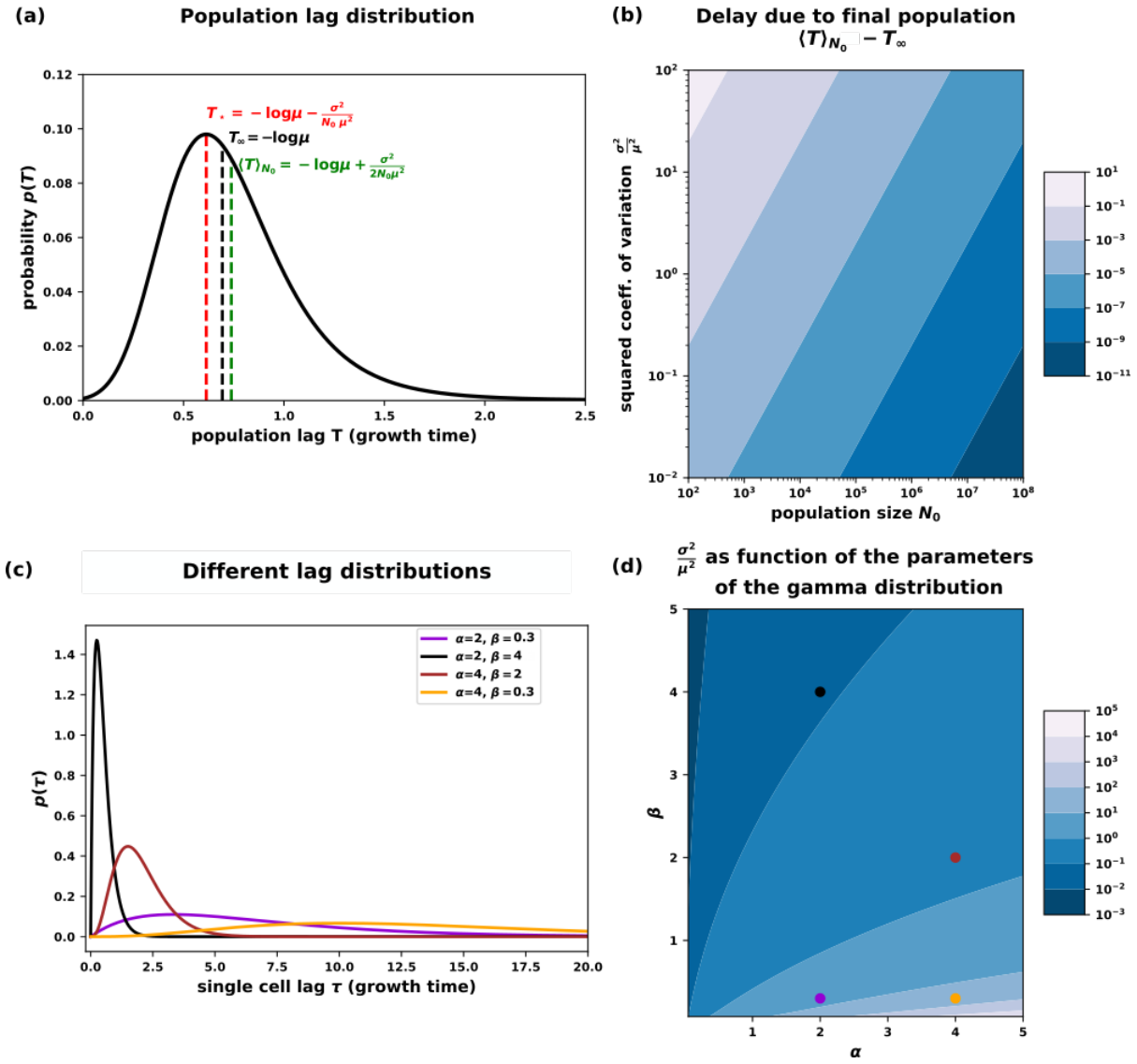

Fig. 2.2 **(a)** The population lag distribution $p(T_{N_0})$ (2.17)) with parameters $\mu = 0.5$, $\frac{\sigma^2}{N_0} = 2.25 \times 10^{-2}$. The red line represents the mode $T^{\star}_{N_0}$ and the green line represents the mean $\langle T_{N_0} \rangle$ of this distribution. The vertical black line is the value $T_{\infty} = -\log \mu$ i.e. the population lag time in the case $N_0 \to \infty$. **(b)** The difference between $\langle T_{N_0} \rangle$ and $T^{\star}_{N_0}$ as a function of N_0 and $\frac{\sigma^2}{\mu^2}$. This quantity represents the expected population lag time delay a finite population has, compared to the infinitely large population scenario. **(c)** Different LD $p(\tau)$ assumed to be Gamma distributed with different shape and rate parameters (α and β). **(d)** The coefficient of variation $\left(\frac{\sigma^2}{\mu^2}\right)$ of the f distribution as function of the parameters α, β of the underling LD (2.28). The four dots represent the values of $\frac{\sigma^2}{\mu^2}$ for the 4 distributions depicted in panel (c). Even if not proven, it seems clear that long tailed LD correspond to high $\frac{\sigma^2}{\mu^2}$.

also a stochastic variable and we are here interested in determining its probability distribution $p(T)$. In order to find this let us write the bacterial size at time t in the exponential phase as

$$N(t) = N_0 e^t f \quad \text{with} \quad f = \frac{1}{N_0}\sum_{i=0}^{N_0} e^{-\tau_i} \tag{2.14}$$

where the random variable $f \in [0,1]$ represents the fraction of the population size one would get compared to the situation without lag. The central limit theorem allows us to approximate its distribution

$$p(f) \propto e^{\frac{N_0(f-\mu)^2}{\sigma^2}} \quad \text{with} \quad \mu = \left\langle e^{-\tau} \right\rangle \quad \text{and} \quad \sigma^2 = \text{Var}\left[e^{-\tau}\right] \tag{2.15}$$

Equation (2.4) defines the population lag time variable

$$T_{N_0} = -\log f \tag{2.16}$$

where T_{N_0} is the bulk lag time for inoculum of size N_0. Using (2.15) we find its distribution[1]

$$p(T_{N_0}) \propto e^{-T_{N_0}} e^{-\frac{N_0}{2\sigma^2}\left(e^{-T_{N_0}}-\mu\right)^2} \tag{2.17}$$

sometime called the exp-normal distribution ExpNorm $\left(\mu, \frac{\sigma^2}{N_0}\right)$, represented in figure 2.2a. For the case $N_0 \to \infty$, this distribution converges to the Dirac delta function (black dotted line in figure 2.2a)

$$p(T_{N_0}) \underset{N_0\to\infty}{=} \delta\left(T_\infty + \log\mu\right) \tag{2.18}$$

where obviously the mean and the mode correspond to the same value

$$T_\infty = -\log\mu \tag{2.19}$$

in agreement with what we developed in the previous section.
For finite N_0, the distribution is positively skewed as shown in figure 2.2a and we can easily compute its mode (red dotted line)

$$T^\star_{N_0} = -\log\left[\frac{\mu}{2}\left(1+\sqrt{1+\frac{4\sigma^2}{\mu^2 N_0}}\right)\right] = -\log\mu - \frac{\sigma^2}{\mu^2 N_0} + \mathscr{O}\left(\frac{1}{N_0^2}\right) \tag{2.20}$$

[1]The complete distribution is given in equation (S.3)

and mean (green dotted line)

$$\langle T_{N_0} \rangle = -\log \mu + \frac{\sigma^2}{2N_0\mu^2} + \mathscr{O}\left(\frac{1}{N_0^2}\right) \tag{2.21}$$

As we will see in the next section, the fact that small inoculums N_0 give rise to skewed and noisy distribution compared to the case $N_0 \to \infty$, has various consequences in bacterial proliferation. Indeed, if N_0 is finite than (figure 2.2a)

$$p\left(T_{N_0} < T_\infty\right) \neq 0 \tag{2.22}$$

i.e. there is a non negligible probability that small populations will have a very short population lag time T_{N_0} compared to T_∞. If this is the case, then the number of descendent coming from the small populations will be exceptionally large due to this advantage. These events have been called "jackpot" events [13] due to their rare but high impact effect. However, if N_0 is finite, then the lag distribution $p(T_{N_0})$ is a long tailed distribution (figure 2.2a) and this has two major consequences. The first is that

$$p\left(T_{N_0} > T_\infty\right) \neq 0 \tag{2.23}$$

therefore there is a non negligible chance that small populations have a growth disadvantage. The second, and more important, is due to the Jensen inequality [18] which guarantees that

$$\langle T_{N_0} \rangle \geq T_\infty \tag{2.24}$$

In order to quantify the impact of N_0 on the $p(T_{N_0})$ distribution, we define the time delay due to finite N_0 as the difference between the population lag mean at finite and infinite inoculums

$$\langle T_{N_0} \rangle - T_\infty = \frac{\sigma^2}{2N_0\mu^2} \tag{2.25}$$

which is depicted in figure 2.2b as a function of N_0 and of the coefficient of variation $\frac{\sigma^2}{\mu^2}$. This quantity is proportional to $\frac{\sigma^2}{\mu^2}$ and inversely proportional to N_0.
It's important to remember that $\frac{\sigma}{\mu}$ is the coefficient of variation of $p(f)$ and not of $p(\tau)$. The following example will clarify the difference between the two. Consider the single cell lag distribution to be gamma distributed

$$p(\tau) = \text{Gamma}(\alpha,\beta) \stackrel{\text{def}}{=} \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau} \quad \text{with} \quad \alpha,\beta > 0 \tag{2.26}$$

represented in figure 2.2c with different values of shape parameter α and rate parameter β. A detailed analysis of this example is out of the scope of this work, but we want the reader to be aware of the difference between the Cv of $p(f)$ i.e. $\left(\frac{\sigma^2}{\mu^2}\right)$ and the coefficient of variation of the single cell lag distribution Cv_τ. For example, we note that the coefficient of variation of a generic gamma distribution

$$\mathrm{Cv}_\tau^2 = \frac{\frac{\alpha}{\beta^2}}{\left(\frac{\alpha}{\beta}\right)^2} = \alpha^{-1} \tag{2.27}$$

is independent on the β parameter. This means that, except the yellow distribution in figure 2.2c which has a smaller Cv_τ, the others all have the same one. However, in figure 2.2d we depicted the coefficient of variation of the $p(f)$ distribution

$$\frac{\sigma^2}{\mu^2} = \left(\frac{(\beta+1)^2}{\beta(\beta+2)}\right)^\alpha - 1 \tag{2.28}$$

which strongly depends on β. Therefore, in the case of a gamma lag time distribution $p(\tau)$, $\frac{\sigma^2}{\mu^2}$ is inversely proportional to Cv_τ and to β. With this in mind, we can now study the potential consequences on evolution all these observations might have.

2.4 The log genotype fraction depends on the initial population size

Modern experimental techniques allow to label single cells with unique DNA barcodes, and such techniques are said to be able to infer adaptative mutations even at very low frequencies [4],[23],[33]. We will show that these measurements might suffer, due to the N_0 dependence of lag time T, a fictitious evolutionary advantage favoring the more abundant genotype. To show this, we assume a wild type bacteria got a neutral mutation and we are interested in assessing the mutant fitness. Clearly, since the two genotypes are indistinguishable from an evolutionary point of view, their fixation probability must be the same. However, in DNA barcodes like experiments, we do not have access to the fixation probability. The only quantity we can measure is the number of wild type and mutants bacteria within the colony. Therefore, their relative fraction in the log space after a growth phase reads

$$\log\left(\frac{N_0^{\mathrm{mut}}}{N_0^{\mathrm{wt}}}\right) \to \log\left(\frac{N_0^{\mathrm{mut}}}{N_0^{\mathrm{wt}}}\right) - \left(T_{N_0^{\mathrm{mut}}} - T_{N_0^{\mathrm{wt}}}\right) \tag{2.29}$$

with $N_0^{\text{wt/mut}}$ and $T_{N_0^{\text{wt/mut}}}$ the initial population size and the population lag of the wild type and the mutant. The additional term

$$s \stackrel{\text{def}}{=} \left(T_{N_0^{\text{mut}}} - T_{N_0^{\text{wt}}} \right) \tag{2.30}$$

is positive if the wild type genotype after a growth phase expanded more than the mutant, and negative if the mutant genotype expanded more than the wild type. We aim to show that, even if the two populations are indistinguishable from an evolutionary point of view, the more abundant genotype has a systematically larger s which is due to the log transformation. Recalling that the population lag variables T are exp-normal distributed

$$T_{N_0 \text{wt/mut}} \sim \text{ExpNorm}\left(\mu, \frac{\sigma^2}{N_0^{\text{wt/mut}}} \right) \tag{2.31}$$

and assuming, without the loss of generality, the wild type is more abundant $N_0^{\text{wt}} \to \infty$

$$p\left(T^{\text{wt}}\right) \sim \delta\left(T^{\text{wt}} + \log \mu\right) \tag{2.32}$$

we can easily find the distribution of s

$$p(s) = \text{ExpNorm}\left(1, \frac{\sigma^2}{\mu^2 N_0^{\text{mut}}} \right) \tag{2.33}$$

where the mean and the variance of this distribution equal to

$$\langle s \rangle = \frac{\sigma^2}{2\mu^2 N_0^{\text{mut}}} \qquad , \qquad \text{Var}[s] = \frac{\sigma^2}{\mu^2 N_0^{\text{mut}}} = 2\langle s \rangle \tag{2.34}$$

This shows that the extra-term s takes values

$$s = \langle s \rangle \pm \sqrt{2\langle s \rangle} \tag{2.35}$$

The case

$$\lim_{N_0^{\text{mut}} \to \infty} \langle s \rangle = 0$$

predicts no systematic deviations favoring one or the other genotype as it is expected to be. However, the case $\langle s \rangle > 0$ predicts the more abundant genotype (here the wild type) to have a systematically larger relative fraction compared to the mutant. Therefore, looking at the s dynamics, one might confer to the more abundant genotype a larger fitness even-tough the two genotypes are indistinguishable. As it has already been observed by *Hallatscheck* [13] in

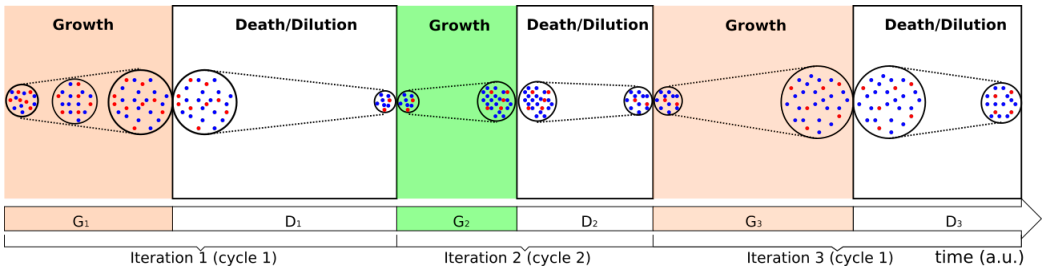

Fig. 2.3 A feast and famine experiment where two genotypes (red and blue) goes through different phases of growth (colored areas) and starvation (white areas). We call a cycle a phase of growth and starvation with a specific growing media. The duration of a cycle equals $G_i + D_i$ where G_i is the duration of the growing phase and D_i is the duration of the starvation phase. The area of the black circle represents the total population size, and the relative fraction between the two genotypes at the begin and at the end of the duration of a growing media reflects the fitness of them in the specific media.

a similar scenario, the neutrality of this process is guarantee by the rare jackpots events of the less abundant genotype. In fact, even if $\langle s \rangle$ seems to be insensitive to these events, the entire dynamic is not and the neutrality of the process guarantee.

2.5 The optimal surviving strategy in fluctuating environments may depend of the colony size

Previously we showed that, in the case of large initial populations $N_0 \to \infty$, selection pressure on the tail of the single cell lag time distribution $p(\tau)$ is weak, and so noisy lag distributions should be expected. Nevertheless, when looking at the bulk lag time distribution $p(T)$, we observed its dependence on the population size (2.17), and especially the fact that $\langle T \rangle_{N_0} \geq T_\infty$. In this section we will show that, in the case of small N_0, the selection pressure on the tail of the lag distribution $p(\tau)$ is non negligible anymore and so long tailed LD should not be expected. To do this, we first have to define a mathematical framework where such quantities come out naturally. We consider a scenario where two different genotypes compete in a feast and famine experiment present in figure 2.3.

The feast and famine experiment In figure 2.3, the different colors represent different growing media i.e. conditions where both genotypes (blue and red) can grow. The time a growing media will last is noted G_i and is only constrained to be long enough to allow both genotype populations to reach the exponential phase. After a period of growth, we initiate a famine period, of duration D_i, where the population loses bacteria through death or dilution

resulting in a shrinkage of the population size. A period of growth and dilution is called a cycle. Fore ease, we neglect the rise of any mutation, except neutral mutation, during the entire experiment duration.
After one cycle of growth and dilution, a genotype in the population will have grown/shrink by a factor

$$e^{r(G_i-T)-\mu D_i} \tag{2.36}$$

where r, T, μ are the growth rate, population lag time and decay rate specific to this cycle and genotype, whereas $G_i,\ D_i$ are the duration of the growing media and famine phase for the i$^{\text{th}}$ iteration. Among r, T, μ we consider only the population lag T as a stochastic variable and, after K concatenations of the same cycle type, the population will have grown/shrink by

$$\prod_{i=1}^{K} e^{r(G_i-T_i)-\mu D_i} = e^{K} e^{r(\langle G\rangle-\langle T\rangle)-\mu\langle D\rangle} \tag{2.37}$$

The growth-adaptation trade-off Now that we mathematically described the feast and famine experiment, let's consider two genotypes growing in this fluctuating environment. It has been shown that the genotype with the largest geometric mean

$$r\left(\langle G\rangle - \langle T\rangle\right) - \mu\,\langle D\rangle \tag{2.38}$$

is the one with more chances to survive [24],[34]. Clearly, the optimal solution would be to adapt as fast as possible to the new environment $\langle T\rangle \to 0$ and to grow as fast as possible $r \to \infty$ inside it. However, since fast growth and fast adaptation has an important energetic cost, no biological system can satisfy both requirements simultaneously and the correct trade-off between them is the key for the organism survival success. Depending on the condition, it may be better for a genotype to optimize either its growth rate r or its expected lag time $\langle T\rangle$ and we are here interested to study the trade-off between growth and adaptation. Consider two genotypes (1 and 2) with the same death rate μ but with different population lags $\langle T_1\rangle$ and $\langle T_2\rangle$ and growth rates r_1 and r_2

$$r_1 = r + \delta_r \qquad \langle T_1\rangle = \langle T\rangle \tag{2.39}$$

$$r_2 = r \qquad \langle T_2\rangle = \langle T\rangle - \delta_T \tag{2.40}$$

with $\delta_r > 0$ and $\delta_T > 0$. Genotype 2 will generate more descendants if

$$r_2\left(\langle G\rangle - \langle T_2\rangle\right) - \mu\,\langle D\rangle > r_1\left(\langle G\rangle - \langle T_1\rangle\right) - \mu\,\langle D\rangle \tag{2.41}$$

$$\Rightarrow \delta_r\left(\langle G\rangle - \langle T\rangle\right) < r\delta_T \tag{2.42}$$

This means that genotype 2 will out-compete genotype 1 only if the extra number of divisions of genotype 2 cells ($r\delta_T$) is larger that the extra number of divisions of genotype 1 cells ($\delta_r(\langle G\rangle - \langle T\rangle)$). A generalization of this simple example to multiple environmental conditions is straightforward but will not be detailed in this work. It's worth noting that the expected lag time $\langle T\rangle$, thus the trade-off (2.42), depends on the size of the population at the begin of the cycle N_0.

2.5.1 The growth-adaptation trade-off depends on the lag noise and on the population size.

In order to study how (2.42) depends on the population size we simulate a feast and famine experiment (supplementary). For ease, we assume the total number of bacteria at the begin of every iteration to be fix and equals to N_0^{tot}. We also assume the two competing genotypes (red and blue) face always the same growth media and they differ only by their lag distribution $p(\tau)$ defined in figure 2.4a. At the begin of the first cycle, the total population is of N_0^{tot} bacteria out of which N_0^{red} are red cells and $N_0^{\text{tot}} - N_0^{\text{red}}$ are blue cells. The lag advantage after the first iteration is

$$\delta_T = \left\langle T^{\text{blue}} \right\rangle - \left\langle T^{\text{red}} \right\rangle \tag{2.43}$$

and in figure 2.4b we plot the theoretically predicted δ_T (S.12) as a function of the total population N_0^{tot} and as function of the fraction of red genotypes at the begin of the first cycle $\rho = N_0^{\text{red}}/N_0^{\text{tot}}$. The region where $\delta_T > 0$ corresponds to the region where the red genotype has a shorter mean bulk lag and therefore a larger geometric mean (or higher fitness) according to (2.42) and vice versa in the region where $\delta_T < 0$. The black line represents the condition $\delta_T = 0$. Figure 2.4b can easily be interpreted as follow: when N_0^{tot} is large the red genotype tends to generate more descendants since there will probably be some red bacteria with a short lag (green area in figure 2.4a) which allows it to out-compete the blue genotype in terms of number of descendants. This is exactly the argument we made in the first section where we showed that, for large colonies, the selection acts only on the head of the lag distribution. However, for small values of N_0^{tot}, the blue genotype has a larger geometric mean (figure 2.4b). This comes from the fact that, for small N_0^{tot}, the chances that short lag red bacteria are present decrease with N_0^{tot} and the tail of the red genotype lag distribution i.e. the long lag cells (blue area in figure 2.4a) has now a deleterious impact on the number of descendants generated.

These considerations are certainly valid if the number of red cell N_0^{red} would remain the same at the beginning of every cycle. However, after the first growth and famine cycle, the

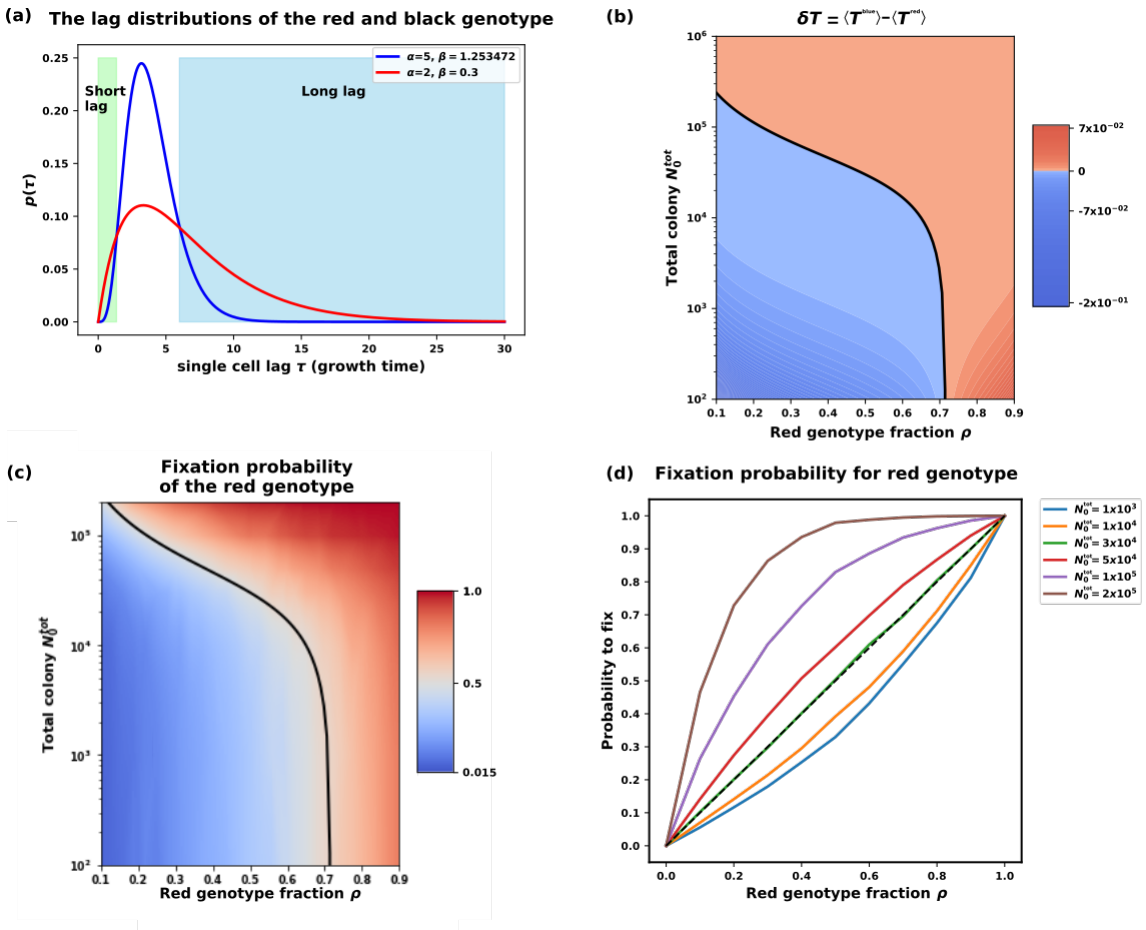

Fig. 2.4 **(a)** The gamma lag distributions for the two competing genotype red and blue with their respective α, β parameters. The green area represents what we call the short lag cells whereas the blue area represents what we call the long lag cells. **(b)** The expected lag advantage δ_T (equation (2.43)) over the first iteration as a function of the total initial population N_0^{tot} and the initial fraction of red genotype cells ρ. In black the condition $\delta_T = 0$. The red area corresponds to the condition where the red genotype has a larger geometric mean (2.42) and vice versa for the blue area. **(c)** The fixation probability for the red genotype in a feast and famine experiment as a function of the initial population size N_0^{tot} and the initial fraction of red cells at the first iteration. In black the same condition $\delta_T = 0$ as in panel (b). The red region is the region where the fixation probability of the red genotype > 0.5 whereas the blue region corresponds to a fixation probability of the red genotype < 0.5. **(d)** The fixation probability of the red genotype as function of the initial red cells fraction ρ for different population sizes N_0^{tot}. The difference between the fixation probability and the black dotted line (genetic drift) gives the advantage (positive) or disadvantage (negative) fixation strength of the red genotype over the blue one. This shows for example that for large N_0^{tot} and small ρ the red genotype has an high selective advantage.

number of red bacteria at the begin of the second cycle is fluctuating since it depends on the growth during the previous iteration and on the impact of the famine period. It is not obvious to theoretically predict the dynamics over the entire feast and famine experiment rely for this on the simulation. In figure 2.4c we show the result of the simulation for the fixation probability of the red genotype depending on the total colony size N_0^{tot} and on the fraction of red cells at the begin of the first cycle ρ. In black we draw the same theoretically computed condition $\delta_T = 0$ as we did in figure 2.4b. Figure 2.4c shows that the dynamic is well described by equation (2.43) due to the similarity between panel (b) and (c) in figure 2.4. Therefore, noisy single cells lag time distributions (the red genotype), or bet-hedging strategies, should be expected for large N_0^{tot} as long as a fraction of the population is well adapted for the new coming environment i.e. short lags cells exists. In the opposite, for small N_0^{tot}, regulated sense and responses strategies (the low lag noise blue genotype) will be preferred due to the absence of long lags bacteria.

In order to quantify the strength of this effect we compare the fixation probability of this phenomena with pure genetic drift. Pure genetic drift would predict [25] that, if no selection is acting, the fixation probability of the red genotype equal its initial fraction ρ. In figure 2.4d we show the fixation probability of the red genotype as function of its initial fraction ρ and for different initial populations sizes N_0^{tot}. The dashed black line represents the genetic drift and the distance from this line quantify the strength i.e. "advantage/disadvantage" one strategy has. By construction the two strategies performs equally well for $N_0^{\text{tot}} = 3 \times 10^4$. However, for $N_0^{\text{tot}} > 3 \times 10^4$ the red genotype has more chance to be fixed than simple genetic drift and vice versa for $N_0^{\text{tot}} < 3 \times 10^4$ as expected. As said, the difference between the actual fixation probability and the genetic drift is a measure of the "strength" of selection. When it is positive the red genotype would be preferred over the blue one and vice versa when it is negative. We observe for example that for large N_0^{tot} and small ρ the advantage the red genotype has is the strongest.

2.6 Discussion

Wild type *E.coli* resume growth stochastically when exposed to new conditions [22][1]. This phenotype may confer to the organism an evolutionary advantage [1][21] and [12] suggested that *E.coli* implement a bet-hedging strategy where, through phenotype randomization, different cells are adapted to different kinds of environments. However, we have shown that, due to the exponential growth of bacterial populations, the time a specific cell needs to exit the lag has an exponential impact on its number of descendants. This observation let us hypothesize that selection is strong on the head of the lag distribution and weak on the tail

i.e. mutations acting on the head of the lag distribution are strongly selected whereas the one acting on the tail contribute less to the mutant fixation probability and so not strongly selected. This translate into the fact that deleterious mutations acting on the tail of the lag distribution are expected to accumulate and so long tailed lag distribution should not be rare to be observed. Therefore, in our interpretation, long tailed lag distribution is not a phenotype which bacteria are actively maintaining but rather an unavoidable trait. We then showed the bulk lag time to also be a stochastic quantity and studied the non trivial relationship between the bulk lag time distribution and the single cell lag time distribution. In particular, we studied the impact of the single cell lag distribution shape and the inoculum size on the expected bulk lag time. We showed that the expected bulk lag time is longer when the initial colony size is small and this effects is stronger when the single cell lag distribution is long tailed. To understand the consequences of this observation on bacteria evolution, we simulated bacterial colonies living in fluctuating environments. As predicted by our theory we showed that noisy lag distributions are effective for large populations as far as a subset of bacteria can adapt fast to the new environment but are inefficient in small populations. This suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environmental changes, small populations will require regulated sense-and-response strategies in order to maximise their survival chances. Last, we studied the potential problems which may arise when we define the log genotype fraction as a measure of fitness. This fitness measure is often used in evolutionary experiments and we show that, due to the colony size dependence of the population lag, one may overestimate the fitness of the more abundant genotype even in cases where no selection is acting. This fictitious selection force is an example of a more general theory developed by [13].

2.7 SUPPLEMENTARY

2.7.1 The moments of the lag distribution

The variable $f \in [0,1]$ is assumed to be Gaussian distributed

$$p(f) = \text{Norm}\, e^{\frac{(f-\mu)^2}{2\tilde{\sigma}^2}} \quad \text{where} \quad \mu = \left\langle e^{-\tau} \right\rangle,\ \tilde{\sigma}^2 = \frac{\text{Var}\left[e^{-\tau}\right]}{N_0} \tag{S.1}$$

and the normalization term Norm equal to

$$\text{Norm} = \frac{\sqrt{2}}{\sqrt{2\pi\tilde{\sigma}^2}\left(\text{erf}\left(\frac{1-\mu}{\sqrt{2}\sqrt{\tilde{\sigma}^2}}\right) + \text{erf}\left(\frac{\mu}{\sqrt{2}\sqrt{\tilde{\sigma}^2}}\right)\right)} \tag{S.2}$$

The population lag variable $T = -\log f$ is therefore distributed as

$$p(T) = \text{Norm}\, e^{-T} e^{-\frac{(e^{-T}-\mu)^2}{2\tilde{\sigma}^2}} \tag{S.3}$$

and the moments of this distribution read

$$\langle T^\alpha \rangle = \text{Norm} \times (-1)^\alpha \int_{-\mu}^{1-\mu} \left(\log\mu + \log(1+\frac{x}{\mu})\right)^\alpha e^{-\frac{x^2}{2\tilde{\sigma}^2}} dx \tag{S.4}$$

To compute this integral we have to realize that the Gaussian term centered in zero and with standard deviation $\tilde{\sigma}$ smaller, by construction to the mean i.e $\tilde{\sigma} < \mu$, is dominating the integration range. This allow us to expand the logarithm and to extend the range of integration to $(-\infty, \infty)$

$$\begin{aligned}
\langle T^\alpha \rangle &\approx \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}(-1)^\alpha \int_{-\infty}^{\infty} \left(\log\mu + \frac{x}{\mu} - \frac{x^2}{2\mu^2}\right)^\alpha e^{-\frac{x^2}{2\tilde{\sigma}^2}} dx \\
&= \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}}(-1)^\alpha \int_{-\infty}^{\infty} \left(\log^\alpha\mu + \alpha\frac{x}{\mu}\log^{\alpha-1}\mu - \alpha\frac{x^2}{2\mu^2}\log^{\alpha-1}\mu \right. \\
&\left. + \frac{\alpha(\alpha-1)}{2}\frac{x^2}{\mu^2}\log^{\alpha-2}\mu\right) e^{-\frac{x^2}{2\tilde{\sigma}^2}} dx \\
&= (-1)^\alpha \left(\delta_{\alpha\geq 1}\log^\alpha\mu - \delta_{\alpha\geq 1}\alpha\frac{\tilde{\sigma}^2}{2\mu^2}\log^{\alpha-1}\mu + \right. \\
&\left. \delta_{\alpha\geq 2}\frac{\alpha(\alpha-1)}{2}\frac{\tilde{\sigma}^2}{\mu^2}\log^{\alpha-2}\mu\right) + \mathscr{O}\left(\frac{\tilde{\sigma}^3}{\mu^3}\right)
\end{aligned} \tag{S.5}$$

where

$$\delta_{i\geq j} = \begin{cases} 1 & \text{if } i \geq j \\ 0 & \text{otherwise} \end{cases} \tag{S.6}$$

Particularly interesting for us are the mean

$$\langle T \rangle = -\log \mu + \frac{\tilde{\sigma}^2}{2\mu^2} + \mathscr{O}\left(\frac{\tilde{\sigma}^3}{\mu^3}\right) \tag{S.7}$$

and the variance

$$\mathrm{Var}[T] = \frac{\tilde{\sigma}^2}{\mu^2} + \mathscr{O}\left(\frac{\tilde{\sigma}^3}{\mu^3}\right) \tag{S.8}$$

2.7.2 Feast and famine experiment

To simulate a feast and famine experiment (figure 2.3) we assume the two competing genotypes (red and blue) face always the same cycle type and they differ only by their lag distribution $p(\tau)$ defined in figure 2.4a. The two lag distribution $p(\tau^{red})$ and $p(\tau^{blue})$ were chosen such that $\delta T = \langle T^{blue} \rangle - \langle T^{red} \rangle = 0$ for $N_0^{\mathrm{tot}} = 3 \times 10^4, \rho = 0.5$ i.e. there is not lag advantage for $N_0^{\mathrm{tot}} = 3 \times 10^4, \rho = 0.5$.
At the begin of the first cycle, the total population is of N_0^{tot} bacteria with a fraction ρ of red cells and $1-\rho$ of blue cells. Therefore, we randomly sampled $N_0^{tot}\rho$ from the "red" $p(\tau^r)$ distribution and $N_0^{tot}(1-\rho)$ from the "blue" lag distribution $p(\tau^b)$. Then we compute the new fraction (ρ') of red bacteria at the end of the growth cycle

$$\rho' = \frac{\sum_{i=0}^{\rho N_0^{tot}} e^{-\tau_i^r}}{\sum_{i=0}^{\rho N_0^{tot}} e^{-\tau_i^r} + \sum_{i=0}^{(1-\rho)N_0^{tot}} e^{-\tau_i^b}} \tag{S.9}$$

The famine cycle has been simulated by considering binomial sampling. This means that at the begin of the next iteration we sample N_0^{tot} cells with probability ρ' to be red and $1-\rho'$ to be blue. We iterate this procedure of growth and famine until one of the two genotype get extinct. By repeating this simulation several times with different N_0^{tot} we can compute the probability for a genotype to get extinct depending on N_0^{tot}.
Note that we can also theoretically compute ρ' after the first iteration

$$\rho \to \rho' = \frac{\rho}{\rho + (1-\rho)e^{T^{\mathrm{red}} - T^{\mathrm{blue}}}} \tag{S.10}$$

where the bulk lag random variables $T^{\text{red/blue}}$ have the distribution (2.17) with parameters

$$\mu = \left(\frac{\beta}{\beta+1}\right)^{\alpha} \quad \text{and} \quad \sigma^2 = \left(\frac{\beta}{\beta+2}\right)^{\alpha} - \mu^2 \tag{S.11}$$

Clearly we can also easily compute the expected lag advantage over the first cycle

$$\delta T \stackrel{\text{def}}{=} \left\langle T^{\text{blue}} \right\rangle - \left\langle T^{\text{red}} \right\rangle \tag{S.12}$$

since

$$\left\langle T^{\text{red}} \right\rangle = -\log\left(\frac{0.3}{0.3+1}\right)^2 + \frac{1}{2N_0^{\text{tot}}\rho}\left(\left(\frac{(0.3+1)^2}{0.3(0.3+2)}\right)^2 - 1\right) \tag{S.13}$$

and

$$\left\langle T^{\text{blue}} \right\rangle = -\log\left(\frac{1.25}{1.25+1}\right)^5 + \frac{1}{2N_0^{\text{tot}}(1-\rho)}\left(\left(\frac{(1.25+1)^2}{1.25(1.25+2)}\right)^5 - 1\right) \tag{S.14}$$

Chapter 3

A Bayesian model to infer the gene expression dynamics.

3.1 Introduction

One of the first method developed to measure cell growth and gene expression at the single cell level consisted in the use of agarose patches, on which cell grow and form microcolonies, combined with quantitative fluorescence time-lapse microscopy [61]. Whereas the cell size was directly visible through microscope images, gene expression was monitored through genetically encoded fluorescent proteins, such as the green fluorescent protein (GFP), which intensity reflects the activity of the promoter studied. Two main problems arose when using such methods. One is that the size of the microcolony grows so quickly that soon most of the colony is out of the microscope field view. The other is that microcolonies growing on agarose patches form multi-layers which make the monitoring of the single cells impossible. Microfluidic devices solved these problems by flushing away the cell progeny and so drastically increasing the observation time. Among the various microfluidic approaches [56][55] we focus on the so-called Mother Machine [56], a device designed to study long-term growth in *E.coli*. As shown in figure 3.1(a) mother machine has several small channels, closed in one side, where bacteria are trapped. Nutrients and other products can diffuse in and out of these channels through the part connected to the main tube in which the medium constantly flows. The growth channels are approximately $20[\mu m]$ long with a cross section of $\sim 1[\mu m] \times [1\mu m]$. Because of that, *E.coli* are stacked one over the other leaving one cell trapped at the bottom of the channel (bottom cell). During the time course, cells divide and push their progeny up until they leave the growth channel. While almost

all the cells can only be observed for a relatively short time before they leave the growth channel, the bottom cell can be monitored during the entire experiment duration.

As an example [22], in figure 3.1b we show a time series of microscope images of a single growth-channel where *E.coli*, that carries a translational lacZ-GFP fusion at the native locus, are exposed to alternate carbon sources (glucose/lactose). To analyse these images, i.e. to automatically segment and track cells and division events, and to quantify the cell size and levels of fluorescence; *Kaiser et al.* developed the MoMa software [22]. In figure 3.1c we show, as an example, some of the MoMa predicted cell volume and fluorescence levels of the previously mentioned experiment. For more details about this experiment, we refer to [22]. Even-tough the MoMa software precisely measures the cells sizes and the levels of the fluorescent proteins, these measurements are not free from measurement noise. In the next sections we will propose two different strategies, both based on kriging, to reduce the effects of measurement noise. First, we will introduce what kriging, or Gaussian process, regression is and how it is used on the MoMa time series data. Then we will introduce a biophysical model which, combined with the kriging technique, allows us to predict the dynamic of some important latent variables[1] and disentangle promoter specific fluctuations from other noise sources.

3.2 Gaussian Processes Regression in general

Gaussian process regression, or kriging, is a regression model widely used in machine learning. We here present the main concepts behind this method and we refer to the vast literature [7][19] for more details.

3.2.1 Gaussian distribution and Gaussian identities

Before explaining what Gaussian processes are and how regression with these models is done, let's remind some basic concepts and relations on Gaussian distributions.

Definition The $n-$dimensional random vector

$$\vec{x} = [x_1, \ldots, x_n]^T \tag{3.1}$$

[1] Latent variables are variables that are not directly observed but are rather inferred (through a mathematical model) from observed variables.

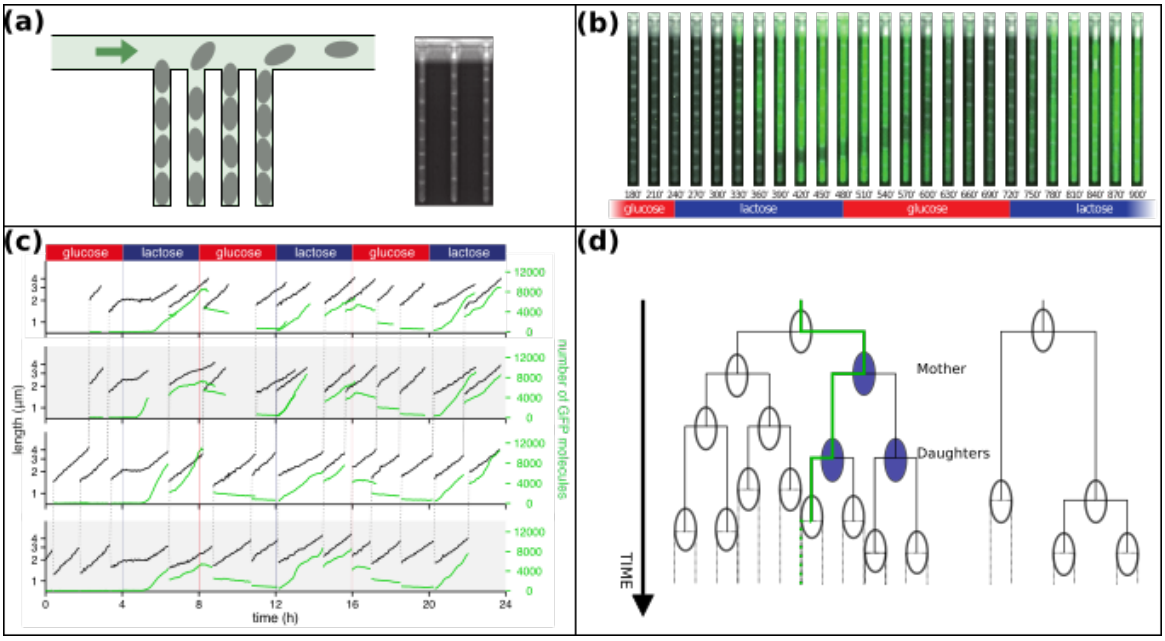

Fig. 3.1 **(a)** On the left a schematic representation of the mother machine design. Bacteria (gray ellipse) are growing one over the other in the four dead-end growth channels. The media (green) constantly flows in the main tube (green arrow) and diffuses inside the channels. On the right a real phase-contrast microscope image of three mother machine growth channels with *E.coli* inside. **(b)** A time series of microscope images of a single growth-channel where *E.coli* strain, that carries a translational lacZ-GFP fusion at the native locus, are exposed to alternate carbon source (glucose/lactose) (figure from [22]). **(c)** Single cells growth and gene expression dynamics for the experiment described in **(b)**. The data are obtained using the MoMa software [22] where the cell size (black, log scale) and the LacZ-GFP expression (green, linear scale) are shown as a function of time. Dashed vertical lines show the lineage of cell division and connects mother cells with their respective daughter cells (figure from [22]). **(d)** A schematic example of cell division inside the mother machine. Two growth channels with their respective off-springs are represented. We define cell lineage the division history of a particular cell (green).

is said to be Gaussian distributed with mean

$$\vec{\mu} = [\langle x_1 \rangle, \ldots, \langle x_n \rangle]^T \tag{3.2}$$

and covariance matrix C

$$C_{ij} = \left\langle (x_i - \vec{\mu}_i)\left(x_j - \vec{\mu}_j\right) \right\rangle \quad \forall i,j = 1,\ldots,n \tag{3.3}$$

if

$$\vec{x} \sim \frac{1}{\sqrt{(2\pi)^n \det C}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})C^{-1}(\vec{x}-\vec{\mu})} \tag{3.4}$$

We note this distribution $\mathcal{N}(\vec{x}|\vec{\mu},C)$.

Let's now list some useful results on Gaussian distributions, others can be found in [7].

Multiplication The multiplication of two Gaussian distributions results in another Gaussian distribution

$$\mathcal{N}(\vec{x}|\vec{a},A)\,\mathcal{N}\left(\vec{x}|\vec{b},B\right) \propto \mathcal{N}(\vec{x}|\vec{c},C) \tag{3.5}$$

where

$$\vec{c} = CA^{-1}\vec{a} + CB^{-1}\vec{b} \tag{3.6}$$

$$C = \left(A^{-1} + B^{-1}\right)^{-1} \tag{3.7}$$

Convolution Let $\vec{x},\vec{y}$ be two generic $n-dimensional$ Gaussian distributed random vectors where $\vec{x} \sim \mathcal{N}(\vec{x}|\vec{a},A)$ and $\vec{y}-\vec{x} \sim \mathcal{N}(\vec{y}-\vec{x}|\vec{b},B)$. Then the convolution of them simply reads

$$\int d\vec{x}\,\mathcal{N}(\vec{x}|\vec{a},A)\,\mathcal{N}(\vec{y}-\vec{x}|\vec{b},B) = \mathcal{N}(\vec{y}|\vec{a}+\vec{b},A+B) \tag{3.8}$$

Propagation Let $\vec{x},\vec{y}$ be two generic $n-dimensional$ Gaussian distributed random vectors where $\vec{y} \sim \mathcal{N}(\vec{y}|\vec{b},B)$ and $\vec{x} \sim \mathcal{N}(\vec{x}|F\vec{y}+\vec{a},A)$. Then the propagation reads

$$\int d\vec{y}\,\mathcal{N}(\vec{x}|F\vec{y}+\vec{a},A)\,\mathcal{N}(\vec{y}|\vec{b},B) = \mathcal{N}\left(\vec{x}|F\vec{b}+\vec{a},A+FBF^T\right) \tag{3.9}$$

Marginal Consider the $2n-dimensional$ Gaussian distributed random vector

$$\begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \middle| \begin{bmatrix} \vec{a} \\ \vec{b} \end{bmatrix}, \begin{bmatrix} A & E \\ E^T & B \end{bmatrix}\right) \tag{3.10}$$

In order to obtain the marginal distribution for the variable $\vec{x}$ starting from the join distribution (3.10) we simply integrate over $\vec{y}$ and obtain

$$\vec{x} \sim \mathscr{N}(\vec{x}|\vec{a},A) \tag{3.11}$$

Conditional From the joint distribution (3.10) it is also easy to obtain the conditional distribution of $\vec{x}$ given $\vec{y}$

$$\vec{x}|\vec{y} \sim \mathscr{N}\left(\vec{x}\middle|\vec{a}+EB^{-1}\left(\vec{y}-\vec{b}\right),A-EB^{-1}E^{T}\right) \tag{3.12}$$

In all the above equations A,B,F,E are $n \times n$ matrices and $\vec{a},\vec{b}$ are $n-dimensional$ vectors.

3.2.2 Gaussian process regression

Gaussian process regression, or kriging, is a method of interpolation for which the interpolated values are modeled by a Gaussian process. Imagine a quantity $x(t) \in \mathbb{R}$ that varies over time t to which we do not know its functional form. In general a regression problem has the objective to learn the function $x(t)$ given a finite series of measurements of this quantity $\mathscr{D} = \{x(t_1),x(t_2),...,x(t_n)\}$. This is an ill-defined problem since there are infinitely many functions that take the same values at $(t_1,t_2,\ldots,t_n)$ but differs elsewhere. To overcome this problem, we have to make some additional assumptions and for the Gaussian processes regression method we assume $\{x(t) : t \in \mathbb{R}\}$ to be a stochastic[2] Gaussian process. A stochastic process $\{x(t) : t \in \mathbb{R}\}$ is said to be Gaussian if and only if any sub-collections $\{x(t_1)\ldots,x(t_n)\}$ is Gaussian distributed. Equivalently, the stochastic process $\{x(t) : t \in \mathbb{R}\}$ with mean function

$$\mu(t) = \langle x(t) \rangle \tag{3.13}$$

and covariance function

$$k(t,s) = \left\langle \left(x(t) - \langle x(t) \rangle\right)\left(x(s) - \langle x(s) \rangle\right) \right\rangle \tag{3.14}$$

is Gaussian, only if any $n-dimensional$ vector $\vec{X} = [x(t_1),x(t_2),...,x(t_n)]^T$ follows

$$\vec{X} \sim \mathscr{N}\left(\vec{X}\middle|\begin{bmatrix} \mu(t_1) \\ \vdots \\ \mu(t_n) \end{bmatrix}, \begin{bmatrix} k(t_1,t_1) & \ldots & k(t_1,t_n) \\ \vdots & \ddots & \vdots \\ k(t_n,t_1) & \ldots & k(t_n,t_n) \end{bmatrix}\right) \tag{3.15}$$

[2] We can think a stochastic process to be represented by numerical values of some system randomly changing over time.

As we will see in the next paragraphs, once we choose a suitable mean function $\mu(\cdot)$ and a suitable covariance function $k(\cdot,\cdot)$, the regression problem becomes tractable.
As an example of valid covariance function we present the squared exponential kernel

$$k(t_i,t_j|\Theta) = \alpha e^{-\frac{(t_i-t_j)^2}{2\gamma}} \tag{3.16}$$

and the rational quadratic kernel

$$k(t_i,t_j|\Theta) = \alpha\left(1+\frac{(t_i-t_j)^2}{2\gamma^2\beta}\right)^{-\beta} \tag{3.17}$$

Many others are listed in [7]. Note that these functions usually depends on some parameters $(\alpha,\beta,\gamma,\dots)$ (hyperparameters) which we grouped under a unique symbol Θ. One of the hardest tasks in using the Gaussian regression method is actually the choice of a "good" mean and a "good" covariance function to describe the data but once this is done the regression is tractable and consists in two main steps i.e. maximise the marginal likelihood and find the posterior.

Maximise the marginal likelihood In this paragraph we will work out the marginal likelihood for the Gaussian process regression problem. Let $\vec{X} = [x(t_1),\dots,x(t_n)]^T$ be the vector collecting all the measurements of the quantity x at time $(t_1,\dots,t_n)$ and let $\mu(\cdot|\Theta)$ and $k(\cdot,\cdot|\Theta)$ be the mean and covariance function we chose to describe this process. Then, if we assume $\{x(t) : t \in \mathbb{R}\}$ to be a Gaussian process, the likelihood of the observed data is given by (3.15) which depends on the hyperparameters Θ. We use the maximum likelihood estimator (MLE) in order to estimate the hyperparameters Θ of (3.15). This means that we consider $\Theta^\star$ to be the optimal parameter set only if it maximise the likelihood i.e. if the gradient is zero (3.15)

$$\left.\frac{\partial p\left(\vec{X}|\Theta\right)}{\partial\Theta_j}\right|_{\Theta^\star} = 0 \tag{3.18}$$

and the Hessian matrix H, where

$$H_{ji} = \left.\frac{\partial^2 p\left(\vec{X}|\Theta\right)}{\partial\Theta_j\partial\Theta_i}\right|_{\Theta^\star} \tag{3.19}$$

must be negative definite.
In this way we find the optimal hyperparameter set $\Theta^\star$ in the case of noiseless observations.

However, real world measurements usually contain measurement errors. For additive and non correlated Gaussian measurement errors, the observed values $y(t_j)$ of the quantity $x(t_j)$ at time t_j can be written as

$$y(t_j) = x(t_j) + \varepsilon(t_j) \tag{3.20}$$

where the measurement error ε is assumed to be Gaussian distributed with mean zero and covariance $\langle \varepsilon(t_i)\varepsilon(t_j)\rangle = \sigma_i\sigma_j\delta_{ij} = \sigma_i^2\delta_{ij}$. The marginal likelihood[3] $P(\vec{Y}|\Theta)$ of the observed values $\vec{Y} = [y(t_1),\ldots,y(t_n)]^T$ can easily be computed if we assume again $\{x(t) : t \in \mathbb{R}\}$ to follow a Gaussian process

$$P(\vec{Y}|\Theta) = \int d\vec{x}\, P(\vec{Y}|\vec{X},\Theta)P(\vec{X}|\Theta) \tag{3.21}$$

which gives (3.8)

$$\vec{Y}|\Theta \sim \mathcal{N}\left(\vec{Y}\,|\vec{\mu}, K+D\right) \tag{3.22}$$

where the components[4] $\mu_j = \mu(t_j)$, $K(t_i,t_j) = k(t_i,t_j)$ and $D_{ij} = \sigma_i^2\delta_{ij}$. Again, we consider the optimal hyper-parameters $\Theta^\star$ as the one which optimize the marginal likelihood (3.22). Note that knowing the marginal likelihood (3.22) also allows us to compare different models i.e. to compare different mean functions $\mu(\cdot)$ and covariance functions $k(\cdot,\cdot)$, among themselves. Indeed, the "best" model will be the one with the highest marginal likelihood.

Predictions The main goal of the kriging, and of any regression in general, is to predict the value of the quantity $x(t^\star)$ at some generic time $t^\star$ having observed $\{x(t_1),\ldots,x(t_n)\}$. More precisely, let's imagine we would like to predict the values of the quantity x at times $t_1^\star,\ldots,t_m^\star$ i.e. the vector $\vec{X}^\star = [x(t_1^\star),\ldots,x(t_m^\star)]^T$, knowing the noisy observations $\vec{Y} = [y(t_1),\ldots,y(t_n)]^T$ of the quantity x at time $(t_1,\ldots,t_n)$. As we will see, since the stochastic process is assumed to be Gaussian, we only need to find the mean and covariance function of the join distribution of $\left[\vec{Y},\vec{X}^\star\right]^T$ in order to do this. First note that

$$\langle x(t_j^\star)\rangle = \mu(t_j^\star) \quad \text{and} \quad \langle y(t_j)\rangle = \mu(t_j) \tag{3.23}$$

$$\text{Cov}\left[y(t_i),x(t_j^\star)\right] = \text{Cov}\left[x(t_i),x(t_j^\star)\right] = k(t_i,t_j^\star) \tag{3.24}$$

$$\text{Cov}\left[y(t_i),y(t_j)\right] = k(t_i,t_j) + \sigma_i^2\delta_{ij} \tag{3.25}$$

[3]The measurment errors σ_j and the hyper-parameters of the mean and covariance function are all contained in the symbol Θ.

[4]We drop the explicit dependence on Θ to keep a more readable notation.

therefore the join distribution reads

$$\left[\begin{array}{c}\vec{Y}\\ \vec{X}^\star\end{array}\right]\Big|\Theta\sim\mathscr{N}\left(\left[\begin{array}{c}\vec{Y}\\ \vec{X}^\star\end{array}\right]\Big|\left[\begin{array}{c}\vec{\mu}\\ \vec{\mu}^\star\end{array}\right],\left[\begin{array}{cc}K+D & K^\star\\ K^{\star T} & K^{\star,\star}\end{array}\right]\right) \tag{3.26}$$

where $\vec{\mu}_i^\star=\mu(t_i^\star)$, $K_{i,j}^\star=k(t_i,t_j^\star)$, $K_{i,j}^{\star\star}=k(t_i^\star,t_j^\star)$ and $D_{ij}=\sigma_j^2\delta_{ij}$. Using the rules for conditioning Gaussians distribution (3.12) we obtain

$$\vec{X}^\star|\vec{Y},\Theta\sim\mathscr{N}(\vec{m},C) \tag{3.27}$$

with

$$\begin{aligned}\vec{m}&=\vec{\mu}^\star+K^{\star T}(K+D)^{-1}(\vec{y}-\vec{\mu})\\ C&=K^{\star\star}-K^{\star T}(K+D)^{-1}K^\star\end{aligned} \tag{3.28}$$

this means that we are able to predict the quantity x at times $(t_1^\star,\ldots,t_m^\star)$ knowing the noisy observations $\vec{Y}$ and the hyperparameters $\Theta^\star$ (estimated through MLE). Now that we described how kriging works in general, let us apply it to the MoMa time series data.

3.2.3 Gaussian processes for single cell time series

The software we develop to treat MoMa time series data can be downloaded at https://github.com/fioriathos/gaussian_smoothing.git. The rest of this section will describe the main concepts of this algorithm and give an example of use (figure 3.2).
Let $\vec{Y}^K=\left[y^k(t_0^k),\ldots,y^k(t_n^k)\right]^T$ be the fluorescent protein level, or the log cell size, estimated by MoMa for the cell k from its birth (t_0^k) to its division (t_n^k). We use the notation $y^k(t_j^k)$ to indicate the j^{th} observation of the noisy quantity x for bacteria k after its division. Let N be the total number of different cells growing in similar conditions during the entire experiment[5] duration i.e. we consider the data-set $\mathscr{D}=\left\{\vec{Y}^1,\ldots,\vec{Y}^N\right\}$. In order to use the Gaussian process regression method explained above we need to choose a suitable mean and covariance function. As suggested by [7], instead of giving an explicit form of the mean function of the process we consider

$$z^k(t_j^k)=y^k(t_j^k)-\bar{y}(t_j)\quad\text{where}\quad\bar{y}(t_j)=\sum_{i=1}^{N}\frac{y^i(t_j^i)}{N} \tag{3.29}$$

[5]If the environment switches we have to consider cells growing in different media belonging to different data-set.

to have mean zero. For the covariance function we assume

$$\left\langle z^k(t_j^k), z^l(t_i^l) \right\rangle = \alpha e^{-\frac{(t_j^k - t_i^l)^2}{2\gamma}} \delta_{lk} \tag{3.30}$$

which assumes an exponential decaying correlation (3.16) within a cell cycle and independence among observations of different cell cycles. Since we consider independence among the cells we can speed up the computations of the log likelihood by making use of the fact that

$$\log P\left(\vec{Z}^1, \ldots, \vec{Z}^n | \Theta\right) = \sum_{i=1}^{N} \log P\left(\vec{Z}^i | \Theta\right) \tag{3.31}$$

where $\vec{Z}^k = \left[z^k(t_0^k), \ldots, z^k(t_n^k)\right]^T$ is the rescaled vector (3.29) and $P\left(\vec{Z}^i | \Theta\right)$ is the likelihood (3.22) with zero mean and covariance given by equation (3.30) with $\Theta = \{\alpha, \gamma\}$. The optimal parameter set $\Theta^\star$ is the MLE of (3.31) and it is compute through the quasi-Newton BFGS algorithm developed in [11]. The gradient of (3.31) has been analytically computed in order to increase the speed of convergence of the BFGS algorithm. The line search algorithm starts from an initial random guess of Θ and updates it until it finds the value $\Theta^\star$ which maximise the likelihood function. This method may converge for local optimum as well and in order to avoid it we repeat the line search from different initial conditions Θ and select the best one ($\Theta^\star$) to be the one with largest likelihood. This should ensure that $\Theta^\star$ is actually a global optimum. Once the best hyperparameters $\Theta^\star$ are known, it is easy to do predictions of the quantity x at any time point using equation (3.27). Being able to predict the noiseless quantity x at any time point is particularly useful for computing time derivatives x'. If, for example, we would like to know the time derivative $x'(t)$ at time t of x, we would consider the mean $\bar{x}(t)$ as the "best prediction" of x at time t. Using the central difference method we find

$$x'(t) = \frac{\bar{x}\left(t + \frac{\Delta}{2}\right) - \bar{x}\left(t - \frac{\Delta}{2}\right)}{\Delta} \tag{3.32}$$

where $\Delta > 0$ is a small time step.

3.3 A biophysical model for the Gaussian process regression

In the previous section we showed how the Gaussian process regression brings new light into the analysis of time lapse microscopy data. However, the previous formulation lacks some

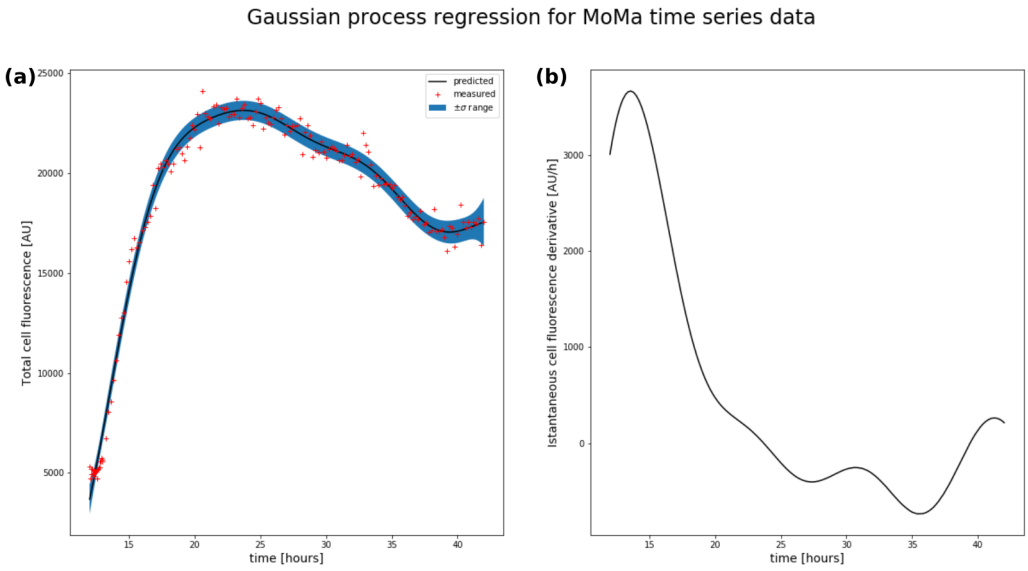

Fig. 3.2 *E.Coli* strain CGSC#6300 with a Yellow fluorescent protein under the control of a synthetic promoter in conditioned media. **(a)** The raw fluorescent amplitude (red) for one *E.Coli* cell from birth to division. The mean (black) and standard deviation (blue) of the predicted fluorescent amplitude distribution. **(b)** The discrete time derivative (3.32) with $\Delta = 1\ [min]$. Preliminary data from Théo Gervais.

important aspects of the modeling of the underlying biophysical model which can lead to serious problems. One of them is the strong assumption of independence between bacteria (3.31) which is clearly unrealistic for binary division. In binary division the "mother cell" divides into two copies and it is hard to believe that the independence assumption holds true for bacteria just few division apart. Moreover, assuming a squared exponential covariance function is simply a way to assume for very smooth regression functions and does not come from any biological model of gene expression or cell growth. This makes very hard to interpret the underlying biology in a meaningful way. Having a realistic model of cell growth and gene expression not only will allow us to better interpret the underlying biology but will also allow us to investigate latent variables, like the mRNA number or the cell growth rate, which are not directly measured in our experiments. For all these reasons we developed a method which combines the Gaussian processes regression with a biophysical model of cell growth and gene expression, to treat the previously mentioned time lapse microscopy data.

3.3.1 The biophysical model

The precise description of all the biochemical reactions involved in the biomass production and gene expression at the single cell level is clearly out of the scope of this section. We here present a simple stochastic model of cell size growth and gene expression which, combined to the previously presented Gaussian process regression method, allows us to better analyze the MoMa data-set i.e. time series data of cell size and fluorescence levels of the target promoter. The key point of this model is to assume the single cell growth rate and the gene expression rate to follow a stationary Gauss–Markov process known as Ornstein-Uhlenbeck process.

Cell size Cell volume growth is the result of many biophysical reactions inside the cell wall and, as shown by [53], bacterial volume growth can reasonably be well described by an exponential function. However, due to the stochastic nature of the biophysical reactions, the cell size deviations from the perfect exponential reflect the stochastic nature of the underlying process. Our goal is to go beyond the assumption of perfect exponential growth and to develop a model which takes into consideration fluctuations of the cell size within a cell cycle. To do this we make use of the Ornstein-Uhlenbeck process, one of the simplest stochastic process which describes random fluctuations of a stochastic variable λ around a fixed value $\bar{\lambda}$. The Ornstein–Uhlenbeck process $\lambda_t = \lambda(t)$ is a stationary Gauss–Markov process defined by the following stochastic differential equation

$$\frac{d\lambda_t}{dt} = -\gamma_\lambda(\lambda_t - \bar{\lambda}) + \sigma_\lambda \eta_1(t) \tag{3.33}$$

or in the integral form

$$\lambda_t = \bar{\lambda}\left(1 - e^{-\gamma_\lambda t}\right) + \lambda_0 e^{-\gamma_\lambda t} + \sigma_\lambda \int_0^t e^{\gamma_\lambda(\tau - t)} \eta_1(\tau) \tag{3.34}$$

where $\gamma_\lambda > 0$, $\sigma_\lambda > 0$, $\bar{\lambda}$ are parameters and $\eta_1(t)$ denotes a Wiener process (Brownian motion). An intuitive way to understand this process is to consider the equation of motion $\left(\frac{d\lambda_t}{dt}\right)$ of a particle transported by a river. Clearly the average speed of this particle is close to the speed of the river $(\bar{\lambda})$ in which the particle is transported. However, it may happen that this particle flows a bit faster or a bit slower than the river, depending on the obstacles it encounters or the particle-particle collisions it does. These random events are model through a Brownian motion $\eta_1(t)$ with strength σ_λ. Note that the speed difference between the particle and the river can not last forever due to viscous force $-\gamma_\lambda(\lambda_t - \bar{\lambda})$. The viscous

force ensures the process to drift forward its mean value $\bar{\lambda}$ with a characteristic time $\frac{1}{\gamma_\lambda}$.

Let's now consider the cell size over time $(s(t))$ to follow an exponential function

$$s(t) = s_0 \exp\left[\bar{\lambda} t\right] \tag{3.35}$$

where s_0 is the cell size at the begin of the cell cycle and $\bar{\lambda}$ the exponential growth rate. Our model assumes that, instead of having a fixed exponential growth rate $\bar{\lambda}$ along the entire cell cycle, we consider it to randomly fluctuate as described in (3.33). Therefore, if no division event occurs between t_0 and t, the log cell size $x_t = \log s(t)$ is simply the time integral from birth, at t_0, to t of the growth rate

$$x_t = x_0 + \int_{t_0}^{t} \lambda_\tau d\tau \tag{3.36}$$

where $x_0 = \log s(t_0)$. Three parameters γ_λ, $\bar{\lambda}$ and σ_λ together with (3.33) and (3.36) allow us to describe the cell size dynamic.

Gene expression Another quantity measured by the MoMa software is the amount of fluorescent proteins produced from the targeting promoter. As we know, gene expression involves many noisy processes, stochastic binding/unbinding of ribosomes and RNAP, burst in transcription, cell to cell variations in ribosomes/RNAP and nutrients, etc. All these introduce noise in the observed proteins copy numbers and a precise model which takes into account all these noise sources is out of our scope. However, we will think the noise affecting the protein copy numbers to belong to two different classes and model them as two independent Ornstein-Uhlenbeck processes. The first class is the one non-specific to the target protein and affecting all gene products equally. An example of non-specific noise is the fluctuation in the ribosomes levels which we expect to affect all the transcripts almost equally. The other class of noise is the one specific to the gene of interest and not affecting the other gene products. Transcription burst for example can be considered as a specific noise term. Since the cell volume scales roughly linear with the protein content [40] we use it as a proxy of the "non-specific" fluctuations and we model the specific fluctuations $q_t = q(t)$ as an independent Ornstein-Uhlenbeck process

$$\frac{dq_t}{dt} = -\gamma_q(q_t - \bar{q}) + \sigma_q \eta_2(t) \tag{3.37}$$

where again $\gamma_q, \bar{q}, \sigma_q$ are parameters and $\eta_2(t_2)$ denotes a Wiener process. The equation governing the fluorescent protein numbers $g_t = g(t)$ at time t reads

$$\frac{dg}{dt} = s_t q_t - \beta g_t \tag{3.38}$$

where β is the term taking into account protein decay and photobleaching. Note that we can interpret (3.38) as follow. Let m_t be the number of transcript of the target gene at time t, R_t be the number of free ribosomes within the cell at time t and α the translation rate. Then translation can be described through the differential equation

$$\frac{dg}{dt} = \alpha R_t m_t - \beta g_t \tag{3.39}$$

If we consider that ribosomes also scale with the cell volume i.e. $R_t \propto s_t$ then q_t is proportional to the mRNA levels. In addition to having a biological interpretation, equation (3.38) allows us to disentangle promoter specific fluctuations from other noise sources.
Equations (3.33), (3.36), (3.37) and (3.38) fully describe the cell size and genetic expression dynamic over the time course and, in the following sections, we will show that these equations, together with the previously presented kriging method, allow us to do precise estimations of them.

3.3.2 The mean and covariance function given by the model

Gaussian processes are defined through suitable means and covariance functions. In this section we will compute the mean and covariance function given by the biophysical model described above. First define the four dimensional cell state vector at time t as

$$\vec{z}_t = \begin{pmatrix} x_t \\ g_t \\ \lambda_t \\ q_t \end{pmatrix} \tag{3.40}$$

i.e. the first component represents the log cell size x_t at time t , the second represents the fluorescent protein levels g_t at time t and the third and forth represents the cell growth rate λ_t and protein production per unit volume (q_t) at time t. We must now find the 4 dimensional mean vector function $\langle \vec{z}_t \rangle$ and the 4×4 covariance matrix function $\langle \vec{z}_t, \vec{z}_s \rangle$ for time $t, s > 0$ given by this process. The rest of this section is dedicated to these computations but before entering the details of such computations let us give some usefult trick when dealing with Gaussian integrals.

Wick theorem with source term

Let $\vec{x}$, $\vec{\mu}$ and $\vec{J}$ be generic n dimensional vectors and C be a generic $n \times n$ symmetric and positive define matrix. The Wick theorem with a source term [57], largely used in physics, allows us to compute

$$Z_{\vec{J}} \stackrel{\text{def}}{=} \int d\vec{x}\, e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})+\vec{J}^T(\vec{x}-\vec{\mu})} = \sqrt{\frac{(2\pi)^n}{\det C}} e^{\frac{1}{2}\vec{J}^T C\vec{J}+\vec{J}^T\vec{\mu}} \tag{3.41}$$

and in particular to compute the expected values of the form

$$\langle e^{J_{k_i}x_{k_i}} x_{k_1} ... x_{k_n} \rangle = \frac{1}{Z_0}\frac{\partial}{\partial J_{k_1}} ... \frac{\partial}{\partial J_{k_n}} Z_{\vec{J}}\Big|_{J_{k_j}=0\ ,j\neq i} \tag{3.42}$$

Let's now give two examples of the use of the Wick theorem. The first is very simple and we go through all the steps

$$\begin{aligned} \langle x_0 \rangle &= \frac{1}{Z_0}\int x_0 e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})} d^n x = \frac{1}{Z_0}\frac{\partial}{\partial J_0}\int e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T C^{-1}(\vec{x}-\vec{\mu})+\vec{J}\vec{x}} d^n x\Big|_{\vec{J}=0} \\ &= \frac{\partial}{\partial J_0} e^{\frac{1}{2}J^T CJ+J\vec{\mu}}\Big|_{\vec{J}=0} = \vec{\mu}_0 \end{aligned} \tag{3.43}$$

Let's now compute a less trivial expected value, for example

$$\begin{aligned} \left\langle x_0 x_1 e^{x_0+3x_1} \right\rangle &= \frac{1}{Z_0}\frac{\partial}{\partial J_0}\frac{\partial}{\partial J_1} Z_{\vec{J}}\Big|_{J_0=1, J_1=3, J_s=0 \text{ for } s\neq 0,1} \\ &= \vec{\mu}_1(\vec{\mu}_0 + 3C_{01} + C_{00}) + C_{01}(\vec{\mu}_0 + 9C_{11} + C_{00} + 1) \\ &+ 3C_{11}(\vec{\mu}_0 + C_{00}) + 3C_{01}^2 \end{aligned} \tag{3.44}$$

where $\vec{\mu}_i$ is the i-th component of the $\vec{\mu}$ vector and C_{ij} the ij component of the C matrix.

Special Gaussian integrals

In the next section we will be dealing with different kinds of uncompleted Gaussian integrals and we here define them in order to have a more readable notation afterwards. For

$a,b,c,t_0,t \in \mathrm{R}$ we define

$$\mathscr{Z}(a,b,c,t,t_0) \stackrel{\text{def}}{=} \int_{t_0}^{t} e^{a\tau^2+b\tau+c} d\tau \tag{3.45}$$

$$\mathscr{O}(a,b,c,t,t_0) \stackrel{\text{def}}{=} \int_{t_0}^{t} \tau e^{a\tau^2+b\tau+c} d\tau \tag{3.46}$$

$$\mathscr{T}(a,b,c,t,t_0) \stackrel{\text{def}}{=} \int_{t_0}^{t} \tau^2 e^{a\tau^2+b\tau+c} d\tau \tag{3.47}$$

$$\mathscr{F}(a,b,c,t,t_0) \stackrel{\text{def}}{=} \int_{t_0}^{t} \tau^3 e^{a\tau^2+b\tau+c} d\tau \tag{3.48}$$

Note that it exists a well known analytical solution of these integrals.

Mean function of the Gaussian process

First, we will compute the mean function

$$\langle \vec{z}_t \rangle = \begin{pmatrix} \langle x_t \rangle \\ \langle g_t \rangle \\ \langle \lambda_t \rangle \\ \langle q_t \rangle \end{pmatrix} \tag{3.49}$$

of the cell state vector at time t in the case where there is no cell division between (t_0,t). Note that all these computations are done in the following order. First, we compute the mean function of the cell state vector $\langle \vec{z}_t | \vec{z}_0 \rangle$ constrained to the initial condition $\vec{z}_0$ and then use the general identity

$$\langle \vec{z}_t \rangle = \int d\vec{z}_0 \, \langle \vec{z}_t | \vec{z}_0 \rangle \, p(\vec{z}_0) \tag{3.50}$$

The cell state vector $\vec{z}_0$ at time $t_0 = 0$ is assumed to be Gaussian distributed with

$$\text{mean } \vec{\mu}^0 = \begin{bmatrix} \bar{x}_0 \\ \bar{g}_0 \\ \bar{\lambda}_0 \\ \bar{q}_0 \end{bmatrix} \quad \text{and covariance} \quad C^0 = \begin{bmatrix} C^0_{xx} & C^0_{gx} & C^0_{\lambda x} & C^0_{qx} \\ C^0_{gx} & C^0_{gg} & C^0_{\lambda g} & C^0_{qg} \\ C^0_{\lambda x} & C^0_{\lambda g} & C^0_{\lambda\lambda} & C^0_{q\lambda} \\ C^0_{gq} & C^0_{qg} & C^0_{\lambda q} & C^0_{qq} \end{bmatrix} \tag{3.51}$$

As an example, let us compute the expected growth rate $\langle \lambda_t \rangle$ at time t. Solving the equation (3.34) we find the constrained mean

$$\begin{aligned} \langle \lambda_t | \vec{z}_0 \rangle &= \bar{\lambda} \left(1 - e^{-\gamma_\lambda t}\right) + \lambda_0 e^{-\gamma_\lambda t} + \sigma_\lambda \int_0^t d\tau e^{\gamma_\lambda (\tau - t)} \underbrace{\langle \eta_1(\tau) \rangle}_{=0} \\ &= \lambda_0 e^{-\gamma_\lambda t} + \bar{\lambda} \left[1 - e^{-\gamma_\lambda t}\right] \end{aligned} \tag{3.52}$$

therefore, using (3.50) and the Wick theorem , the unconstrained version reads

$$\langle \lambda_t \rangle = \int d\vec{z}_0 \, \langle \lambda_t | \vec{z}_0 \rangle \, p(\vec{z}_0) = \bar{\lambda}_0 e^{-\gamma_\lambda t} + \bar{\lambda} \left[1 - e^{-\gamma_\lambda t}\right] \tag{3.53}$$

In a similar way we find

$$\langle q_t \rangle = \bar{q}_0 e^{-\gamma_q t} + \bar{q} \left[1 - e^{-\gamma_q t}\right] \tag{3.54}$$

The expected log cell size at time t constrained to the initial conditions reads

$$\begin{aligned} \langle x_t | \vec{z}_0 \rangle &= x_0 + \int_0^t \langle \lambda_\tau | \vec{z}_0 \rangle d\tau \\ &= x_0 + \lambda_0 \left[\frac{1 - e^{-\gamma_\lambda t}}{\gamma_\lambda}\right] + \bar{\lambda} \left[\frac{e^{-\gamma_\lambda t} - (1 - \gamma_\lambda t)}{\gamma_\lambda}\right] \end{aligned} \tag{3.55}$$

and again using (3.50) and the Wick theorem we can find the unconstrained expected log cell size at time t

$$\langle x_t \rangle = \bar{x}_0 + \bar{\lambda}_0 \left[\frac{1 - e^{-\gamma_\lambda t}}{\gamma_\lambda}\right] + \bar{\lambda} \left[\frac{e^{-\gamma_\lambda t} - (1 - \gamma_\lambda t)}{\gamma_\lambda}\right] \tag{3.56}$$

For the protein level the general solution of the differential equation (3.38) reads

$$g_t = e^{-\beta t} g_0 + e^{-\beta t} \int_0^t d\tau e^{\beta \tau} e^{x(\tau)} q(\tau) \tag{3.57}$$

and, in order to have analytical solutions for the expected protein levels $\langle g_t \rangle$ at time t, we have to linearise the non linear term $e^{x(\tau)}$ inside the integrand. To do this, we consider no fluctuations of the log size growth between $t_0 = 0$ and τ i.e. $x(\tau) = x_0 + \lambda_0 \tau$ and so

$$g_t \approx e^{-\beta t} g_0 + e^{-\beta t} \int_0^t d\tau e^{\beta \tau} e^{x_0 + \lambda_0 \tau} q(\tau) \tag{3.58}$$

This approximation together with the fact that we consider fluctuations in q to be independent from fluctuations in x or λ allows us to solve the unconstrained expected protein level at time t. In-fact, by first integrating over $\vec{z}_0$ and then over τ we find

$$
\begin{aligned}
\langle g_t \rangle &= \int d\vec{z}_0 P(\vec{z}_0) \left(e^{-\beta t} g_0 + e^{-\beta t} \int_0^t d\tau e^{\beta\tau} e^{x_0+\lambda_0\tau} \left(q_0 e^{-\gamma_q \tau} + \bar{q} \left[1 - e^{-\gamma_q \tau} \right] \right) \right) \\
&= \bar{g}_0 e^{-\beta t} + (\bar{q}_0 + C^0_{xq} - \bar{q}) \int_0^t e^{\frac{C^0_{\lambda\lambda}}{2}\tau^2 + \left(\beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q\right)\tau - \beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}} d\tau \\
&+ C^0_{\lambda q} \int_0^t \tau\, e^{\frac{C^0_{\lambda\lambda}}{2}\tau^2 + \left(\beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q\right)\tau - \beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}} d\tau \\
&+ \bar{q} \int_0^t e^{\frac{C^0_{\lambda\lambda}}{2}\tau^2 + \left(\beta + \bar{\lambda}_0 + C^0_{x\lambda}\right)\tau - \beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}} d\tau \\
&= \bar{g}_0 e^{-\beta t} + (\bar{q}_0 + C^0_{xq} - \bar{q}) \mathscr{L}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0 \right) \\
&+ C^0_{\lambda q} \mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0 \right) \\
&+ \bar{q} \mathscr{L}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda}, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0 \right)
\end{aligned}
\tag{3.59}
$$

The covariance function

In a similar way we can compute the 4×4 covariance matrix function[6]

$$
\langle \vec{z}_t, \vec{z}_t \rangle = \begin{pmatrix}
\mathrm{Var}[x_t] & \mathrm{Cov}[x_t, g_t] & \mathrm{Cov}[x_t, \lambda_t] & \mathrm{Cov}[x_t, q_t] \\
\mathrm{Cov}[g_t, x_t] & \mathrm{Var}[g_t] & \mathrm{Cov}[g_t, \lambda_t] & \mathrm{Cov}[g_t, q_t] \\
\mathrm{Cov}[x_t, \lambda_t] & \mathrm{Cov}[g_t, \lambda_t] & \mathrm{Var}[\lambda_t] & \mathrm{Cov}[\lambda_t, q_t] \\
\mathrm{Cov}[x_t, q_t] & \mathrm{Cov}[g_t, q_t] & \mathrm{Cov}[\lambda_t, q_t] & \mathrm{Var}[q_t]
\end{pmatrix}
\tag{3.60}
$$

in the case where the cell does not divide between (t_0, t). Note that, again, all these computations are done in the following order. First, we compute $\langle \vec{z}_t, \vec{z}_t | \vec{z}_0 \rangle_{ij}$ i.e. the i, j component of the covariance matrix at time t constrained to the initial condition $\vec{z}_0$. Then we use the general identity

$$
\langle \vec{z}_t, \vec{z}_t \rangle_{ij} = \int d\vec{z}_0 p(\vec{z}_0) \left(\langle \vec{z}_t, \vec{z}_t | \vec{z}_0 \rangle_{ij} + \left(\langle \vec{z}_t | \vec{z}_0 \rangle_i - \langle \vec{z}_t \rangle_i \right) \left(\langle \vec{z}_t | \vec{z}_0 \rangle_j - \langle \vec{z}_t \rangle_j \right) \right)
\tag{3.61}
$$

[6] We could also compute $\langle \vec{z}_t, \vec{z}_s \rangle$ in a similar way but equations become less readable and anyway we will not make use of them.

Again we consider (3.51) the initial cell state vector $\vec{z}_0$ to be Gaussian distributed with mean $\vec{\mu}^0$ and covariance C^0 . As an example we can easily compute

$$\begin{aligned}\mathrm{Cov}\left[\lambda_t,\lambda_s|\vec{z}_0\right] &= \sigma_\lambda^2\int_0^t\int_0^s d\tau d\tau' e^{\gamma(\tau+\tau'-t-s)}\underbrace{\left\langle\eta_1(\tau)\eta_1(\tau')\right\rangle}_{=\delta(\tau-\tau')}\\ &= \frac{\sigma_\lambda^2}{2\gamma_\lambda}\left(e^{-\gamma_\lambda|t-s|}-e^{-\gamma_\lambda(t+s)}\right)\end{aligned} \tag{3.62}$$

and so

$$\mathrm{var}[\lambda_t|z_0] = \frac{\sigma_\lambda^2}{2\gamma_\lambda}\left(1-e^{-2\gamma_\lambda t}\right), \tag{3.63}$$

Then using (3.61) and the Wick theorem, we calculate the unconstrained covariance

$$\begin{aligned}\mathrm{Var}[\lambda_t] &= \int dz_0 P(z_0)\left[\mathrm{var}[\lambda_t|z_0]+\left(\lambda_0-\bar{\lambda}_0\right)^2 e^{-2\gamma_\lambda t}\right]\\ &= \frac{\sigma_\lambda^2}{2\gamma_\lambda}\left(1-e^{-2\gamma_\lambda t}\right)+C^0_{\lambda\lambda}e^{-2\gamma_\lambda t}\end{aligned} \tag{3.64}$$

Another easy case to show is the computation of $\mathrm{Var}[\mathrm{x_t}]$. First, we compute the constrained variance

$$\begin{aligned}\mathrm{var}[x(t)|\vec{z}_0] &= \iint_0^t \mathrm{Cov}\left[\lambda_\tau,\lambda_{\tau'}\right]d\tau d\tau'\\ &= \frac{\sigma_\lambda^2}{2\gamma_\lambda^3}\left(2\gamma_\lambda t-3+4e^{-\gamma_\lambda t}-e^{-2\gamma_\lambda t}\right)\end{aligned} \tag{3.65}$$

and once again with (3.61) and the Wick theorem, we find the unconstrained version

$$\begin{aligned}\mathrm{var}[x_t] = \int d\vec{z}_0 P(\vec{z}_0)\Bigg[&\mathrm{var}[x_t|\vec{z}_0]+(\lambda_0-\bar{\lambda}_0)^2\left(\frac{1-e^{-\gamma_\lambda t}}{\gamma_\lambda}\right)^2\\ &+2(x_0-\bar{x}_0)(\lambda_0-\bar{\lambda}_0)\frac{(1-e^{-\gamma_\lambda t})}{\gamma_\lambda}+(x_0-\bar{x}_0)^2\Bigg]\\ = \frac{\sigma_\lambda^2}{2\gamma_\lambda^3}&\left(2\gamma_\lambda t-3+4e^{-\gamma_\lambda t}-e^{-2\gamma_\lambda t}\right)\\ &+C^0_{\lambda\lambda}\left(\frac{1-e^{-\gamma_\lambda t}}{\gamma_\lambda}\right)^2+2C^0_{x\lambda}\frac{(1-e^{-\gamma_\lambda t})}{\gamma_\lambda}+C^0_{xx}\end{aligned} \tag{3.66}$$

In a similar way we obtain

$$\mathrm{Var}[q_t] = \frac{\sigma_q^2}{2\gamma_q}\left(1-e^{-2\gamma_q t}\right) + C_{qq}^0 e^{-2\gamma_q t} \tag{3.67}$$

$$\mathrm{Cov}(\lambda_t, q_t) = C_{\lambda q}^0 e^{-(\gamma_\lambda+\gamma_q)t} \tag{3.68}$$

$$\mathrm{Cov}(\lambda_t, x_t) = \frac{\sigma_\lambda^2}{2\gamma_\lambda^2}\left(1-e^{-\gamma_\lambda t}\right)^2 + C_{\lambda\lambda}^0 e^{-\gamma_\lambda t}\left(\frac{1-e^{-\gamma_\lambda t}}{\gamma_\lambda}\right) + C_{x\lambda}^0 e^{-\gamma_\lambda t} \tag{3.69}$$

$$\mathrm{Cov}(x_t, q_t) = C_{\lambda q}^0\left(\frac{1-e^{-\gamma_\lambda t}}{\gamma_\lambda}\right)e^{-\gamma_q t} + C_{xq}^0 e^{-\gamma_q t} \tag{3.70}$$

All the terms involving g_t are less straightforward to compute manually but we implemented a Mathematica® procedure to help computing them. The technique is always the same. First, find the covariance function using (3.61) together with the approximation (3.58) for the protein level. Remember that we always consider fluctuations in q to be independent from fluctuations in x and λ. Then integrate over $\vec{z}_0$ using the Wick theorem and last integrate over time. For example using (3.61) and (3.58) we find

$$\begin{aligned}\mathrm{Cov}(g_t, \lambda_t) = &\int P(\vec{z}_0)\,\langle g_t|\vec{z}_0\rangle\,\langle \lambda_t|\vec{z}_0\rangle\, d\vec{z}_0 - \langle g_t\rangle\,\langle\lambda_t\rangle = \int P(\vec{z}_0)\\ &\times\left(e^{-\beta t}g_0 + e^{-\beta t}\int_0^t e^{\beta\tau+x_0+\lambda_0\tau}\left(q_0 e^{-\gamma_q\tau} + \bar{q}\left[1-e^{-\gamma_q\tau}\right]\right)d\tau\right)\\ &\times\left(\lambda_0 e^{-\gamma_\lambda t} + \bar{\lambda}\left[1-e^{-\gamma_\lambda t}\right]\right)d\vec{z}_0 - \langle g_t\rangle\,\langle\lambda_t\rangle\end{aligned} \tag{3.71}$$

The means $\langle g_t\rangle, \langle\lambda_t\rangle$ have been computed in the previous section. We are only left to solve the integrals which is done by first integrating over $\vec{z}_0$ using the Wick theorem and then integrating over τ. The solution reads

$$
\begin{aligned}
\mathrm{Cov}(g_t,\lambda_t) = {} & (\bar\lambda_0 C^0_{\lambda q} + \bar q_0 C^0_{\lambda\lambda} + C^0_{\lambda\lambda}C^0_{xq} - C^0_{\lambda\lambda}\bar q + C^0_{\lambda q}C^0_{x\lambda} - C^0_{\lambda q}\bar\lambda)\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& + (\bar\lambda_0\bar q_0 + \bar\lambda_0 C^0_{xq} - \bar\lambda_0\bar q + \bar q_0 C^0_{x\lambda} - \bar q_0\bar\lambda + C^0_{\lambda q} + C^0_{x\lambda}C^0_{xq} - C^0_{x\lambda}\bar q - C^0_{xq}\bar\lambda + \bar\lambda\bar q)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -bt + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& + (\bar q_0\bar\lambda + C^0_{xq}\bar\lambda - \bar\lambda\bar q)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) + C^0_{\lambda\lambda}C^0_{\lambda q}\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& + C^0_{\lambda q}\bar\lambda\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) + (\bar\lambda_0\bar q + C^0_{x\lambda}\bar q - \bar\lambda\bar q)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& + C^0_{\lambda\lambda}\bar q\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) + \bar\lambda\bar q\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) \\
& + \bar g_0\bar\lambda_0 e^{-(t(\beta+\gamma_\lambda))} - \bar g_0\bar\lambda e^{-(t(\beta+\gamma_\lambda))} + \bar g_0\bar\lambda e^{-bt} + C^0_{g\lambda}e^{-(t(\beta+\gamma_\lambda))} - \langle g_t\rangle\langle\lambda_t\rangle
\end{aligned}
\tag{3.72}
$$

The other components are computed in a similar way and we here only give the final results for completeness.

$$
\begin{aligned}
\mathrm{Cov}(x_t,g_t) = {} & \left(\frac{\bar\lambda_0 C^0_{\lambda q}}{\gamma_\lambda} + \frac{\bar q_0 C^0_{\lambda\lambda}}{\gamma_\lambda} + \bar q_0 C^0_{x\lambda} + \bar x_0 C^0_{\lambda q} + \frac{C^0_{\lambda\lambda}C^0_{xq}}{\gamma_\lambda} - \frac{C^0_{\lambda\lambda}\bar q}{\gamma_\lambda} + \frac{C^0_{\lambda q}C^0_{x\lambda}}{\gamma_\lambda} + C^0_{\lambda q}C^0_{xx} - \frac{C^0_{\lambda g}\bar\lambda}{\gamma_\lambda} + C^0_{\lambda q}\bar\lambda t + C^0_{x\lambda}C^0_{xq} - C^0_{x\lambda}\bar q\right) \\
& \mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) + \left(-\frac{\bar\lambda_0 C^0_{\lambda q}}{\gamma_\lambda} - \frac{\bar q_0 C^0_{\lambda\lambda}}{\gamma_\lambda} - \frac{C^0_{\lambda\lambda}C^0_{xq}}{\gamma_\lambda} + \frac{C^0_{\lambda\lambda}\bar q}{\gamma_\lambda} - \frac{C^0_{\lambda q}C^0_{x\lambda}}{\gamma_\lambda} + \frac{C^0_{\lambda q}\bar\lambda}{\gamma_\lambda}\right) \\
& \times\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) + \mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right)\times\left(\frac{\bar\lambda_0\bar q_0}{\gamma_\lambda} + \frac{\bar\lambda_0 C^0_{xq}}{\gamma_\lambda}\right. \\
& \left. - \frac{\bar\lambda_0\bar q}{\gamma_\lambda} + \bar q_0\bar x_0 + \frac{\bar q_0 C^0_{x\lambda}}{\gamma_\lambda} + \bar q_0 C^0_{xx} - \frac{\bar q_0\bar\lambda}{\gamma_\lambda} + \bar q_0\bar\lambda t + \bar x_0 C^0_{xq} - \bar x_0\bar q + \frac{C^0_{\lambda q}}{\gamma_\lambda} + \frac{C^0_{x\lambda}C^0_{xq}}{\gamma_\lambda} - \frac{C^0_{x\lambda}\bar q}{\gamma_\lambda} + C^0_{xq}C^0_{xx} - \frac{C^0_{xq}\bar\lambda}{\gamma_\lambda} + C^0_{xq}\bar\lambda t + C^0_{xq} - C^0_{xx}\bar q + \frac{\bar\lambda\bar q}{\gamma_\lambda} - \bar\lambda\bar q t\right) \\
& + \left(-\frac{\bar\lambda_0\bar q_0}{\gamma_\lambda} - \frac{\bar\lambda_0 C^0_{xq}}{\gamma_\lambda} + \frac{\bar\lambda_0\bar q}{\gamma_\lambda} - \frac{\bar q_0 C^0_{x\lambda}}{\gamma_\lambda} + \frac{\bar q_0\bar\lambda}{\gamma_\lambda} - \frac{C^0_{\lambda q}}{\gamma_\lambda} - \frac{C^0_{x\lambda}C^0_{xq}}{\gamma_\lambda} + \frac{C^0_{x\lambda}\bar q}{\gamma_\lambda} + \frac{C^0_{xq}\bar\lambda}{\gamma_\lambda} - \frac{\bar\lambda\bar q}{\gamma_\lambda}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& - \frac{\bar g_0\bar\lambda}{\gamma_\lambda}\left(e^{-\beta t} + e^{-(t(\beta+\gamma_\lambda))} + \gamma_\lambda t e^{-\beta t}\right) + \left(\frac{C^0_{\lambda\lambda}C^0_{\lambda q}}{\gamma_\lambda} + C^0_{\lambda q}C^0_{x\lambda}\right)\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) \\
& - \frac{C^0_{\lambda\lambda}C^0_{\lambda q}\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right)}{\gamma_\lambda} + \left(\frac{\bar\lambda_0\bar q}{\gamma_\lambda} + \bar x_0\bar q + \frac{C^0_{x\lambda}\bar q}{\gamma_\lambda} + C^0_{xx}\bar q - \frac{\bar\lambda\bar q}{\gamma_\lambda} + \bar\lambda\bar q t\right) \\
& \mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) + \left(-\frac{\bar\lambda_0\bar q}{\gamma_\lambda} - \frac{C^0_{x\lambda}\bar q}{\gamma_\lambda} + \frac{\bar\lambda\bar q}{\gamma_\lambda}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right) \\
& + \left(\frac{C^0_{\lambda\lambda}\bar q}{\gamma_\lambda} + C^0_{x\lambda}\bar q\right)\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2}, t, 0\right) - \frac{C^0_{\lambda\lambda}\bar q\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar\lambda_0 + C^0_{x\lambda}, -\beta t + \bar x_0 + \frac{C^0_{xx}}{2} - \gamma_\lambda t, t, 0\right)}{\gamma_\lambda} \\
& + \frac{\bar g_0\bar\lambda_0 e^{-\beta t}}{\gamma_\lambda} - \frac{\bar g_0\bar\lambda_0 e^{-(t(\beta+\gamma_\lambda))}}{\gamma_\lambda} + \bar g_0\bar x_0 e^{-\beta t} + \frac{C^0_{g\lambda}e^{-\beta t}}{\gamma_\lambda} - \frac{C^0_{g\lambda}e^{-(t(\beta+\gamma_\lambda))}}{\gamma_\lambda} + C^0_{gx}e^{-\beta t} - \langle g_t\rangle\langle x_t\rangle
\end{aligned}
\tag{3.73}
$$

$$
\begin{aligned}
\mathrm{Cov}(q_t, g_t) = &\left(\bar{q}_0^2 + 2\bar{q}_0 C^0_{xq} - 2\bar{q}_0\bar{q} + C^0_{qq} + C^{02}_{xq} - 2C^0_{xq}\bar{q} - \frac{\sigma_q^2}{2\gamma_q} + \bar{q}^2\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right) \\
&+ (2\bar{q}_0 C^0_{\lambda q} + 2C^0_{\lambda q}C^0_{xq} - 2C^0_{\lambda q}\bar{q})\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right) + \left(\bar{q}_0\bar{q} + C^0_{xq}\bar{q} - \bar{q}^2\right) \\
&\times \mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda}, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right) + \left(\bar{q}_0\bar{q} + C^0_{xq}\bar{q} - \bar{q}^2\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0\right) \\
&+ C^{02}_{\lambda q}\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right) + C^0_{\lambda q}\bar{q}\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda}, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right) \\
&+ C^0_{\lambda q}\bar{q}\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} - \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0\right) + \frac{\sigma_q^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda} + \gamma_q, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2} - \gamma_q t, t, 0\right)}{2\gamma_q} \\
&+ \bar{q}^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta + \bar{\lambda}_0 + C^0_{x\lambda}, -\beta t + \bar{x}_0 + \frac{C^0_{xx}}{2}, t, 0\right) + \bar{g}_0\bar{q}_0 e^{-(t(\beta+\gamma_q))} - \bar{g}_0\bar{q}e^{-(t(\beta+\gamma_q))} + \bar{g}_0\bar{q}e^{-\beta t} + C^0_{gq}e^{-(t(\beta+\gamma_q))} - \langle q_t\rangle\langle g_t\rangle
\end{aligned}
\tag{3.74}
$$

$$
\begin{aligned}
\mathrm{Var}(g_t) = {} & \mathscr{F}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), t, 0\right) C^{02}_{\lambda q} - \mathscr{F}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) C^{02}_{\lambda q} \\
& - \frac{2\bar{q}\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2\bar{x}_0+2C^0_{xx}-(2\beta+\gamma_q)t, 2t, t\right) C^0_{\lambda q}}{\gamma_q} + \frac{2\bar{q}\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t+\gamma_q t, 2t, t\right) C^0_{\lambda q}}{\gamma_q} \\
& + 2C^0_{g\lambda}\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+C^0_{x\lambda}-\gamma_q, \bar{x}_0+\frac{C^0_{xx}}{2}-2\beta t, t, 0\right) C^0_{\lambda q} + \left(\bar{g}_0^2+C^0_{gg}\right)e^{-2\beta t} + 2C^0_{g\lambda}\bar{q}\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+C^0_{x\lambda}, \bar{x}_0+\frac{C^0_{xx}}{2}-2\beta t, t, 0\right) \\
& + \frac{\bar{q}(2C^0_{\lambda q}+\gamma_q\bar{q})\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2(\bar{x}_0+C^0_{xx}-\beta t), t, 0\right)}{\gamma_q} + 2(\bar{q}_0C^0_{g\lambda}+C^0_{xq}C^0_{g\lambda}-\bar{q}C^0_{g\lambda}+\bar{g}_0C^0_{\lambda q}+C^0_{\lambda q}C^0_{xg}) \\
& \times \mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+C^0_{x\lambda}-\gamma_q, \bar{x}_0+\frac{C^0_{xx}}{2}-2\beta t, t, 0\right) + \frac{1}{\gamma_q}\left(\gamma_q\bar{q}_0^2+4C^0_{xq}\gamma_q\bar{q}_0-2\gamma_q\bar{q}\bar{q}_0+\gamma_q\bar{q}^2+4C^{02}_{xq}\gamma_q+C^0_{qq}\gamma_q-2C^0_{\lambda q}\bar{q}-4C^0_{xq}\gamma_q\bar{q}\right) \\
& \times \mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), t, 0\right) - \bar{q}^2\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) \\
& - \frac{\sigma_q^2\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t, t, 0\right)}{2\gamma_q} + \frac{\sigma_q^2\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t, 2t, t\right)}{2\gamma_q} \\
& + \left(-\bar{q}_0^2-4C^0_{xq}\bar{q}_0+2\bar{q}\bar{q}_0+4C^0_{\lambda q}t\bar{q}_0-4C^{02}_{xq}-\bar{q}^2-C^0_{qq}+4C^0_{xq}\bar{q}+8C^0_{\lambda q}C^0_{xq}t-4C^0_{\lambda q}\bar{q}t\right)\mathscr{O}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) \\
& + (2\bar{q}_0C^0_{\lambda q}+4C^0_{xq}C^0_{\lambda q}-2\bar{q}C^0_{\lambda q})\mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), t, 0\right) + \left(2tC^{02}_{\lambda q}-2\bar{q}_0C^0_{\lambda q}-4C^0_{xq}C^0_{\lambda q}+2\bar{q}C^0_{\lambda q}\right) \\
& \times \mathscr{T}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) + (2\bar{g}_0\bar{q}+2C^0_{xg}\bar{q})\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+C^0_{x\lambda}, \bar{x}_0+\frac{C^0_{xx}}{2}-2\beta t, t, 0\right) \\
& + \left(-\frac{2\bar{q}^2}{\gamma_q}+\frac{2\bar{q}_0\bar{q}}{\gamma_q}+\frac{4C^0_{xq}\bar{q}}{\gamma_q}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2(\bar{x}_0+C^0_{xx}-\beta t), t\right) \\
& + (2\bar{g}_0\bar{q}_0+2C^0_{xg}\bar{q}_0+2C^0_{gq}+2\bar{g}_0C^0_{xq}+2C^0_{xg}C^0_{xq}-2\bar{g}_0\bar{q}-2C^0_{xg}\bar{q})\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+C^0_{x\lambda}-\gamma_q, \bar{x}_0+\frac{C^0_{xx}}{2}-2\beta t, t, 0\right) \\
& + \left(\frac{2\bar{q}^2}{\gamma_q}-\frac{2\bar{q}_0\bar{q}}{\gamma_q}-\frac{4C^0_{xq}\bar{q}}{\gamma_q}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), t, 0\right) \\
& + \frac{\sigma_q^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2\bar{x}_0+2C^0_{xx}-2\beta t, t, 0\right)}{2\gamma_q^2} + \frac{\sigma_q^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2\bar{x}_0+2C^0_{xx}-2\beta t, 2t, t\right)}{2\gamma_q^2} \\
& + 2\bar{q}^2t\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) + \left(\frac{2\bar{q}^2}{\gamma_q}-\frac{2\bar{q}_0\bar{q}}{\gamma_q}-\frac{4C^0_{xq}\bar{q}}{\gamma_q}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}, 2\bar{x}_0+2C^0_{xx}-(2\beta+\gamma_q)t, 2t, t\right) \\
& - \frac{\sigma_q^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t, t, 0\right)}{2\gamma_q^2} - \frac{\sigma_q^2t\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t, 2t, t\right)}{\gamma_q} \\
& + \left(2t\bar{q}_0^2+8C^0_{xq}t\bar{q}_0-4\bar{q}t\bar{q}_0+8C^{02}_{xq}t+2\bar{q}^2t+2C^0_{qq}t-8C^0_{xq}\bar{q}t\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2(\bar{x}_0+C^0_{xx}-\beta t), 2t, t\right) \\
& + \left(-\frac{2\bar{q}^2}{\gamma_q}+\frac{2\bar{q}_0\bar{q}}{\gamma_q}+\frac{4C^0_{xq}\bar{q}}{\gamma_q}\right)\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}-\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t+\gamma_q t, 2t, t\right) \\
& - \frac{\sigma_q^2\mathscr{Z}\left(\frac{C^0_{\lambda\lambda}}{2}, \beta+\bar{\lambda}_0+2C^0_{x\lambda}+\gamma_q, 2\bar{x}_0+2C^0_{xx}-2\beta t-2\gamma_q t, 2t, t\right)}{2\gamma_q^2} - \langle g_t\rangle^2
\end{aligned}
\tag{3.75}
$$

The components of the mean and covariance matrix function computed until now consider no cell division between t_0 an t. We now have to take into account the case of cell division in order to have a meaningful description of our data-set.

Cell division

Right now we computed the unconstrained mean and covariance function of the cell state vector $\vec{z}_t$ if there is no cell division between time $t_0 = 0$ and time t. But what if the cell divides within this time lapse? This simple question may be harder than one may think. At cell division bacteria have to form the so called Z ring i.e. the septum which will allows cytokinesis. This process involves the recruitment of several proteins [38] and probably during this phase the cell physiology is different compared to the rest of the cell cycle. This has a potential impact on the cell growth rate or on the expression of the target protein. Even though all these questions are interesting and relevant we will not model these complicate dynamics at cell division and we only assume that, after division, the two daughter cells will have half the volume and half the number of proteins than their mother. That is, if the cell divide "precisely" at time t, then we assume

$$\begin{aligned} p\left(x_t^{\text{daughter}} \middle| x_t^{\text{mother}}\right) &= \mathcal{N}\left(x_t^{\text{daughter}} \middle| x_t^{\text{mother}} - \log 2, \sigma_{dx}^2\right) \\ p\left(g_t^{\text{daughter}} \middle| g_t^{\text{mother}}\right) &= \mathcal{N}\left(g_t^{\text{daughter}} \middle| \frac{g_t^{\text{mother}}}{2}, \sigma_{dg}^2\right) \end{aligned} \tag{3.76}$$

where the superscript "mother" stands for the case just before division and "daughter" just after division. The fact that cell division is not perfect is modeled through the two parameters σ_{dx}, σ_{dg}. In matrix form, we can rewrite these two equations as

$$p\left(\vec{z}_t^{\text{daughter}} \middle| \vec{z}_t^{\text{mother}}\right) = \mathcal{N}\left(\vec{z}_t^{\text{daughter}} \middle| F z_t^{\text{mother}} + \vec{f}, D_d\right) \tag{3.77}$$

with

$$D_d = \mathbf{diag}\left[\sigma_{dx}^2, \sigma_{dg}^2, 0, 0\right] \quad F = \mathbf{diag}\left[1, \frac{1}{2}, 1, 1\right] \quad \vec{f} = [-\log 2, 0, 0, 0]^T \tag{3.78}$$

Note that we have no precise information on when the cell exactly divide between time t_0 and time t and a more sophisticated models should take this into consideration. However, here we assume cell division always takes place exactly and instantaneously at the observation time t. Therefore, if we want to compute the cell state vector distribution $p\left(\vec{z}_t^{\text{daughter}}\right)$ at time t knowing that the cell divide between t_0 and t, we first compute the cell state vector at

time t as if no division happened $p\left(\vec{z}_t^{\text{mother}}\right)$, and then consider division to happen exactly at t (3.77).

Using equation (3.9), it is easy to solve

$$\begin{aligned} p\left(\vec{z}_t^{\text{daughter}}\right) &= \int p\left(\vec{z}_t^{\text{daughter}} \middle| \vec{z}_t^{\text{mother}}\right) p\left(\vec{z}_t^{\text{mother}}\right) d\vec{z}_t^{\text{mother}} \\ &= \mathcal{N}\left(\vec{z}_t^{\text{daughter}} \middle| F\left\langle \vec{z}_t^{\text{mother}} \right\rangle + \vec{f}, D_d + F\left\langle \vec{z}_t^{\text{mother}}, \vec{z}_t^{\text{mother}} \right\rangle F^T\right) \end{aligned} \tag{3.79}$$

where, obviously, the mean and covariance matrix of the "mother" cell are given by the previously compute equations (3.49) and (3.60). We now have all the ingredients to apply the previously discussed kiring method.

3.3.3 Gaussian process regression

Once the mean and covariance functions are given, we can apply the Gaussian process regression method previously developed. As we will see, this will allow us to predict the cell state vector $\vec{z}(t)$ at time t given the measurements of the cell size and protein levels along the entire experiment. Let us denote $\vec{Z}_{\vec{J}}$ the cell state vector at time t_j and assume it follows a Gaussian process

$$\begin{bmatrix} \vec{z}_0 \\ \vdots \\ \vec{z}_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \vec{z}_0 \\ \vdots \\ \vec{z}_n \end{bmatrix} \middle| \begin{bmatrix} \langle \vec{z}_0 \rangle \\ \vdots \\ \langle \vec{z}_n \rangle \end{bmatrix}, \begin{bmatrix} \langle \vec{z}_0, \vec{z}_0 \rangle & \cdots & \langle \vec{z}_0, \vec{z}_n \rangle \\ \vdots & \ddots & \vdots \\ \langle \vec{z}_n, \vec{z}_0 \rangle & \cdots & \langle \vec{z}_n, \vec{z}_n \rangle \end{bmatrix} \right) \tag{3.80}$$

where $\langle \vec{z}_i \rangle$ and $\langle \vec{z}_i, \vec{Z}_{\vec{J}} \rangle$ are the mean and covariance matrix functions previously computed. Ideally, we would use (3.22) and (3.27) in order to compute the likelihood and predict the cell state vector $\vec{z}(t)$ at any time t. Unfortunately, we can not use (3.22) and (3.27) directly since the previously developed Gaussian process regression method did not take into consideration the presence of latent variables[7]. In this section we discuss how the likelihood (3.22) and the posterior distribution (3.27) are computed in this situation. For simplicity we first consider an unique observation[8] $\mathscr{D} = \left\{(x_0^m, g_0^m)\right\}$ of the cell size x_0^m and protein level g_0^m at time t_0. The generalisation in the case of multiple observations is then straightforward. Let's consider the initial state vector

[7] Latent variables are variables to which we do not have direct observations and in our case are λ_t and q_t.

[8] We use the superscript m to denote observations/measured quantities.

$$\vec{z}_0 = \begin{pmatrix} x_0 \\ g_0 \\ \lambda_0 \\ q_0 \end{pmatrix} = \begin{pmatrix} [\vec{z}_0]_0 \\ [\vec{z}_0]_1 \\ [\vec{z}_0]_2 \\ [\vec{z}_0]_3 \end{pmatrix} \sim \mathcal{N}\left(\vec{z}_0 \middle| \langle \vec{z}_0 \rangle, \langle \vec{z}_0, \vec{z}_0 \rangle\right) \tag{3.81}$$

to be Gaussian distributed with mean $\langle \vec{z}_0 \rangle$ and covariance matrix

$$\begin{aligned} \langle \vec{z}_0, \vec{z}_0 \rangle &= \left(\begin{array}{cc|cc} \mathrm{Var}[x_0] & \mathrm{Cov}[x_0, g_0] & \mathrm{Cov}[x_0, \lambda_0] & \mathrm{Cov}[x_0, q_0] \\ \mathrm{Cov}[g_0, x_0] & \mathrm{Var}[g_0] & \mathrm{Cov}[g_0, \lambda_0] & \mathrm{Cov}[g_0, q_0] \\ \hline \mathrm{Cov}[x_0, \lambda_0] & \mathrm{Cov}[g_0, \lambda_0] & \mathrm{Var}[\lambda_0] & \mathrm{Cov}[\lambda_0, q_0] \\ \mathrm{Cov}[x_0, q_0] & \mathrm{Cov}[g_0, q_0] & \mathrm{Cov}[\lambda_0, q_0] & \mathrm{Var}[q_0] \end{array}\right) \\ &:= \begin{pmatrix} K_0 & K_1 \\ K_1^T & K_2 \end{pmatrix} \end{aligned} \tag{3.82}$$

Likelihood In order to compute the likelihood we first have to model the measurement noise. We consider the measurement errors in the cell size and in the protein levels to be Gaussian distributed

$$\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \begin{bmatrix} [\vec{z}_0]_0 \\ [\vec{z}_0]_1 \end{bmatrix}, D\right) \tag{3.83}$$

where

$$D = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_g^2 \end{pmatrix} \tag{3.84}$$

The likelihood then simply reads

$$\begin{aligned} p\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \Theta\right) = \int d\vec{z}_0 \mathcal{N}\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \begin{bmatrix} [\vec{z}_0]_0 \\ [\vec{z}_0]_1 \end{bmatrix}, D\right) \\ \times \mathcal{N}\left(\vec{z}_0 \middle| \langle \vec{z}_0 \rangle, \langle \vec{z}_0, \vec{z}_0 \rangle\right) \end{aligned} \tag{3.85}$$

where $\Theta = \left\{\bar{\lambda}, \gamma_\lambda, \sigma_\lambda^2, \bar{q}, \gamma_q, \sigma_q^2, \sigma_x^2, \sigma_g^2, \sigma_{dx}^2, \sigma_{dg}^2\right\}$ are the hyperparameters of the model. Note that, to keep a more readable notation, we sometimes omit to explicit write Θ in the distributions. In order to solve the integral (3.85) we first have to integrate over $[\vec{z}_0]_2$ and

$[\vec{z}_0]_3$. Using the property (3.11) of Gaussian distributions this is straightforward and gives

$$p\left(\left[\begin{array}{c} x_0^m \\ g_0^m \end{array}\right]\middle|\Theta\right) = \int d\,[\vec{z}_0]_1\, d\,[\vec{z}_0]_0\, \mathcal{N}\left(\left[\begin{array}{c} x_0^m \\ g_0^m \end{array}\right]\middle|\left[\begin{array}{c} [\vec{z}_0]_0 \\ [\vec{z}_0]_1 \end{array}\right], D\right) \times \mathcal{N}\left(\left[\begin{array}{c} [\vec{z}_0]_0 \\ [\vec{z}_0]_1 \end{array}\right]\middle|\left[\begin{array}{c} \langle\vec{z}_0\rangle_0 \\ \langle\vec{z}_0\rangle_1 \end{array}\right], K_0\right) \tag{3.86}$$

This integral is straightforward to solve once we realize it has the same shape as the integral we computed in (3.22)

$$p\left(\left[\begin{array}{c} x_0^m \\ g_0^m \end{array}\right]\middle|\Theta\right) = \mathcal{N}\left(\left[\begin{array}{c} x_0^m \\ g_0^m \end{array}\right]\middle|\left[\begin{array}{c} \langle\vec{z}_0\rangle_0 \\ \langle\vec{z}_0\rangle_1 \end{array}\right], K_0 + D\right) \tag{3.87}$$

which gives the likelihood of the measurement (x_0^m, g_0^m) given the hyperparameters Θ.

Predictions Similar as when we derived (3.27), we would like to predict $\vec{z}_{t^\star}$ at time $t^\star$ given the noisy observation (x_0^m, g_0^m). First note that the vector $\left[x_0^m, g_0^m, \vec{z}_{t^\star}\right]^T$ is distributed as

$$\left[\begin{array}{c} x_0^m \\ g_0^m \\ \vec{z}_{t^\star} \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} x_0^m \\ g_0^m \\ \vec{z}_{t^\star} \end{array}\right]\middle|\left[\begin{array}{c} \langle\vec{z}_0\rangle_0 \\ \langle\vec{z}_0\rangle_1 \\ \langle\vec{z}_{t^\star}\rangle \end{array}\right], \left[\begin{array}{cc} K_0 + D & \tilde{K} \\ \tilde{K}^T & \langle\vec{z}_{t^\star}, \vec{z}_{t^\star}\rangle \end{array}\right]\right) \tag{3.88}$$

where $\tilde{K}$ is the 2×4 matrix defined as

$$\tilde{K} = \left[\begin{array}{cccc} \langle x_0, x_{t^\star}\rangle & \langle x_0, g_{t^\star}\rangle & \langle x_0, \lambda_{t^\star}\rangle & \langle x_0, q_{t^\star}\rangle \\ \langle g_0, x_{t^\star}\rangle & \langle g_0, g_{t^\star}\rangle & \langle g_0, \lambda_{t^\star}\rangle & \langle g_0, q_{t^\star}\rangle \end{array}\right] \tag{3.89}$$

Using the property (3.12) of Gaussian distributions we immediately find

$$\vec{z}_{t^\star}\left|\left[\begin{array}{c} x_0^m \\ g_0^m \end{array}\right] \sim \mathcal{N}(\vec{m}, C)\right. \tag{3.90}$$

with

$$\begin{aligned} \vec{m} &= \langle\vec{z}_{t^\star}\rangle + \tilde{K}^T (K_0 + D)^{-1} \left[\begin{array}{c} x_0^m - \langle\vec{z}_0\rangle_0 \\ g_0^m - \langle\vec{z}_0\rangle_1 \end{array}\right] \\ C &= \langle\vec{z}_{t^\star}, \vec{z}_{t^\star}\rangle - \tilde{K}^T (K_0 + D)^{-1} \tilde{K} \end{aligned} \tag{3.91}$$

Generalize to multiple observations The generalization (3.87) and (3.90) to the case of multiple observations is straightforward. The only thing we have to pay attention is

to stack the vectors in a good way in order to keep computations simple. Consider $\mathscr{D} = \left\{(x_0^m, g_0^m), \ldots, (x_n^m, g_n^m)\right\}$ then the likelihood reads

$$p\left(\left[\begin{array}{c} x_0^m \\ \vdots \\ x_n^m \\ g_0^m \\ \vdots \\ g_n^m \end{array}\right] \middle|\middle| \Theta\right) = \mathscr{N}\left(\left[\begin{array}{c} x_0^m \\ \vdots \\ x_n^m \\ g_0^m \\ \vdots \\ g_n^m \end{array}\right] \middle|\middle| \left[\begin{array}{c} \langle \vec{z}_0 \rangle_0 \\ \vdots \\ \langle \vec{z}_n \rangle_0 \\ \langle \vec{z}_0 \rangle_1 \\ \vdots \\ \langle \vec{z}_n \rangle_1 \end{array}\right], K_0 + D\right) \tag{3.92}$$

where

$$D = \mathbf{Diag}\left[\underbrace{\sigma_x^2, \ldots, \sigma_x^2}_{n}, \underbrace{\sigma_g^2 \ldots, \sigma_g^2}_{n}\right] \tag{3.93}$$

and

$$[K_0]_{ij} = \begin{cases} \mathrm{Cov}[x_i, x_j] & \text{if } i, j < n \\ \mathrm{Cov}[x_i, g_j] & \text{if } i < n,\ j \geq n \\ \mathrm{Cov}[g_i, x_j] & \text{if } i \geq n,\ j < n \\ \mathrm{Cov}[g_i, g_j] & \text{if } i \geq n,\ j \geq n \end{cases} \tag{3.94}$$

whereas for the prediction of $\vec{z}_t^{\star}$ given $\mathscr{D}$ the generalized form of (3.90) reads

$$\vec{z}_{t^\star} \middle|\middle| \left[\begin{array}{c} x_0^m \\ \vdots \\ x_n^m \\ g_0^m \\ \vdots \\ g_n^m \end{array}\right] \sim \mathscr{N}(\vec{m}, C) \tag{3.95}$$

with

$$\begin{aligned} \vec{m} &= \langle \vec{z}_{t^\star} \rangle + \tilde{K}^T (K_0 + D)^{-1} \left[\begin{array}{c} x_0^m - \langle \vec{z}_0 \rangle_0 \\ \vdots \\ x_n^m - \langle \vec{z}_n \rangle_0 \\ g_0^m - \langle \vec{z}_0 \rangle_1 \\ \vdots \\ g_n^m - \langle \vec{z}_n \rangle_1 \end{array}\right] \\ C &= \langle \vec{z}_{t^\star}, \vec{z}_{t^\star} \rangle - \tilde{K}^T (K_0 + D)^{-1} \tilde{K} \end{aligned} \tag{3.96}$$

where D and K_0 are give by (3.93) and (3.94) whereas the $\tilde{K}$ matrix equal to

$$\left[\tilde{K}\right]_j = \begin{cases} \left[\langle x_j, x_{t^\star}\rangle \ \langle x_j, g_{t^\star}\rangle \ \langle x_j, \lambda_{t^\star}\rangle \ \langle x_j, q_{t^\star}\rangle\right] & \text{if } j < n \\ \left[\langle g_j, x_{t^\star}\rangle \ \langle g_j, g_{t^\star}\rangle \ \langle g_j, \lambda_{t^\star}\rangle \ \langle g_j, q_{t^\star}\rangle\right] & \text{if } j \geq n \end{cases} \tag{3.97}$$

for the j^{th} line. We now have all the ingredients to predict[9] $\vec{z}_t$ at any time t, by first finding the MLE $\Theta^\star$ and then using (3.95). Note that this involves the non trivial computation of $\langle \vec{z}_t, \vec{z}_s \rangle$ for any time $t, s > t_0$ and the computationally expensive inversion of $2n \times 2n$ dimensional matrix. In order to avoid this we use the markovian property of this process.

The markovian property and the likelihood computation In the previous paragraphs, we theoretically showed how to apply the Gaussian process regression method together with our biophysical model to the MoMa time series data. However, computing the full covariance matrix function is not trivial and, as said, Gaussian regression involves the computationally expensive matrix inversion. For these reasons, we developed a recursive way to treat the MoMa time series data based on the markovian property of the process. This means that if $(\vec{z}_0, \vec{z}_1, \ldots, \vec{z}_n)$ is a series of cell state vectors at times $(t_1, \ldots, t_n)$ over a cell lineage (figure 3.1d) then the markovian property gives

$$p\left(\vec{z}_{n+1} | \vec{z}_n, \vec{z}_{n-1}, \ldots, \vec{z}_0\right) = p\left(\vec{z}_{n+1} | \vec{z}_n\right) \tag{3.98}$$

i.e. the probability to be in the cell state $\vec{z}_{n+1}$ at time t_{n+1} only depends on the cell state $\vec{z}_n$ at time t_n. This allows to apply the Gaussian process regression machinery in a more simple way. Let $\mathscr{D} = \left\{(x_n^m, g_n^m), \ldots, (x_0^m, g_0^m)\right\}$ be a series of measurements of the log cell size and fluorescent protein molecules over one cell lineage and let's assume we know the initial cell state distribution

$$p\left(\vec{z}_0\right) = \mathscr{N}\left(\vec{z}_0 | \langle \vec{z}_0 \rangle, \langle \vec{z}_0, \vec{z}_0 \rangle\right) \tag{3.99}$$

The likelihood of the first observation is given by (3.87)

$$p\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \Theta\right) = \mathscr{N}\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \begin{bmatrix} \langle \vec{z}_0 \rangle_0 \\ \langle \vec{z}_0 \rangle_1 \end{bmatrix}, \begin{bmatrix} \langle \vec{z}_0, \vec{z}_0 \rangle_{00} + \sigma_x^2 & \langle \vec{z}_0, \vec{z}_0 \rangle_{01} \\ \langle \vec{z}_0, \vec{z}_0 \rangle_{10} & \langle \vec{z}_0, \vec{z}_0 \rangle_{11} + \sigma_g^2 \end{bmatrix}\right) \tag{3.100}$$

[9] In a similar way we can compute quantities like $p(\vec{z}_t, \vec{z}_s | \mathscr{D})$.

whereas the posterior over the initial cell state is given by (3.90)

$$p\left(\vec{z}_0 \middle| \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) = \mathcal{N}(\vec{z}_0 | \vec{m}, C) \tag{3.101}$$

where

$$\begin{aligned}
\vec{m} &= \langle \vec{z}_0 \rangle + \tilde{K}^T \left(\begin{bmatrix} \langle \vec{z}_0, \vec{z}_0 \rangle_{00} + \sigma_x^2 & \langle \vec{z}_0, \vec{z}_0 \rangle_{01} \\ \langle \vec{z}_0, \vec{z}_0 \rangle_{10} & \langle \vec{z}_0, \vec{z}_0 \rangle_{11} + \sigma_g^2 \end{bmatrix} \right)^{-1} \begin{bmatrix} x_0^m - \langle \vec{z}_0 \rangle_0 \\ g_0^m - \langle \vec{z}_0 \rangle_1 \end{bmatrix} \\
C &= \langle \vec{z}_0, \vec{z}_0 \rangle - \tilde{K}^T \left(\begin{bmatrix} \langle \vec{z}_0, \vec{z}_0 \rangle_{00} + \sigma_x^2 & \langle \vec{z}_0, \vec{z}_0 \rangle_{01} \\ \langle \vec{z}_0, \vec{z}_0 \rangle_{10} & \langle \vec{z}_0, \vec{z}_0 \rangle_{11} + \sigma_g^2 \end{bmatrix} \right)^{-1} \tilde{K} \\
\tilde{K} &= \begin{bmatrix} \langle \vec{z}_0, \vec{z}_0 \rangle_{00} & \langle \vec{z}_0, \vec{z}_0 \rangle_{01} & \langle \vec{z}_0, \vec{z}_0 \rangle_{02} & \langle \vec{z}_0, \vec{z}_0 \rangle_{03} \\ \langle \vec{z}_0, \vec{z}_0 \rangle_{10} & \langle \vec{z}_0, \vec{z}_0 \rangle_{11} & \langle \vec{z}_0, \vec{z}_0 \rangle_{12} & \langle \vec{z}_0, \vec{z}_0 \rangle_{13} \end{bmatrix}
\end{aligned} \tag{3.102}$$

In order to compute the conditional prior distribution for the next cell state vector

$$p\left(\vec{z}_1 \middle| \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) = \mathcal{N}\left(\vec{z}_1 \middle| \langle \vec{z}_1 \rangle, \langle \vec{z}_1, \vec{z}_1 \rangle\right) \tag{3.103}$$

we have to compute the conditional mean

$$\langle \vec{z}_1 \rangle = \int d\vec{z}_0 p\left(\vec{z}_0 \middle| \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \langle \vec{z}_1 | \vec{z}_0 \rangle \tag{3.104}$$

and the conditional covariance matrix

$$\langle \vec{z}_1, \vec{z}_1 \rangle_{ij} = \int d\vec{z}_0 p\left(\vec{z}_0 \middle| \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \left(\langle \vec{z}_1, \vec{z}_1 | \vec{z}_0 \rangle_{ij} + (\langle \vec{z}_1 | \vec{z}_0 \rangle_i - \langle \vec{z}_1 \rangle_i)(\langle \vec{z}_1 | \vec{z}_0 \rangle_j - \langle \vec{z}_1 \rangle_j) \right) \tag{3.105}$$

which is exactly what we developed in section 3.3.2. So we know how to obtain the prior distribution (3.103). We are in a similar situation as in (3.99) and we can use equation (3.100) to compute the likelihood

$$p\left(\begin{bmatrix} x_1^m \\ g_1^m \end{bmatrix} \middle| \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \tag{3.106}$$

and (3.101) for the posterior

$$p\left(\vec{z}_1 \middle| \begin{bmatrix} x_1^m \\ g_1^m \end{bmatrix}, \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \tag{3.107}$$

of the next observation $\left(x_1^m, g_1^m\right)$. It is obvious that we can iterate this procedure over the entire cell lineage data-set leaving us with a series of conditional likelihoods

$$p\left(\begin{bmatrix} x_j^m \\ g_j^m \end{bmatrix} \middle| \begin{bmatrix} x_{j-1}^m \\ g_{j-1}^m \end{bmatrix}, \ldots, \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \tag{3.108}$$

and posteriors

$$p\left(\vec{Z}_{\vec{J}} \middle| \begin{bmatrix} x_j^m \\ g_j^m \end{bmatrix}, \ldots, \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) \tag{3.109}$$

In order to find the total likelihood over one cell lineage we use the general relation

$$\log p\left(\mathscr{D}|\,\Theta\right) = \sum_{j=1}^{n} \log p\left(\begin{bmatrix} x_j^m \\ g_j^m \end{bmatrix} \middle| \begin{bmatrix} x_{j-1}^m \\ g_{j-1}^m \end{bmatrix}, \ldots, \begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix}, \Theta\right) + \log p\left(\begin{bmatrix} x_0^m \\ g_0^m \end{bmatrix} \middle| \Theta\right) \tag{3.110}$$

To find the optimal parameter set $\Theta^\star$ we sum all the log likelihoods coming from all the cell lineages presents in the experiment. Then we search for the MLE $\Theta^\star$ by maximizing the total likelihood. Once $\Theta^\star$ is known, predictions of the cell state vectors $\vec{Z}_{\vec{J}}$ are done using and algorithm similar to the backward-forward algorithm as described below.

Backward-forward algorithm In the previous paragraph we showed how to compute the total likelihood of a MoMa data-set and how to estimate the optimal parameter set $\Theta^\star$. Once this is known, we would like to predict the posterior distribution

$$P(\vec{Z}_{\vec{J}}|\mathscr{D}, \Theta^\star) \tag{3.111}$$

where $\mathscr{D}$ is one of the cell lineage where the cell state vector $\vec{Z}_{\vec{J}}$ is present. To do this we make use of the backward-forward algorithm and an additional observation. First, let's write

$$\mathscr{D} = \{(x_n^m, g_n^m), \ldots, (x_0^m, g_0^m)\}$$

as

$$\mathscr{D} = \mathscr{D}_j \cup \mathscr{D}^{j+1}$$

where

$$\mathscr{D}_j = \left\{(x_j^m, g_j^m), \ldots, (x_0^m, g_0^m)\right\} \tag{3.112}$$
$$\mathscr{D}^{j+1} = \left\{(x_n^m, g_n^m), \ldots, (x_{j+1}^m, g_{j+1}^m)\right\} \tag{3.113}$$

i.e. we time separate the data before and after t_j. The backward-forward algorithm allow us to write

$$P(\vec{Z}_{\vec{J}}|\mathscr{D}) = \frac{P(\vec{Z}_{\vec{J}}\big|\mathscr{D}_j)P(\mathscr{D}^{j+1}\Big|\vec{Z}_{\vec{J}})}{P(\mathscr{D})} \tag{3.114}$$

and by making use of some basic probability rules, we can write

$$P(\vec{Z}_{\vec{J}}|\mathscr{D}) \propto \frac{P(\vec{Z}_{\vec{J}}\big|\mathscr{D}_j)P(\vec{Z}_{\vec{J}}\big|\mathscr{D}^{j+1})}{P(\vec{Z}_{\vec{J}})} \tag{3.115}$$

If we consider the prior distribution $P(\vec{Z}_{\vec{J}})$ to be uniform, we are only left to find $P(\vec{Z}_{\vec{J}}\big|\mathscr{D}_j)$ and $P(\vec{Z}_{\vec{J}}\big|\mathscr{D}^{j+1})$ and use the Gaussian multiplication property (3.5). Note that the first term has been computed in the previous paragraph (3.109), so we are only missing to find $P(\vec{Z}_{\vec{J}}\big|\mathscr{D}^{j+1})$. To do this let us consider the time backward process

$$\frac{d\tilde{\lambda}_t}{dt} = -\gamma_\lambda(\tilde{\lambda}_t + \bar{\lambda}) + \sigma_\lambda \eta_1(t) \tag{3.116}$$
$$\frac{d\tilde{x}_t}{dt} = \tilde{\lambda}_t \tag{3.117}$$
$$\frac{d\tilde{q}_t}{dt} = -\gamma_q(\tilde{q}_t + \bar{q}) + \sigma_q \eta_2(t) \tag{3.118}$$
$$\frac{d\tilde{g}_t}{dt} = e^{\tilde{x}_t}\tilde{q}_t + \beta\tilde{g}_t \tag{3.119}$$

i.e. we consider that, if we go backward in time, we will see the cell volume shrinking with a negative rate $-\bar{\lambda}$ and similar for the protein production rate $-\bar{q}$ and bleaching rate $-\beta$. This said, we can now easily compute the time backward cell state vector

$$\vec{w}_t = \begin{pmatrix} \tilde{\lambda}_t \\ \tilde{x}_t \\ \tilde{q}_t \\ \tilde{g}_t \end{pmatrix} \tag{3.120}$$

distribution. Indeed, if we consider the initial cell state distribution

$$p\left(\vec{w}_n\right) = \mathcal{N}\left(\vec{w}_n \middle| \left\langle \vec{w}_n \right\rangle, \left\langle \vec{w}_n, \vec{w}_n \right\rangle\right) \tag{3.121}$$

we can use the exact same procedure we used to compute (3.109) to find

$$p\left(\vec{w}_j \middle| \begin{bmatrix} x_{j+1}^m \\ g_{j+1}^m \end{bmatrix}, \ldots, \begin{bmatrix} x_n^m \\ g_n^m \end{bmatrix}\right) \tag{3.122}$$

keeping in mind to change the parameters

$$\begin{pmatrix} \bar{\lambda} \\ \bar{q} \\ \beta \end{pmatrix} \to \begin{pmatrix} -\bar{\lambda} \\ -\bar{q} \\ -\beta \end{pmatrix} \tag{3.123}$$

and to consider cell division in time backward i.e. to consider the division matrices (3.78) to become

$$D_d = \mathbf{diag}\left[\sigma_{dx}^2, \sigma_{dg}^2, 0, 0\right] \quad F = \mathbf{diag}\left[1,2,1,1\right] \quad \vec{f} = \left[+\log 2, 0, 0, 0\right]^T \tag{3.124}$$

Once the Gaussian distribution

$$p\left(\vec{w}_j \middle| D^{j+1}\right) = \mathcal{N}\left[\vec{w}_j \middle| \vec{\tilde{c}}, \tilde{C}\right] \tag{3.125}$$

is known, we find the backward distribution

$$p\left(\vec{Z}_{\vec{j}} \middle| D^{j+1}\right) = \mathcal{N}\left[\vec{Z}_{\vec{j}} \middle| \vec{c}, C\right] \tag{3.126}$$

by simply changing

$$\begin{aligned} &\vec{c}_\lambda = -\vec{\tilde{c}}_\lambda, \quad \vec{c}_q = -\vec{\tilde{c}}_q, \quad C_{\lambda x} = C_{x\lambda} = -\tilde{C}_{x\lambda} \\ &C_{\lambda g} = C_{g\lambda} = -\tilde{C}_{g\lambda}, \quad C_{qx} = C_{xq} = -\tilde{C}_{xq} \\ &C_{gq} = C_{qg} = -\tilde{C}_{gq}, \quad \vec{c}_j = \vec{\tilde{c}}_j \text{ and } C_{ij} = \tilde{C}_{ij} \text{ for the rest} \end{aligned}$$

The entire procedure described above to compute the likelihood, maximize it and find the posteriors had been written in Python and published on https://github.com/fioriathos/biophysical_gaussian_process_regression.git. In the next chapter we will see an application of this procedure on real data.

Chapter 4

The dynamic of the bacterial growth

Athos Fiori[1], Erik van Nimwegen[1,*]

[1] Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland.
[*] to whom correspondence should be addressed: erik.vannimwegen@unibas.ch

4.1 Introduction

A clear understanding of the stochasticity in the cellular growth is central for the understanding of phenomena like phenotypic diversity and cell size homeostasis. Metabolism and growth rate are often considered constant for a given condition [8][15] due to the large number of metabolic reactions, which average, should reduce fluctuations to undetectable levels. However, time-lapse microscopy data show that the fluctuations of the instantaneous growth rate of single cells can not be considered constant even within a cell cycle [26][47]. In order to quantify these deviations, we combined a biophysical stochastic model of cell growth with a dedicated Bayesian regression method. With the high resolution provided by our method we can investigate the dynamic of the growth rate during the cell cycle.

4.2 Model

It has been shown that, at a first approximation, the size of a single *E.coli*, grows almost exponentially [22][26] during the cell cycle. However, due to the stochastic nature of the cell elongation, the cell size fluctuates within a cell cycle and the deviations from the perfect exponential growth are not only due to measurement errors. Some heuristic attempts [26][47] have been done to infer the cell size dynamic with sub-cell cycle resolution but, to our knowledge, none of them is justified through a biophysical model. Moreover, due to the non negligible measurement errors, most of these methods give non reliable results. Here we propose a Bayesian model, based on a simple Langevin equation, which allows us to decouple measurement noise from biological fluctuations. Through this model we can then obtain reliable estimation of the growth parameters within the cell cycle.
A simple stochastic model which describes random fluctuations of a stochastic variable $\lambda_t = \lambda(t)$ around a fixed value $\bar{\lambda}$ is the Ornstein-Uhlenbeck process

$$\frac{d\lambda_t}{dt} = -\frac{1}{\tau}(\lambda_t - \bar{\lambda}) + \eta(t) \tag{4.1}$$

where $\langle \eta(t_2), \eta(t_1) \rangle = \sigma\delta(t_2 - t_1)$ is the stochastic part representing the Gaussian white noise of strength σ, whereas the drift term $\frac{1}{\tau}(\lambda_t - \bar{\lambda})$ ensures the process to drift forward its mean value $\bar{\lambda}$ with a characteristic time τ. The Ornstein-Uhlenbeck process is a stationary Gauss–Markov random process with mean $\bar{\lambda}$ and noise to mean ratio (coefficien of variation)

$$\mathrm{Cv}[\lambda] = \sqrt{\frac{\tau}{2}}\frac{\sigma}{\bar{\lambda}} \tag{4.2}$$

Instead of considering the cell size of one *E.coli* cell over time $s(t)$ as a simple exponential function

$$s(t) = s_{t_0} \exp\left[\bar{\lambda} t\right] \tag{4.3}$$

where s_{t_0} is the cell size at the begin of the cell cycle and $\bar{\lambda}$ the exponential growth rate, our model assumes the growth rate λ_t to follow an Ornstein-Uhlenbeck process. If no division event occurs between t_0 and t, the log cell size $x_t = \log s_t$ is then simply the time integral from birth (t_0) to t of the growth rate

$$x_t = x_0 + \int_{t_0}^{t} \lambda_\tau d\tau \tag{4.4}$$

where $x_0 = \log s_0$. Let us define the two dimensional cell state vector $\vec{z}_t$ at time t as

$$\vec{z}_t = \begin{pmatrix} x_t \\ \lambda_t \end{pmatrix} \tag{4.5}$$

i.e. the vector which components are the log cell size and growth rate at time t. Then, the Gaussian nature of equations (4.1) and (4.4) allow to easily compute the cell state vector probability distribution at any time in future $t > t_0$

$$p(\vec{z}_t | \vec{z}_{t_0}, \Theta) = \mathcal{N}(\vec{a} + F\vec{z}_{t_0}, C) \tag{4.6}$$

where $\Theta = \left(\lambda, \tau, \sigma^2\right)$ and $\vec{a}, F, C$ are defined in equations S.2,S.3.
We assume that, through time lapse microscopy, we are able to measure the log cell size of a cell from its birth $x^m(t_0)$ to its division $x^m(t_d)$. We will now show how, through (4.6), we can predict the instantaneous growth rate and cell size constrained to these measurements. For simplicity we consider the data-set consists of measurements of the log cell size during the cell cycle of one mother cell and one of its daughter cells only

$$D = \left\{ (x_i^m, t_i) \,|\, x_i^m \text{ is the measured log size of one cell at time } t_i,\ i = 0, \ldots, d, d+1, \ldots, n \right\} \tag{4.7}$$

where t_d is the time at which the only division event happen. Generalizing this to more division events is then straightforward.
The method we will present in the next paragraphs consists of three main steps. First it computes the necessary probability distributions at the initial conditions $\vec{z}_0$. Then updates these distributions for the next observation time (t_1). Finally it generalize the procedure through the entire data-set with a recursive relation.

Initial conditions We here compute the likelihood of the first observation (x_0^m) and the posterior of the initial cell state vector $\vec{z}_0$ given this observation.
The initial state vector is assumed to be Gaussian distributed

$$p(\vec{z}_0) = \mathscr{N}(\vec{s}, S) \tag{4.8}$$

with given mean $\vec{s}$ and covariance matrix S. The measurement[1] error in the log cell size is considered to be Gaussian distributed with mean zero and variance σ_ε^2

$$p(x_j^m|x_j) = \mathscr{N}\left(x_j, \sigma_\varepsilon^2\right) \tag{4.9}$$

The likelihood of the first measurement x_0^m reads

$$p(x_0^m|\tilde{\Theta}) = \int p\left(x_0^m|\vec{z}_0\right) p(\vec{z}_0) d\vec{z}_0 = \mathscr{N}\left(s_0, S_{00} + \sigma_\varepsilon^2\right) \tag{4.10}$$

with $\tilde{\Theta} = \{\Theta, \sigma_\varepsilon^2\}$ and s_0, S_{00} the components of the mean and covariance matrix. To compute the posterior probability distribution of the vector $\vec{z}_0$ given the observation x_0^m we use the Bayes theorem and we find

$$p\left(\vec{z}_0|x_0^m, \tilde{\Theta}\right) = \frac{p(x_0^m|\vec{z}_0)p(\vec{z}_0)}{p(x_0^m)} \propto \mathscr{N}(\vec{b}, B) \tag{4.11}$$

where the mean and covariance matrices $(\vec{b}, B)$ are defined in (S.13).

Update equation In the previous paragraph we saw how, starting from the prior distribution (4.8), we were able to find the likelihood (4.11) and the posterior distribution (4.10) if we considered Gaussian measurement errors (4.9). Finding the prior distribution at the next time point i.e.

$$p\left(\vec{z}_1|x_0^m, \tilde{\Theta}\right) \tag{4.12}$$

allows the procedure to be iterated.

Two possible scenarios can happen between the two observations at time t_0 and t_1 i.e. either the cell only grows, or the cell grows and divide.
If there is no division event between time t_0 and t_1 then through equation (4.6) we can easily find the prior distribution at time t_1

$$p\left(\vec{z}_1|x_0^m, \tilde{\Theta}\right) = \int p\left(\vec{z}_1|\vec{z}_0, \Theta\right) p\left(\vec{z}_0|x_0^m, \tilde{\Theta}\right) d\vec{z}_0 = \mathscr{N}\left(\vec{s}', S'\right) \tag{4.13}$$

[1]The superscript m stands for "measured quantity".

where $(\vec{s}', S')$ are the mean and covariance matrix defined in (S.6).
If the cell divide in this time lapse then the prior distribution of the cell state vector of the daughter cell $\left(\vec{z}_1^{\text{daugther}}\right)$ reads

$$\begin{aligned} p\left(\vec{z}_1^{\text{daugther}}|x_0^m,\tilde{\Theta}\right) = \int d\vec{z}_0^{\text{mother}} z_1^{\text{mother}} p\left(\vec{z}_1^{\text{daugther}}|\vec{z}_1^{\text{mother}}\right) \times \\ p\left(\vec{z}_1^{\text{mother}}|\vec{z}_0^{\text{mother}}\right) p\left(\vec{z}_0^{\text{mother}}|x_0^m,\tilde{\Theta}\right) \\ = \mathcal{N}\left(\vec{s}'', S''\right) \end{aligned} \tag{4.14}$$

where we consider the division event to split the cells into two identical parts

$$p\left(\vec{z}_1^{\text{daugther}}|\vec{z}_1^{\text{mother}}\right) = \mathcal{N}\left(z_1^{\text{mother}} - \log 2, \sigma_d^2\right) \tag{4.15}$$

The full shape of the mean $\vec{s}''$ and covariance matrices S'' and more details about the assumptions made are given in the supplementary material (S.9).

Recursive relation and total likelihood Equation (4.13) or (4.14) predict the prior distribution of the cell state vector at the next time point $\vec{z}_1$ given the masurement x_0^m

$$p\left(\vec{z}_1|x_0^m,\tilde{\Theta}\right) = \mathcal{N}\left(\vec{s}, S\right) \tag{4.16}$$

where $\vec{s}$ and S are the mean and covariance matrix given in (4.13) or (4.14) depending if the cell divides or not.
The likelihood of the next measurement x_1^m is computed in the same way as in (4.10), and reads

$$p\left(x_1^m|x_0^m,\tilde{\Theta}\right) = \mathcal{N}\left(s_0, S_{00} + \sigma_\varepsilon^2\right) \tag{4.17}$$

Similar the posterior is computed in the same way as in (4.11), and reads

$$p\left(\vec{z}_1|x_1^m,x_0^m,\tilde{\Theta}\right) \propto \mathcal{N}\left(\vec{b}, B\right) \tag{4.18}$$

It is clear that the procedure can be iterated over the entire data-set $\mathscr{D}$ to find

$$p\left(x_{i+1}^m|x_j^m,\ldots,x_0^m,\tilde{\Theta}\right) \text{ and } p\left(\vec{z}_j|x_j^m,\ldots,x_0^m,\tilde{\Theta}\right) \tag{4.19}$$

Summary In the previous paragraphs we show how to compute

$$p\left(x_0^m|\tilde{\Theta}\right), \quad p\left(x_{i+1}^m|x_j^m,\ldots,x_0^m,\tilde{\Theta}\right) \text{ and } p\left(\vec{z}_j|x_j^m,\ldots,x_0^m,\tilde{\Theta}\right) \tag{4.20}$$

We can easily compute the total likelihood using the basic probability rule

$$p(\mathscr{D}) = \prod_{i=0}^{n-1} p\left(x_{i+1}^m | x_i^m, \ldots, x_0^m, \tilde{\Theta}\right) p\left(x_0^m | \tilde{\Theta}\right) \tag{4.21}$$

We consider the best parameter set $\tilde{\Theta}_\star$ as the one maximising the total likelihood (4.21). Therefore, once $\tilde{\Theta}_\star$ is found, predictions of the cell state vector $\vec{z}_j$ are given by

$$p\left(\vec{z}_j | x_j^m, \ldots, x_0^m, \tilde{\Theta}_\star\right) \tag{4.22}$$

A generalization of this procedure to the case where $\mathscr{D}$ contains more division events is clearly straightforward. Note that if $\mathscr{D}$ contains two cells which do not share a known common ancestor, we consider the two cell genealogies as independent. This imply that the respective log likelihood contributions simply sums.

4.3 Results

All the following results and images can be generated using the ipython notebook in https://github.com/fioriathos/dynamic_of_bacterial_growth.git. The public available data [59] are obtained using an integrated microfluidics and time-lapse microscopy approach to quantitatively characterize growth and division in parallel across many lineages of single *E. coli* cells, both in slow and fast growth conditions. The different conditions are M9 minimal media supplemented with glycerol, glucose or glucose and eight amino acids (rich media), resulting in doubling times of 89, 53 and 41 min. These measurements allowed us to quantify each single cell cycle by a number of variables such as the growth rate, the sizes at birth and at division and the time between birth and division. As done in [59], we assume the cell radius is constant and use the cell length as a proxy for the cell volume. Since we can follow cells over multiple generations, we can also measure quantities that span multiple division cycles such as the long term auto correlation function. Some basic statistics are given in figure S.1 where the growth rate has been computed by assuming perfect exponential cell volume growth together with the least square method. In the next paragraphs we show how the previously developed model applies to these data-sets and focus on new results this method can bring.

Inference For the three different data-sets we compute the total log-likelihood using (4.21) and infer the parameter set $\tilde{\Theta}^\star$ which maximise it. The inferred parameters are shown in table 4.1 and we refer to section 4.5.5 for more details.

Media	$\bar{\lambda}[min^{-1}]$	τ [min]	$\frac{\sigma_\varepsilon}{\langle x \rangle}$ [%]	Cv$[\lambda]$[%]
glucose	$1.246 \times 10^{-2} \pm 9 \times 10^{-5}$	82 ± 4	3.61 ± 0.02	24 ± 1
glycerol	$7.30 \times 10^{-3} \pm 3 \times 10^{-5}$	105 ± 4	2.91 ± 0.02	22 ± 1
rich media	$1.386 \times 10^{-2} \pm 5 \times 10^{-5}$	50 ± 5	3.10 ± 0.03	20 ± 4

Table 4.1 The maximum likelihood parameters inferred for the three different media. Note that the measurement error is re-scaled with the mean log cell size $\langle x \rangle$ and the fluctuation in growth rate σ are expressed through the more meaningful quantity Cv$[\lambda]$.

Predictions Once the optimal parameter set $\tilde{\Theta}^\star$ is found, it is easy to compute the posterior distribution of the cell state vector $\vec{z}_j$ at any time point t_j (4.22). An example is given in figure 4.1 where a random cell and its daughter, growing in glycerol, are tracked from birth to division (orange points). The predicted log cell size and growth rate are computed through (4.22) and we represent in blue the mean and standard of this distribution. Clearly we here represent only two cells but we apply this on all the cells presents in the data-sets.

Cell cycle dynamic We can now compute quantities with a resolution which was not possible to obtain before. As an example, we compute the growth rate auto-correlation function with a resolution of a few minutes (figure 4.2). Not only this function scales with the doubling time, as already shown [26], but correlations are not decaying perfectly exponentially (black curve). Indeed, within the first cell cycle, the correlation drops faster than the exponential function and the trend seems to be inversely proportional to the growth rate. Another interesting observation is the dynamic of the growth rate during the cell cycle. In all conditions the growth rate is at the minimum ($\approx 4\%$ less than average growth rate) between the 30% and the 40% of the cell cycle, and reaches its maximum at the end of the cell cycle. Note that a similar growth rate dynamic pattern been observed [47] in *B.subtilis*.

4.4 Discussion

Metabolic and growth rates fluctuations have been often neglected in studies of bacterial growth. Even tough some more recent studies [22][26] have considered fluctuations of these quantities with a cell cycle resolution, none has never presented a justified method enabling to quantify these fluctuations with sub cell cycle time scale. To account for this, we presented a simple stochastic model for describing instantaneous cell growth. This model not only is capable to theoretically predict quantities like the growth rate correlation time (τ) or the instantaneous growth rate fluctuation ($Cv[\lambda]$) but, if combined with the Gaussian process regression method, it allows us to precisely estimate the instantaneous cell size and growth

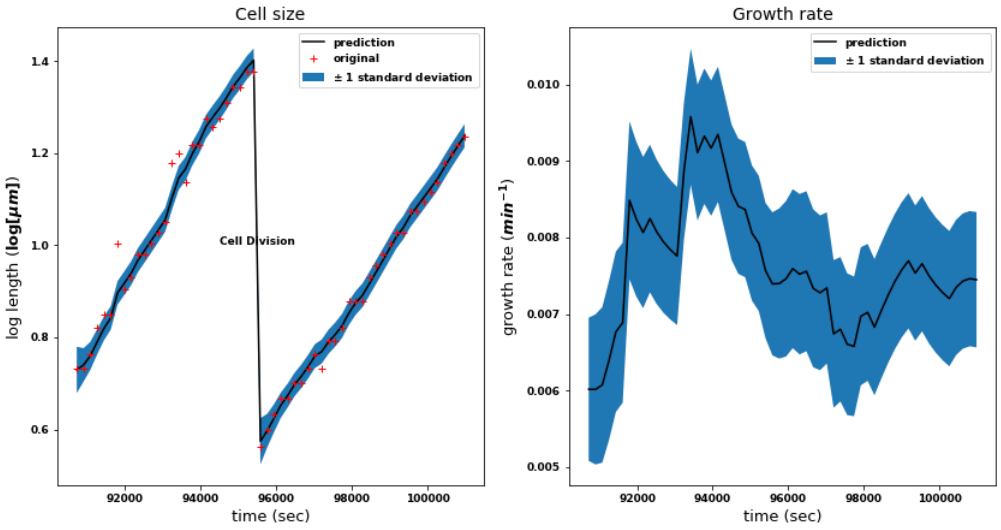

Fig. 4.1 **(left)** The log cell size of a random *E.Coli* cell, before and after division (red), growing in M9 minimal media supplemented with glycerol. The predicted log cell length distribution is given by (4.22) with $\tilde{\Theta}_\star$ given in table 4.1. The plot represents the mean (black) and standard deviation (blue) of the distribution. **(right)** The respective predicted growth rate distribution mean (black) and standard deviation (blue).

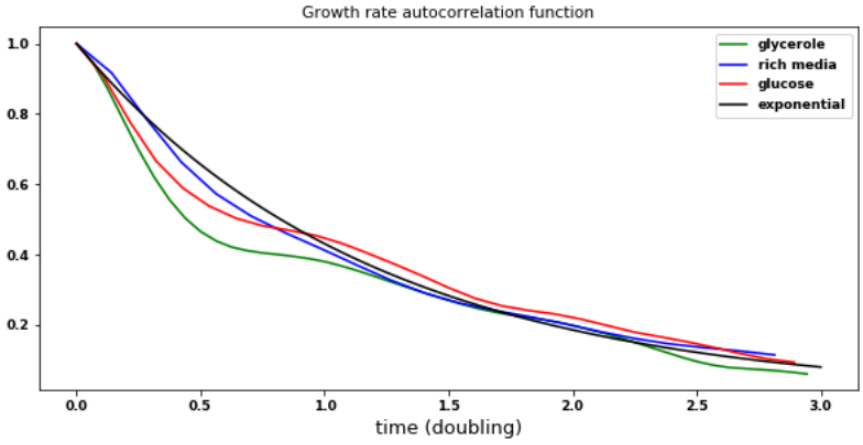

Fig. 4.2 Growth rate autocorrelation function computed with growth rates predicted by (4.22). The black line represent the exponential function $e^{-2.52\times t}$.

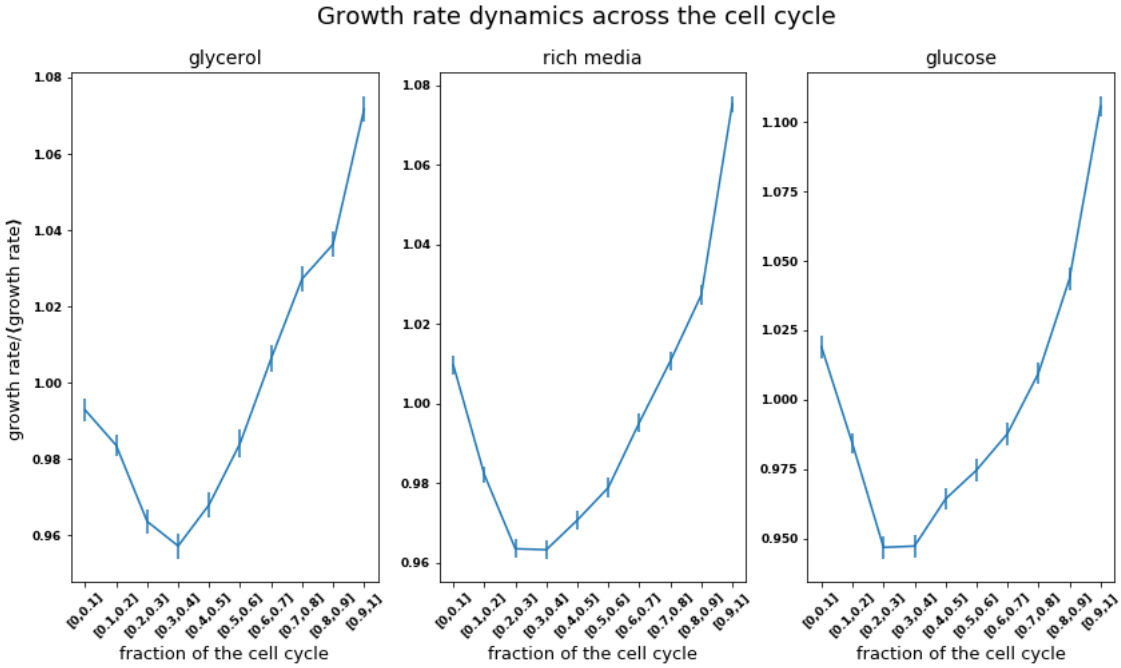

Fig. 4.3 Growth rate binned depending on the fraction of the cell cycle. The bin mean and standard error of the growth rate, divided by the total mean growth rate, in every condition is plotted. In all conditions the growth rate has the same dynamic. First, it decreases until reaching its minimum ($\approx 4\%$ less between $[30\%, 40\%]$ of the cell cycle) then starts to increase again.

rate. We show how these predictions bring new light in the study of the metabolic and growth rate dynamics and, as an example, we focus in computing the growth rate autocorrelation function and the dynamic of the growth rate over the cell cycle. The first shows that the growth rate correlation does not follow a perfect exponentially decaying function. Indeed, over the first cell cycle, the growth rate correlation first decays very fast and then reaches a plateau. This phenomena seems to be proportional to the doubling time. The second shows that the growth rate dynamic has a specific pattern over the cell cycle. For the three conditions, the growth rate is decreasing in the first 40% of the cell cycle and then increases again. Clearly, in this study we focus on the method and the potential application we can do with it more than on the final results. More conditions and more data should be collected in order to confirm the observed autocorrelation function and growth rate dynamics.

4.5 Supplementary Material

4.5.1 Prior distribution

First of all let's work out the shape of the matrices a, F and C used in equation (4.8). Notice that the Langevin equation (4.1) and its time integral generate Gaussian distributions, so it is sufficient to compute the mean and covariance function in order to find

$$p\left(\vec{z}_{j+k}|\vec{z}_j\right) \tag{S.1}$$

For notation simplicity we will consider $z_j = z_0$ i.e. the cell state vector at time $t_0 = 0$ and $z_{j+k} = z_{dt}$ i.e. the cell state vector at time $t_{j+k} = dt$. From the properties of the Wiener processes it is easy to find the expectation value of $\vec{z}_{dt}$ conditioned on $\vec{z}_0$

$$\begin{aligned}\langle \lambda_{dt}|\lambda_0\rangle &= \lambda_0 e^{-\gamma dt} + e^{-\gamma dt}\int_0^{dt} d\tau e^{-\gamma\tau}\bar{\lambda} \\ &= \lambda_0 e^{-\gamma dt} + \bar{\lambda}\left[1 - e^{-\gamma dt}\right] \\ \langle x_{dt}|\lambda_0, x_0\rangle &= x_0 + \int_0^{dt}\langle\lambda_\tau|\lambda_0\rangle d\tau = x_0 + \lambda_0\left[\frac{1-e^{-\gamma dt}}{\gamma}\right] \\ &\quad + \bar{\lambda}\left[\frac{e^{-\gamma dt} - (1-\gamma dt)}{\gamma}\right]\end{aligned}$$

and so

$$\langle\vec{z}_{dt}|\vec{z}_0\rangle = \begin{pmatrix}\langle x_{dt}|\lambda_0, x_0\rangle \\ \langle\lambda_{dt}|\lambda_0\rangle\end{pmatrix} = \underbrace{\begin{pmatrix}\bar{\lambda}\left[\frac{e^{-\gamma dt}-(1-\gamma dt)}{\gamma}\right] \\ \bar{\lambda}\left[1-e^{-\gamma dt}\right]\end{pmatrix}}_{\vec{a}} + \underbrace{\begin{pmatrix}1 & \frac{1-e^{-\gamma dt}}{\gamma} \\ 0 & e^{-\gamma dt}\end{pmatrix}}_{F}\vec{z}_0 \tag{S.2}$$

Similarly we can find the covariance matrix

$$C = \begin{pmatrix}\frac{\sigma^2}{2\gamma^3}\left[2\gamma dt - 3 + 4e^{-\gamma dt} - e^{-2\gamma dt}\right] & \frac{\sigma^2}{2\gamma^2}\left[1-e^{-\gamma dt}\right]^2 \\ \frac{\sigma^2}{2\gamma^2}\left[1-e^{-\gamma dt}\right]^2 & \frac{\sigma^2}{2\gamma}\left[1-e^{-2\gamma dt}\right]\end{pmatrix} \tag{S.3}$$

Now consider the cell state distribution at time t_j to be Gaussian with mean vector $\vec{b}$ and covariance matrix B

$$p\left(\vec{z}_j\right) = \frac{1}{\sqrt{(2\pi)^2\mathbf{Det}\left[B\right]}}\exp\left[-\frac{1}{2}\left(\vec{z}_j - \vec{b}\right)B^{-1}\left(\vec{z}_j - \vec{b}\right)\right] = \mathscr{N}\left(\vec{b}, B\right) \tag{S.4}$$

If no division event happens between time t_j and time t_{j+k}, the prior distribution at time t_{j+k} reads

$$p(\vec{z}_{j+k}) = \int d\vec{z}_j p\left(\vec{z}_{j+k}|\vec{z}_j\right) p\left(\vec{z}_j\right) = \mathscr{N}\left(\vec{s}', S'\right) \tag{S.5}$$

where

$$\vec{s}' = \vec{a} + F\vec{b} \quad \text{and} \quad S' = C + FBF^T \tag{S.6}$$

However, if a division event occurs between time t_j and time t_{j+k} then we have to consider the cell to get split into two almost identical parts within this time period. Let's name "mother" the cell before the division event occurs and "daughter" one of the two half after division. Consider the cell division distribution to be Gaussian, centered at half the size of the mother and with an asymmetry factor of σ_d which represents how much asymmetrically the cells divide

$$p\left(x_{j+k}^{\text{daughter}}|x_{j+k}^{\text{mother}}\right) = \frac{1}{\sqrt{2\pi\sigma_d^2}} e^{-\frac{\left(x_{i+k}^{\text{daughter}} - x_{i+k}^{\text{mother}} + \log 2\right)^2}{2\sigma_d^2}} \tag{S.7}$$

Clearly we do not know exactly when the division happened between t_j and t_{j+k} and if division affects the growth rate in a systematic way. To make it simple we considered that the cell grows as if no division had happened until t_{j+k} (S.5), and then we assume division happens exactly at the observed time t_{j+k}. With these assumptions we can easily find the prior distribution after division

$$p(\vec{z}_{j+k}^{\text{daughter}}) = \int d\vec{z}_{j+k}^{\text{mother}} p\left(\vec{z}_{j+k}^{\text{daughter}}|\vec{z}_{j+k}^{\text{mother}}\right) p\left(\vec{z}_{j+k}^{\text{mother}}\right) = \mathscr{N}\left(\vec{s}'', S''\right) \tag{S.8}$$

where

$$\vec{s}'' = \vec{s}' - \begin{pmatrix} \log 2 \\ 0 \end{pmatrix} \text{ and } S'' = S' + \begin{pmatrix} \sigma_d^2 & 0 \\ 0 & 0 \end{pmatrix} \tag{S.9}$$

4.5.2 Posterior distribution

Let's consider the prior distribution on the cell state vector $\vec{z}_k$ at time t_k to be Gaussian distributed with mean $\vec{s}$ and covariance matrix S

$$p\left(\vec{z}_k\right) = \mathscr{N}\left(\vec{s}, S\right) \tag{S.10}$$

Then if we assume the measured log size x_k^m at time t_k to be Gaussian distributed with error σ_ε

$$p(x_k^m|x_k) = \mathscr{N}\left(x_k, \sigma_\varepsilon^2\right) \tag{S.11}$$

the posterior probability distribution reads

$$p(\vec{z}_k|x_k^m) = \frac{p(x_k^m|\vec{z}_k)p(\vec{z}_k)}{p(x_k^m)} \propto \mathcal{N}\left(\vec{b}, B\right) \tag{S.12}$$

where

$$\vec{b} = \begin{pmatrix} \frac{s_0\sigma_\varepsilon^2 + S_{00}x_k^m}{\sigma_\varepsilon^2 + S_{00}} \\ s_1 - \frac{S_{01}(s_0 - x_k^m)}{\sigma_\varepsilon^2 + S_{00}} \end{pmatrix} \text{ and } B = \begin{pmatrix} \frac{S_{00}\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + S_{00}} & \frac{S_{01}\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + S_{00}} \\ \frac{S_{01}\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + S_{00}} & S_{11} - \frac{S_{01}^2}{\sigma_\varepsilon^2 + S_{00}} \end{pmatrix} \tag{S.13}$$

Clearly, s_j is the j^{th} component of the vector $\vec{s}$ and S_{ij} the i^{th} line and j^{th} column of the matrix S.

4.5.3 Computing basic statistics

We consider cells from birth t_0 to division t_d to grow exponentially

$$s(t) = s_{t_0} e^{\lambda t}, \quad t \in [t_0, t_d] \tag{S.14}$$

and, given the measured cell volumes, we use the least square method to infer s_{t_0} and λ. The autocorrelation function is computed as follow. For every cell i find its growth rate λ_i and the growth rate of the two respective daughter cells λ_1^i and λ_2^i (λ_k^l is the daughter cell k from the mother cell l). Then build the $2 \times n$ dimensional matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & \lambda_1^1 \\ \lambda_1 & \lambda_2^1 \\ \vdots & \vdots \\ \lambda_{n-2} & \lambda_1^{n-2} \\ \lambda_{n-2} & \lambda_2^{n-2} \end{bmatrix} \tag{S.15}$$

and the pearson correlation is computed as

$$\rho = \frac{\mathrm{Cov}\left[[\Lambda]_0 [\Lambda]_1\right]}{\sqrt{\mathrm{Var}\left[[\Lambda]_0\right] \mathrm{Var}\left[[\Lambda]_1\right]}} \tag{S.16}$$

where $[\Lambda]_j$ is the j^{th} column of Λ. Similar we do for the grand daughter. These are used to compute the statistics in figure S.1.

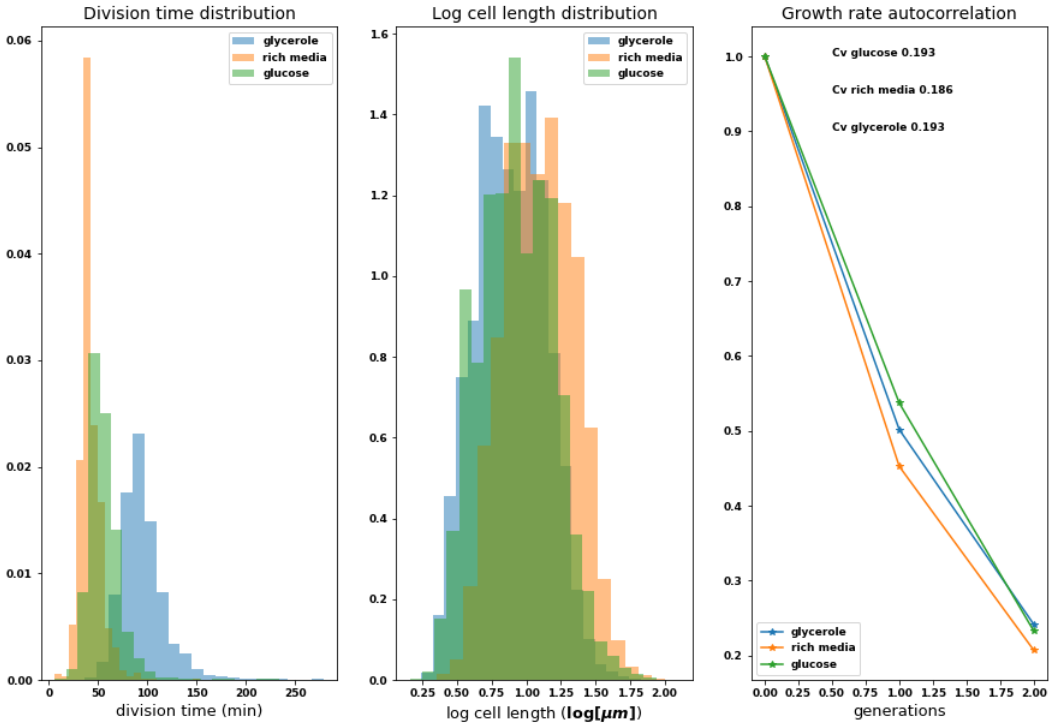

Fig. S.1 *E.Coli* growing in M9 minimal media supplemented with glycerol, glucose and glucose plus 8 amino acids (rich media) . **(left)** The distribution's time from birth to division for *E.coli* growin in these three conditions. **(middle)** The log cell size distribution in the three conditions.**(right)** The single cell growth rate λ is computed using the least square method together with the assumption that cells grow exponentially $\left(\propto e^{\lambda t}\right)$. We report the coefficient of variation (standard deviation divided by mean) of the growth rate distribution for the three conditions (≈ 0.19 in all conditions). Moreover, we plot the pearson correlation coefficient between the growth rate of the mother cells with their daughter cells (1 generation) and of the mother cells with their grand daughter cells (2 generations). The growth rate correlate ≈ 0.5 after one division and drop to ≈ 0.2 after two divisions.

4.5.4 Cell growth dynamic simulation

Let's show how to generate the cell growth dynamics, to mimic the mother machine data, once we know the parameter set

$$\tilde{\Theta} = \{\bar{\lambda}, \tau, \sigma, \sigma_d, \sigma_\varepsilon\}$$

First remember that if we only wish to simulate the standard Brownian motion $\eta(t)$ at one fixed value t, then we only need to generate a unit normal $Z \sim \mathcal{N}(0,1)$ and set $\eta(t) = \sigma\sqrt{t}Z$. For the Ornstein-Uhlenbeck process we use the discrete version of (4.1)

$$\lambda_{t+\Delta t} = \lambda_t - \frac{1}{\tau}(\lambda_t - \bar{\lambda})\Delta t + \sigma \Delta t^{\frac{3}{2}} Z$$

in order to generate the growth rate dynamics. Once the series of growth rates $(\lambda_0, \lambda_{\Delta t}, \ldots, \lambda_{n\Delta t})$ is generated, we need to find cell size dynamics. Remember the log cell size is defined to be the time integral of the growth rate

$$x_{k\Delta t} = x_0 + \sum_{i=1}^{k} \lambda_i \Delta t$$

but cell division and measurement errors must be considered in order to mimic mother machine data. Measurement errors are easy to simulate since we just consider

$$x_j^m \sim \mathcal{N}\left(x_j, \sigma_\varepsilon\right)$$

The cell division is implemented trough the adder model [53] i.e. every time the quantity $\sum_{i=1}^{k} \lambda_{i\Delta t}$ reach the threshold value ΔV, we consider the cell to divide. This means that, if the threshold is reached after k steps, the cell divides in half and the daughter cell will start with a volume

$$x_0^{\text{daughter}} \sim \mathcal{N}\left(x_{k\Delta t} - \log 2, \sigma_d\right)$$

We continue this procedure to form an entire genealogy which should simulate the data observed in the mother machine.

4.5.5 Inference with correlated measurement error

For the three different data-set we compute the total log-likelihood using (4.21) and infer the parameter set $\tilde{\Theta}^\star$ which maximise it (the maximum likelihood estimator (MLE)). As we will see, due to the correlation in measurement noise, we can not apply the inference method

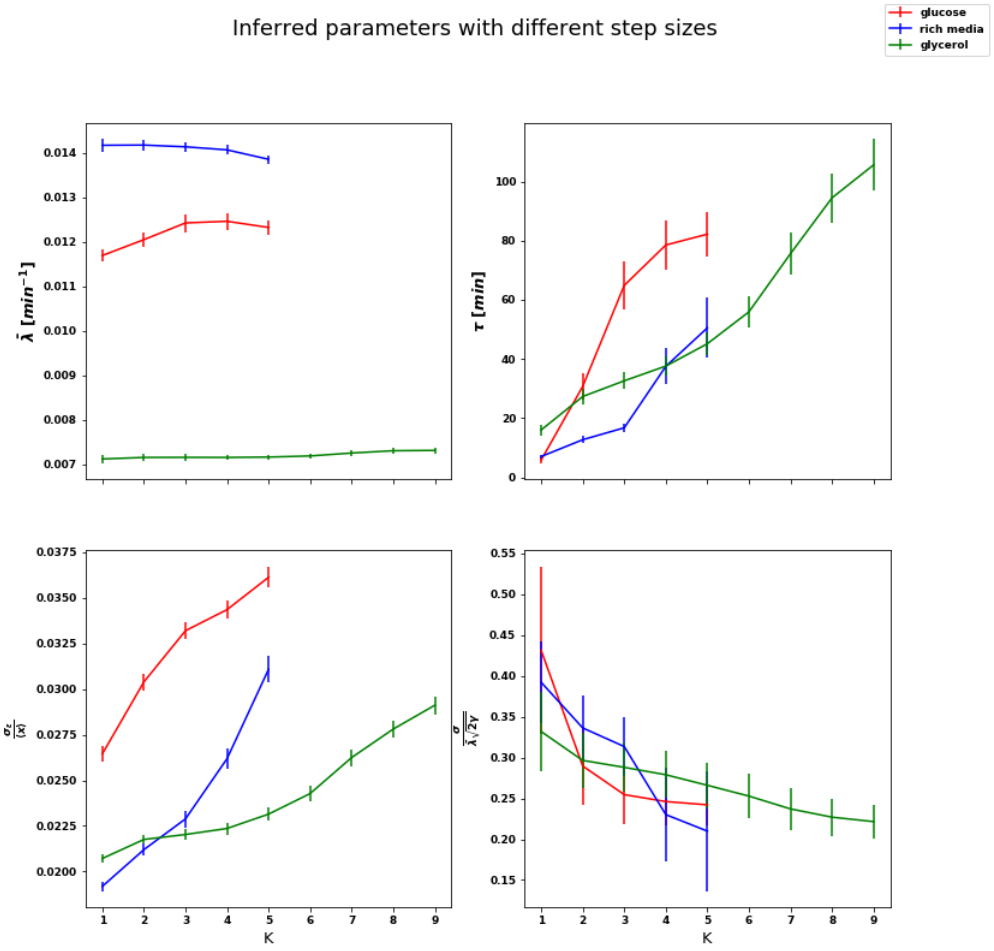

Fig. S.2 The MLE $\tilde{\Theta}_\star$ in the three different conditions using the data-set D_K given in (S.19). The case $K = 1$ is the case where nothing has been applied to the data. Measurement errors are the absolute values of the diagonal elements of the log-likelihood inverse Hessian matrix. **(Top left)** The mean growth rate $\bar{\lambda}$ for the three conditions for different D_K.**(Bottom left)**The relative error $\frac{\sigma_\varepsilon}{\langle x \rangle}$ for the three conditions for different D_K. **(Bottom right)** The coefficient of variation (4.2) for the three conditions for different D_K. Errors are in this case computed with the propagation error formula.

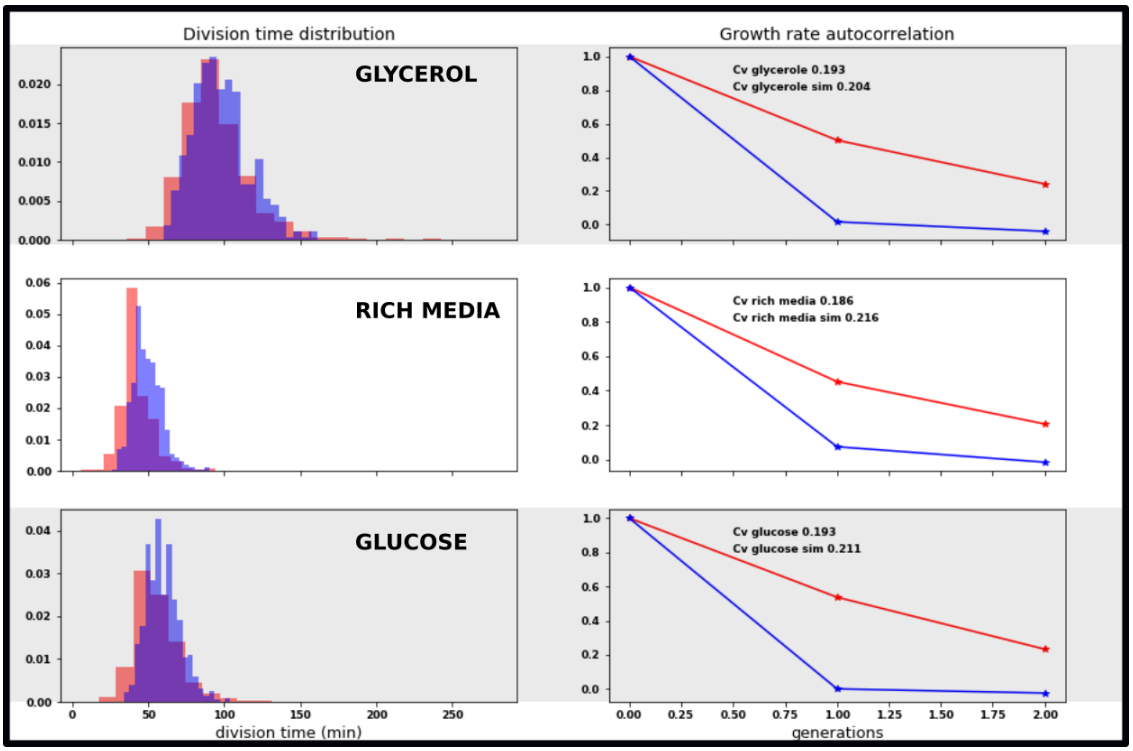

Fig. S.3 In red the basic statistics on the real data already shown in figure S.1. In blue the same statistics computed on synthetic data (section 4.5.4) generated using the inferred parameters given in figure S.2 with $K = 1$. Every line represents a condition. Whereas the division times of the simulations and biological data mostly agree, this is not true for the growth rate autocorrelation function.

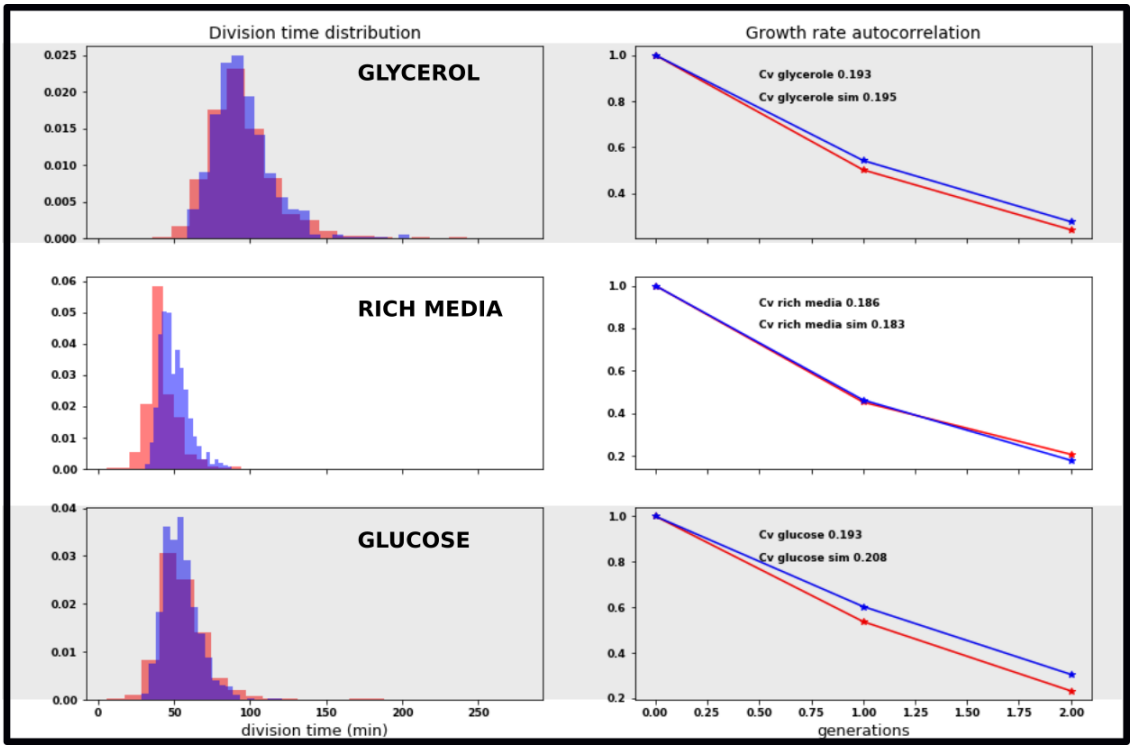

Fig. S.4 In red the basic statistics on the real data already shown in figure S.1. In blue the same statistics computed on synthetic data (section 4.5.4) generated using the inferred parameters given in figure S.2 with $K = 9$ for glycerol, $K = 4,5$ for glucose and rich media respectively. Every line represents a condition. The division times and the growth rate autorcorrelation functions of the simulations and biological data mostly agree and so we assume the inferred parameters $\tilde{\Theta}$ are now reliable to describe cell growth.

directly but we have to uncorrelate the measurement noise first.
Let start by finding the MLE of the data-set $\mathscr{D}$ directly i.e. compute (4.21) exactly as described in the method section. The inferred parameters corresponds to the case $K = 1$ in figure S.2. For example, in glucose, the inferred mean growth rate $\bar{\lambda} = 1.169 \times 10^{-2} \pm 7 \times 10^{-5}[min^{-1}]$, the correlation time $\tau = 5.8 \pm 0.5[min]$, the measurement error $\frac{\sigma_\varepsilon}{\langle x \rangle} = 2.24\% \pm 0.01\%$ and the fluctuations in growth rate (4.2) $\mathrm{Cv}[\lambda] = 43\% \pm 5\%$. As usual, the error bars are the absolute values of the diagonal elements of the inverse Hessian matrix of the log-likelihood function

$$\left.\frac{\partial^2 p\left(\mathscr{D}\middle|\tilde{\Theta}\right)}{\partial^2 \tilde{\Theta}_i}\right|_{\tilde{\Theta}_\star} \tag{S.17}$$

In order to test whether these parameters well describe the cell growth dynamics, we run a computer simulation (section 4.5.4) which mimic cell growth and division. If the simulated traces generated with the inferred parameters $\tilde{\Theta}_\star$ consistently describe cell growth and division we should obtain similar statistics as the one we compute on the biological data. In figure S.3 we compare the division time distributions and the growth rate auto-correlation function for the original data (red) and the simulated data (blue). The most striking observation is the disagreement in the auto-correlation function. Indeed, in the biological data we have a correlation higher than .5 after one generation which is clearly not observed in the simulated traces. This is due to the relatively short correlation time τ inferred, which is several fold smaller than the division time, and makes correlations longer than a cell cycle vanishing. We think the algorithm infers short time scales correlations τ due to the naive assumption of uncorrelated Gaussian noise. Indeed, when looking to the time lapse microscopy images, we realise that cells wiggle inside the channels making the measurement noise correlated. We think that the inferred time scale τ is the time scale of the correlation in the measurement noise and in order to test this hypothesis we apply the following technique. Consider to split the original data-set

$$D = (\{(x_0^m, t_0), (x_1^m, t_1) ..., (x_N^m, t_N)\}) \tag{S.18}$$

into K different data-sets i.e. for $j = 0, 1, .., K$

$$D_j = \left\{(x_j^m, t_j), (x_{j+K}^m, t_{j+K}), (x_{j+2K}^m, t_{j+2K}) ..., (x_{N-K+j}^m, t_{N-K+j})\right\} \tag{S.19}$$

and consider these K data-sets as independent. If the measurement noise is not correlated, the optimal parameter set $\tilde{\Theta}^\star$ would be the same regardless of K. We apply this procedure on our data with $K = 1, \ldots, 5$ for glucose and rich media and $K = 1, \ldots, 9$ for glycerol. All the inferred parameters converge, within error bars, suggesting that the measurement noise is not correlated anymore. We again test these parameters trough the previously described

simulation (section 4.5.4) and summarize the statistics in figure S.4. The statistics of the simulations and the real data matches almost perfectly the biological statistics suggesting that the inferred parameters well describes the cell growth in these conditions.

Chapter 5

Conclusion

In the first chapter we theoretically demonstrated that due to the exponential growth of bacterial populations, the time a specific cell needs to exit the lag has an exponential impact on its number of descendants. This imply that the contributions to the final population given by the fast adapting cells is exponentially larger than the one given by slow adapting cells. This observation let us hypothesize that selection is strong on the head of the single cell lag time distribution and weak on the tail. Meaning that a mutations acting on fast adapting cells (the head of the LD) have an high impact on the reproductive success of the colony and thus are strongly selected. On the contrary, mutations acting on slow adapting cells (the tail of the LD) have a low impact on the final colony implying a weak selection pressure. Thus, we expect deleterious mutations acting on the tail of the LD to accumulated. According to this observation and in addition to any potential advantage a long tailed LD might have, we showed that heterogeneous LD should be observed in wild *E. Coli* just due to the weak selection pressure on the tail. Another important observation we did in this study is to realize that the non trivial relationship between the population lag time and the single cell lag time distribution depends on the initial size of the colony. Indeed, we showed that the expected population lag time is longer when the initial colony size is small and this effects is stronger when the LD is heterogeneous. In order to understand the consequences of this observation we simulated bacterial colonies living in fluctuating environments. As predicted by our theory we showed that heterogeneous LD are effective for large populations as far as a subset of bacteria can adapt fast to the new environment but are inefficient in small populations. This suggests that, while large populations can employ bet-hedging strategies to deal with unexpected environmental changes, small populations will require regulated sense-and-response strategies in order to ensure a short population lag. Last, we also studied the potential problems which may arise when we define the log genotype fraction as a measure of fitness. This fitness measure is often used in evolutionary experiments and we show that, due

to the colony size dependence of the population lag, one may overestimate the fitness of the more abundant genotype even in cases where no selection is acting. This fictitious selection force is an example of a more general theory developed by [13]. In the second part of the thesis we focus our attention on the analysis of the MoMa [22] time series data for cell size and gene expression. We first presented a general Bayesian method used to treat gene expression data and provided the necessary informations in order to understand the package developed in https://github.com/fioriathos/gaussian_smoothing.git. However, this model lacks a biological interpretation of the underlying variables. We therefore developed a biophysical model of cell growth and gene expression based on the simple Ornstein-Uhlenbeck stochastic process. We then showed how to combined the biophysical model of growth and gene expression with the Gaussian regression process method. In the last chapter we apply this method on time series data of cell growth https://github.com/fioriathos/dynamic_of_bacterial_growth.git. We showed that bacteria growing in different conditions have similar dynamics. Among others we showed the growth rate correlates with a time scale proportional to itself and loses correlation faster than exponentially over the first cell cycle. Due to the high resolution of the latent variables provided by this method, new studies of the single cells dynamics with a sub cell cycle resolution are now possible.

References

[1] N. Q. Balaban, J. Merrin, R. Chait, L. Kowalik, and S. Leibler. Bacterial persistence as a phenotypic switch. *Science (New York, N.Y.)*, 305(5690):1622–1625, Sept. 2004. ISSN 1095-9203. doi: 10.1126/science.1099390.

[2] n. Baranyi. Comparison of Stochastic and Deterministic Concepts of Bacterial Lag. *Journal of Theoretical Biology*, 192(3):403–408, June 1998. ISSN 1095-8541. doi: 10.1006/jtbi.1998.0673.

[3] N. Barkai and B.-Z. Shilo. Variability and robustness in biomolecular systems. *Molecular Cell*, 28(5):755–760, Dec. 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.11.013.

[4] J. R. Blundell and S. F. Levy. Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer. *Genomics*, 104 (6, Part A):417–430, Dec. 2014. ISSN 0888-7543. doi: 10.1016/j.ygeno.2014.09.005. URL http://www.sciencedirect.com/science/article/pii/S0888754314001827.

[5] J. J. Bull. EVOLUTION OF PHENOTYPIC VARIANCE. *Evolution; International Journal of Organic Evolution*, 41(2):303–315, Mar. 1987. ISSN 1558-5646. doi: 10.1111/j.1558-5646.1987.tb05799.x.

[6] J. N. Carey and M. Goulian. A bacterial signaling system regulates noise to enable bet hedging. *Current genetics*, 65(1):65–70, Feb. 2019. ISSN 0172-8083. doi: 10.1007/s00294-018-0856-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6291380/.

[7] C. E. R. Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. ISBN 0-262-18253-X.

[8] D. Fell. *Understanding the control of metabolism*. Portland, 1997.

[9] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467 (7312):167–173, Sept. 2010. ISSN 1476-4687. doi: 10.1038/nature09326.

[10] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297(5584):1183–1186, Aug. 2002. ISSN 1095-9203. doi: 10.1126/science.1070919.

[11] Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open source scientific tools for Python, 2001. URL http://www.scipy.org.

[12] O. Fridman, A. Goldberg, I. Ronin, N. Shoresh, and N. Q. Balaban. Optimization of lag time underlies antibiotic tolerance in evolved bacterial populations. *Nature*, 513 (7518):418–421, Sept. 2014. ISSN 1476-4687. doi: 10.1038/nature13469.

[13] O. Hallatschek. Selection-Like Biases Emerge in Population Models with Recurrent Jackpot Events. *Genetics*, 210(3):1053–1073, Nov. 2018. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.118.301516. URL https://www.genetics.org/content/210/3/1053.

[14] M. Hashimoto, T. Nozoe, H. Nakaoka, R. Okura, S. Akiyoshi, K. Kaneko, E. Kussell, and Y. Wakamoto. Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences of the United States of America*, 113 (12):3251–3256, Mar. 2016. ISSN 0027-8424. doi: 10.1073/pnas.1519412113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4812751/.

[15] M. J. Herrgård, M. W. Covert, and B. Palsson. Reconstruction of microbial transcriptional regulatory networks. *Current Opinion in Biotechnology*, 15(1):70–77, Feb. 2004. ISSN 0958-1669. doi: 10.1016/j.copbio.2003.11.002. URL http://www.sciencedirect.com/science/article/pii/S0958166903001812.

[16] G. Hornung, R. Bar-Ziv, D. Rosin, N. Tokuriki, D. S. Tawfik, M. Oren, and N. Barkai. Noise-mean relationship in mutated promoters. *Genome Research*, 22(12):2409–2417, Dec. 2012. ISSN 1549-5469. doi: 10.1101/gr.139378.112.

[17] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, June 1961. ISSN 0022-2836. doi: 10.1016/S0022-2836(61)80072-7. URL http://www.sciencedirect.com/science/article/pii/S0022283661800727.

[18] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906. ISSN 0001-5962, 1871-2509. doi: 10.1007/BF02418571. URL https://projecteuclid.org/euclid.acta/1485887155.

[19] T. C. Jian Qing Shi. *Gaussian Process Regression Analysis for Functional Data*. CRC Press, 2011. ISBN 978-1-4398-3773-3.

[20] T. Julou, D. Blank, A. Fiori, and E. v. Nimwegen. Subpopulations of sensorless bacteria drive fitness in fluctuating environments. *bioRxiv*, page 2020.01.04.894766, Jan. 2020. doi: 10.1101/2020.01.04.894766. URL https://www.biorxiv.org/content/10.1101/2020.01.04.894766v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[21] A. Jõers, N. Kaldalu, and T. Tenson. The frequency of persisters in Escherichia coli reflects the kinetics of awakening from dormancy. *Journal of Bacteriology*, 192(13): 3379–3384, July 2010. ISSN 1098-5530. doi: 10.1128/JB.00056-10.

[22] M. Kaiser, F. Jug, T. Julou, S. Deshpande, T. Pfohl, O. K. Silander, G. Myers, and E. van Nimwegen. Monitoring single-cell gene regulation under dynamically controllable conditions with integrated microfluidics and software. *Nature Communications*, 9(1): 212, 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02505-0.

[23] R. Kalhor, P. Mali, and G. M. Church. Rapidly evolving homing CRISPR barcodes. *Nature Methods*, 14(2):195–200, Feb. 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4108. URL https://www.nature.com/articles/nmeth.4108. Number: 2 Publisher: Nature Publishing Group.

[24] S. Karlin and U. Lieberman. Random temporal variation in selection intensities: Case of large population size. *Theoretical Population Biology*, 6(3):355–382, Dec. 1974. ISSN 0040-5809. doi: 10.1016/0040-5809(74)90016-1. URL http://www.sciencedirect.com/science/article/pii/0040580974900161.

[25] M. KIMURA. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

[26] D. J. Kiviet, P. Nghe, N. Walker, S. Boulineau, V. Sunderlikova, and S. J. Tans. Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379, Oct. 2014. ISSN 1476-4687. doi: 10.1038/nature13582.

[27] S. Klumpp and T. Hwa. Growth-Rate-Dependent Partitioning of RNA Polymerases in Bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–20250, Dec. 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0804953105.

[28] S. Klumpp and T. Hwa. Traffic Patrol in the Transcription of Ribosomal RNA. *RNA Biology*, 6(4):392–394, Sept. 2009. ISSN 1547-6286. doi: 10.4161/rna.6.4.8952.

[29] R. Kolter, D. A. Siegele, and A. Tormo. The Stationary Phase of the Bacterial Life Cycle. *Annual Review of Microbiology*, 47(1):855–874, 1993. doi: 10.1146/annurev.

mi.47.100193.004231. URL https://doi.org/10.1146/annurev.mi.47.100193.004231. _eprint: https://doi.org/10.1146/annurev.mi.47.100193.004231.

[30] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science (New York, N.Y.)*, 309(5743):2075–2078, Sept. 2005. ISSN 1095-9203. doi: 10.1126/science.1114383.

[31] Z. Kutalik, M. Razaz, and J. Baranyi. Connection between stochastic and deterministic modelling of microbial growth. *Journal of Theoretical Biology*, 232(2):285–299, Jan. 2005. ISSN 0022-5193. doi: 10.1016/j.jtbi.2004.08.013.

[32] I. Levin-Reisman, O. Gefen, O. Fridman, I. Ronin, D. Shwa, H. Sheftel, and N. Q. Balaban. Automated imaging with ScanLag reveals previously undetectable bacterial growth phenotypes. *Nature Methods*, 7(9):737–739, Sept. 2010. ISSN 1548-7105. doi: 10.1038/nmeth.1485.

[33] S. F. Levy, J. R. Blundell, S. Venkataram, D. A. Petrov, D. S. Fisher, and G. Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519 (7542):181–186, Mar. 2015. ISSN 1476-4687. doi: 10.1038/nature14279. URL https://www.nature.com/articles/nature14279. Number: 7542 Publisher: Nature Publishing Group.

[34] R. C. Lewontin and D. Cohen. On Population Growth in a Randomly Varying Environment. *Proceedings of the National Academy of Sciences*, 62(4):1056–1060, Apr. 1969. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.62.4.1056. URL https://www.pnas.org/content/62/4/1056.

[35] L. F. Liu and J. C. Wang. Supercoiling of the DNA template during transcription. *Proceedings of the National Academy of Sciences*, 84(20):7024–7027, Oct. 1987. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.84.20.7024. URL https://www.pnas.org/content/84/20/7024. Publisher: National Academy of Sciences Section: Research Article.

[36] M. T. Madigan, J. M. Martinko, and J. Parker. *Brock biology of microorganisms*. 2000. ISBN 978-0-13-081922-2 978-0-13-085264-9. OCLC: 41400792.

[37] R. S. Makman and E. W. Sutherland. Adenosine 3',5'-Phosphate in Escherichia coli. *Journal of Biological Chemistry*, 240(3):1309–1314, Mar. 1965. ISSN 0021-9258, 1083-351X. URL http://www.jbc.org/content/240/3/1309. Publisher: American Society for Biochemistry and Molecular Biology.

[38] W. Margolin. FTSZ AND THE DIVISION OF PROKARYOTIC CELLS AND ORGANELLES. *Nature reviews. Molecular cell biology*, 6(11):862–871, Nov. 2005. ISSN 1471-0072. doi: 10.1038/nrm1745. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4757588/.

[39] P. B. Medawar. *An unsolved problem of biology*. College, 1952.

[40] R. Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, 35(12):1050, Dec. 2013. doi: 10.1002/bies.201300066. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3910158/. Publisher: Wiley-Blackwell.

[41] J. Monod. Recherches sur la croissance des cultures bactériennes. 1942. URL https://agris.fao.org/agris-search/search.do?recordID=US201300336259. Publisher: Hermann.

[42] J. Monod. The Growth of Bacterial Cultures. *Annual Review of Microbiology*, 3(1):371–394, 1949. doi: 10.1146/annurev.mi.03.100149.002103. URL https://doi.org/10.1146/annurev.mi.03.100149.002103.

[43] S. Moreno-Gámez, D. J. Kiviet, C. Vulin, S. Schlegel, K. Schlegel, G. S. v. Doorn, and M. Ackermann. Wide lag time distributions break a trade-off between reproduction and survival in bacteria. *Proceedings of the National Academy of Sciences*, 117(31):18729–18736, Aug. 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2003331117. URL https://www.pnas.org/content/117/31/18729. Publisher: National Academy of Sciences Section: Biological Sciences.

[44] F. C. Neidhardt, P. L. Bloch, and D. F. Smith. Culture medium for enterobacteria. *Journal of Bacteriology*, 119(3):736–747, Sept. 1974. ISSN 0021-9193. doi: 10.1128/JB.119.3.736-747.1974.

[45] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, June 2006. ISSN 1476-4687. doi: 10.1038/nature04785.

[46] G. W. Niven, J. S. Morton, T. Fuks, and B. M. Mackey. Influence of environmental stress on distributions of times to first division in Escherichia coli populations, as determined by digital-image analysis of individual cells. *Applied and Environmental Microbiology*, 74(12):3757–3763, June 2008. ISSN 1098-5336. doi: 10.1128/AEM.02551-07.

[47] N. Nordholt, J. H. van Heerden, and F. J. Bruggeman. Biphasic Cell-Size and Growth-Rate Homeostasis by Single Bacillus subtilis Cells. *Current Biology*, 30(12):2238–2247.e5, June 2020. ISSN 0960-9822. doi: 10.1016/j.cub.2020.04.030. URL http://www.sciencedirect.com/science/article/pii/S0960982220305443.

[48] O. Patange, C. Schwall, M. Jones, C. Villava, D. A. Griffith, A. Phillips, and J. C. W. Locke. Escherichia coli can survive stress by noisy growth modulation. *Nature Communications*, 9(1):5333, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07702-z.

[49] P. A. Pavco and D. A. Steege. Elongation by Escherichia coli RNA polymerase is blocked in vitro by a site-specific DNA binding protein. *The Journal of Biological Chemistry*, 265(17):9960–9969, June 1990. ISSN 0021-9258.

[50] C. Pin and J. Baranyi. Single-cell and population lag times as a function of cell age. *Applied and Environmental Microbiology*, 74(8):2534–2536, Apr. 2008. ISSN 1098-5336. doi: 10.1128/AEM.02402-07.

[51] S. C. Sleight and R. E. Lenski. Evolutionary adaptation to freeze-thaw-growth cycles in Escherichia coli. *Physiological and biochemical zoology: PBZ*, 80(4):370–385, Aug. 2007. ISSN 1522-2152. doi: 10.1086/518013.

[52] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, Oct. 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.162041399. URL https://www.pnas.org/content/99/20/12795. Publisher: National Academy of Sciences Section: Biological Sciences.

[53] S. Taheri-Araghi, S. Bradde, J. T. Sauls, N. S. Hill, P. A. Levin, J. Paulsson, M. Vergassola, and S. Jun. Cell-size control and homeostasis in bacteria. *Current biology: CB*, 25(3):385–391, Feb. 2015. ISSN 1879-0445. doi: 10.1016/j.cub.2014.12.009.

[54] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science (New York, N.Y.)*, 329(5991):533–538, July 2010. ISSN 1095-9203. doi: 10.1126/science.1188308.

[55] G. Ullman, M. Wallden, E. G. Marklund, A. Mahmutovic, I. Razinkov, and J. Elf. High-throughput gene expression analysis at the level of single proteins using a microfluidic turbidostat and automated cell tracking. *Philosophical Transactions of the Royal*

Society of London. Series B, Biological Sciences, 368(1611):20120025, Feb. 2013. ISSN 1471-2970. doi: 10.1098/rstb.2012.0025.

[56] P. Wang, L. Robert, J. Pelletier, W. L. Dang, F. Taddei, A. Wright, and S. Jun. Robust growth of Escherichia coli. *Current biology : CB*, 20(12):1099–1103, June 2010. ISSN 0960-9822. doi: 10.1016/j.cub.2010.04.045. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2902570/.

[57] G. C. Wick. The Evaluation of the Collision Matrix. *Physical Review*, 80(2):268–272, Oct. 1950. doi: 10.1103/PhysRev.80.268. URL https://link.aps.org/doi/10.1103/PhysRev.80.268. Publisher: American Physical Society.

[58] G. C. Williams. Pleiotropy, Natural Selection, and the Evolution of Senescence. *Evolution*, 11(4):398–411, 1957. ISSN 0014-3820. doi: 10.2307/2406060. URL https://www.jstor.org/stable/2406060.

[59] G. Witz, E. van Nimwegen, and T. Julou. Initiation of chromosome replication controls both division and replication cycles in E. coli through a double-adder mechanism. *eLife*, 8. ISSN 2050-084X. doi: 10.7554/eLife.48063. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6890467/.

[60] L. Wolf, O. K. Silander, and E. van Nimwegen. Expression noise facilitates the evolution of gene regulation. *eLife*, 4:e05856, June 2015. ISSN 2050-084X. doi: 10.7554/eLife.05856. URL https://doi.org/10.7554/eLife.05856. Publisher: eLife Sciences Publications, Ltd.

[61] J. W. Young, J. C. W. Locke, A. Altinok, N. Rosenfeld, T. Bacarian, P. S. Swain, E. Mjolsness, and M. B. Elowitz. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature protocols*, 7(1):80–88, Dec. 2011. ISSN 1754-2189. doi: 10.1038/nprot.2011.432. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161363/.

Acknowledgements

First, I would like to thank my advisor, Prof. Erik van Nimwegen, for giving me the chance to do my thesis in his group and for its support throughout these years. I would also like to thank Prof. Zoltan Kutalik and Prof. Richard Neher for accepting to be part of my PhD Committee. Also special thanks to Dorde, Arantxa, Luca, Gwendoline, Dani, Pascal, Thomas, Théo and all the other members who helped and sustained me during these years. An enormous thank to Adélaïde who supported me during this difficult period. Your help and your advises have had a big impact on my thesis and on my life in general and just a thank you is probably not enough. Last but not least, I would like to thank all my family. Only thanks to their help and their efforts I could have done what I did. By the way, you were right mum, doing a PhD "l'é come fa un rosct".