IN VIVO STUDY
OF THE COUPLING BETWEEN
TWO CELLULAR OSCILLATORS

Memòria

Autora
Lara Sofia ESCUAIN i POOLE

Tutors
Josep Lluís FONT GARCÍA
Jordi GARCIA OJALVO

Convocatòria
Setembre de 2013 (pla 2003)
He has made everything beautiful in its time.  
He has also set eternity in the human heart; yet no one can fathom what God has done from beginning to end. 
Ecc. 3:11
12 Mathematical model

12.1 A reminder of the model ................................................. 93
12.2 The parameters of the model at 27°C and 30°C .................... 94
12.3 Analysis of the results of the simulations ............................ 94
   12.3.1 Qualitative comparison of the length and number initiations . 94
   12.3.2 Quantitative analysis of the number of initiations ............... 96
   12.3.3 Further steps for this study ....................................... 97

V CONCLUSIONS

13 Conclusions

13.1 Analysis of the experiments ........................................... 101
13.2 Simulations of the model ............................................... 103

REFERENCES

VI APPENDICES

A Statistical analysis of the variables
   A.1 Detailed description of the plots ................................... 115
   A.1.1 Scatter plots ..................................................... 115
   A.1.2 Histograms ....................................................... 116
   A.1.3 Trending plots ................................................... 116
   A.2 Fitting the growth rate ............................................ 116
   A.3 Statistical significance of differences between the two temperatures . 118
      A.3.1 Significance of the difference between the growth rates ........ 118
      A.3.2 Significance of the difference between the final lengths .......... 119
   A.4 P-P plots for the fitting of cell length data ........................ 120

B Analytical solution of the model

C Code for RK4
   C.1 First version ....................................................... 127
   C.2 Improvements ....................................................... 131
List of Figures

5.1 *Escherichia coli* bacteria (public domain) ........................................... 31
5.2 Theodor Escherich (public domain) ......................................................... 32
5.3 Schematic illustration of a section of the cell membrane structure ............. 33
5.4 Drawings of the inside of the cell ............................................................. 35
5.5 A ribosome synthesising proteins .............................................................. 36
5.6 Schematic drawing of the cell cycle ............................................................ 37
5.7 Schematic representation of chromosomal replication ............................... 38

7.1 A phase (top) and a fluorescence (bottom) image of one of the films .......... 47
7.2 Frames of one of the films ............................................................................ 48
7.3 Preparation of the bacterial culture for observation under microscope ....... 49

8.1 Screenshot of CCSim simulating a nutritional shift-up ............................... 54

9.1 The process of clicking the cells illustrated in screenshots ....................... 57
9.2 A length vs time graph of a growing cell ................................................... 58
9.3 Three of the frames after applying the threshold ....................................... 60
9.4 A frame in which two distinct dots can be observed in one bright region.... 60
9.5 First attempt at counting a cell's number of dots in each frame .......... ....... 61
9.6 Montserrat: part of the mountain crests on the South side ....................... 61
9.7 One of the cells as a contour map ............................................................... 62
9.8 A frame of one of the videos that were used for the dot count ................ 63
9.9 Cell length $\times 10^{-1}$ and number of dots at 27°C .................................... 64

10.1 Simulated cycle time values ................................................................. 71
10.2 Simulation of the cycle times ................................................................. 72
10.3 Length and dots for different values of $\delta$ ........................................... 73
10.4 Length and dots for different values of $k_i$ ............................................. 74
10.5 Histograms for $k_u=1.5$ min$^{-1}$ .............................................................. 74

11.1 Error bars for the temperature of 27°C ................................................. 77
11.2 Trending plots for the temperature of 30°C ........................................... 78
11.3 Probability density functions of the growth rates at 27°C and 30°C ......... 79
11.4 Probability density functions of the cycle times at 27°C and 30°C . . . 80
11.5 Maximum, minimum and mean cell length for each of the cycles at 27°C 81
11.6 P-P plot for the log-normal distribution of the cell length . . . . . . 81
11.7 Probability density functions of the cell lengths at 27°C and 30°C . . . 82
11.8 Global averages of cell size . . . . . . . . . . . . . . . . . . . . . . . . . 82
11.9 Frequency of the number of oriCs in the cell at 27°C and 30°C . . . . 83
11.10 Interinitiation times at 30°C . . . . . . . . . . . . . . . . . . . . . . . . . 85
11.11 Growth rates vs cycle times . . . . . . . . . . . . . . . . . . . . . . . . . 86
11.12 Growth rates and cycle times vs cell length for both temperatures . . 87
11.13 Interinitiation time vs growth rate . . . . . . . . . . . . . . . . . . . . . 88
11.14 Length vs dots and linear regression at 27°C . . . . . . . . . . . . . . 89
11.15 Length vs dots and linear regression at 30°C . . . . . . . . . . . . . . 89
11.16 All the data for length and dots at 27°C . . . . . . . . . . . . . . . . . . 91
12.1 Simulated length vs dots at 27°C . . . . . . . . . . . . . . . . . . . . . . 95
12.2 Simulated length vs dots at 30°C . . . . . . . . . . . . . . . . . . . . . . 95
12.3 Histogram for the number of dots at 27°C . . . . . . . . . . . . . . . . . 96
12.4 Simulated and experimental numbers of dots . . . . . . . . . . . . . . . 97
A.1 Experimental growth rates . . . . . . . . . . . . . . . . . . . . . . . . . . 117
A.2 P-P plots of the growth rates at 27°C and 30°C . . . . . . . . . . . . . 118
A.3 P-P plots for three distributions of the cell length at 27°C . . . . . . . 121
List of Tables

10.1 Symbols in the model and their meaning ........................................... 67
10.2 Parameters for the distribution of the growth rates .......................... 71
11.1 Size of the samples ............................................................................. 77
12.1 Values of the constants of the model for 27°C and 30°C ................. 94
A.1 Edge values for the histogram bins and their normalised values ........ 117
A.2 Grouped counts .................................................................................... 118
A.3 Example of the table used for the calculations for the P-P plot .......... 120
Preface

Motivation

Once people know you are writing a project, they will immediately want to know what it's about. Equally immediate is the surprised question "How come is an Engineering student doing a project on bugs?".

Firstly, and at a personal level, the Engineering student in question has always had a soft spot for the things of life, especially those that are very small. Ever since I could read, and especially since I started studying scientific subjects, sentences such as "To the present moment, mechanisms governing [...] are not yet fully understood" and "Although much is known about [...], several essential aspects of it remain to be elucidated" jumped out of the page to me. I found that they are ubiquitous in any document that explores a subject to any respectable depth, and this has always fascinated me.

Secondly, although the main object of the present study are two cellular oscillators, the approach that has been made to learning more about their coupling is solidly based on the knowledge and familiarity with various procedures and concepts, all of which I have acquired during my degree in Engineering. This has helped me prove a suspicion that has formed slowly over the past four and a half years, which is that all we have been given is a bunch of invaluable tools, and now we are to choose what we are to do with them.

I am positive, therefore, that the question in the first paragraph has a sensible answer: in all the life sciences there is a growing need for abstraction and formalisation. We, as humans, like to get to grips with what is in front of us, and we like to be able to model things, to predict them... in short, to understand them. The relevance of mathematics and technology in this situation is obvious, which is why I believe an engineer can have far to go in the study of all things "bio-".
Practical remarks

A few important notes that are useful for reading the whole report. Aspects of the introduction that are key in understanding what has been done in this work, and why, are written in paragraphs with an exclamation mark at the beginning, like this one here. Also, text written in this type has been used to highlight the names of MATLAB functions or programmes.

It is of extreme importance to understand that this work is not experimental. It is a theoretical study of data obtained elsewhere. I am grateful to Lorena —referred to as Dr. Espinar in the rest of the report— for the data from her thesis, and also for good advice and useful remarks even before the subject of this work was chosen.

Acknowledgements and dedication

First and foremost, my deepest gratitude to my supervisors, Jordi Garcia Ojalvo and Josep Lluís Font. They both placed their trust in me when they agreed to tutor my project, and this alone should be a reason to strive for excellence.

Nevertheless, there is more. The contagious energy and passion of prof. Ojalvo have been an essential driving force and the backbone of this project. Ever since those late-evening third-semester Physics III lessons, his unbeatable enthusiasm and his love for what he does, as well as his caring personality, have been an example for me to follow. I have obtained much more from doing this project with him than just delving into a couple of issues about bacterial division.

Prof. Font “took me in” when I began my studies a month later than is usual, and has been my tutor ever since. The door of his office was always held open for me, and there I always found a listening ear and the best advice on whatever decision I had to make. It wasn’t long before he became, rather than a Physics professor, a teacher of many good and useful things, the greatest among which is that Life is non-linear.

My parents and my siblings witnessed the first stages of the project. They also happen to have been there from the very beginning, not only of the project but of the whole degree, and all the way before this. My parents have always been my reference in all aspects of life, and providers of constant love, support, sound advice, and frozen meals. Thank you.

My husband agreed to marry a crazy student in the last months of her project and I’m thankful that he did. His love, patience, understanding and support are well beyond anything that can be reasonably expected, and without him this project would be very far from being what it is. (And so would the rest of my life.)
And, finally, here goes to my "bestest friends" during these past years at university: Marc Pau, Albert Martí, Marc Freixas and Eric Martínez, promising engineers and, above all, fantastic people. It would have been impossible for me to get to this point without them, and for sure the journey wouldn’t have been half as fulfilling and fun. All the best for this new stage in our lives!

I would like to dedicate this work to my sons, Pau and Lluc. I have recently acquired the shared responsibility of instilling in them the passion for the unknown and the thirst for knowledge (gifts that I myself received from those who’ve come before me). They are already lovely people and I hope they will grow up to push back the limits of what is known.
Part I

PRELIMINARY
1. Object

The object of this project is to study the coupling between two cellular oscillators (the cell cycle and the chromosome replication cycle) \textit{in vivo} in the bacteria \textit{Escherichia coli}, and to establish a simple model of this phenomenon using experimental data.
2. Scope

In this project, the following will be done:

• The biological fundamentals and the methodologies used in the project will be explained in detail.

• The study will be conducted from the point of view of individual cells.

• *In vivo* data from existing films of *Escherichia coli* bacteria at two different temperatures will be processed.

• A statistical analysis of the biological data will be performed. (Emphasis will be laid on the oscillations of cell length and of the number of initiation sites at every moment.)

• Programmes will be developed using the MATLAB environment in order to make the collection and analysis of data possible.

• The data from the two temperatures will be compared and conclusions will be drawn.

• A theoretical model will be developed to simulate the behaviour of the main characteristics of the bacteria and the number of replication initiation sites, using the data obtained from the experimental study.

The following will not be done:

• Experiments will not be carried out expressly for this project.

• A programme to collect the data automatically will not be written.

• A complex model describing the evolution in time of all the biological aspects of cell division will not be implemented.
3. Justification

*Escherichia coli* is one of the most well-known bacterial species. Its processes of growth and replication have been under extensive study over the past decades, due to its role as a model organism\(^1\). Also, *Escherichia coli* are bacteria with a high rate of growth; cultures of it can be grown with ease and with a low cost, since they proliferate on inexpensive substrates; they can grow in the presence of oxygen or without it; they are highly competent as host organisms; and their complete genome is known since 1997 [1]. These qualities make it ideal for large-scale industrial processes.

The perspectives and uses of *E. coli* range from synthesising biofuel [2] to solving mathematics problems such as the Hamiltonian Path Problem [3]. Nevertheless, it is in the biotechnology industry that this bacterium has found its most important role.

Since the birth of biotechnology in the early 70s [4, 5], *E. coli* has played a major role in the biopharmaceutical industry. The production of recombinant proteins is among the most attractive and most massively wide-spread uses [6, 7]. Recombinant proteins are proteins that are derived from recombinant DNA, which is DNA that includes genetic material from multiple sources and that is obtained in the laboratory. The main attractive of the production of recombinant protein production processes is that higher-order eukaryotic proteins are expressed in prokaryotic organisms, since recombinant DNA techniques enable organisms to express proteins that they could not code for in their wild-type state.

One of the first and most notable recombinant proteins is human insulin, which, contrarily to the insulin hitherto obtained by other methods, presents no incompatibility problems for use in human patients [8]. The human growth hormone [9, 10] is also among the most important products obtained this way.

Industrial scale production of these very necessary hormones by bacteria that have been genetically modified seems an attractive solution, if compared to extracting these substances from pigs and cadavers. (Before recombinant protein production, insulin was obtained from livestock, mostly pigs, and the human growth hormone was extracted from the pituitary gland of cadavers.) Many more therapeutic

\(^1\)A model organism is a non-human species that is easy to grow and handle, and whose study yields important information on particular biological phenomena, also providing insight into how other organisms work.
substances are obtained from recombinant protein processes, and almost 30% of the approved recombinant therapeutics are produced in *E. coli* [11].

*Escherichia coli* is not only useful for producing pharmaceuticals. Many other industrial processes can rely on *Escherichia coli* as an industrial platform for bio-based chemical production [12]. (This includes processes such as fermentation, production of bioethanol and advanced biofuels, organic acids, amino acids, sugar alcohols, biopolymers and biomonomers, enzymes...) This would reduce the strain on the environment, decreasing carbon dioxide emissions and several other pollutants in the production of bulk and fine chemicals.

As has been mentioned before, *E. coli* is chosen for these industrial processes because of its reliability, stability, and rapid growth and replication. In fact, current efforts are being devoted to obtaining bacteria with even higher growth rates [13].

Nevertheless, and the wide industrial use of the bacteria notwithstanding, the mechanisms by which *Escherichia coli* replicate their genetic information and divide, yielding two identical "daughter" cells, are not yet fully understood. Particularly, one of the mysteries that remains to be elucidated is when, in relation to its size, a cell "decides" to initiate replication of its chromosome.

The efforts to shed some light on this matter started in the late 1960s, when W. Donachie [14] and Cooper and Helmstetter [15] proposed the concept of a critical mass, the "initiation mass", which, once attained, would trigger duplication of the cell's DNA. These works laid the foundations of the investigation that is being carried out to this date.

However, these first publications, and many more afterwards, studied the bacteria at a population average level and/or *in vitro*. Population averages, on the one hand, do not allow for individual tracking of the cell lineages, so it is made difficult to analyse the change with time of the characteristics of individual cells. On the other hand, *in vitro* studies might be suitable for finding out any information that is a permanent characteristic of the cell (its genome, for instance), but certainly not for observation of live cells' behaviour.

Films offer a powerful way to address these shortcomings [16]. By genetically modifying bacteria, it is possible to obtain strains that express fluorescent proteins; these can be attached genetically to any protein of interest of their chromosome(s), or to its DNA promoter sequence, in such a way that the fluorescence protein is expressed either together with the protein of interest, or whenever that protein of interest is expressed from its natural gene. This, combined with automated time lapse microscopy, makes it possible to observe the cells' growth, replication, and division—in short, their evolution with time—and it also offers the technological possibility of tracking one cell's lineage individually from beginning to end.

*Lara Sofia Escuain i Poole*  
24  
*Barcelona, September 2013*
Despite the attractiveness of these techniques, wide-spread use of them remains to be implemented. Although many studies have been dedicated to investigating the coupling of chromosome replication initiation to cell size over the years, very few have made use of films in vivo. In 2012, Chen et al. performed a similar study on the bacterium Synechococcus elongatus, and Athale and Chaudhari wrap up their research on cell length variability and nucleoid numbers in E. coli by suggesting further investigation on individual cell lineage tracking [17].

Because non-operative basic investigation is still necessary to fully understand the processes that are used at an industrial level, this project aims to delve in the issue of cellular growth by attempting to clarify the coupling between cell size and initiation of DNA replication, by making the most of modern cell tracking and imaging techniques.
4. Basic specifications

The basic specifications of any engineering project ensure that the project will meet certain requirements of time, cost, and quality. This project is a theoretical study and, as such, no installation or device is manufactured, used, or designed.

The specification for the time in which this study had to be completed was seven months. The start date was the beginning of the semester, in February, and the deadline for the presentation was in mid-September.

Given the nature of this study, considerations of cost do not apply. With respect to quality, it was requested that a model be proposed to simulate the variables of interest, in addition to analysing the experimental data.
Part II

INTRODUCTION
5. Bacterial biology fundamentals

The veiled mysteries shrouding many aspects of bacterial life notwithstanding, Escherichia coli remains the most well known living organism to date. Details of the characteristics of this common bacterium are presented here, as well as a summary of what is known about its division and the replication of its genetic material.

5.1 Escherichia coli bacteria

Escherichia coli is a Gram-negative facultative anaerobic rod-shaped prokaryote bacterium. The following sections focus on describing this organism.

5.1.1 General properties

Figure 5.1 shows an electron microscopy image of Escherichia coli (from now on, E. coli). This bacterium is typically 2.0 µm long and 0.5 µm in diameter; consequently, the cell volume ranges mainly between 0.6 and 0.7 µm³. Its preferred growth temperature is 37°C, but some strains have been shown to thrive at temperatures as high as 49°C [18] and as low as 7.5°C [19].

!(An important note: constant cell density and cell width are assumed in this work. Therefore, throughout the project, the terms "cell size", "cell mass", and "cell width" will be used indistinctly when referring to proportionalities or critical thresholds.)

---

1 Gram staining is usually the first step used in classifying bacteria. It divides almost all organisms in two groups, Gram-negative and Gram-positive, according to the physical response of their membrane to chemical staining.

2 Facultative anaerobic bacteria feature aerobic respiration in the presence of oxygen, but can switch to fermentation in an atmosphere where oxygen is denied.

3 Prokaryote cells, as opposed to eukaryotes, are those which do not have a membrane-bound, defined nucleus (karyon) enclosing the genetic material.
Theodor Escherich (1857-1911) discovered *E. coli* in the faeces of healthy newborn babies in 1885 and gave the species the name of *Bacterium coli commune*, due to the fact that it is found in the colon and to its shape and motility\(^4\) [20]. The bacterium was later renamed to its current name, after its original discoverer.

Harmless strains of *E. coli* and other related organisms amount to 0.1% of the intestinal flora in human beings [21]. Other strains of *E. coli* are virulent pathogens, related to severe enteric and extra-enteric diseases, and can lead to death [22].

Because of its long history of culture and ease of manipulation, *E. coli* has become a model organism and one of the most widely used prokaryotic systems in modern industrial microbiology and biological engineering [23]. Indeed, due to its rapid growth rate and to the fact that the genome is fully sequenced [1], amongst other reasons, more is known about *E. coli* than any other living being.

### 5.1.2 Physiology

**Shape**

*Escherichia coli* is of the shape of a cylinder with hemispherical caps on each end. In his beautifully literary book dating back as far as 1917 [24], Thompson put forward the idea that bacterial form is subject to physical principles. Almost 70 years later, A. Koch suggested a physical resemblance of the cell wall to a cylindrical soap bubble [25]. A cylindrical soap bubble happens to break spontaneously in the centre when its length reaches twice its radius multiplied by \(\pi\), much like cell division. In the case of cell division, if there was only hydrostatic pressure to account for, the cell would remain in a compact, spheric shape; elongation would not occur and division would be impossible. In *E. coli*, a surface tension-like force can be considered responsible for allowing cell elongation and, therefore, division [26].

---

\(^4\)The term motility applies to unicellular and simple multicellular organisms. It denotes the power of moving actively and spontaneously, consuming energy in the process.
Cell membrane

The bacterial cell is enclosed in a murein (peptidoglycan) sacculus, which is the exoskeleton bacteria need to withstand the internal cytoplasmic turgor pressure [27]. The sacculus of *E. coli* is one giant\(^5\) peptidoglycan macromolecule with a molecular mass of more than 3\(\times\)10\(^9\) Da. (This is of the same order as the mass of its chromosome, 2.32\(\times\)10\(^9\) Da.) This peptidoglycan structure is relatively thin in *E. coli* and the other organisms in the Gram-negative group (see page 31). Several models have been proposed for the structure of the cell membrane [28, 29, 30], but for the scope of this project it is sufficient to have in mind the idea depicted in figure 5.3.

![Figure 5.3 Schematic illustration of a section of the cell membrane structure](image)

The pink structures in the drawings represent phospholipids, which consist of a hydrophilic head (circle) and a hydrophobic tail (wiggly line), and form a lipid bilayer both on the outer membrane layer and the inner cell membrane. Two of the main functions of the lipid bilayers are maintaining an ion gradient between the inner and the outer cell media, and regulating the passage of proteins and other components across the cell membrane, as necessary. Finally, the cell membrane as a whole performs the obvious, yet fundamental, task of defining the boundaries and the shape of the cell.

\(^{5}\)As heavy as about 5 fg, or 5\(\times\)10\(^{-15}\) grammes.
Cytoplasm and cytosol

The term "cytoplasm" refers, in prokaryotes, to those parts of the cell that are enclosed within the cell membrane. Its main component is a fine, granular, colloidal emulsion, the cytosol, which is the major environment in all bacterial cells.

Although the composition of ions, proteins, metabolites, and several significant metabolic processes in the cytosol are well understood, there is a general lack of understanding as to the structure and the dynamical nature of the cytosol solution [31]. This is because the concentration of these elements is so high that the cytoplasm differs significantly from a dilute aqueous solution, which is frequently used for \textit{in vitro} biochemistry.

In a unique and beautiful paper, D. S. Goodsell presents three drawings that illustrate the interior of a living cell [32]. Page 35 shows the fourth page of the paper, which displays general and detailed views of the cytoplasm\textsuperscript{6}.

Several complicated models, each one better than the previous, intend to accurately describe the nature of the cytosol [33, 31, 34]. For the scope of this project, the ideas already presented in this paragraph will suffice.

Nucleoid

Bacteria, as a prokaryote, does not possess a separate nucleus, where the DNA is enclosed in a membrane and secluded from the rest of the cell. Instead, the genetic material freely floats in the cytoplasm in the shape of a highly compacted circular chromosome\textsuperscript{7} [38], which is also precisely oriented spatially within the cell [39]. (It is interesting to note that the length of the chromosome, around 1.6 mm, greatly exceeds that of the cell, which is approximately 2 µm.) The condensed chromatin —i.e. the genetic material—, nevertheless, occupies a functional pseudo-compartment (hence the name \textit{nucleoid}), which forms a distinct, separate region that is characterised by the absence of ribosomes [40].

Ribosomes

Ribosomes are molecular machines that synthesise proteins in the cell. They consist of three RNAs (or ribonucleic acids), in prokaryotes, and more than 50 proteins that assemble to form two large subunits of different size [41].

\textsuperscript{6}In order to obtain the necessary information for these drawings, the author determines information from the fine, sub-cellular structure seen with electron microscopy, on the one hand, and on the other hand he learns from X-ray crystallography and classical biochemistry. He then finds a mid-way point that describes the structure of the cytoplasm.

\textsuperscript{7}Some interesting species with multiple circular or linear chromosomes are described in references [35], [36], and [37].
Figure 1. If we could magnify a cell one million times, making molecules the size of everyday objects, what would we see? Three portions of a typical E. coli cell are magnified one million times. A schematic of the cell at 50,000 times magnification shows the location and size of each 100 nm window with respect to the whole cell and the key identifies the macromolecular components. Although only three examples are shown in the key, proteins come in many shapes and sizes.
(a) The cytoplasm, showing all macromolecular components.
(b) Close-up of one portion of the cytoplasm, showing all molecules, including water (circles), small molecules (dark outlines) and a small portion of a protein.
(c) The cell wall, showing all macromolecular components.
(d) The nuclear region, showing all macromolecular components.

Figure 5.4 Drawings of the inside of the cell [32]
A simple drawing of a ribosome and its function is presented in figure 5.5. The ribosome acts as a host to the process of protein transcription. The messenger RNA (mRNA) bears a negative duplicate of the DNA that codes for certain proteins. The transport DNA (tDNA), which holds an aminoacid, places itself on the mRNA where appropriate, according to the base sequences. Then, the peptide bonds are formed, which join together the aminoacids (the red balls in figure 5.5) to form the correct protein [42]

Reference [43] presents one of the first papers on ribosomes. Although by now outdated by far, it features a valuable collection of beautiful plates that illustrate the presence of ribosomes. (The cells that are represented are eukaryotes, but analogies may be drawn for prokaryotes where suitable.)

Plasmids

Plasmids are heritable fragments of extrachromosomal DNA that are present in the cytoplasm of the cell. They are double-stranded and they replicate independently from the chromosome. Plasmids are mostly present in prokaryotes and they give bacteria genetic variability, which can result in resistance to antibiotics or to adverse environmental conditions.

Flagellum

Some strains of E. coli possess one or more flagella and can therefore swim. Flagella are thin helical filaments which are each driven at the base by a tiny, reversible rotary motor, which is powered by an ion flux. (Although much is known about structure, components, genetics, and assembly of this motor, so far there is no general understanding as to how it actually works [44].)

Other organelles

Contrarily to what has been hitherto generally accepted by authorities such as Campbell’s Biology [45] and Tortora et al. [46], prokaryotes do have organelles

---

8 Organelles are any membrane-enclosed compartment within the cell with a dedicated biochemical function [47].
Size and replication initiation in *E. coli* [48, 47]. These organelles perform various functions related to cell structure, metabolism, motility, and reproduction.

### 5.2 Bacterial reproduction

Cell reproduction is arguably the main event in the vegetative life cycle of any bacterial cell [49]. It is the process by which a cell passes on to its offspring its genetic information, equitably and without variations (with the exception of infrequent mutations). By cell division, a bacterial species has its continuity ensured, as the daughter cells possess an accurate and complete copy of their mother’s DNA.

#### 5.2.1 Overview of the cell cycle

The cell cycle comprises all the events that occur in a cell that lead to division. In other words, it is all that happens between one division of the cell, when it is "born", and the next, when it divides in turn.

During the course of a cycle, a bacterial cell doubles its mass, replicate their DNA and separates the two newly formed chromosomes. It then septates between the chromosomes, in the middle of the cell, constricts, and finally divides into two identical newborn cells.

**Figure 5.6** Schematic drawing of the cell cycle. The yellow dot corresponds to the origin of replication, and the blue dots represent the ribosomes.

The following sections detail the two most relevant aspects of cell division. First, the duplication of the chromosome and second, the division of the cytoplasm, or cytokinesis.

---

9 Or forms a septum, a biological membrane that separates the two new cells.
5.2.2 Chromosomal replication

Replication initiation

Chromosomal replication in bacteria always begins at a certain location of the chromosome, denominated "origin of replication", known as oriC in literature (and, from now on, in this work). The opposite locus of the chromosome, where replication terminates, is called terminus, or ter.

Replication initiation is mediated by a single initiator protein, DnaA, which is segregated constitutively (that is to say, proportionally to the volume of the cell) and which binds selectively to certain DNA sequences in oriC, or boxes [50]. Indeed, it can be said that replication initiation depends on the concentration and availability of DnaA and oriC, respectively [51].

In *E. coli*, oriC has a length of approximately 250 base pairs\textsuperscript{10}, or bp. In oriC there are multiple 9-bp sequences, termed DnaA boxes, that specifically interact with DnaA, the initiator protein [50]. DnaA induces the local separation of the two strands of DNA. DnaC mediates the further separation of the helices to obtain single strands, so that the DnaB helicases\textsuperscript{11} can be loaded and the replisome\textsuperscript{12} machinery assembled [52].

Once the replisome is completely loaded, replication begins. It starts at oriC and progresses along each strand bidirectionally until the replication forks meet at the terminus ter [53]. In figure 5.7 the whole process is sketched. Once the chromosomes are complete, the oriC regions move away towards opposite ends of the cell and the ter regions leave the replication site. This process is known

\textsuperscript{10}Base pairs are the building blocks of DNA. They are nitrogenous compounds that bond to create a "step" in the DNA helix ladder.

\textsuperscript{11}The enzyme which unwinds and separates the DNA strands.

\textsuperscript{12}The complex molecular machine that carries out DNA replication.
as segregation\textsuperscript{13} and marks the end of chromosomal replication, giving way to the division of the cytoplasm, or cytokinesis.

**Regulation of replication**

Necessarily, replication of the chromosome must occur once and only once per cell cycle\textsuperscript{59}. (Notice that this does not imply that a cell will not harbour several rounds of replication at once\textsuperscript{60, 61}, which is confirmed in the present work.) In order to achieve this, three regulatory mechanisms inhibit the initiation of new rounds of replication at the same origin\textsuperscript{59, 51, 62}.

**Inhibition of DnaA** Inactivation of the DnaA protein occurs by a mechanism called RIDA, or Regulatory Inactivation of DnaA. RIDA is a replication-coupled negative feedback mechanism. Prior to replication initiation, the concentration of ATP-bound DnaA (the active form of DnaA) is high within the cell. Just after replication, however, the Hda protein hydrolyses DnaA-ATP into inactive DnaA-ADP\textsuperscript{63}. Hda works in cooperation with the $\beta$-clamp subunit of DNA polymerase III, which is the primary holoenzyme\textsuperscript{14} involved in replication. This implies that DnaA is inactivated for re-initiation immediately after the loading of the clamp\textsuperscript{64}.

Research shows that this is the primary mechanism for regulation of initiation\textsuperscript{64}. It is also the mechanism that has been studied and modelled in this work. In section 10.1 of page 67, the equations are put forth that describe the behaviour of RIDA, and reference is made to the present paragraphs.

**Titration of free DnaA** oriC is not the only region in the chromosome with a high affinity to DnaA. There are many other regions which can bind the protein, especially the datA locus. datA has the capacity of attracting eight times as many molecules of DnaA as oriC\textsuperscript{62}. Since datA is fairly close to the oriC region, it is reasonable to suppose that datA can act as a sink for excess DnaA and reduce its concentration in the cell. This mechanism is known as titration.

**Sequestration of the oriC** The third regulatory mechanism focuses on inhibiting the oriC rather than the DnaA protein. There are 11 GATC\textsuperscript{15} sequences throughout oriC that are fully methylated in normal conditions. Nevertheless, newly synthesised GATC sequences present a transient hemimethylated state. The SeqA

---

\textsuperscript{13}Although a number of reviews suggest various drivers for this [54, 55, 56, 57], the most satisfying answer that has been found up to date is that segregation is due simply to entropy [58].

\textsuperscript{14}A complete enzyme made up of subunits. It retains some function even without some of its subunits.

\textsuperscript{15}A, C, G, and T are the four nitrogenous bases that, in pairs, form the DNA helix.
protein has a high affinity to these hemimethylated sequences and binds to them. This, in combination with the fact that there is a high density of these sites at oriC, results in sequestration of newly replicated oriC regions and prevents re-initiation.

In the images that have been analysed in this study, a yellow fluorescent protein (from now on, YFP) was attached to SeqA. Therefore, fluorescence microscopy images showed each oriC region as a shiny dot and enabled the initiation count that is the basis of part of the study (see figures 7.1 and 7.2 on pages 47 and 48).

5.2.3 Cytokinesis

Cytokinesis occurs after complete chromosomal replication and is therefore not a point of particular interest for this project. Nevertheless, the fundamentals of the process will be explained.

After chromosomal segregation, septation and constriction occur to divide the elongated "mother" cell, with two chromosomes, into two "daughter" cells with one chromosome each. This requires assembly of a complex molecular machine, and the key to this machine is the FtsZ protein [65]. This protein assembles into a ring-like structure, the Z ring, which is a cytoskeleton element. The Z ring then mediates the recruitment of those proteins necessary for the divisome, the molecular machine that performs bacterial division. After this comes the synthesis of a septal wall and constriction of the "mother" cell at the replication site, followed by posterior hydrolysis of the septum in order to enable separation.

William Margolin [66] provides excellent and detailed explanations on cytokinesis and its regulation. Veiga and Pinho [53], with a noticeable twelve years’ difference, also give a comprehensive explanation.
6. Mathematical fundamentals

6.1 Statistical tools and methods

To analyse the data, many different statistical resources were used. A brief overview of them is presented here. Only a theoretical description is given: for details on how they were used, see section A.1 (appendix A on page 115); the development of necessary calculations can also be seen in this appendix.

The necessary information to write this section, as well as for the calculations in the appendices and all other statistical analyses of the data, have been extracted from class notes of the course Statistics Methods in Engineering, imparted by Prof. Pepió from the 2nd of February to the 31st of May, 2010.

6.1.1 Theoretical distributions

A statistical distribution is the description of the quantity of times a possible outcome will be obtained in a given number of trials. In other words, it gives the probability of getting a certain result from the total number of possibilities.

Statistical distributions come in all shapes and sizes, but some are more common than others in different situations. Some are discrete, and other continuous. Below is the description of two that are among the most common in Biology, and that have been used in this project: both are continuous, because they were used to describe variables that can take any value.

Normal distribution

The normal distribution, also called Gaussian or bell curve, is the most common in all kinds of phenomena. It is a symmetrical and asymptotic law. It is centred at the mean and its width is decided by the variance of the data $X$, and the area beneath the curve is always unitary.

This law is used to describe phenomena whose randomness is due to multiple and independent causes, of very small and additive effects on the variable $X$. Therefore, it is ubiquitous in biological studies.

Although there are an infinite number of bell curves (as many as different means and variances exist), a certain curve is often used in analysis. This is the standard normal curve, which has a mean of 0 and a variance of 1. The probabilities for
this standard distribution have been tabulated, which makes calculations easier. The values of the standard distribution are not called \( X \) any more, but they have a special name, \( Z \), or Z score. It is possible to go from a variable \( X \) to its standard homologue in the following way:

\[
Z = \frac{X - m}{\sigma}; \quad Z \sim N(0; 1),
\]

(6.1)

where \( m \) is the mean of the original variable and \( \sigma \) its variance. The right part of the expression indicates that the standard variable \( Z \) follows a normal distribution with mean \( m = 0 \) and variance \( \sigma = 1 \).

**Log-normal distribution**

The log-normal distribution is not centred, but skewed. Its main attractive is that it is the natural logarithm of a variable that follows a normal distribution. This makes it possible, once standardised, to calculate probabilities, just like for a normal distribution.

Log-normal distributions are common in cases of low mean values, large variances, and where variables must always be positive. As with the normal law, in this distribution the variables are influenced by many independent factors, but with a multiplicative effect (and not an additive effect, as with the normal law). Also, they are very common of quantities that grow exponentially.

**6.1.2 Graphs**

Graphs are a very convenient way of synthesising information from experimental data. Inspecting well-made graphs, many conclusions can be drawn from the study of the data.

**Histograms**

Histograms are a discrete way of representing data. The values are each assigned an interval within which they fall, from a conveniently chosen number of equidistant bins or intervals. The fraction of occupancy of each one of the bins (called “frequency” in this thesis), and on one occasion the count, is represented in XY plots.

They have been used, in this study, to analyse probability density distributions of continuous variables for which, nevertheless, only discrete measurements were available. Also, they have been used to compare quantities of a given variable at different conditions.

*Lara Sofia Escuain i Poole*  
Barcelona, September 2013
Scatter plots

Scatter plots represent one set of data vs another of the same size and place a visible mark in the point they determine. They are mostly used to evaluate dependence of one variable with another or to see the evolution with time of a single variable.

Trending plots

These graphs give an idea of the mean and standard deviation of a set of data. Previous calculations of these two parameters are first done, and the error bars are placed in the graph with the centre of the bar indicating the mean and the width of the bar indicating the variance. They have been used to examine the variation in time of means and standard deviations of significant variables (see section 11.1 on page 77).

P-P plots

P-P plots are an important tool for data fitting. They are a quick and simple, albeit a little subjective, way of evaluating the goodness of fit of experimental data to a given theoretical distribution. The key to this is the transformation of the distribution function so that it will present a linear dependence of the variable; once represented, the alignment of the experimental points can be evaluated subjectively by comparing the alignment of the experimental points on the theoretical straight line.

To make this graph, a table must be first filled in with the necessary values for representation. The contents of this table are the following:

- The data $x_i$ are listed in a first column, for reference.
- $x_i$ is sorted in ascending order.
- A third column $i$ is filled in with numbers from 1 to the size of the sample $n$, from the first until the last row.
- $\hat{F}_i$ is calculated in the fourth row. $\hat{F}_i$ gives the expected values of the order statistics\(^1\).
- The fifth column contains the values of the distribution for which it gives a probability of $\hat{F}_i$. This last column (and only this last one) is different according to the theoretical distribution.

---

\(^1\)According to the modified Kaplan-Meyer estimator, $\hat{F}_i = \frac{i - 0.5}{n}$
For the exponential distribution, this fifth column contains the results calculating $-\ln(1 - \hat{F}_i)$. This is because of the transformation that has already been mentioned, which is necessary to make the function linearly dependent on the variable:

$$F(x) = 1 - e^{-\lambda x} \rightarrow -\ln(1 - F(x)) = \lambda x$$  \hspace{1cm} (6.2)

In the graph, the fifth column, $-\ln(1 - \hat{F}_i)$, is represented vs the sorted experimental data in the second column.

For the normal distribution, the fifth column simply contains the Z score that has the probability given by $\hat{F}_i$. The Z values are represented vs the sorted data in the second column.

For the log-normal distribution the fifth column contains the Z score also, as for the normal distribution. The difference is that, for this distribution, the Z score is represented vs the logarithm of the sorted data.

An example of these plots can be seen in figure 11.6 (page 81). The plot shown is a log-normal P-P plot showing good fit of the data.

6.2 Numerical integration via the fourth-order Runge-Kutta method

This problem is an initial value problem, that is to say, all the functions have an initial value and it is desired to find the value of the functions at some final point, or at some list of discrete points. The basic idea of all such methods is to add to the functions small increments corresponding to derivatives (the right-hand side of the equations) multiplied by stepsizes. The integrating method that has been used in this work is the fourth-order Runge-Kutta. The Runge-Kutta family comprises several variations: of them all, the fourth-order Runge-Kutta, also known as “classic” or RK4 method, has been applied here. This is an explicit iterative method that is among those most widely used.

The fundamental of the RK4 method is that the approximation for the value of the function at the next integration step $n+1$ is the value of the function at the present step $n$, plus the weighed average of four increments. These increments are the product of the step size times an estimation of the slope, specified by the function on the right-hand side of the equation. Equations (6.3) to (6.6) give the expressions for the increments, while equation (6.7) shows the final estimation for the function at the next step. There are many equivalent but different ways of writing these expressions; the following equations have been taken from reference [90].
$k_1 = hf(x_n, y_n)$ \hspace{1cm} (6.3)

$k_2 = hf(x_n + \frac{h}{2}, y_n + \frac{k_1}{2})$ \hspace{1cm} (6.4)

$k_3 = hf(x_n + \frac{h}{2}, y_n + \frac{k_2}{2})$ \hspace{1cm} (6.5)

$k_4 = hf(x_n + h, y_n + k_3)$ \hspace{1cm} (6.6)

$y_{n+1} = y_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} + O(h^5)$ \hspace{1cm} (6.7)

The following is an explanation for the terms in these equations.

- $k_1$ is the increment based on the slope at the beginning of the interval.
- $k_2$ uses $k_1$ to calculate the increment based on the slope at mid-point.
- $k_3$ uses $k_2$ to calculate, again, the increment based on the slope at mid-point.
- $k_4$ uses $k_3$ to calculate the increment based on the slope at the end of the interval.

This is the numerical algorithm that will be used in this project to integrate the theoretical model of chromosomal replication initiation. Details of the implementation can be found in chapter 10 (page 67), and its results in chapter 12 (page 93).
7. Data gathering

The necessary data for the analysis in this project was kindly provided by Dr. Espinar from her thesis work in 2012 [67]. Section 7.1 in this chapter describes the form of the data, while sections 7.2 and 7.3 summarise the processes for growing the bacterial strains, inoculating the desired genes in them and preparing the films. This information is extracted from chapter 2 of Dr. Espinar's thesis (reference [67]).

7.1 Description of the data

The data came in the form of a great quantity of microscopy pictures of several cell colonies for two different temperatures (30°C and 27°C). These pictures are, in actual fact, frames of the films that were produced of living bacteria. From these images, the length of the cell and the number of DNA replication forks were obtained.

An important note: it is believed that the SeqA protein, because of its predisposition towards hemimethylated GATC sequences throughout the entire genome [68], follows closely after the replication forks as they slide down the chromosome [69]. Newly replicated DNA is soon fully methylated to prevent an untimely reinitiation of replication [70]. Therefore, it is proposed that the expression of the fluorescent protein reveals a replication fork at every shiny dot in the image.

The quantity that has the most biological entity is the number of chromosomes, not of replication forks. Nevertheless, since the chromosomes are not fully replicated for most of the cell cycle, the number of dots will be referred to as the number of initiation sites throughout the whole project. (The number of replication forks and of initiation sites remains unchanged during chromosome replication.)
Figure 7.1 shows one of the frames that were used for the analysis: the top frame is a phase image and the bottom frame is a fluorescence image, both of the same frame, or instant in time. Phase images show the opaque bacteria; as will be explained later on, phase images were used to determine the length of the cell. Fluorescence images show the expression of the fluorescent protein attached to the chromosome.

The underlying principle in the obtention of phase images is the contrast between cell and medium. The fluorescence images, however, rely on the fusion of a protein of interest (in this case, SeqA) to the coding sequence of a fluorescent protein (YFP in this study).

Figure 7.2 shows one of the bacterial colonies at every fifteen minutes. In these frames, phase and fluorescence images are combined.

Figure 7.2 Frames of one of the films. A photograph of the colony is shown at every fifteen minutes

7.2 The bacterial culture

7.2.1 Origin of the strains and their growth conditions

For the making of the films, three strains were used which share the genetic background of the K-12 MG1655 strain. The bacteria were grown on an overnight culture at temperatures of 27°C and 30°C and fed on LB.

\(^{1}\)LB stands for *lysogeny broth* [71] and is one of the most frequent culture media. It is used when the experiment requires no-stress conditions. It is rich in nutrients and it can be prepared by reconstituting a commercially available powder.
### 7.2.2 Genetic engineering for the expression of a fluorescent protein

*Escherichia coli* cells do not naturally possess a fluorescent protein that marks the presence of replication fork in its expression. In order for this to happen, several processes of genetic engineering were undertaken, which are explained in what follows.

First of all, the desired genetic sequence was amplified with the **PCR process**\(^2\) in order to obtain a number of copies of it. This DNA will be cut open at specific sites with **enzymatic restriction**, so that it will have free ends compatible with the free ends of a cloning vector (such as a plasmid). After this comes purification by **electrophoresis** in order to remove unwanted DNA fragments. The last step is **ligation and transformation**, so that in the end, genetically modified and competent bacterial strains are obtained. (For a broader explanation of the processes in bold type and how they were carried out, see reference [67]).

### 7.3 Making the videos

In order to prepare the culture for observation under the microscope, an overnight culture was dissolved at a ratio of 1:100 in LB with a 0.7% arabinose supplement as an inducer. This solution was incubated for 3 hours at a temperature of 37°C with antibiotics for bacterial selection. During this time, the agarose sandwich seen in figure 7.3 was prepared.

\(^2\)PCR stands for Polymerase Chain Reaction. It consists of separating the two DNA chains and binding a primer, or initiator, for the elongation and completion of the chain. When this process is repeated cyclically, the amount of desired DNA fragments rises exponentially.

---

*Barcelona, September 2013*

*49*  
*Lara Sofia Escuain i Poole*
pad. Then, the agarose pads were allowed to dry for about an hour. After this, they were placed on a WillCo plate and sealed; the WillCo plate was then allowed to rest for an hour, in order to become vertically balanced on the microscope.

The videos were made with the aid of temporised fluorescence microscopy. A photograph of the colony was taken every five minutes, and with these photographs the videos were made. This makes *in vivo* observation of a culture possible at a single cell level.

Dr. Espinar used a Nikon TE2000 microscope and a Prior slide with motorised XYZ movement. The photographs were taken with a Hamamatsu ORCA-ER camera. The whole system was automated by the Nis-Elements AR software. The temperature was maintained constant for the culture by a controllable incubator.
8. State of the art

The efforts devoted to understanding the various aspects of cell growth and division are countless. This chapter focuses on past works and investigation on the coupling of replication initiation to cell size in *E. coli* over the past 40 years.

8.1 Experimental evidence

In 1968, both William D. Donachie [14] and Helmstetter and Cooper [72] put forth that the mass at which a round of initiation takes place is always an integral multiple of a particular size. Ten years later, Koppes et al. [73] studied, always at a population level, the cells' length at diverse cycle events, such as initiation and termination of chromosome replication, constriction, and separation; they also studied the relationship between length and age, the rate of DNA replication... Of particular interest is the section on cell size at DNA replication.

Donachie, in his excellent review of the cell cycle in *E. coli* [60], suggests that replication is initiated once the cell has grown to a critical size, which is twice the unit volume. He also poses three questions and suggests answers for them.

The first question is how initiation takes place at every copy of oriC at each successive doubling of a fixed cell mass. This could be explained by the existence of an initiator substance that would be produced proportionally to cell size, that initiation takes place once a critical initiator concentration is reached, and that this initiator is destroyed after initiation. It is now known that this substance is DnaA, although it is not destroyed after initiation, but inactivated (see "Inhibition of DnaA" on page 39).

The second question posed by W. Donachie is how each copy of oriC is replicated only once, during a period of initiation. Hemimethylation (see "Sequestration of the oriC" on page 39) is the explanation for this refractory period.

Thirdly, if initiation is blocked, initiation potential accumulates; when replication is allowed to resume, initiations take place successively, until the normal proportion of oriC to cell mass has been restored. It is now known that the switching between the two forms of DnaA is responsible for this.

Some of the aspects discussed in this review of 1993 were not much more than reasonable suppositions. Nevertheless, subsequent studies later showed them to
be true (although much remains to be explained) and they are fundamental to the model proposed in this work.

Ten years later, Donachie and G. W. Blakely [74] reaffirm the constitutive segregation of DnaA; its conversion to DnaA-ATP; the conversion of DnaA-ATP into DnaA-ADP after replication during an eclipse period (hemimethylation); and binding of DnaA to other DnaA boxes throughout the chromosome, especially at the datA locus (see "Titration of free DnaA" on page 39).

The model proposed in this work relies heavily on the conclusions extracted from this paper. Also, numerical values for the ratios of DnaA-ATP to DnaA-ADP both after division and at replication initiation have been obtained from this paper.

In a very controversial paper, Bates and Kleckner [75] proposed in 2005 that initiation is not coupled to cell mass; instead, cell division in itself would "license" the beginning of a new replication round. In 2008, Haeusser and Levin [76] wrote a paper on the coordination of cell cycle events during growth and division. They too questioned the need for a specific cell mass to trigger replication. Although this idea has been challenged, it is enough to be able to assert that the timing of DNA replication initiation is more complex than initially proposed by Donachie in 1968 [14]. Additionally, Haeusser and Levin suggest the necessity of a size homoeostasis regulator; the first report of this regulator came in 2007 [77].

Chien et al. [78] confirm in 2012 the suitability of DnaA as a candidate for the initiation regulating substance that Donachie mentions —initially suggested by Katayama et al. in 2010 [59] and Løbner-Olesen et al. in 1989 [79], among others. Also, because abnormalities in initiation timing are compensated by changes in the timing and duration of successive cycle events, they argue that there must be an independent homoeostatic mechanism responsible for cell size that can even correct changes in initiation mass.

Following a series of studies on how perturbations in DNA affect the cell cycle, Hill et al. [80] studied, for the first time, the exact reverse of this: how perturbation of cell size affects replication. An interesting finding is that replication initiation not only depends on the concentration of the initiator DnaA, but rather on the total amount of this molecule. This is taken into account in the model developed in this work (see table 10.1).

8.2 Theoretical studies

More recently, due to the rise in computational power, theoretical biology has also set an eye on the issue of coupling replication initiation to cell size.
In 1995, Keasling et al. [81] used a Monte Carlo method to determine whether the cell had enough "potential" (be it cell mass, volume, cell wall or number of initiation proteins) to initiate replication. Based on previous works [15, 14], at each step they compare the cell mass divided by the number of chromosome origins to the critical mass, $Mc$:

\[
\frac{\text{Cell mass}}{\text{Number of origins}} = Mc \tag{8.1}
\]

If the equation proves true, then one or more origins are allowed to initiate replication. One of the interesting ideas of this paper is the study of the synchronisation between multiple replication forks. The methods used for this simulation, nevertheless, are not comparable to the set of differential equations used for ours.

A. Zaritsky et al. [82], much more recently, introduce a programme in a very instructive paper, which was developed on the fundamentals of the Cooper-Helmstetter model [83], also known as "The Central Dogma of the Bacterial Cell Division Cycle" (BCD):

\[
M_{\text{avg}} = \ln 2 M_i 2^{C+D \over T_d} \tag{8.2}
\]

In this equation, $M_i$ is the threshold value of the mass over which replication is initiated. $C$ is the time of replication and $D$ is the time lapse between replication and cell division. ($C$ and $D$ are constant.) $T_d$ is the doubling time, the time needed to double the cell mass.

In their simple yet rigorous programme, CCSim, Zaritsky and co-workers describe bacterial cells under various conditions with only four parameters: replication, division and inter-division times, and cell mass at replication initiation. They discuss various cases, four of which are presets in the programme, and compare the programme to reality while studying the coupling between cell size and replication initiation, between replication and division, and also a special emphasis on the eclipse\(^1\).

Nevertheless, as can be seen in figure 8.1, these simulations do not take into account the intrinsic variability of the growth rates, the cycle time, the replication time... Although useful for studying many phenomena, this programme is not completely faithful to reality.

In 2012, Zhang and Shi [84] created a model that was based on two previous works [85, 86]. This model works with the probabilities of cell cycle events happening. The authors take into account the processes of DnaA synthesis, DnaA

\(^1\)The eclipse is the minimal possible distance between one replisome and the preceding one on the same chromosome.
Size and replication initiation in *E. coli*

Figure 8.1 *Screenshot of CCSim simulating a nutritional shift-up*

autorepression in transcription, DnaA titration by DnaA boxes, DnaA inactivation and reactivation, and replication initiation controlled by DnaA. Zhang and Shi use the experimental data from Speck et al. [87] for more updated results. They conclude that initiation mass is invariant, in agreement with Donachie [14]. Also, that *E. coli* can coordinate chromosomal replication initiation with cell growth by regulating DnaA: they propose that this can be achieved mainly by adjusting the concentration of free effective RNA polymerase.

Lastly, Creutziger et al. [88] introduce a novel way of dealing with replication in fast-growing bacteria. They analyse 128 different models that make assumptions about the unknown molecular processes that regulate replication. By comparing with available experimental data, they single out 34 of these models which are in agreement with the experiments. From these it was concluded that the cell size and the chromosomal replication are necessarily linked, either by coupling replication to division or to the amount of cell mass. After this, suggestions are made for new experiments which would further reduce the number of models which agree with experimental data. One of the positive aspects of this study is the use of noise to simulate the inherent stochasticity in cell cycle events, even under constant growing conditions.
Part III

METHODOLOGY
9. Data processing and analysis

The length and position of the cells was determined with iFilli, a MATLAB programme written by Dr. Jordi Garcia Ojalvo. The number of DNA replication initiation sites, however, was determined mainly visually and manually, with the aid of self-written code.

9.1 Clicking the cells

Clicking a cell consisted of tracking one cell from the last frame all the way back to the first frame, through several generations. iFilli is an interactive programme that enables the clicker to select a cell’s region, one frame at a time. Figure 9.1 shows an example of what clicking a cell looks like on the screen. When the contrast of the image was good enough, it was possible to fill in the cell in just one click, as in the top image. (A cell was filled in by detecting its contour and filling in the inner region.) When the contrast of the image was not good enough, however, the area of the cell had to be selected manually, much as one would select a region manually in any image processing programme, as seen in the bottom image. The middle image shows a little magenta square, which is the programme’s

Figure 9.1 The process of clicking the cells illustrated in screenshots
“guess” as to where the next cell to fill in is, according to its position in the previously analysed frame.

Figure 9.2 shows a length vs time graph of a clicked cell. As can be seen, one cell undergoes several cycles in the time duration of a video. A few frames have been included in the image, which show the cells at certain times of their growing cycle.

A total of 54 cells were clicked, 30 for the temperature of 30°C and 24 for the temperature of 27°C. At an average of more than 90 frames per cell, this amounts to almost 5000 frames that were analysed (although a small number was not used because of posterior corruption of files). iFilli then gathers and saves the information of each cell in a *.mat file.

The information from the *.mat file was accessed and transferred to an Excel file for correction of obviously impossible values for cell length. This incorrect data consisted of values of cell lengths that were inferior to both the values immediately previous and immediately subsequent to a certain point: to fix this, a linear interpolation was calculated for the defective value. Although this could be considered falsification of experimental data, at some level, the defective values are the result of mistaken filling in of the cell area by iFilli; the interpolation value is of course much more accurate than the defective value, from a physical and biological point of view. And, also, it allowed analysis of cell cycle durations, growth rates, maximum lengths, etc. (This is because analysis of these variables was performed on the basis that all local minima in cell lengths were the first instants of cell cycles.)

9.2 Counting the dots

The process of computing the number of dots in the videos was a very decisive and complicated step. After all, it was the way of gathering half of the essential information that was needed to carry out this study. This section describes the frustrated attempt of automatising the dot count at every frame and the posterior simpler manual solution.
9.2.1 What didn’t work

A great deal of time was spent attempting to automatise the dots count of the frames. In order to carry this out, a programme would have to isolate the desired cell and detect in the fluorescence image those white dots that the regions of interest.

The image is, after all, a matrix of numbers representing intensity values from black to white (or from 0 to 1). Knowing this, it was attempted to find a way of selecting the dots as regions that were distinct and significantly brighter than the rest, within the isolated cell. Also, the goal was established to find the size of the said regions, and to obtain an average value for their brightness. This would enable the plotting of histograms in order to draw possible conclusions from the brilliance of the dots.

Isolating the cell was no problem, thanks to the data stored in the *.mat file after clicking. Some lines of code singled out those pixels within the region of the cell of interest and filled in the rest with zeroes.

It was seen that, due to various possible reasons (mostly the image capture processes, probably), the cells of interest were not always within the same brightness values; in some frames, they were all darker or lighter than in other frames. This meant that, if the same criteria were to be applied to every image, some dots would appear weaker or stronger than others in other frames, while being equivalent in reality.

In order to correct that, the images were normalised to range between the same two brightness values: the maximum and the minimum of the whole cell. After that the image was softened using the MATLAB function conv2, in order to remove any noise deriving from the contour hardness, and then normalised again, repeating the first step. This way, images were obtained that had less noise and could be treated the same way as all the rest.

After these preliminaries, the first thing that was done was to establish a threshold under which all the pixels were given a value of zero. This was done to eliminate noise from posterior image manipulation. The threshold was given an initial value

---

1 Early in the analysis of the frames, it was suspected that very bright dots could be caused by two origins, one on top of the other in the Z dimension, or very close to one another. A distinctly bimodal histogram would have confirmed this hypotheses and this would have been taken into account in the process of counting the dots.

2 conv2 goes over all the pixels of the image, and places in each one the mean value of a number of the pixels around it. The larger the number of pixels it uses, the softer the image is, but information can be lost in the process. In this case, each pixel was assigned the value of the 5x5 matrix of which it was the centre. Using a bigger matrix meant losing much information, to the point that two dots were taken for one.
of 0.4 and tweaked taking into account the great majority of the frames, trying to find a compromise between eliminating as much noise as possible and not losing any valuable information. The final value for the threshold was taken to be 0.52.

**Figure 9.3** Three of the frames after applying the threshold, shown as an intensity image with a colour map. These images were created with a threshold of 0.6 for another purpose.

Once the images showed only the regions that were candidates to being counted as brighter spots (intensity peaks or just peaks, from now on), several attempts were made to count the dots automatically.

The first function that was tested for suitability to the purpose was the MATLAB function `imregionalmax`, which finds local maxima in a two-dimensional matrix. Looking at figure 9.3 will be enough to see why it didn’t work. The first cell has two very clear bright dots, and even this function would have got it right; it was images like the third that posed a real problem. Two tiny darker spots can be seen above the dark red spot at the right, and a light green area above the red spot at the middle. These three, at least, would have been counted as replication origins: they are, after all, local maxima.

Therefore, this first attempt was ruled out. Figure 9.5 shows the result of this way of calculating the number of peaks. As can be seen, it is no use even to bother looking for a periodicity or for any sort of coupling.

The MATLAB function `imextendedmax` is another very interesting function that was thought would solve the problem. It computes the extended-maxima transform. This would single out those regions with a higher intensity value that the surrounding areas, and then yield a binary image with 1s where there was a maximum region and 0s where there wasn’t. The intention was to later go back to

**Figure 9.4** A frame in which two distinct dots can be observed in one bright region.
the original image and calculate the average values of the extended maxima in the real image. (This would also allow for correction of any irregularities.) The biggest problem found with this was that it wasn’t effective when there were two separate dots in one bright region, like in figure 9.4 at the bottom of the image.

Seeing this, two different thresholds were proposed: one for very bright dots and another for less intense spots. This idea was immediately set aside, because no threshold could be found high enough to distinguish two very close bright dots from one another which wouldn’t, at the same time, ignore many other very bright spots that were too bright for the lower threshold!

![Figure 9.6 Montserrat: part of the mountain crests on the South side](image)

**Figure 9.5** First attempt at counting a cell’s number of dots in each frame
All this time it was imagined that the cell was like a landscape with valleys and mountains: an analogy of the intensity of the dots’ brightness was made to the height or depth of natural accidents, as seen in figure 9.6. With this came the idea of displaying the image as a contour map; some low hills were not to be taken into account, while higher peaks were to be marked and analysed independently.

The contour function was used in order to visualise the cell as a contour map (figure 9.7). It was hoped that some sort of criteria could be applied to count peaks; for instance, if the curves were closer or not, how separate they were... However, it was soon seen that the struggle was with the colour of the lines, which meant that, once again, the issue was all about the intensity values of the thresholds.

After many weeks’ work, the goal was abandoned to compute the whole bright regions or even to automatise the dot count; after all, it was enough, for the scope of the project, to know how many initiation sites there were in a cell at a given moment. This realisation led to find another simpler, but much more effective, solution.

9.2.2 What did work

After many failed attempts at automatising the dot count, it was decided to count the number of shiny regions manually. An alternative method was devised that first displayed four different views of the cell (see figure 9.8). The views that were most used were the fluorescence images, naturally, and the coloured intensity map below, as an aid to quick counting. The other two were used for reference.

At every frame of the video (each frame of the video corresponding to a frame of the original film), an automatised dot count (calculated with imregionalmax, as explained in the previous subsection) was suggested, and if necessary it had to be corrected. The correct figures were then written onto an Excel file, side by side with the length of the cell in that frame and the corresponding time value.

This Excel file was modified where necessary. (Sometimes the cell displayed, for instance, four dots on one frame, three on the next, and four on the following; this is clearly an artefact. See section 9.1, where it explains how the same procedure was used for cell lengths, and footnote 1 on page 59.)
Size and replication initiation in *E. coli*

9.3 Analysis

Different aspects of cell growth and replication (and auxiliary checkings) needed different approaches to analysis; prior to this, though, all the data needed to be stored conveniently for ease of access. Several types of graphs were produced, according to the requirements of the data. Again, MATLAB was the software that was used for this purpose.

9.3.1 Data storage

All the values of interest were stored in two single *.mat* files, one for each temperature. The values stored in the Excel file (length, number of dots and the corresponding time) were used to perform all the calculations needed to obtain the other variables of study listed below. These calculations include exponential regressions—to obtain the growth rate of the cells at each cycle—, and analysis of local maxima and minima in the length of the cell—for the purpose of breaking the data into cell cycles—. Figure 9.9 summarises the basics of extracting the other data from cell length and dots.

The use of only two *.mat* files makes it easier to analyse the data than it would have been for 54 files: because the data were conveniently grouped in one file for each temperature, it was easier to analyse the information of interest for all the cells with just one click.

*Figure 9.8 A frame of one of the videos that were used for the dot count. From top to bottom and from left to right: the fluorescence image; the phase image with the cell of interest in white; the coloured intensity map with the threshold applied; and the maxima calculated automatically*

*Barcelona, September 2013 63*  
*Lara Sofia Escuain i Poole*
Figure 9.9 Cell length $\cdot 10^{-1}$ and number of initiation sites at 27°C. This graph reflects the information that can be extracted from the length and number of dots.

The *.mat files contain the following data:

- $T_{div}$: the cycle time for each of the cycles, in one vector
- $T_{div\text{mean}}$: the calculated mean of the cycle times, in one scalar
- $T_{div\text{sigma}}$: the calculated variance of the cycle times, in one scalar
- $gr$: the growth rate for each of the cycles, in one vector
- $gr\text{mean}$: the calculated mean of the growth rates, in one scalar
- $gr\text{sigma}$: the calculated variance of the growth rates, in one scalar
- $llargs$: the length of the cell at every instant, in one cell array
- $llargs\text{max}$: the maximum lengths of the cell at each cycle, in one vector
- $punts$: the number of dots at every instant, in one cell array
- $times$: the times at every instant, grouped in cycles, in one cell array

---

3The means and deviations of the cycle times and the growth rates were calculated with the MATLAB `normfit` function, which returns the mean and the variance of a set of values following a normal distribution.

4Each cell contains the data for one cycle. This applies to all the data stored in cell arrays.

5Given that the photographs were captured once every five minutes, the term "times" refers to the time that had really passed when the photograph was taken, while the term "instant" can denote the number of the frame, or the sampling time.
9.3.2 Plots

The plots for analysis were drawn with the aid of MATLAB. The main types of plots used were scatter plots, histograms, and bar plots; trending plots were used for one specific purpose. The MATLAB functions used were `scatter`, `hist`, `bar`, and `errorbar`, respectively. More information on the procedure can be found in section A.1 of appendix A.
10. Implementation of the model

10.1 Equations of the model

The following equations constitute the model, with its variables and constants defined in Table 10.1. The actors of the play have already been introduced in chapter 5, where the primary regulatory mechanism for replication is described (see subsection 5.2.2 in page 39). The equations proposed are mainly based on reference [74].

\[
\begin{align*}
\frac{dD^*}{dt} &= \alpha \frac{L}{N} - k_i H D^* - k_u D^* \\
\frac{dD}{dt} &= k_i H D^* - k_u D \\
\frac{dH}{dt} &= -\delta H \\
\frac{dL}{dt} &= \gamma L
\end{align*}
\]

(10.1) (10.2) (10.3) (10.4)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Variable</th>
<th>Symbol</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D^*)</td>
<td>Number of DnaA-ATP molecules bound to oriC</td>
<td>(\alpha)</td>
<td>DnaA production rate</td>
</tr>
<tr>
<td>(D)</td>
<td>Number of DnaA-ADP molecules bound to oriC</td>
<td>(k_i)</td>
<td>DnaA-ATP hydrolysation rate</td>
</tr>
<tr>
<td>(H)</td>
<td>Number of Hda molecules bound to oriC</td>
<td>(k_u)</td>
<td>DnaA unbinding rate</td>
</tr>
<tr>
<td>(L)</td>
<td>Volume of the cell</td>
<td>(\delta)</td>
<td>Hda decay rate</td>
</tr>
<tr>
<td>(N)</td>
<td>Number of initiation sites</td>
<td>(\gamma)</td>
<td>Growth rate</td>
</tr>
</tbody>
</table>

Equation 10.1 models the variation of the number of DnaA-ATP molecules in the cell. It is assumed that as soon as DnaA is produced, it is immediately attached to ATP and bound to the oriC region. Therefore, the number of DnaA-ATP molecules grows proportionally with the length (constitutive segregation) at a constant rate \(\alpha\) (the larger the cell, the more RNA polymerases, ribosomes, etc., and larger the rate of production). At the same time, it decreases with the number of initiation sites, because the more copies of the chromosome there are in a cell, the more oriC regions that have to share the available DnaA-ATP in the cell.
In the same equation it is seen that DnaA-ATP decreases proportionally to $k_i$ and $H$. This is because DnaA is hydrolysed by Hda at a rate $k_i$. (This is also reflected in the first term of equation (10.2).) The last term of the two first equations answers to the fact that both DnaA-ATP and DnaA-ADP unbind oriC at the same rate, $k_u$.

Equation 10.3 describes the number of Hda molecules in the cell. Hda leaves oriC by sliding down the chromosome behind the replisome, and so it decays exponentially at a rate $\delta$. Lastly, the cell length (equation (10.4)) also follows an exponential pattern and increases at a rate $\gamma$.

Three clarifications should be made at this point. Firstly, and most importantly, the main size characteristic of a cell is its volume. Also, it is proportionally to volume, not to length, that DnaA is segregated. Nevertheless, for technical reasons, it was the cell length that was available from the experiments, and the data from the experiments was necessary to obtain and adjust the parameters of the model. Therefore, and only due to this limitation, cell length is the variable used in the model, and not volume, which would make more sense biologically. However, as has already been mentioned, if constant length is assumed, proportionality with respect to volume or length is indifferent.

The second clarification to be made is that the growth mode of *E. coli* was not known for quite some time, and most references took it to be either linear or exponential. Trueba et al. showed it to be exponential in 1998 [89].

Lastly, the third aspect to be taken into account is that these equations do not describe concentrations, as one might expect in this kind of chemical kinetic equations. They are a continuous approximation of the molecule number.

An analytical solution has been attempted to find for this model, or system of ordinary differential equations (see the calculations in appendix B). Nevertheless, the system proved to be solvable only numerically.

### 10.2 Numerical integration

The base of the programme are equations (10.1) to (10.4), described above. The programme will be described in detail in the following subsections, and the code is annexed in appendix C.

### 10.2.1 Preliminary calculations

At the beginning of the programme, some necessary parameters are defined, such as the step size and the duration of the simulation. Then, a matrix of predefined
Size and replication initiation in *E. coli*

A matrix of size (*m* rows, for all the integration steps, and 4 columns, one for each function) is created to store the four outputs, with a separate vector for the time values at each step.

The four functions were notated as \( y(n,1) \), \( y(n,2) \), \( y(n,3) \), and \( y(n,4) \), for \( D^* \), \( D \), \( H \), and \( L \), respectively. \( n \) denotes the row in the matrix, or the integration step.

The integration routine was performed at every step and the results saved in the matrix. This was done by placing the equations for RK4 in a for loop which started at 1 and ended at \( m-1 \).

It was not enough, of course, with the RK4 integration of the functions. Two disturbances had to be introduced artificially —namely, the increase in the number of initiation sites when fitting (subsection 10.2.3) and the division of the cell at the end of a cycle (subsection 10.2.3). Otherwise, the cell grew *ad infinitum* for as long as the simulation time allowed it!

### 10.2.2 Initial values

The initial values for both the functions and the constants, wherever possible, were first chosen based on reasonable guesses. They were later corrected with the information obtained by running the simulation until it reached a steady state (see subsection 10.2.4, where it is described how the constants were adjusted).

In 1991, Hansen et al. suggested that the DnaA content of the cell was a function of the growth rate [91]. This was a valuable tool for the estimation of the initial content of DnaA. Since it was only for a first guess, an average of the average growth rates for the two temperatures was calculated,

\[
\frac{(0.0119 + 0.0164) \text{ min}^{-1} \times 60 \text{ min}}{1 \text{ h}} = 0.8494 \text{ h}^{-1},
\]

and then used to obtain the number of molecules from the function obtained by a linear regression of the data in table 2 of reference [91]. This yielded a total number of DnaA molecules of 416. An even better guess could be obtained by respecting the 0.2 ratio of DnaA-ATP to DnaA-ADP suggested by Donachie and Blakely [74], so the initial guesses were left at 70 molecules of DnaA-ATP and 350 molecules of DnaA-ADP.

Su’etsugu et al. state that the level of Hda is about 100 molecules per cell [92]; this is the starting value for \( H \). The starting length of the cell, \( L \), was calculated as a rounded figure for the mean value of the average lengths for both temperatures.

---

\(^1\)\text{m} \text{ is the duration of the simulation divided by the step size.}
The threshold ratio of DnaA-ATP over DnaA-ADP that was required in order to initiate replication (R in the code) was taken from the literature to be 0.8 [74].

10.2.3 Imposed disturbances

Replication initiation

At every integration step it was checked whether the ratio of DnaA-ATP to DnaA-ADP (in practice, the ratio of \( y(n,1) \) to \( y(n,2) \)) was equal or higher than \( R \), the threshold value. If it wasn’t, nothing was done; if it was, the number of initiation sites was increased by one, and the number of Hda molecules (H, or \( y(n,3) \)) returned to its initial level.

Cell division

The duration of a cycle is decided in the previous cycle. Cell division is carried out if the present instant of time is past the cycle time and if and only if the number of initiation sites is greater than 1.

To impose cell division, three of the model variables (\( D^* \), \( D \), and \( H \)) are forced to adopt the same value as in the previous instant. The length becomes in \( n+1 \) half of its value at \( n \), naturally, and the number of initiation sites in the new cell is also half the number in the "mother" cell.

By definition, the number of initiation sites must always be an integer. Because there is sometimes an odd number of initiation sites present in the cell, though, it was a problem to always divide by 2 to equally distribute the oriC regions between the "daughter" cells. (For instance, a cell containing 5 oriC regions could give either 2 or 3 of them to either of the new cells, so it could give either 2 or 3 origins to the new cell of interest.) A solution was found to decide at random how many sites the newborn cell would hold. A random number \( r \) was generated between 0 and 1. If this number \( r \) was smaller than 0.5, the result of dividing the uneven number of oriCs by 2 was rounded down to the nearest integer; if it was equal to or bigger than 0.5, the number of origins for the new cell was rounded up to the nearest integer.

After division, and under the same conditions, the growth rate and the cycle time for the next cycle are decided. Also, a new reference respect to which calculate the cycle time for the following cycle is established. The new reference is the last step of the present cycle.

A whole section has been devoted to describe how the growth rates and the cycle times were simulated. This is no trivial matter, and of high importance, since it introduces in this model the variability that is lacking in other proposed models to this date (see section 8.2).
Simulating the growth rates and the cycle times  

The growth rates influence the length of the cell before dividing. The cycle times, however, determine the length of the cycles.

The first problem to consider is the possibility that they might be linked; after all, it would be reasonable to expect that the cell will divide sooner if the cell grows faster, and later if it grows slower. In order to eliminate this possibility, the growth rates for each cycle vs the corresponding cycle times were represented. Figure 11.11 on page 86 shows a cloud-like distribution for both temperatures; therefore, it is safe to generate the two variables independently. (For more details, see section 11.3.1.)

The growth rate was suspected to follow a normal distribution. The calculations are developed in chapter A.2 of appendix A. Indeed, as can be seen, the growth rate follows a normal distribution with the parameters shown in table 10.2, which were obtained from the experiments (section A.1.2).

<table>
<thead>
<tr>
<th>Temperature</th>
<th>27 °C</th>
<th>30 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu )</td>
<td>0.0119</td>
<td>0.0164</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.0015</td>
<td>0.0020</td>
</tr>
</tbody>
</table>

Simulating a cycle time that resembles reality was not trivial. In this case, there was no known distribution that fitted the data. The problem was, again, with the fact that, even though a sample of 170-180 points is mostly sufficient in industrial statistics, in biology it can still be too small. (This is due to the enormous variability in biological phenomena.)

For this reason, it was interesting to avoid choosing a distribution and generating random values from it. Instead, another method was devised that uses no intermediate assumptions. In order to justify the validity of this method qualitatively, a histogram was prepared with a comparable amount of simulated data, which can be seen in figure 10.1.

Figure 10.2 shows a graphic example of the method (for a temperature of 30 °C) to accompany this explanation. To begin with, the histogram was plotted for the...
cycle times data, with as many bins as possible to favour precision. Then, the cumulative probability function was also drawn and placed on top. Random numbers between 0 and 1 were generated; these were taken to be the cumulative probability values. Then, the real experimental time corresponding to that value was found.

The main point in this method is that, for an equiprobable distribution of Y values for the cumulative probability function (equidistant horizontal lines), the result is a non-equiprobable distribution of cycle time values (non-equidistant vertical lines).

As can be seen in the figure, the random numbers that are generated still make the majority of the obtained time values coincide with the bulk of the real-life distribution.

10.2.4 Adjusting the parameters

Adjusting the parameters was a mainly qualitative task, based on the expectation that the simulated results should resemble the experiments as closely as possible. For all the parameters, the graphs showing the number of initiation sites and the cell length were used. Wherever possible, other more quantitative ways were applied.

In this example, and in order to explain the procedure, the parameters were adjusted using the data from the temperature of 27°C. After adjusting for 27°C, the constants were readjusted for the temperature of 30°C. The results can be seen in table 12.1 on page 94.
Size and replication initiation in *E. coli*

(a) \( \delta = 1 \text{ min}^{-1} \)

(b) \( \delta = 10^{-6} \text{ min}^{-1} \)

(c) \( \delta = 0.1 \text{ min}^{-1} \)

Figure 10.3 *Cell length (burgundy, px) and number of dots (mustard) for different values of \( \delta \)*

The parameters are discussed one by one, in the order in which they were adjusted. In figures 10.3 and 10.4, the cell length has been divided by 10 to suit graphic representation.

**Hda decay rate \( \delta \)**

This was the first constant that was adjusted, because the function for Hda concentration is independent from the rest. \( \delta \) was adjusted qualitatively, resulting a very narrow range of values between what was too big (figure 10.3a) and what was too small (figure 10.3b). **0.1 min\(^{-1} \) is an acceptable and the most correct value found for \( \delta \) (figure 10.3c).**

**DnaA unbinding rate \( k_u \)**

The rate at which DnaA unbinds oriC was next to be adjusted. This is because it is a time constant which influences the duration of the interinitiation time (the time lapse between one initiation of replication and the next) in the histogram for N.

The two histograms, the simulated and the experimental values, were compared for the time during which a given number of oriC regions were present in the cell to the experimental data. The histograms were studied for 2, 3, and 4 oriC regions in the cell.

Also, a quantitative evaluation for the average squared differences in the histograms for N values of 2, 3, and 4 initiation sites was made. The average of the three results was then calculated, which has been used as an indicator of the goodness of fit between the experimental data and the simulated data. With an average difference of 0.03 between the two histograms (see figure 10.5), **the value of 1.5 min\(^{-1} \) is the final value accepted for \( k_u \).**
Size and replication initiation in *E. coli*

(a) $k_i = 5 \text{ min}^{-1}$  
(b) $k_i = 0.01 \text{ min}^{-1}$  
(c) $k_i = 0.26 \text{ min}^{-1}$

**Figure 10.4** Cell length $\cdot 10^{-1}$ (burgundy, px) and number of dots (mustard) for different values of $k_i$

**DnaA-ATP hydrolysation rate $k_i$**

$k_i$ was adjusted manually, like $\delta$. The procedure was very similar, figure 10.4 showing impossible values in figures 10.4a and 10.4b; the most accurate value for $k_i$ was found to be $0.26 \text{ min}^{-1}$ and can be seen in figure 10.4c.

**DnaA segregation rate $\alpha$**

For the adjustment of the rate at which DnaA is segregated constitutively, the same graph was used as with $k_i$ and $\delta$. It was observed that the cell achieved abnormally high values of length with very low values of $\alpha$, and that over a value of around 400 or 500 min$^{-1}$ the graphs resemble those seen in figures 10.3b and 10.4a. **The value of 200 min$^{-1}$ was chosen for $\alpha$, because it is the value which takes the graph as close as possible to the experiments.**

**Figure 10.5** Histograms for $k_U = 1.5 \text{ min}^{-1}$
Part IV

RESULTS AND DISCUSSION
11. Experimental analysis

The data collection focused mainly on cell length and number of initiation sites, as has already been explained (chapter 9). Nevertheless, from these two quantities—and knowing the times at which the measurements were taken—, several other variables were obtained, such as cycle times and growth rates. This enabled the control of the growth conditions and the deeper study of the variables and the couplings between them.

Table 11.1 shows the size of the samples studied. Since some variables were taken for each time point (cell length, number of dots. . . ) and others characterise a cell cycle (growth rate, cycle time. . . ), the size of the samples varies according to the variable type. The noticeable difference in the size of the instant variables is due to the fact that the cycles are much longer for the temperature of 27°C (see section 11.2.2).

11.1 Process control

Figures 11.1 and 11.2 show that the growth rate, the cycle time, and the lengths have no visible trend with time (this applies to their means and standard deviations).

![Figure 11.1](image-url)

**Figure 11.1** Error bars for the temperature of 27°C
This enabled to ensure the constant growth conditions during all the time the video was being made.

11.2 The observables

The two quantities whose information was extracted directly from the data —cell length and number of initiation sites— and those derived from them are described here in detail. (Although they are listed as “the observables” alone, in some cases their evolution with temperature is also studied.)

11.2.1 Growth rate

The growth rate was analysed on a single cycle basis. The data were represented as probability density functions (one for each temperature), fitted to a statistical distribution, and compared according to temperature. Exponential growth rates were assumed (see footnote on page 67).

In section A.2 of appendix A, the calculations for the fitting of the growth rate were performed. It was concluded that the growth rate follows a normal distribution with means $\mu_{27} = 0.0119$ and $\mu_{30} = 0.0164$ and standard deviations $\sigma_{27} = 0.0015$ and $\sigma_{30} = 0.0020$.

The means of the growth rates at each of the temperatures resulted to be different from one another at a glance. A study of the significance of the difference was performed (see section A.3.1, appendix A). It was concluded that the means of the difference is significant—the small difference in the temperatures notwithstanding—, which makes the growth rates different according to the temperature. Specifically, the growth rate at 30 °C is higher than the growth rate at 27 °C, as it is seen in figure 11.3.

Many attempts to describe the effect of the temperature on bacterial growth rate have been made, some dating as far back as 60 years ago [93]. Some models use
Arrhenius’ law\(^1\), reasoning that bacterial growth is based on chemical reactions just like any other. Other literature refers to the Eyring equation as a major extension to the Arrhenius equation \([94]\). Nevertheless, these equations were found to be insufficient for the complexity and non-linearity of these processes. Ratkowsky et al. suggested, in 1982, a linear dependence of the square root of the growth rate with temperature \([95]\):

\[
\sqrt{\gamma} = A(T - T_0)
\]  

(11.1)

This expression is beautifully simple and it fits experimental data well. Nevertheless, it has received criticism due to its lack of theoretical foundation, as well as the limited temperature range within which it is accurate. Present efforts to effectively describe such relationships include extensions, modifications, and combinations based on the models already described \([96]\).

Regardless of the function of this dependence, the explanation of this reality is based on the fact that, after all, cellular growth is based on chemical reactions. It is well known that lower temperatures make enzymatic mediation and movement of the molecules difficult, as well as increasing the viscosity of the intracellular medium. High temperatures, however, favour the speeding up of cellular reactions and

\(^1\)Arrhenius’ law is expressed as

\[\gamma = A e^{\frac{E_a}{RT}}\]

\(A\) being the frequency factor (or the number of collisions between reacting species), \(E_a\) the activation energy, \(R\) the gas constant, 8.314 J/(mol·K), and \(T\) the temperature in Kelvin. The equation was proposed in 1889 and is remarkable for its description of the dependences of reaction rates.

Figure 11.3 Probability density functions of the growth rates at 27°C and 30°C
the rapid increase in cell size. (Temperatures that are too high will also contribute to denaturalisation of proteins and, eventually, to a halt in growth and/or cell death.)

11.2.2 Cycle time

The study of the cycle time presented more complications than that of the growth rate. No known statistical distribution was found that fitted the data, and all the references to cell growth and cycle in literature were found to be relative to the growth rate. What's most important can be perceived just by eyeing the histograms in figure 11.4: that the cycles are longer for a temperature of 27°C. As has been mentioned in section 10.2.3, one would expect the growth rate and the cycle time to be coupled; nevertheless, no such relationship was found.

![Histogram of cycle times at 27°C and 30°C](image)

**Figure 11.4** Probability density functions of the cycle times at 27°C and 30°C

It is known that, in biology, variables often behave in a normal way. Also, if a distribution is found to be not normal, a log-normal distribution can be the next guess. In this case, the data were tested for both distributions manually (using first the cycle time and then its natural logarithm) following the method described in section A.2, and they proved negative. After that, a statistics programme, EasyFit® was used to find other possible solutions, such as the logistic distribution. All efforts were in vain.

The impact on this project of fitting the data to a known distribution, however, is not so large. It must be taken into account that, in biology, samples of ~200 cycles are often not sufficiently large. It can be expected or hoped, therefore, that if the means of obtaining more data had been provided, the outcome would have been
Figure 11.5 Maximum, minimum and mean cell length for each of the cycles at 27°C. This graph displays information for every cycle, not every instant of time: hence the absence of values on the horizontal axis. Additionally, three lines are displayed as the averages of the three lengths shown.

more accurate. An alternative way of using the cycle time distribution has been employed where necessary (section 10.2.3).

11.2.3 Cell length

This section focuses on the distributions of the absolute cell lengths, regardless of their evolution with time. For the variation of cell length with time, see section 11.2.1.

Figure 11.5 shows the maximum, minimum and mean cell length for each of the cycles, and also the average values of the three lengths. The purpose of this figure is to show the variability of the length values for each of the cycles, by comparing all the data for each of the temperatures.

Fittings developed in section A.4 (page 120) show that the best fit is that of a log-normal distribution (figure 11.6), which is in accord with what might have intuitively been expected looking at figure 11.7. The skewness in the histograms rules out...
the normal possibility for the distribution and already hints at the possibility of a log-normal distribution.

Findings on the distribution of cell length are, surprisingly, very scarce in literature. References [97] and [98] mention that cell length has or could have a log-normal distribution, while reference [99] shows a similar graph for cell length to graph 11.7. Nevertheless, properties of the log-normal distribution make it very adequate for describing cell length [100] and fit experimental observations well.

The most direct observation regarding cell size during the collection of the data was that there were many more very long cells than there were short cells. This is perfectly in accord with the long tail present in log-normal histograms. Limpert et al. describe how skewed distributions are particularly common when mean values are low, variances large, and values can never be negative [100]. These three aspects are characteristic of cell size.

Moreover, log-normal patterns are present in processes which are influenced by many factors acting independently, in a multiplicative and not additive way (the latter would be more typical of normal distributions). This is also the case in cell
growth process and all the molecular machinery, enzymatic activity and metabolic processes involved. Finally, log-normal distributions are typical of positive, real-valued quantities that grow exponentially. This perfectly fits the case of cell size variation.

Scatter plots such as the one in figure 11.5 (page 81) are not very informative, other than for realising the immense variability that is present in these biological processes. Figure 11.8 shows the summarised information of these two graphs. It is of great importance, because it shows a difference in the cell lengths according to the temperature.

It is interesting to observe the noticeable difference in cell length given one temperature or another, even with differences as small as 3°C. (The calculations that justify this assertion can be found in section A.3.2.) The reasons for this, as has already been explained, can be related to the difference in the growth rates, which are higher at higher temperatures.

### 11.2.4 Number of initiation sites and synchronisation

It is important to remember that the number of initiation sites in the microscopy images, the variable that was counted and has been referred to since as "dots", are visible thanks to the protein YFP, which is attached to SeqA, which in turn presents great affinity to hemimethylated GATC sequences in the chromosome (section 5.2.2). Thus, the number of dots is the number of replication initiations being carried out at a given moment.

![Figure 11.9](image)

**Figure 11.9** Frequency of the number of oriCs in the cell at 27°C and 30°C

Figure 11.9 shows the histograms for the number of initiation sites present in a cell at a given moment. The distinct bimodality gives an idea of the synchronicity (or asynchronicity) of chromosome replication. Indeed, two dots in a figure indicate
that the one chromosome in the cell is initiating replication—the two dots being the two new oriCs. If replication was perfectly synchronised, it would be reasonable to expect that the two new oriC regions would initiate replication simultaneously, passing directly from two dots to four dots. Therefore, there would be no occurrence of three dots in the pictures, and the histogram would only show those values belonging to a series of the type $2^n$.

Nevertheless, this is not the case. The graphs show bimodality with modes at two and four, but there is still an important component in the three-dot bar. (Only the values of two, three, and four dots are used for reference, because further numbers of replication initiation sites are not to be taken into account in terms of normal cell behaviour.)

This presence of three dots in the cell might serve as an indicator of replication initiation synchronicity. With an expression of the form

$$s = \frac{N_4}{N_3 + N_4},$$

(11.2)

where $N_4$ is the number of times a cell exhibits four dots and $N_3$ is the analogue for three dots, it is possible to evaluate or to at least compare the overall degree of synchronicity of a cell population at two different temperatures. The point of this equation is the comparison between the times a cell replicates both of its origins simultaneously, going from two to four dots, and the times there is a transition of three dots. (This is counted by the occurrences of 4 and 3 dots.) The parameter $s$, therefore, is an indicator of the synchronicity at a given temperature.

For total synchronicity (that is to say, the transition from 2 oriC regions to 4 is direct),

$$N_3 = 0 \rightarrow s = 1.$$  

(11.3)

For total asynchronicity (i.e. in no case does the number of replication sites double simultaneously),

$$N_4 = 0 \rightarrow s = 0.$$  

(11.4)

For the obtained data:

$$s_{27} = \frac{N_{4_{27}}}{N_{3_{27}} + N_{4_{27}}} = \frac{717}{374 + 717} = 0.6572$$

(11.5)

$$s_{30} = \frac{N_{4_{30}}}{N_{3_{30}} + N_{4_{30}}} = \frac{472}{312 + 472} = 0.6020$$

(11.6)
Given the boundaries that have been explained, it can be seen that the degree of synchronicity at 27°C is higher than that at 30°C. This could be caused by lower growth rates and longer cycles at 27°C, which would allow slower replication rates and thus better intracellular coordination.

Little or no reference is made in literature to this asynchronicity. Rather, the exponential model for chromosome number seems to be widely accepted. (Also, studies of replication based on fluorescence imaging, and tracking cells’ lineage at an individual level, are uncommon to this date.)

Chen et al., working with *Synechococcus elongatus*, suggested in 2012 a decoupling of DNA replication and cell division [101], supported by previous ideas of a constant DNA synthesis rate over time in *S. elongatus* [102]. If there were nothing else to take into consideration, the findings presented in this work would be in agreement with this thesis. Indeed, the quantity of experimental data relative to the exact number of oriC regions per cell at any given instant can very well shake the foundations of the $2^n$ theory in *S. elongatus*, but the case of *Escherichia coli* is fairly complex and will be the object of further discussion.

To wrap these considerations up, figure 11.10 shows the number of occurrences of a given number of dots during a given amount of time (i.e. how many observations of 4 dots during 5 minutes, 4 dots during 10 minutes, etc.). It can be seen that, although the sum of dots during different lengths of time favour the presence of even numbers of oriC regions, the most common occurrences are 3 oriC regions (or the transition from 3 to four initiation sites) during interinitiation times of 10 and 15 minutes.

**Figure 11.10** Description and occurrence of interinitiation times at 30°C. The colour code is in accord with the size of the marks and indicates the count of the occurrences of an amount of dots during a length of time.
11.3 Correlations between the observables

Correlations between the different observables were checked for by using scatter plots. This made it possible to see at a glance whether there was any correlation between any two features of the division and replication dynamics.

11.3.1 Growth rate and cycle time

In this graph, the scatter plots for the growth rates vs the cycle times at the two temperatures can be seen. Intuition would say that a negative correlation would be expected, i.e. shorter cycle times for higher growth rates and vice versa. Although the cloud-like distribution seen in figure 11.11 discourages any further attempts at coupling the two variables, the correlation coefficient\(^2\) has been calculated for both temperatures and is shown in the graph. The previous intuition that it should be negative is confirmed, but, as expected, it is so low that it is hardly possible to speak of any correlation at all.

This lack of correlation has, as a consequence, a large variability in the length of the cells (see figure 11.5). If, as previously expected, slower growth rates had resulted in longer cycle times, the cell length would have been more conserved.

11.3.2 Cell length and growth rate

Figure 11.12a shows no coupling between cell length and growth rate at any of the temperatures. Little or no references are made in literature to this fact. On the one hand, _E. coli_ as a species would tend to always maintain a cell size and shape

---

\(^2\)The correlation coefficient calculated here and in the following sections is the Pearson product-moment correlation coefficient, or Pearson’s \(r\). It gives a value from -1 to 1, ranging from perfect negative linear correlation to perfect positive linear correlation. Values of 0 or close to 0 indicate no linear correlation.

It is used in this work because the goal is to quantify the linear dependence of one variable vs another, and not the goodness of a model —as the coefficient of determination \(R^2\) would have done. (This is especially applicable in section 11.3.5.)
in accord with environmental conditions, nutrient availability and cell cycle requirements (see sections 5.1.2 and 5.2.1). In order to attain this, regulation mechanisms in the cell would control the growth rate and the cell mass conveniently.

On the other hand, the variability in cell size is immense (figure 11.5). Only \( \sim 180 \) experimental points of growth rate vs final cell length have been obtained for each of the temperatures, which makes up a somewhat small sample. It is to be expected that the results would vary if more data had been available.

### 11.3.3 Cell length and cycle time

Similar considerations apply for cycle time as for the growth rate (see section 11.3.2 and figure 11.12b). It was also found that the coupling was practically non-existent.

In the case of the results at \( 27^\circ C \), the correlation coefficient is significantly higher than that at \( 30^\circ C \), \( r_{27} = 0.4093 \) vs \( r_{30} = 0.1636 \). This high (or higher) correlation is in agreement with what was to be expected, that final length is greater with longer cycle times.

### 11.3.4 Growth rate and replication initiation

The graph in figure 11.13 shows the time lapse before each replication initiation (that is to say, the duration of the transition between one replication initiation and the next, or interinitiation time) plotted against the growth rate of that cycle. At a first glance, it is seen that there exists no correlation between the duration of these lapses and the growth rate; on the other hand, though, it can also be seen that this is still dependent on temperature. This can imply, other considerations notwithstanding, that initiation time is still coupled to growth rate and/or cell size.
Several references indicate a coupling between growth rate and replication initiation. In any case, this coupling is always indirect, although there are different explanations.

The first is that initiation mass, proposed by W. Donachie as a trigger for replication [14], increases with the growth rate [103]. This is supported by the findings of the present project, where it was seen that both growth rate and cell size increase with temperature (see figures 11.3 and 11.8). Although which depend on which remains somewhat unclear, it is still true that replication initiation is triggered earlier in faster growing cells.

Secondly, an increasing growth rate causes a decrease in the C period [104]. (The C period is the time lapse between cell division and replication of the chromosome.)

Thirdly, DnaA concentration is suggested to be proportional to cell growth [91, 105, 106]. This assumption can be reasoned with the fact that segregation of DnaA is constitutive (i.e. DnaA is segregated proportionally to cell size). Thus, larger cell sizes will result in more DnaA presence in the cell, and larger cell sizes have been obtained in conditions in which higher growth rates have also been found.\(^3\)

11.3.5 Cell length and number of initiation sites

The coupling of cell size with the number of initiation sites is one of the most important aspects of study of this work. To observe the relationship between the two variables, a scatter plot has been graphed both for 27°C and for 30°C. They can be seen in figures 11.14 and 11.15, respectively.

A correlation between cell size and number of replication initiation sites is observable at a glance. The fitted straight line is present in the graph, as well as its equation and the correlation coefficient (Pearson’s r) in both cases.

It was a long time ago, in 1968, that W. Donachie suggested a coupling of replication initiation to cell size [14]. W. Donachie worked at a population level and

\(^3\)The big problem with these ideas in this work is the lack of correlation between cell size and growth rate (figure 11.12a). It is still true that larger cells are obtained with higher growth rates, but only because larger cells and higher growth rates are obtained at higher temperatures.
based his research on estimations of the shape $2^n$ from the work of Cooper and Helmstetter [15]. At the time these studies came to light, modern imaging and analysis tools that have been used in this project were not available.

Thanks to such tools, bacteria at a single cell level have been studied. Tracking individual cells over time allows for the collection of data relative to cell size, and the possibility of counting the number of oriC regions at every frame exists as well. This enabled to come to new conclusions about replication initiation coupling to other processes in the cell. Nevertheless, many other more recent publications on the matter still based their work on the concept of a critical mass for initiation [80, 74, 107].
In section 11.2.4, new ideas were put forth about initiation control. Studies on *Synechococcus elongatus* suggested the possibility of continuous segregation of DNA; multiple copies of the chromosome normally present in the cell; and, most importantly, complete independence of replication initiation from cell cycle events, being initiation a function solely of cell size [101, 102, 108].

The case is different with *E. coli*. Firstly, recent findings couple replication initiation strongly to cell cycle [109], and Bates and Kleckner go as far as suggesting that chromosome replication is triggered by previous cell division [75]. Secondly, and perhaps more intuitively, *E. coli* has only one chromosome per cell *per se*. If there are more copies at a given moment, it’s because higher growth rates cause the replication rounds to overlap [110, 15]. Therefore, if only one chromosome has to be present in cells dividing exponentially with base 2 —the $2^n$-shaped evolution that has already been mentioned—, it logically follows that, ideally, the replication will also follow a $2^n$ pattern.

In addition, a reinforcement of this idea has been seen in the experimental data. On the one hand, a distinctly bimodal density distribution was observed when counting the number of initiation sites in the cell at each instant (figure 11.9). On the other hand, it is reasonable to suppose that a better correlation would have been obtained in figures 11.14 and 11.15 if it was indeed true that initiation depends solely on cell length, and not also on the $2^n$ pattern. References [111] and [61] emphasise the necessity of a synchronisation in successive replication rounds.

Figure 11.16 shows the available cell length and number of dots at 27 °C, a total of 2235 experimental points. (Note that the graphs are not plotted against time, because this graph contains data from different videos that were filmed simultaneously.) It is a good way of graphically seeing how, mostly the number of oriC regions oscillates between 2 and 4, especially for values of length closer to the global average. (The presence of 2 initiation sites implies one chromosome undergoing replication, and 4 initiation sites a second round of replication. This is the most common —and most ideal— situation.)

(It is true, nevertheless, that an inevitable asynchrony exists to a certain degree. This is why there is an actual presence of 3 sites (or any number of sites out of the $2^n$ series) in multiple occasions, and why the calculations in section 11.2.4 have been performed. Explanations for this asynchrony could be suboptimal conditions of initiation, or interference with the sequestration apparatus [50].)

The conclusion that has been reached, up until this point, is that the number of initiation sites is correlated to cell length. Nevertheless, the cell length is not the only explanation, or the sole governor, for the number of oriC regions present in the cell; or rather, for replication initiation.
Size and replication initiation in *E. coli*

As a consequence, another agent is needed to trigger cell replication: a substance whose presence in the cell is necessarily proportional to its size, but whose activity will depend on multiple other factors other than cell size. This initiator is the DnaA protein⁴ [84, 80, 112, 59, 62, 50, 52].

One of the most enlightening studies in this respect is that of Hill et al. [80]. Recently, in 2012, these authors published the first study on the effect of cell size on DNA replication (and not the reverse, DNA replication changing cell size, which has been widely documented).

Hill et al. studied the consequences of altering cell size on chromosomal replication in the model bacteria *Bacillus subtilis* and *Escherichia coli*. In contrast to *B. subtilis*, small mutant⁵ *E. coli* cells delayed replication until wild-type cell initiation size was achieved, *even though the DnaA concentration was the same as in wild-type cells*. Furthermore, when the DnaA content in the cell was increased, the delay was shortened.

---

⁴No experiments have been done in this work, or for this work, to obtain any data relative to DnaA. The present discussion, as well the design of the model, are based on extensive bibliographic research.

⁵Two cell-size mutants were employed, on the basis of two studies which showed that these mutants were ~25% smaller in size.

One of the mutants had a loss-of-function mutation in *pgm:kan*, the gene encoding phosphoglucomutase. These mutants are partially blocked in the metabolism of glucose 1-phosphate and showed consequent alterations in the cell size, reaching only 70% of the size in wild-type cells [113].

The other of the mutants exhibited a mutation in the cell division gene *ftsA*. Changing one amino acid in *ftsA* results in *ftsA*. This mutation accelerates the assembly of the Z ring, causing division at shorter lengths and decreasing mean cell length by 27% [114].

---

**Figure 11.16** Cell length and number of oriC regions at 27°C. All the available experimental points are shown
This, and the previous considerations, lead to believe that it is the amount of DnaA, and not its concentration (as it was earlier suggested [51]), that acts as a trigger to replication initiation. The gate, or the critical step, that a cell has to overcome before licensing a new round of replication is the adequate quantity of DnaA present in the cell.

The defence of the idea of a substance whose critical amount in the cell mediates replication initiation does not mean to rule out the work of previous investigators, who laid the foundations with the concept of a critical mass for initiation [14, 15]. Rather, the fact that DnaA is segregated proportionally to cell size allows for the conciliation of the two ideas [79]: replication initiation begins at a certain critical mass, which in turn is determined by the quantity of DnaA protein in the cell.
12. Mathematical model

The mathematical model presented here is a simple set of non-linear differential equations. These equations are based on the role of DnaA as the initiator of replication, and describe its synthesis and its conversion from active DnaA-ATP to inactive DnaA-ADP, as well as indicating the behaviour of Hda (which acts as a negative regulator for DnaA-ATP) and cell length.

12.1 A reminder of the model

The model aims to describe the main aspects of cell growth and chromosome replication on the basis of DnaA-triggered replication initiation and RIDA as the regulation of DnaA (see section 5.2.2). The equations of the model can be found on page 67, and a short reminder of the processes that the model describes follows:

• The DnaA protein, which is the replication initiator, is present in two forms: an active ATP-bound form, DnaA-ATP, and an inactive form bound to ADP, DnaA-ADP. In the model, they are referred to as $D^*$ and $D$, respectively.

• Segregation of DnaA is constitutive—that is to say, its production is dependent on the cell length—at a rate $\alpha$. It is assumed that immediately after segregation, DnaA is bound to ATP; this implies that DnaA becomes active immediately on production.

• The Hda protein $H$ hydrolyses DnaA-ATP into its inactive form, DnaA-ADP. This is done at a rate $k_i$.

• Both forms of DnaA, active and inactive, leave the origin of replication at the same rate. The rate of oriC unbinding is $k_u$.

• Hda unbinds the origin of replication exponentially at a rate $\delta$, sliding down the chromosome behind the replisome.

• Finally, the length of the cell $L$ grows exponentially at a rate $\gamma$. (This equation is necessary in itself, but it also affects the quantity of DnaA in the cell.)
The model works by solving equations (10.1) to (10.4) at every integration step. Before, two conditions have to be checked.

The first condition is for replication initiation. If the ratio of DnaA-ATP to DnaA-ADP (or $D^*:D$) is superior to 0.8, initiation is triggered [74]. The level of Hda is forced back to its original value. The rest of the variables are solved for with the equation system.

If the time instant is, at a given integration step, equal or superior to cycle time, the length of the cell is divided by two, and likewise with the number of initiation sites $N$. (For more details on how this was programmed, see section 10.2.3 on page 70.) The levels of DnaA and Hda are maintained from the previous step. Finally, the growth rate and the cycle time were established for the next generation (see section 10.2.3 on page 71.)

If neither of the conditions apply, the cell undergoes no major event. The system is then solved for by only taking into account equations (10.1) to (10.4).

### 12.2 The parameters of the model at 27°C and 30°C

It can be easily noticed that all the rates (the constants that have been adjusted) have a higher value at 30°C than at 27°C (table 12.1). In section 11.2.1 it was reasoned why the growth rate was higher at 30°C than at 27°C: namely, that these biological rates are still rates of chemical reactions, however complicated they might be. Like any other reaction, a rise in temperature —provided it is not as high as to cause protein denaturalisation or the death of the bacteria— favours the process by allowing more efficient enzymatic catalysis, speedier movement of the molecules, and a more fluid cytoplasm.

### 12.3 Analysis of the results of the simulations

#### 12.3.1 Qualitative comparison of the length and number initiations

The model was run systematically in order to adjust the experimental constants, as explained in section 10.2.4. The results of the assays can be found in table 12.1 (top). The constant $\gamma$ can be simulated directly from the experimental data; it is nor-

---

**Table 12.1 Values of (top) and for (bottom) the constants of the model for 27°C and 30°C**

<table>
<thead>
<tr>
<th>Temperature</th>
<th>27°C</th>
<th>30°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$ (min$^{-1}$)</td>
<td>200</td>
<td>250</td>
</tr>
<tr>
<td>$k_i$ (min$^{-1}$)</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td>$k_u$ (min$^{-1}$)</td>
<td>1.50</td>
<td>2.00</td>
</tr>
<tr>
<td>$\delta$ (min$^{-1}$)</td>
<td>0.10</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\gamma$ (min$^{-1}$)</th>
<th>27°C</th>
<th>30°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0.0119</td>
<td>0.0164</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.0015</td>
<td>0.0020</td>
</tr>
</tbody>
</table>
Normally distributed, with different mean and variance depending on the temperature (table 12.1, bottom).

Below are two graphs which are representative of the results. These show the simulated cell length and the number of initiation sites at each of the temperatures.

**Figure 12.1** Simulated results of cell length $\cdot 10^{-1}$ vs number of initiation sites at 27°C

**Figure 12.2** Simulated results of cell length $\cdot 10^{-1}$ vs number of initiation sites at 30°C

Figure 12.1 shows a result of running the model for a temperature of 27°C, while figure 12.2 shows the same for 30°C. On page 64, figure 9.9 displays a representation of the experimental data for length and number of initiation sites. Although the latter is for a cell that is amongst the most regularly behaved, it is interesting to qualitatively see how well the model adjusts to reality in this sense. Among all the
graphs obtained from experimental points, many graphs showed more irregularities in cell size and the number of dots at the different cycle stages. (For a more general comparison, refer to figure 11.16 (page 91).)

12.3.2 Quantitative analysis of the number of initiations

First results

Simulations were ran with the code in section C.1 to obtain an equivalent number of cycles as in the experimental data, and *.mat files were written for the results with the same structure. The same methods that were used to analyse the experimental data were used for the simulated data. The following discussion is based on information obtained therewith.

It has been concluded that the cell length \( L \) is merely an exponential function governed by a growth rate \( \gamma \), which is distributed normally with parameters \( \mu \) and \( \sigma \). Therefore, cell length need not be the object of further discussion.

The other fundamental variable —of this project in general and of the simulations in particular— is the number of initiation sites in the cell, \( N \). The number of \( oriC \)s is a significantly more complicated issue, since it depends on the length of the cell and on the concentrations of the two forms of DnaA.

Figure 12.3 is a representative histogram for the number of dots in the cell, resulting from a run of the model. (This run was at 27°C, but the results are analogue at 30°C.) As can be seen, it is very far from the bimodality seen in the experimental results in figure 11.9. (It could be adventured that the research on *Synechococcus elongatus* [101], discussed in section 11.2.4, could have shown a similar distribution of initiation sites in the cell.) Moreover, the data fit no known distribution.

This is a very serious shortcoming of the model, because its goal is to describe the coupling between cell size and replication initiation. If the model does not simulate the number of initiation sites correctly, this renders it useless for the present purpose.
The reason for this behaviour of the model is that it does not make any distinction among the oriC regions. Instead, it treats the cell as a whole. This means that, if there is enough global potential for the cell to initiate replication, it will, on any of its replication initiation origins.

**Subsequent analysis**

The obvious issues of the model were addressed by treating the initiation sites as independent entities which evolve in time, each according to its own circumstances. This was done by solving the model with as many sets of equations as oriC regions there are in the cell at any given instant of time.

Since the variability of the model made adjusting the model parameters exceedingly difficult, the variability due to the growth rate distribution and the cycle times was removed from the model. Figure 12.4a shows the outcome. Its chief characteristic is the almost total absence of occurrences of three dots, which is an indicator of the prevalence of the $2^n$ pattern in the replication of the chromosome.

Once the model was adjusted, variability was again introduced in the shape of variable growth rates, cycle times, and concentration of Hda after replication. The result can be seen in figure 12.4b, which exhibits a good agreement between the model and the experimental observations.

![Comparison between the histograms for simulated and experimental numbers of dots](image)

**Figure 12.4** Comparison between the histograms for simulated and experimental numbers of dots

### 12.3.3 Further steps for this study

The histograms for simulated and experimental results shown in figure 12.4b bear a remarkable resemblance to each other. Nevertheless, it must be indicated that,
once the model simulates the variability found in the experiments, it becomes highly dependent on parameter values. Further work on the model will involve finding the source of such parameter sensitivity and correcting it. The aim is to obtain a reliable, sturdy model that will faithfully simulate the behaviours analysed from the experiments. Specifically, one of the first steps to undertake will be a study of the behaviour of the set of non-linear differential equations near the stationary state (which has been found to be the null solution to the equations).
Part V

CONCLUSIONS

I hear you've divided already! Aren't you excited?!

Oh, yes! I'm BESIDE MYSELF!!
13. Conclusions

The goal of this project was to shed light on the temporal coupling between the division and the chromosomal replication cycles in *Escherichia coli*. This was done, firstly, by analysing experimental data provided by Dr. Ojalvo’s host laboratory; and secondly, by proposing a simple mathematical model to describe the evolution with time of the two variables.

13.1 Analysis of the experiments

The data was obtained as frames of several films, which included two sets for each of the colonies. The first set contained the phase contrast images of the colony, in which the outer shape of the bacteria could be seen. The other set showed a fluorescence dot for every binding site of SeqA by a fluorescent protein attached to it. Data for two different temperatures, 27°C and 30°C, were available.

The binding sites of this protein were taken to be the freshly replicated regions immediately after the replication forks. Since the number of replication forks and of initiation sites is identical, this enabled to compute the number of initiation sites present in the cell.

Therefore, the number of initiation sites was counted, tracking each lineage frame by frame for several cells at an individual level. Likewise, the length of the cells was measured at every point in time. This was done for 54 cells.

From the data for length and number of initiation sites data was obtained, other derived information was also available. Among this, the cycle time—or the length of the cell cycles—and the growth rates of the cells at each cycle are the most important.

The variables were studied individually and, where appropriate, by pairs. This was done with the aid of self-written code in MATLAB and by applying statistical analysis methods.

For the variables studied individually, a discussion of their effects was made. Their evolution with time and the changes that the two different temperatures caused were examined.

The growth rate was found to be distributed normally. The parameters of the distribution were obtained from the experimental samples. After the corresponding
hypothesis test, it was concluded that the growth rate is higher at 30°C than at 27°C. This is because cellular growth is a series of chemical reactions, which are favoured by increases in temperature within biologically reasonable bounds.

For the cycle time, no distribution was fitted. Nevertheless, it was seen that the cycles are longer at 27°C. This suggested a negative coupling between the growth rate and the cycle time (see below). It is straightforward to suppose that the faster the cells reproduce, the shorter the cycles will be, and vice versa.

Maximum cell length (just prior to division) was found to be log-normally distributed. This indicates a larger presence of unusually long cells than of unusually short cells (therefore, the long cells are not so unusual). Moreover, average cell length is higher at 30°C than at 27°C; it can be reasoned that the effect of the higher growth rate on the cells’ length is more powerful than the shorter cycle time.

The probability distribution of the presence of replication initiation sites showed a distinct bimodality for the values of 2 and 4 initiation sites. This implies that the most common scenario was the presence of one replicating chromosome, followed by a second round of replication. Nevertheless, a certain asynchronicity was detected, which was higher at 30°C than at 27°C. This could be because slower growth rates at a lower temperature would allow for better intracellular coordination.

The variables were then grouped by pairs. In order to observe any possible correlation, they were plotted against each other. The growth rate and the cycle time show only an extremely faint correlation. Nevertheless, the correlation is negative, which is in accord with the previous reasoning that the slower the cells grow, the longer they take to complete a cycle.

Growth rates and cycle times show no coupling with cell length. Cells are continuously adapting to the environment and the requirements of their own cycle, so they are continually adjusting their growth parameters in order to respond to the circumstances. Also, cell length variability is very high, and it is thought that if the samples had been much larger, the results would have been different.

A small degree of correlation, however, was observed for the cell length and the cycle time at 27°C. The correlation is positive, which is consistent with the fact that cell length is proportional to cycle time.

There was no direct coupling observed between growth rates and interinitiation time (the time lapse between two consecutive initiations of replication). However, it was observed that the data are dependent on temperature. Because growth rates and cell size are higher at higher temperatures, this suggests that replication initiation is indirectly dependent on cell size.

Cell length and the number of initiation sites present in the cell at a given moment are clearly coupled. Because the correlation is not perfect, nevertheless, it is
Size and replication initiation in *E. coli*

proposed that cell size is not the only trigger for replication initiation. The best candidate for this trigger is a substance which regulates initiation but whose presence in the cell is dependent on cell size. It is thought that this substance is the DnaA protein, which is segregated constitutively.

Therefore, the coupling of replication initiation to cell size takes on a new dimension. With DnaA as an intermediary, replication initiation is triggered by a certain amount of DnaA in the cell, which in turn depends on cell size.

### 13.2 Simulations of the model

A simple model was developed to simulate experimental growth and replication conditions. The model is a set of ordinary non-linear differential equations that was solved numerically with a fourth-order Runge-Kutta integration method.

These equations describe the constitutive segregation of DnaA and the control of replication initiation, mainly by regulatory inactivation of DnaA (RIDA). The model was integrated with two conditioned perturbations: the start of chromosome replication when the quantity of cellular DnaA exceeded a threshold, and cellular division when the integration time exceeded the cycle time.

Simulations of this model were very satisfactory qualitatively, when only the cell length and dots graphs vs time were eyed. However, more quantitative examination of the probability density function of the number of initiation sites revealed distributions that are far from the bimodal histogram that was expected.

Indeed, it was realised that the model did not make any distinction between the different origins of initiation, and therefore there was no synchrony in the successive replication rounds. A further step was taken to implement a routine that would treat each initiation site individually, assigning a set of equations (the same equations of the model) to each of them.

Although this enabled for better control and synchrony of replication initiation by treating each initiation site differently, the model became very unstable. As a consequence, further work will entail finding the reasons for this behaviour and correcting it, in order to obtain a reliable model that is faithful to the experimental data.
Bibliography


Size and replication initiation in *E. coli*


Size and replication initiation in *E. coli*


Size and replication initiation in *E. coli*


Size and replication initiation in *E. coli*


[70] K. Skarstad; N. Torheim; S. Wold; R. Lurz; W. Messer; S. Fossum; T. Bach. The *Escherichia coli* SeqA protein binds specifically to two sites in fully and hemimethylated oriC and has the capacity to inhibit DNA replication and affect chromosome topology. *Biochimie*, 83:49–51, 2001.


Part VI

APPENDICES
A. Statistical analysis of the variables

A.1 Detailed description of the plots

A.1.1 Scatter plots

The scatter plots were used for qualitative analysis, mostly of length, number of initiation sites and time. They were plotted using the MATLAB scatter function, which represents one given vector vs another.

Scatter plot for one temperature only

This graph is very conclusive because it displays the cell size and the corresponding number of initiation sites at every instant of time, for all the cells at one of the temperatures. Due to its shape, it gives an idea of the correlation between these two variables.

Scatter plots to compare the two temperatures

This graph is highly significant- It shows the global averages for the lengths at each temperature, i.e. one average maximum, one average minimum and one average mean for each of the two temperatures.

Weighed scatter plots

This third type of plot is the one which presents the highest complexity. It was used to determine the number of occurrences of the time lapses during which a cell exhibited a constant number of oriC regions. That is to say, it counts the number of times a cell shows three dots for 10 minutes, the number of times a cell shows three dots for 15 minutes, and so on. Therefore, a two-dimensional graph was not enough, so each dots-duration point was smaller or bigger, and of a different colour, according to its number of occurrences.

The complexity in this type of graph lies in the need to count occurrences of two-dimensional points on a plane. For this a vector was created to describe the number of repetitions; this vector was used to represent the size and colour of the different points, and all this without mixing up the information.
Firstly, a two-column matrix (for dots and duration) was sorted first according to
the number of dots and then, respecting the first sorting, according to the durations.
Then, each row of the matrix was examined from top to bottom. For every new point,
a new vector obtained a new position with the value of 1; for repeated points, the
number of the last position of the vector was incremented by 1. Finally, only one
row of each was kept in the two-column matrix, using the unique function: this
resulted in a two-column matrix of unique values and a vector of the same length.
The latter is the vector that was used to determine the size and colour of each point.

This graph can be seen in figure 11.10 on page 85. A similar graph was obtained
for the data at 27°C.

A.1.2 Histograms

Histogram plots have no complication. They are all based on the hist function, to
group the data in the specified (or not) bins (i.e. make the actual histogram), and
the bar function, which provides flexibility when displaying the graph and defining
the axes values.

A.1.3 Trending plots

These graphs display the mean value of a variable and its standard deviation at
each point. In this case it was necessary to check whether three variables showed
any significant trend with time. The data for these three variables was therefore
grouped according to the cycle number. Then the mean and the variance were
calculated for each group, because they are two necessary arguments for the errorbar function, which was then used to display the required graphs.

A.2 Fitting the growth rate

The growth rate was determined to follow a normal distribution after performing the
Chi Square method. The Chi Square method compares the histogram of the ex-
perimental data with the theoretical histogram, which is obtained with the statistical
law (in this case, the normal distribution). Then a hypothesis test is performed, with
the null hypothesis being the good fit of the data to the distribution.

The following example uses the data for 30°C. (The same method was per-
formed on the temperature of 27°C.) Figure A.1 displays the histogram for the ex-
perimental data.

Let $e_i$ be the experimental observations recorded for each class. $t_i$ is the mean
of the values of the normal function in each class.
The theoretical data were obtained by normalising the edges of the bins and calculating the probabilities for every interval. To normalise, the following expression is used:

\[ Z = \frac{X - \mu}{\sigma} \]  

A.1

In this case, \( \mu = 0.0164 \) and \( \sigma = 0.0020 \) were taken as the mean and deviation from the samples. The values of the edges of the bins and their normalised values \( Z \) are shown in table A.1.

<table>
<thead>
<tr>
<th>Edges</th>
<th>0.0100</th>
<th>0.0110</th>
<th>0.0120</th>
<th>0.0130</th>
<th>0.0140</th>
<th>0.0150</th>
<th>0.0160</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Z )</td>
<td>-3.1729</td>
<td>-2.6793</td>
<td>-2.1856</td>
<td>-1.6919</td>
<td>-1.1983</td>
<td>-0.7046</td>
<td>-0.2110</td>
</tr>
</tbody>
</table>

The theoretical probability for every interval was calculated using the MATLAB `normcdf` function, which calculates the cumulative probability at a given point. There are five classes that contain less than 5 theoretical counts. Therefore, pooling is necessary in order to comply with the restrictions of the method. Table A.2 shows the experimental data, \( e \), and the expected data, \( E \), after pooling. The result is a sample of size \( n \) grouped in \( k = 9 \) classes. The statistic for the test is:

\[ W = \sum_{i=1}^{k} \frac{e_i^2}{t_i} - n \rightarrow W \sim \chi^2 \]  

A.2

In this case,

\[ W = \sum_{i=1}^{13} \frac{e_i^2}{t_i} - 172 = 4.33 \]  

A.3

The law is only functionally specified: the number of degrees of freedom is

\[ \nu = k - p - 1 = 9 - 2 - 1 = 6 \]  

A.4

where \( p \) is the number of unknown parameters.
The threshold value is
\[ \chi^2_{v} = \chi^2_{6} = 12.59 \]  
(A.5)

Comparing the threshold value with the statistic it can be seen that \( W < \chi^2_{v} \). It is immediate that the null hypothesis for the normal distribution of the growth rates can be accepted.

It must be indicated that MATLAB has a function to perform this hypothesis test: \texttt{chi2gof}. Nevertheless, it was thought better to calculate each of the steps manually for better control of the process.

Moreover, the normal P-P plots of the growth rate have both been graphed with the MATLAB function \texttt{probplot}. (The method is explained in chapter A.4.) With figure A.2 it is confirmed that the normal distribution can be safely accepted.

![Figure A.2 P-P plots of the growth rates at 27°C and 30 °C](image)

**A.3 Statistical significance of differences between the two temperatures**

**A.3.1 Significance of the difference between the growth rates**

A hypotheses test was conducted to test for the significance of the difference between the means of the growth rates. The two growth rates follow independent normal distributions and the variance of the samples is known. The general expression for these cases would be
Size and replication initiation in *E. coli*

\[
Z = \frac{\bar{X}_{27} - \bar{X}_{30} - \alpha}{\sqrt{\frac{\sigma^2_{27}}{n_{27}} + \frac{\sigma^2_{30}}{n_{30}}}}, \quad Z \sim N(0; 1). \tag{A.6}
\]

In this particular case, the test was conducted for the assumptions of equal vs different growth rates:

\[
H_0 : m_{30} - m_{27} = \alpha \tag{A.7}
\]

\[
H_1 : m_{30} - m_{27} \neq \alpha, \text{ with } \alpha = 0 \tag{A.8}
\]

Because \( \alpha = 0 \) (being \( \alpha = m_{30} - m_{27} \)), though, it can directly be used that

\[
\bar{X}_{30} - \bar{X}_{27} \sim N \left( m_{30} - m_{27}; \frac{\sigma^2_{30}}{n_{30}} + \frac{\sigma^2_{27}}{n_{27}} \right) \tag{A.9}
\]

Nevertheless, it is still easier to use the formula in equation (A.6), because the table for the standardised normal distribution is easily accessible [115]. Thus:

\[
Z = \frac{\bar{X}_{30} - \bar{X}_{27}}{\sqrt{\frac{\sigma^2_{30}}{n_{30}} + \frac{\sigma^2_{27}}{n_{27}}}} = \frac{0.0164 - 0.0119}{\sqrt{\frac{0.0020^2}{172} + \frac{0.0015^2}{177}}} = 23.3454 \tag{A.10}
\]

This value exceeds 1.96 by far, so it is well in the critical region for a risk of 5% (because engineers don’t believe in bad luck). This means that the null hypothesis is to be rejected, so it can safely be asserted that the means of the growth rates at 27°C and at 30°C are significantly different from one another.

### A.3.2 Significance of the difference between the final lengths

This analysis needs some prior considerations. It has been found that the distribution of the final cell lengths is not normal, so this hypothesis test would be invalid at a first glance. Nevertheless, the distribution of the cell lengths is log-normal, which means that the natural logarithms of the samples are distributed normally. As a consequence, after a previous step of taking the natural logarithms of the data, the hypothesis test can be performed as usual.

Once the previous transformation had been made, the calculations were done analogously to the growth rates. Again,
Size and replication initiation in *E. coli*

\[ H_0 : m_{30} - m_{27} = a \]  \hspace{1cm} (A.11)

\[ H_1 : m_{30} - m_{27} \neq a, \text{ with } a = 0 \]  \hspace{1cm} (A.12)

and also

\[ Z = \frac{\bar{X}_{27} - \bar{X}_{30} - a}{\sqrt{\frac{\sigma^2_{27}}{n_{27}} + \frac{\sigma^2_{30}}{n_{30}}}} \sim N(0; 1), \]  \hspace{1cm} (A.13)

with \( a = 0 \). Therefore, substituting,

\[ Z = \frac{\bar{X}_{30} - \bar{X}_{27}}{\sqrt{\frac{\sigma^2_{30}}{n_{30}} + \frac{\sigma^2_{27}}{n_{27}}}} = \frac{4.8441 - 4.7843}{\sqrt{\frac{0.2373^2}{172} + \frac{0.2765^2}{177}}} = 2.1701 \]  \hspace{1cm} (A.14)

Again, the value exceeds that of the threshold value for a risk of 5%, but it is the threshold value for a risk of 3%. Nevertheless, the risk that can be taken is of 5%, so the null hypothesis is rejected and the significance of the difference between the means is proved.

### A.4 P-P plots for the fitting of cell length data

Some of the fitting has been done with an alternative method: P-P plots. In this section, a distribution for the cell length was found with this graphical procedure. The calculations here are for the cell lengths at 30°C but the results obtained at 27°C are equivalent.

Table A.3 shows the first 15 rows of the table used to calculate the figures necessary for the P-P plot of the exponential distribution. Figure A.3 shows the three resulting graphs of evaluating the exponential distribution (figure A.3a), the normal distribution (figure A.3b), and the log-normal distribution (A.3c).

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( x_{i, \text{sorted}} )</th>
<th>( i )</th>
<th>( F_i )</th>
<th>( -\ln(1 - F_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>69</td>
<td>1</td>
<td>0.0028</td>
<td>0.0028</td>
</tr>
<tr>
<td>134</td>
<td>80</td>
<td>2</td>
<td>0.0085</td>
<td>0.0085</td>
</tr>
<tr>
<td>115</td>
<td>81</td>
<td>3</td>
<td>0.0141</td>
<td>0.0142</td>
</tr>
<tr>
<td>108</td>
<td>81</td>
<td>4</td>
<td>0.0198</td>
<td>0.0200</td>
</tr>
<tr>
<td>97</td>
<td>81</td>
<td>5</td>
<td>0.0254</td>
<td>0.0258</td>
</tr>
<tr>
<td>81</td>
<td>82</td>
<td>6</td>
<td>0.0311</td>
<td>0.0316</td>
</tr>
<tr>
<td>91</td>
<td>83</td>
<td>7</td>
<td>0.0367</td>
<td>0.0374</td>
</tr>
<tr>
<td>104</td>
<td>84</td>
<td>8</td>
<td>0.0424</td>
<td>0.0433</td>
</tr>
<tr>
<td>88</td>
<td>84</td>
<td>9</td>
<td>0.0480</td>
<td>0.0492</td>
</tr>
<tr>
<td>213</td>
<td>84</td>
<td>10</td>
<td>0.0537</td>
<td>0.0552</td>
</tr>
<tr>
<td>207</td>
<td>84</td>
<td>11</td>
<td>0.0593</td>
<td>0.0612</td>
</tr>
<tr>
<td>155</td>
<td>85</td>
<td>12</td>
<td>0.0650</td>
<td>0.0672</td>
</tr>
<tr>
<td>121</td>
<td>86</td>
<td>13</td>
<td>0.0706</td>
<td>0.0732</td>
</tr>
<tr>
<td>128</td>
<td>88</td>
<td>14</td>
<td>0.0763</td>
<td>0.0793</td>
</tr>
<tr>
<td>94</td>
<td>88</td>
<td>15</td>
<td>0.0819</td>
<td>0.0855</td>
</tr>
<tr>
<td>88</td>
<td>88</td>
<td>16</td>
<td>0.0876</td>
<td>0.0916</td>
</tr>
<tr>
<td>80</td>
<td>88</td>
<td>17</td>
<td>0.0932</td>
<td>0.0979</td>
</tr>
</tbody>
</table>
The calculations for the normal and the log-normal distributions were performed on MATLAB, although each of the steps was programmed manually. This was done because the OpenOffice Calc® software did not present a satisfactory way of calculating inverse error functions (needed for the study of normal and log-normal distributions), and searching the table manually for the Z score of 177 probability points was considered an unnecessary waste of time. MATLAB has a very convenient function for this, zscore.

Figure A.3c shows a very well aligned set of dots, as opposed to other possible distributions. Therefore, it can be concluded that the final cell length is log-normally distributed.
B. Analytical solution of the model

The mathematical model of chromosomal replication studied in this project, and already explained in the main text, is:

\[
\begin{align*}
\frac{dD^*}{dt} &= \alpha \frac{L}{N} - k_i H D^* - k_u D^* \\
\frac{dD}{dt} &= k_i H D^* - k_u D \\
\frac{dH}{dt} &= -\delta H \\
\frac{dL}{dt} &= \gamma L
\end{align*}
\]  \hspace{1cm} (B.1-4)

The model is a set of ordinary differential equations. Although the model has been integrated numerically, the analytical solution to it was also sought following this procedure:

• Solve for \(H\) and \(L\)

• After inserting the solutions for \(H\) and \(L\) in the equation \(D^*\), solve the equation that is obtained

• Solve for \(D\)

The equations for \(H\) and \(L\) are separable first-order linear ordinary differential equations and are therefore trivial. In the case of \(H\),
\[ \frac{dH}{dt} = -\delta H \]  

(B.5)

\[ \frac{dH}{H} = -\delta dt \]  

(B.6)

\[ \int \frac{dH}{H} = -\delta \int dt \]  

(B.7)

\[ \ln H = -\delta t + C \]  

(B.8)

\[ e^{\ln H} = e^{-\delta t + C} \]  

(B.9)

\[ H = k e^{-\delta t} \]  

(B.10)

Likewise,

\[ L = k' e^{\gamma t} \]  

(B.11)

Therefore, substituting the expressions of equations (B.10) and (B.11) in the first equation in the model, which describes the behaviour of the number of DnaA-ATP molecules bound to oriC, the following is obtained:

\[ \frac{dD^*}{dt} = \frac{k' e^{\gamma t}}{N} - k_1 k e^{-\delta t} D^* - k_u D^* \]  

(B.12)

After rearranging, in-homogeneous, non-autonomous ordinary differential equation results:

\[ \frac{dD^*}{dt} + (k_1 k e^{-\delta t} + k_u) D^* = \frac{\alpha k'}{N} e^{\gamma t} \]  

(B.13)

the general expression of which is

\[ y' + P(x) y = Q(x) \]  

(B.14)

To solve this equation, the method taught by Prof. Magaña to his Calculus III class on the 6th of October, 2009, will be used.

As instructed by Prof. Magaña, this ODE admits an integrating factor of the form \( \mu(x) \). This integrating factor is obtained by

\[ \mu(x) = e^{\int P(x) \, dx} \]  

(B.15)

Solving for this case makes the integrating factor

\[ \mu(x) = e^{\int (k_1 k e^{-\delta t} + k_u) \, dx} \]  

(B.16)
The general solution for this type of equations with integrating factors is of the type

\[ y = e^{-\int P(x) \, dx} \left( \int e^{\int P(x) \, dx} Q(x) \, dx + C \right), \quad (B.17) \]

so, substituting,

\[ D^* = e^{-\int (k_i k e^{-\delta t} + k_u) \, dx} \left( \int e^{\int (k_i k e^{-\delta t} + k_u) \, dx} \frac{\alpha k'}{N} e^{\gamma t} \, dx + C \right) \quad (B.18) \]

This is the final expression for the function. As can be seen, inside the braces there is the integral of an exponential function whose exponent is another integral of an exponential function. Given that the integral of an exponential function is always another exponential function, this would become the integral of an exponential function whose exponent is another exponential function, as in

\[ \int a e^{b e^{c x}} \, dx \quad (B.19) \]

Unfortunately, this has no analytical solution. Therefore, the model cannot be solved analytically, so the study should be carried out integrating numerically.
C. Code for RK4

C.1 First version

This first version of the code treats the cell as a whole and does not make a distinction between the different oriC regions. This programme features three main parts and each step of the code is explained in detail in the comments in green.

clear all
close all

% -----------------------------------------------------------------------
% INITIALISATION
% -----------------------------------------------------------------------

% Selects the path and the file for the necessary information to simulate
% growth rates and division times
[FileName,PathName] = uigetfile(‘*.mat’,’Select MightyMat’);
cd(PathName)
load(FileName)
Tdiv=Tdiv+5; % Corrects an error in the processing of the data
e3=min(Tdiv):5:max(Tdiv); % Defines a vector of bins for the histogram
q=hist(Tdiv,e3); % Histogram of the experimental division times
cs=cumsum(q); % Cumulative sum of the histogram
cs=cs/max(cs); % The above divided by the number of observations to
% obtain a cumulative density function

% Simulation parameters
disp(‘ ‘)
lsim=input(‘Length of simulation (min)? [Default: 360] ‘); % Enter the
% desired time length of simulation
disp(‘ ‘)
h=input(‘Time step (min)? [Default: 0.01 min] ‘); % Enter the desired
% stepsize

if isempty(lsim)
    lsim=360; % Default value for the length of the simulation time
end
if isempty(h)
    h=0.01;  % Default value for the stepsize
end

m=lsim/h+1;  % Length of the vectors for time and solutions

a=0.00002;  ki=2;  ku=1.08;  d=0.1;  R=0.8;  H0=100;  % Values of the constants

% Creates the empty matrix for the functions and the empty vector for the
% number of initiation sites
y=zeros(m,4);
N=zeros(m,1);

% Creates the vector for the time and initialises the time reference for
% division
t=0:h:lsim;
ti=0;

% Necessary values to get the simulation started

In the initialisation, the necessary experimental data are obtained from the *.mat
files. Then, the simulation length and the step size are defined, and the variables
in the code are allocated a name and a size.

for n=1:m-1
    % Increases N when the ratio of ATP--DnaA to ADP--DnaA is above the
    % threshold R
    if y(n,1)/y(n,2)>=R
        N(n+1)=N(n)+1;
        y(n+1,3)=H0;
    elseif y(n,1)/y(n,2)<R
        N(n+1)=N(n);
    end

Lara Sofia Escuain i Poole

Barcelona, September 2013
% Use the new value of N because this is for a new point of the
% function
% k1
k1=h*[(a+y(n,4)/N(n+1)−ki*y(n,3)*y(n,1)−ku*y(n,1))...
   +ki*y(n,3)*y(n,1)−ku*y(n,2)...
   −d*y(n,3)...
   +g*y(n,4)];

% k2
k2=h*[a*(y(n,4)+k1(4)/2)/N(n+1)−ki*(y(n,3)+k1(3))/2]*...
   +(y(n,1)+k1(1)/2)−ku*(y(n,1)+k1(1)/2)...
   +ki*(y(n,3)+k1(2)/2)−(y(n,1)+k1(1)/2)−ku*(y(n,2)+k1(2)/2)...
   −d*(y(n,3)+k1(3)/2)...
   +g*(y(n,4)+k1(4)/2)];

% k3
k3=h*[a*(y(n,4)+k2(4)/2)/N(n+1)−ki*(y(n,3)+k2(3))/2]*...
   +(y(n,1)+k2(1)/2)−ku*(y(n,1)+k2(1)/2)...
   +ki*(y(n,3)+k2(2)/2)−(y(n,1)+k2(1)/2)−ku*(y(n,2)+k2(2)/2)...
   −d*(y(n,3)+k2(3)/2)...
   +g*(y(n,4)+k2(4)/2)];

% k4
k4=h*[a*(y(n,4)+k3(4))/N(n+1)−ki*(y(n,3)+k3(3))]*...
   +(y(n,1)+k3(1))−ku*(y(n,1)+k3(1))...
   +ki*(y(n,3)+k3(2))−(y(n,1)+k3(1))−ku*(y(n,2)+k3(2))...
   −d*(y(n,3)+k3(3))...
   +g*(y(n,4)+k3(4))];

% y(n+1,:)
y(n+1,:)=y(n,:) + k1/6 + k2/3 + k3/3 + k4/6;

% Corrects the value of H in case the integration squashed it
if y(n,1)/y(n,2)>>R
   y(n+1,3)=H0;
end

% Cell division
if n+1>=ti+(simTdiv/h+1) && N(n+1)>1
   % L
   y(n+1,:)=[y(n,1) y(n,2) y(n,3) y(n,4)/2]; % Makes the cell divide
   % N
   r=rand(1); % Gives the adequate number of initiation sites to the
   % new cell
   if r<0.5
\begin{verbatim}
N(n+1)=floor(N(n)/2);
end
if r>=0.5
    N(n+1)=ceil(N(n)/2);
end

% Decides about the next cycle
ti=n+1; % Sets the new reference respect of which to calculate
        % the division time of the next cycle
g=normrnd(grmean,grsigma); % Sets the growth rate for the next
        % cycle
simTdiv=e3(find(cs>=rand(1),1)); % Sets the division time for the
        % next cycle

end
end

% =======================================================================
The integration is the main part of the code. A fourth-order Runge-Kutta integration
routine is performed. Two external disturbances condition it: replication initiation
and division. The programme checks for these two conditions at every integration
step.

% =======================================================================
% SELECTION OF RESULTS
% =======================================================================
% Creates new vectors for t and y with samples every five minutes, in
% order to resemble the experiments, in case it is wanted
tvec=0:5:lsim;
yvec=zeros(length(tvec),4);
Nvec=zeros(length(tvec),1);

for k=1:length(tvec)
    yvec(k,:)=y(t==tvec(k),:);
    Nvec(k)=N(t==tvec(k));
end

% =======================================================================

This third part selects the desired precision of the integration results for further pro-
cessing, if necessary. It can also be used to make the simulated results resemble
the graphs from the experiments in shape.

Lara Sofia Escuain i Poole

Barcelona, September 2013
\end{verbatim}
C.2 Improvements

This code features the improvements that enable the treatment of each oriC region individually. It is a far more complex programme because it implements the integration routine as many times at each instant as oriC regions there are in the cell. Nevertheless, the structure is analogous to that of the first programme.

clear all
close all

% INITIALISATION

% Selects the file from which to obtain the necessary experimental data in function of the desired temperature

t=input('Temperature: ');
FileName=sprintf('MightyMat%i.mat',t);
load(FileName)

Tdiv=Tdiv+5; % Corrects an error in the processing of the data
for rr=1:length(llargs) % Previous calculation for the initial length
    lrr=llargs{1,rr};
    lmin(rr)=lrr(1);
end

% Simulation parameters

t=input('Length of the simulation (min) [Default: 360 min]: ');
% Enter the desired time length of simulation

t=input('Time step (min) [Default: 0.01 min]: '); % Enter the desired % stepsize

if isempty(lsim)
    lsim=360; % Default value for the length of the simulation time
end

if isempty(h)
    h=0.01; % Default value for the stepsize
end

% Definition of the constants and parameters

Barcelona, September 2013
Lara Sofia Escuain i Poole
if t==27,
    a=200;
    ki=0.26;
    ku=1.5;
    d=1;
    R=0.8;
    H0=100;
elseif t==30,
    a=0.025;
    ki=0.75;
    ku=0.09215;
    d=100;
    R=0.8;
    H0=100;
end

v=input('Variability (%) [Default: none]: ');
if isempty(v)
    v=0; % Default value for the variability
end

% Preliminaries for the calculation of the cycle time
e3=min(Tdiv):5:max(Tdiv);
l=hist(Tdiv,e3);
cs=cumsum(l);
cs=cs/max(cs);

% Length of the vectors for time and solutions
m=lsim/h+1;
t=0:h:lsim;

% Definition of sizes, values and parameters for the first step
N0=2; % Number of oriC regions
w=N0; % 3D dimension of the solutions matrix (number of layers)
np=N0; % Number of dots in the first instant
L0=mean(lmin); % Initial length
lo=L0; % Parameter for the calculation of the length
ti=0; % First instant of time

% Preallocation of size for the matrices for the solutions
y=zeros(m,3,w); % 3D matrix for the solutions
N=zeros(m,w); % Matrix for the number of dots
L=zeros(m,1); % Vector for the lengths
NN=zeros(m,1); % Independent and exterior vector for the dots

% Initialisation of the matrices for the solutions
for ww=1:np, % Layers for oriCs
Size and replication initiation in \textit{E. coli}

\begin{verbatim}
y(1,:,ww)=...
    [70 350 randi([int16(H0*(1-v/100)), int16(H0*(1+v/100))],1)];
end
L(1)=L0; % Length
NN(1)=N0; % Number of oriCs

% Calculation of the experimental parameters for the first cycle
g=normrnd(grmean,grsigma); % Growth rate
simTdiv=e3(find(cs>=rand(1),1)); % Cycle time

% TO REMOVE ALL VARIABILITY, PLEASE UNCOMMENT THIS SECTION
% .................................................................
% v=0; % Removes the variability in the concentrations of Hda after
% % replication initiation
%
% oo=length(Tdiv); % Prepares for the removal of variability in length
% for rr=1:oo
%     lrr=llargs{1,rr};
%     lmin(rr)=lrr(1);
%     lmax(rr)=lrr(end);
% end
% TT=log(mean(lmax)/mean(lmin))/grmean; % Constant value for cycle times
% simTdiv=TT;
% g=grmean; % Constant value for growth rates
% .................................................................
% .................................................................

The initialisation part selects the necessary experimental data, defines the necessary parameters for simulation, and creates the necessary structures for the solutions. Moreover, it provides the option of introducing more variability —in the concentration of Hda after initiation, specifically—and also of completely removing all the simulated biological variability if necessary.

\end{verbatim}
for z=1:w % Calculations for the present oriC

% k1
k1=h*[(a+L(n)/NN(n)−ki*y(n,3)*y(n,1)−ku*y(n,1))...
ki*y(n,3)*y(n,1)−ku*y(n,2)...
−d*y(n,3)];

% k2
k2=h*[a*(L(n)/2)/NN(n)−ki*(y(n,3)+k1(3)/2)*(y(n,1)+k1(1)/2)−...
kii*(y(n,1)+k1(1)/2)...
ki*(y(n,3)+k1(2)/2)*(y(n,1)+k1(1)/2)−ku*(y(n,2)+k1(2)/2)...
−d*(y(n,3)+k1(3)/2)];

% k3
k3=h*[a*(L(n)/2)/NN(n)−ki*(y(n,3)+k2(3)/2)*(y(n,1)+k2(1)/2)−...
kii*(y(n,1)+k2(1)/2)...
ki*(y(n,3)+k2(2)/2)*(y(n,1)+k2(1)/2)−ku*(y(n,2)+k2(2)/2)...
−d*(y(n,3)+k2(3)/2)];

% k4
k4=h*[a*(L(n))/NN(n)−ki*(y(n,3)+k3(3))*(y(n,1)+k3(1))−...
kii*(y(n,1)+k3(1))...
ki*(y(n,3)+k3(2))*(y(n,1)+k3(1))−ku*(y(n,2)+k3(2))...
−d*(y(n,3)+k3(3))];

% Solution for a certain oriC at a certain step
y(n+1,:,z)=y(n,:,z)+k1/6+k2/3+k3/3+k4/6;

end % of the integration routine for this oriC

% CONDITIONED INITIATION
if y(n,1,z)/y(n,2,z)>R, % Condition for initiation (DD/D>0.8)
    np=np+1; % Initiation occurs
    y(n+1,:,z)=[70 350 randi([int16(H0*(1−v/100)),...
    int16(H0*(1+v/100))],1)]; % Resets the values of the
    % concentrations for immediately after replication
elseif y(n,1,z)/y(n,2,z)<R,
    N(n+1,z)=NN(n); % The number of oriCs is the same as before
end % of initiation

% Escape valve: this is in case the calculations crash MATLAB
if np>1000,
    disp('I am too tired to continue.')
end

Lara Sofia Escuain i Poole 134 Barcelona, September 2013
% CONDITIONED DIVISION
if n+1>=ti+(simTdiv/h+1) && N(n+1)>1 % Division (not allowed if there
  % is only one division site)
    L(n+1)=L(n)/2; % Divides the length by two
    lo=L(n)/2; % New parameter for length calculation
    r=rand(1); % Divides the number of oriC rounding up at random
    if r<0.5
        np=floor(NN(n)/2); % Rounds up the resulting division
        y(:,;,np:end)=[]; % Eliminates the corresponding layers of y
    elseif r>=0.5
        np=ceil(NN(n)/2); % Rounds up the resulting division
        y(:,;,np:end)=[]; % Eliminates the corresponding layers of y
    end
    y(n+1,:,:)=y(n,1,:); y(n,2,:); y(n,3,:)); % Resets the
    % concentration values
    ti=n+1; % Sets the new reference respect of which to calculate
          % when the following cycle ends
    g=normrnd(grmean,grsigma); % Growth rate for the next cycle
    simTdiv=e3(find(cs>=rand(1),1)); % Cycle time for the next cycle
end % of division

NN(n+1)=np; % Assigns the new value to the independent dots vector
hoho=NN(n);
y(:,;,hoho+1:np)=zeros(m,3,np=hoho); % Creates new layers in the
  % solutions matrix for the new dots
for ww=hoho+1:np,
    y(n+1,:,ww)=70 350 randi([int16(H0*(1-v/100)),...
                              int16(H0*(1+v/100))],1)); % Assigns initial values to the new
    % layers of the concentrations matrix
end
end % of the calculations for this instant

% =----------------------------------------------------------

The integration part has an additional loop with respect to the first version. Not only does it perform the integration routine at every step, but it must do so for each initiation site within the same time step.

There are two new aspects of this programme worth mentioning: one is that a layer is added to the 3D matrix of solutions for every new initiation site, which progresses independently from the others; and the other is that an escape had to
be provided to prevent MATLAB from crashing. This is because the growth of the number of dots is exponential, while division only divides it by two, and the number of oriC regions often exceeded the computer’s capacity while adjusting the model’s parameters.

% VISUALISATION OF THE RESULTS
% Length and dots graph

F=figure;
maxfig(F,1);
plot(t,L)
xlabel('Time (min)')
ylabel('Length')
xlim([0 lsim]);
hold on
plot(t,NN*10,'g') % N is multiplied by 10!!!

% Selection of simulated values to match the precision of the experimental data

\[
\text{tvec} = 0:5:lsim;
\]
\[
\text{Nvec} = \text{zeros}(|\text{tvec}|,1);
\]

for \(k=1:\text{length(tvec)}\)
\[
\text{Nvec}(k)=\text{NN}(t==\text{tvec}(k));
\]
end
\[
\text{p} = \text{Nvec};
\]
\[
\text{psim} = \text{p};
\]

% Calculation of interinitiation times for simulated data

\[
n = \text{length(p)};
\]
\[
e = 1;
\]
\[
\text{huhus}(e) = 1;
\]

for \(k=2:n\)
\[
\text{if } \text{p}(k) == \text{p}(k-1)
\]
\[
\text{huhus}(e) = \text{huhus}(e) + 1;
\]
\[
\text{else}
\]
\[
e = e+1;
\]
\[
\text{huhus}(e) = 1;
\]
\[
\text{end}
\]
end
Size and replication initiation in *E. coli*

\[ \text{xsim} = \text{huhus} \times 5; \]

% Calculation of interinitiation times for experimental data

```matlab
load(FileName)
O = length(punts);
p = punts{1,1};
for i = 2:O
    p = vertcat(p, punts{1,i});
end
P = length(p);
pexp = p;
e = 1;
huhu(e) = 1;
for k = 2:P
    if p(k) == p(k-1)
        huhu(e) = huhu(e) + 1;
    else
        e = e + 1;
        huhu(e) = 1;
    end
end
huhu = huhu \times 5;
```

% Bar chart for simulated and experimental interinitiation times

```matlab
abs = min(huhu):5:max(huhu); % X vector
% Matrix with the simulated and experimental histograms for comparison
mm = [hist(xsim, abs)/length(xsim); hist(huhu, abs)/length(huhu)];
```

F = figure(3);
maxfig(F, 1);
bar(abs', mm);
legend('sim', 'exp');

SS = sum((mm(:, 1) - mm(:, 2)).^2) % Shows how different the histograms are

% Bar chart for simulated and experimental numbers of oriCs

MAX = max(max(psim), max(pexp));
MIN = min(min(psim), min(pexp));
oabs = MIN:MAX; % X vector
% Matrix with the simulated and experimental histograms for comparison
hnb = [hist(psim, oabs)/length(psim); hist(pexp, oabs)/length(pexp)];
```

F = figure(4);
maxfig(F, 1);
This last part, as before, collects the data obtained from the simulation. This time, however, it also loads the file of interest that contains the experimental data, in order to make comparisons possible. These comparisons are made with the aid of two bar charts, which show the interinitiation times and the number of initiation sites for both simulated and experimental data.