# Modelling, analysis and control for systems biology : application to bacterial growth models

Alfonso Carta

**UNIVERSITY OF NICE-SOPHIA ANTIPOLIS**

# DOCTORAL SCHOOL STIC

**INFORMATION AND COMMUNICATION SCIENCES AND TECHNOLOGIES**

# PhD THESIS

A dissertation submitted in partial fulfilment of the requirements for the degree of

**DOCTOR OF SCIENCE**

from University of Nice-Sophia Antipolis

Specialized in: **Control, Signal and Image Processing**

prepared and presented by

Alfonso CARTA

# Modelling, analysis and control for systems biology: application to bacterial growth models

*Supervised by* Jean Luc Gouzé (Inria)

*Co-supervised by* Madalena Chaves (Inria)

Date of defence, 22-05-2014

**Jury Members**

| | | |
|---|---|---|
| Jean-Luc Gouzé | DR Inria Sophia Antipolis | Advisor |
| Madalena Chaves | CR Inria Sophia Antipolis | Co-Advisor |
| Andreas Kremling | Pr., Technische Universität München, Munich | Reviewer |
| Lorenzo Farina | Pr., University of Roma - La Sapienza, Rome | Reviewer |
| Vincent Fromion | DR INRA Jouy en Josas | Examiner |
| Delphine Ropers | CR Inria Grenoble | Examiner |

**UNIVERSITE DE NICE-SOPHIA ANTIPOLIS**

# ECOLE DOCTORALE STIC

SCIENCE ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

# THESE

pour obtenir le titre de

**Docteur en Sciences**

de l'Université de Nice-Sophia Antipolis

Mention: **Automatique, Traitement du Signal et des Images**

présentée et soutenue par

Alfonso CARTA

# Modélisation, analyse et contrôle pour la biologie des systèmes : application à des modèles de croissance de bactéries

*Thèse dirigée par* Jean Luc Gouzé (Inria)

*Co-encadrée par* Madalena Chaves (Inria).

Date de soutenance, 22-05-2014

**Composition du jury:**

| | | |
|---|---|---|
| Jean-Luc Gouzé | DR Inria Sophia Antipolis | Directeur |
| Madalena Chaves | CR Inria Sophia Antipolis | Co-Directrice |
| Andreas Kremling | Pr., Technische Universität München, Munich | Rapporteur |
| Lorenzo Farina | Pr., University of Roma - La Sapienza, Rome | Rapporteur |
| Vincent Fromion | DR INRA Jouy en Josas | Examinateur |
| Delphine Ropers | CR Inria Grenoble | Examinateur |

*"Per aspera sic itur ad astra."*
Seneca, *Hercules furens*, 437

## Abstract

This thesis deals with modelling, analysis and control of gene regulatory networks in the bacterium *E. coli*, with tools of Control Theory. Different mathematical methodologies (qualitative/quantitative, deterministic/stochastic) have been used to best describe the different biological systems under investigation. Notably, in the first part of the thesis we mainly addressed the problem of controlling the growth rate of bacterial cells. Growth control is essential in industrial biotechnology and fundamental research of this kind could pave the way to novel types of antimicrobial strategies. To this aim we developed new qualitative mathematical formalisms, derived from piecewise linear systems, to couple gene expression with growth rate. We applied these formalisms to small *E. coli* synthetic gene circuit models (conceived with our collaborators from Ibis, Inria Grenoble) implementing both open and closed loop configurations. By means of phase plane analysis and bifurcation diagrams we showed that the proposed qualitative control strategies, which act on the gene expression machinery (GEM), can mathematically control the cell growth rate. Moreover, in order to identify the key components of GEM that mostly determine the bacterial growth rate, we also tested several growth rate models using Boolean computational tools. In the second part of the thesis, we developed a coarse-grained, but quantitative, ODE model of *E. coli* GEM whose parameter values have been identified from published experimental data at different steady state growth rate values. This coarse-grained *E. coli* model may be used, in the future, as starting point for the design of synthetic genetic manipulations, which implement desired controls of the bacterial GEM. In the third part, we moved from population cell models to single cell models. In particular we addressed the problem of stochastic state estimation for gene regulatory networks at the level of single cells. We took the Chemical Master Equation (CME) as a reference modelling approach, and investigated the use of stochastic differential model approximations for the construction of practical real-time filters. To this aim, we considered a Square-Root Unscented Kalman Filter built on a Chemical Langevin Equation approximation of the CME. State estimation is interesting *per se* for the reconstruction of gene network variables that cannot be measured directly; in addition it can be used as an intermediate step for identification, and plays a central role toward model-based control.

# Résumé

Cette thèse porte sur la modélisation, l'analyse et le contrôle de réseaux de régulation génétique dans la bactérie E. Coli, avec les outils de la Théorie du Contrôle. On utilise plusieurs formalismes (qualitatif/quantitatif, déterministe/stochastique) pour décrire les différents systèmes. Dans la première partie de la thèse, on considère le problème du contrôle du taux de croissance pour les bactéries. Le taux de croissance est une caractéristique essentielle pour l'industrie des biotechnologies, et cette recherche peut ouvrir la voie à de nouvelles stratégies antimicrobiennes. Nous avons développé de nouveaux formalismes qualitatifs, basé sur les systèmes affines par morceaux différentiels, qui couplent l'expression des gènes et la croissance. Nous appliquons ces formalismes à de petits modèles de circuits génétiques synthétiques (conçus avec nos collaborateurs de Ibis, Inria Grenoble), et étudions des boucles de contrôle ouvertes ou fermées. Par une étude du portrait de phase et des bifurcations , nous montrons que la stratégie qualitative de contrôle proposée, qui agit sur la machinerie cellulaire globale, permet de contrôler le taux de croissance. Pour trouver les composants les plus représentatifs de cette machinerie cellulaire, nous testons plusieurs modèles de taux de croissance, avec des outils de calcul booléens. Dans la seconde partie de la thèse, nous développons un modèle simplifié de la machinerie cellulaire globale chez E. Coli, basé sur des équations différentielles, et dont les paramètres sont identifiés à partir de données de la littérature pour plusieurs taux de croissance. Ce modèle pourra être utilisé comme une base pour tester des stratégies de contrôle de la machinerie cellulaire, par des techniques de biologie synthétique. Dans la troisième partie, nous passons des modèles de population de cellules à des modèles de cellule individuelle ; nous considérons le problème de l'estimation stochastique de l'état dans des réseaux de gènes pour une seule cellule. L'équation maitresse de la chimie est prise comme modèle de référence, et nous étudions l'utilisation d'approximation par des modèles différentiels stochastiques pour la construction de filtres efficaces en temps réel. Nous considérons pour cela une version non-linéaire du filtre de Kalman basée sur une approximation de l'équation maitresse par l'équation de Langevin chimique. L'estimation de l'état est intéressante en soi, car elle permet de reconstituer des variables non mesurables dans des réseaux génétiques ; de plus, c'est une étape nécessaire pour l'identification, et pour le contrôle basé sur un modèle.

# Contents

# Chapter 1

# Introduction

During the last century research in molecular and cell biology has deeply studied and investigated how cellular components work and interact among each other. For examples, biologist have examined the biochemistry of different molecules, the structure of DNA, RNA and proteins, the principles governing DNA replication, as well as transcription and translation.

At the beginning of the $21^{st}$ century, however, the research in biology has moved towards a more integrated understanding of molecular and cellular systems. In fact, we are increasingly observing how the integration of different fields of biology, natural sciences and engineering allows researchers to better understand how cells work, how cellular process are regulated, and how cells respond to environmental stresses or even anticipate those changes. This is in part due to the fact that new technologies provide faster, cheaper, more accurate and comprehensive measurements of gene expression levels [26, 111]. The larger and always more complex amount of data, that has become available in this way, is however not suited to be analysed manually. Hence, mathematical and computational tools turn out to be indispensable to pre-process, analyse, and eventually extract meaning and insights from such experimental data.

Concomitant with all these advances and the potential technological impact due to these revolutionary biological developments, leading scientists in life science have recognized that a new system-level approach is required. This approach has been termed *systems biology*. Notably, "system biology" can be defined as an holistic molecular analysis of cellular networks, relying on the integration of different experiments, data processing, and modelling [17, 109–111]. Eventually, the ambitious goal of systems biology is the understanding of an entire biological system by modelling, predicting, and controlling the behaviour of all its components.

## 1.1 Motivations

A major topic of systems biology is the modelling and analysis of networks: gene networks, protein interaction networks, signalling networks, metabolic networks, etc [111]. In this regard, we developed the main content of this thesis, which focuses on modelling, analysis and control of gene regulatory networks (GRN) using different tools from Control Theory.

Models of GRN provide significant insights into the underlying molecular mechanisms, and could lead to important biotechnological applications. They allow researchers to dissect the role of components of a given system, so as to give precise definitions of the functions exerted by system components and their interactions. Moreover, they can be easily used for system simulations to test different scenarios that are not accessible by experiments. However, we have to take into account that biological systems have particular constraints, which often give rise to new mathematical problems for which classical methods are not applicable [158]. In fact, the type of data available for the system is one of the first aspects to consider in modelling and analysing of GRN. Currently, there are many different biochemical experimental techniques, which range from qualitative (e.g.. micro-array data) to quantitative (e.g reporter genes or fluorescence proteins). Therefore, according to the type/quality of data, modellers have to choose the more suitable mathematical formalisms to describe the given biological process: for instance a choice between discrete/continuous, deterministic/stochastic (see Chapter 3). Moreover, a second aspect which has to be considered is the purpose of the models, that is the particular phenomenon or aspect of the biological system that is intended to study. Hence, the mathematical part can be kept as simple as possible to facilitate the implementation and the analysis, so as to have faster and more comprehensible results. Or the system can be modelled in much more detail in order to have a very realistic model, but then it can be very complicated to deal with. Of course, none of the considerations mentioned above makes a model wrong or right, but they determine whether a model is able to solve the problem for which it has been developed [111]. In addition, model analysis is strongly related to the particular mathematical formalism employed to formulate the model. This is because the mathematical framework will also determine the types of methods that can be used to analyse the model.

Along with measurements (outputs), there are also external actions (inputs) that can be applied to the system. As we have seen for outputs, also inputs (terminology taken from Control Theory) strongly depend on and are limited by the experimental techniques currently available. For instance, possible inputs that can be applied to a biological system are chemicals gradients, mechanical and electrical stimulations, optical excitations, and genetic type inputs (inducible engineered promoters) [117]. Input-output

approaches play a fundamental role when the objective is that of controlling biological systems. Intracellular functions can basically be controlled in two fashions: by open-loop or closed-loop approaches. Open-loop methodology is employed by researchers by altering the functions of the cells, for instance through inducers, drugs, toxins, and observing the corresponding outputs; there is no "closing" of the systems by feeding back the response to determine a new input for further system alterations [76, 96, 160]. The main goal of open-loop approaches in biology is that of testing intracellular functions with respect to altered dynamics. Moreover, open-loop gene circuits have strongly contributed to the emerging of the *synthetic biology* discipline. Synthetic biology aims at designing and constructing gene networks from basic biochemical components [21, 107, 131]. The most difficult and intriguing challenge in applying control theory to biology is not simply to understand how a biological system controls itself, but also to figure out how to exert dynamic control on a system. Implementing a closed-loop control in cells is a very challenging task because it requires real-time measurements, possibly without interfering too much with the biological systems, and controllers that, using the readouts, modify the biological system state accordingly [117]. Examples of recent closed-loop controls applied to GRNs are [128, 169].

Along these lines, we applied modelling, analysis and control techniques to the gene expression machinery (GEM) of the bacterium *Escherichia coli* (*E. coli*), with the aim of dynamically linking gene expression and cell growth rate so as to develop possible (theoretical) controllers for the bacterial growth [41]. *E. coli* is a model organism that is easy to manipulate and, being one of the most widely used biological models, much knowledge is available about its regulatory networks [118]. *E. coli*, in the presence of a carbon source—glucose being the preferred sugar—grows in an exponential manner until it exhausts the nutrient sources, and then enters a stationary phase with practically zero growth. The wild-type bacteria grow at different rates in the presence of carbon sources of different types: fastest growth on glucose and a large range of slower growth rates on alternative carbon sources. While many experimental and theoretical studies have addressed the regulation of the growth rate [24, 59, 74, 136], no attempts have been made to modify these control mechanisms in a directed way.

From the biological point of view, the topics of the physiological regulation of cellular growth and the functioning of the gene expression machinery have been extensively studied in the literature (e.g., [29]). However, the dynamic intertwining of the two, and the redesign of the system to externally control the growth rate, do not seem to have been much explored. Several modelling attempts opened the way for understanding the relation between the gene expression machinery and the control of cellular growth in a fixed environment. Some models focus on specific components of the machinery, for instance,

the growth-dependent changes of RNA polymerase partitioning [30, 113] or the relationship between ribosomal RNA (rRNA) synthesis and the growth rate [59]. Recent models have provided a more global view of this system, with for instance a metabolic network reconstruction of the transcriptional and translational machinery [162] and a biophysical description of the gene expression machinery [159]. All these models show a relation between steady state exponential growth on the one hand, and the free RNA polymerase concentration and the synthesis rate of rRNAs on the other hand. However, they do not consider the possibility of externally controlling some of the parameters affecting growth. The present thesis intends to fill in this gap, by constructing minimal dynamical models of the multiple connections between growth control and the functioning of the gene expression machinery. We will not develop our models from scratch: we will build upon the existing work mentioned above and also, take into account preliminary (unpublished) results from partners (Ibis team, Inria Grenoble).

To recap, the main objectives of this thesis are those of developing new modelling formalisms as well as new control strategies for GRN, by applying or adapting classical tools from Control Theory to Systems and Synthetic Biology.

## 1.2 Contributions

After two introductory chapters on fundamentals of molecular cell biology (Chapter 2) and mathematical modelling formalisms of GRNs (Chapter 3), we mainly address—in the first part of the thesis—the problem of theoretically controlling the cell growth rate. Notably, we developed two novel qualitative formalisms to model and dynamically link gene expression and bacterial growth rate. These qualitative formalisms are derived from piecewise linear systems (PL, see Chapter 3 and references therein), from which we kept the use of step functions to model the interactions between the elements of the GRN, and added (and tested) two mathematical descriptions of growth rate to model the dilution effect.

Notably, in Chapter 4 we develop and analyse an open-loop model of a minimal synthetic gene circuit, that describes part of the gene expression machinery (GEM) in *Escherichia coli*. Here we assume that the bacterial growth rate is proportional to the concentration of a component of the GEM (CGEM) (e.g., RNAP, ribosome). This model is a piecewise non-linear system with two variables (the concentrations of CGEM and another protein) and an input (an inducer) which controls the CGEM expression. We study the qualitative dynamics of the model and the bifurcation diagram with respect to the input. Moreover, an analytic expression of the growth rate during the exponential phase

as function of the input is derived, and an identifiability analysis of its parameters is pursued using artificial data.

In Chapter 5 we continue studying and developing the synthetic gene circuit model introduced in Chapter 4, for which also a closed-loop configuration is developed. This closed-loop model, achieved by appropriately rewiring and designing the gene interactions, mathematically reproduces the *inverse diauxie* phenomenon in an appropriate range of the input values. This means that the designed feedback law theoretically allows modified *E. coli* to grow slower on glucose than on a less preferred carbon source (e.g.. maltose), inversely to the normal (wild-type) bacterial behaviour.

In Chapter 6, we further improve the qualitative modelling formalism by introducing a new expression for the growth rate. That is, we now model the growth rate as the weighted minimum of two limiting gene products responsible for bacterial growth (RNAP and ribosome). The resulting system is a switched system with two piecewise quadratic (PQ) modes. Moreover, we propose and analyse a bi-dimensional SPQ open-loop model—describing the bacterial gene expression machinery—in which the growth rate of cells can be controlled by an external input which tunes the synthesis of one of the variables.

In Chapter 7 we use some recent Boolean tools to compute attractors of Boolean GRNs, with the aim of testing several bacterial growth rate models depending on the GEM. Notably, some of the growth rate models tested here are Boolean versions of those growth rate functions introduced in Chapters 4-5-6, whose plausibility have been confirmed. The discrete Boolean framework permits easier testing of different combinations of biochemical interactions, leading to hypotheses elimination and model discrimination, and thus providing useful insights for the construction of a more detailed dynamical growth model.

Then, we moved from qualitative to quantitative modelling of GRNs. In Chapter 8 we present an ODE model of *E. coli* gene expression machinery developed with our collaborators in Grenoble (Inria Ibis team). Such ODE model has been calibrated using parameter values from the literature, and estimating the remaining ones from published experimental data. A reduced version of the model, obtained by assuming quasi-steady state equilibrium of mRNA species, is also presented. In fact, a low-dimensional, but still quantitative model, for which the parameters are known, can be used as starting point to design and study—*in silico*—possible growth rate control strategies. In this regard, the model can be easily extended considering the network motifs and dynamical growth rate expressions of Chapters 4-5-6, which qualitatively describe possible synthetic gene modifications, allowing the control the growth rate of the cells.

Finally, in Chapter 9 we switched from population based models, which use deterministic formalisms, to single cell models, where a stochastic approach may be required to better describe the biochemical reactions when species concentrations become too low. Here we address the state estimation problem for gene regulatory networks at the level of single cells. In fact, state estimation is an important tool for many aspects: it allows reconstruction of state variables that cannot be measured directly, it can be used as an intermediate step for identification [33], and plays a central role toward model-based control [169]. Notably, we consider models that include both intrinsic noise, in terms of stochastic dynamics, and extrinsic noise, in terms of random parameter values. We take the Chemical Master Equation (CME) with random parameters as a reference modelling approach, and investigate the use of stochastic differential model approximations for the construction of practical real-time filters. To this aim we consider a Square-Root Unscented Kalman Filter built on a Chemical Langevin Equation approximation of the CME. Then, using arabinose uptake regulation in *Escherichia coli* bacteria as a case study, we show that performance is comparable to that of a (computationally heavier) particle filter built directly on the CME, and that the use of information about parameter uncertainty allows one to improve state estimation performance.

Conclusions to these research works as well as some perspectives can be found in Chapter 10.

# Chapter 2

# Notes on Molecular Cell Biology

In this chapter we briefly introduce the most important concepts on molecular cell biology that we used throughout the thesis. Notably, we have mainly focused on gene expression, its regulation, and on some techniques used to measure gene products. For more details on the molecular biology of the cell we remand to [12, 13].

## 2.1 The Cell

All living organisms are made of **cells**. Cells are small units (mostly $1-100\ \mu m$), enclosed by a membrane and filled with a concentrated aqueous solution of chemicals. Each cell posses the same genetic information of the parent organism and this information, stored in DNA, is passed on to the daughter cells during cell division.

Organisms may consist of just one cell, and they are called *unicellular*, or they may be *multicellular*. Multicellular organisms are typically organized into tissues, which are groups of similar cells arranged so as to perform a specific function in addition to the *housekeeping* processes common to all cells.

In this thesis we will not address cell differentiation, i.e. formation of cell types in a multicellular organism, but we will only discuss the general (housekeeping) aspects of cell components and functions.

### 2.1.1 Prokaryotes and Eukaryotes

Cells are divided into two categories depending on the way the genetic material (DNA) is organized within them.

The first category is composed of *prokaryotes* which—by definition—are organisms whose cells do not have a nucleus nor other well-defined compartments (see Figure 2.1). Most prokaryotes are single-celled organisms, although some join together to form chains, clusters or other multicellular structures [12]. In prokaryote cells DNA is stored in the cytoplasm in an area called *nucleoid*, but it is not enclosed within a separate nuclear envelope.

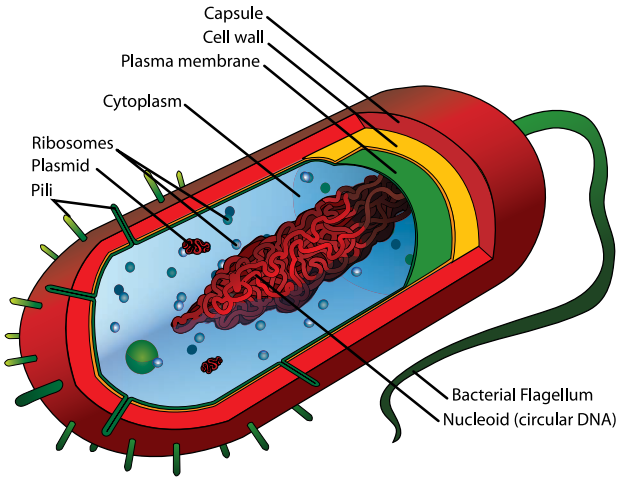



FIGURE 2.1: **Schematic of a prokaryote cell**. In a prokaryotic cell, all their intracellular components (proteins, DNA and metabolites) are located together in the same volume enclosed by the cell membrane. Many prokaryotes (bacteria) are able to move in a fluid-like environment using flagella, which are also used as sensors to detect concentration gradients and other signals. (Picture taken from [1]).

*Eukaryotes* belong to the second category and they can be defined as organisms whose cells have a nucleus. Eukaryotic cells, in general, are bigger and more elaborate than prokaryotes. They range from unicellular yeast to plants and animals, which are very complex multicellular organisms with billions of cells. Eukaryotes, in addition to a nucleus, have other organelles, sub-cellular structures that carry out specialized functions (see Figure 2.2). For examples, mitochondria are responsible for energy production through metabolism, and containing a very small amount of DNA; chloroplasts (plants) for photosynthesis; ribosomes serve as machinery for protein synthesis, and made up themselves of proteins and RNAs; endoplasmatic reticulum; and so forth. The cytoskeleton, made up of micro-tubules and filaments, controls cell shape, drives and guides cell movements and plays a role in intra-cell substance transport.

Since in this thesis I mainly focus on bacteria, in what follows I will introduce the bacterium *E. coli*, which is considered by biologists as the model organism for prokaryotic cells and we will mainly concentrate on prokaryotic cell functions.

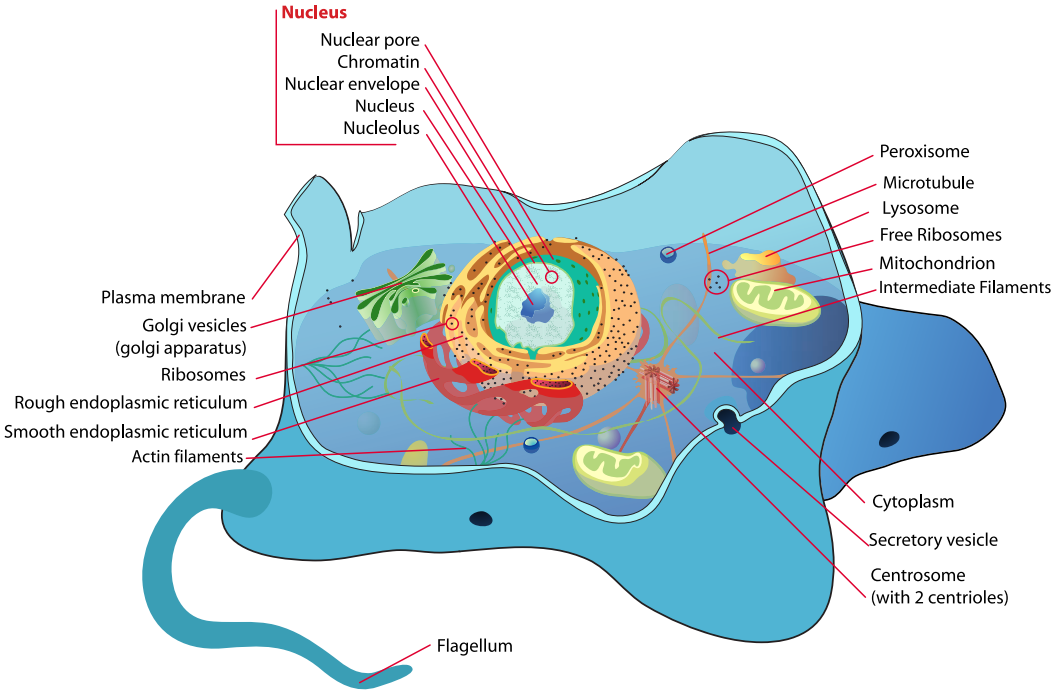



FIGURE 2.2: Eukaryotic (animal) cell. The nucleus is the most prominent organelle in the cell and contains chromosomes (the storage sites of DNA). Mitochondria produce chemical energy (ATP) for the cell. Centriolis are involved in nuclear division during cell division. Ribosomes, the endoplasmatic reticulum and the Golgi apparatus work together in the synthesis of proteins. (Picture taken from [2]).

### 2.1.2 *E. coli* as model organism

It is thought that all cells descended from a common ancestor [12]. Hence, the knowledge gained from the study of one organism allows us to better understand others, even ourselves. But some organisms are more convenient than others to study in the laboratory. This is because some are easier to genetically manipulate and reproduce faster; others are multicellular but transparent and so biologists can easily watch the development of their tissues and organs.

Molecular biologists have focused on *Escherichia coli* (*E. coli* for short) as a model organism for prokaryotic cells. *E. coli* is a small, rod-shaped bacterium that normally lives in the gut of humans and other vertebrates, but it can be grown easily in a simple nutrient broth in a culture bottle. *E. coli* is able to grow in variable chemical conditions and it reproduces rapidly (approximately one generation in 20 minutes). The bacterium *E. coli* was one of the first organisms to have its complete genome sequenced [25]. Its genetic information is stored in a single, circular double-stranded molecule of DNA, approximately 4.6 million nucleotide pairs long, and it makes 4300 different proteins.

The molecular functioning of *E. coli* is better understood than any other organisms and most of our knowledge of the fundamental mechanisms of life (how cells replicate their DNA, how they decode these genetic instructions to make proteins, etc.) has come from studies on it. In fact, although human cells are eukaryotic cells, subsequent research has confirmed that basic molecular processes occur in the same way both in human and in *E. coli* cells [12].

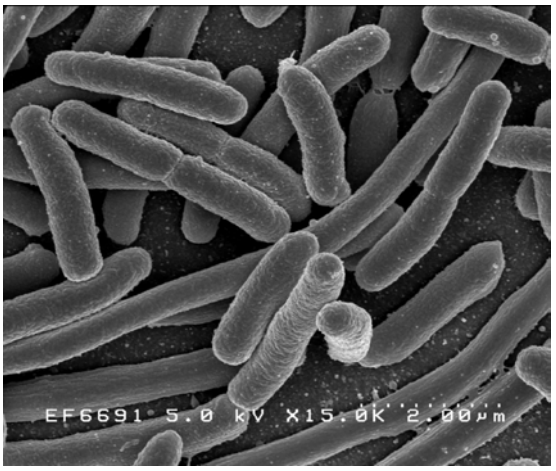



FIGURE 2.3: **Escherichia coli**: Scanning electron micrograph of *Escherichia coli*, grown in culture and adhered to a cover slip. (Picture taken from [3]).

## 2.2 Gene expression: from DNA to Protein

The *central dogma of molecular biology* says: "DNA makes RNA, RNA makes protein, and proteins make the cell" [50]. This key paradigm of molecular biology states that the flow of information in gene expression is from genes encoded by DNA to mRNA by transcription and from mRNA to protein by translation (see Figure 2.4). At any given time, and in any given cell of an organism, thousands of genes and their products (RNA, proteins) actively participate in an orchestrated fashion to generate the macromolecular machinery for life.

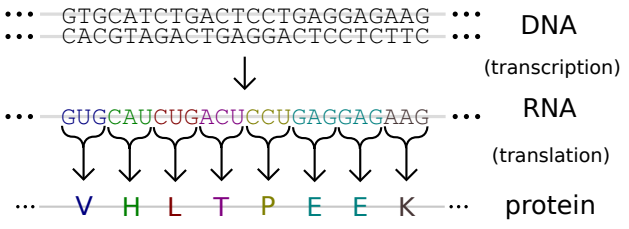



FIGURE 2.4: Diagram of the central dogma, DNA to RNA to protein, illustrating the genetic code. (Picture taken from [4]).

### 2.2.1 Transcription: from gene to RNA

The *genome*, i.e. the genetic information of an individual, describes all the proteins that are potentially present in every cell of a given organism. This information is encoded in the DNA molecule, which is a double-stranded helix made of alternating sugars (deoxyribose) and phosphate groups (related to phosphoric acid), with the nucleobases (guanine(G), adenine(A), thymine(T), and cytosine(C)) attached to the sugars (see Figure 2.5).

The sugar-phosphate backbones of the two DNA strands form a uniform helix, with strands placed in opposite directions. The strands are held together by hydrogen bonds between opposing bases according to the base pair rule: A is always paired with T and G is always paired with C.

Within cells, DNA is organized into long structures called chromosomes. During cell cycle these chromosomes are duplicated in the process of DNA replication, providing each cell its own complete set of chromosomes.

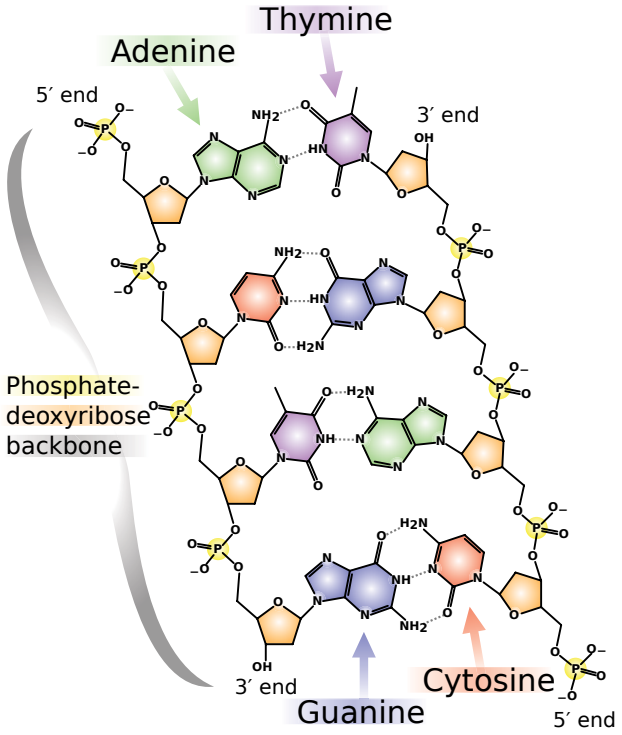



FIGURE 2.5: Chemical structure of DNA, with coloured label identifying the four bases as well as the phosphate and deoxyribose components of the backbone. (Picture taken from [5]).

The first step in the synthesis of protein is **transcription** and it consists in copying the nucleotide sequence of a gene into RNA (*ribonucleic acid*). Like DNA, **RNA** is a polymer made of four different nucleotides . It differs from DNA in three respects:

1. whereas DNA is always a double-stranded helix, RNA is single stranded;
2. the nucleotides in RNA are *ribonucleotides*, i.e they contain the sugar ribose rather than deoxyribose;
3. although, like DNA, RNA contains the bases adenine (A), guanine (G), and cytosine (C), it contains uracil (U) instead of thymine (T) found in DNA.

All of the RNA in a cell is made by transcription. The enzyme that carries out transcription is called **RNA polymerase** (RNAP). RNAP, to begin transcription, must be able to recognize the start of a gene, called **promoter**, and bind steadily to the DNA at this site. Then, RNAP moves stepwise along the DNA, unwinding the DNA double helix to expose the bases on each DNA strand. As RNAP progresses, it adds nucleotides one by one to the RNA chain using an exposed DNA strand as a template. Chain elongation continues until RNAP meets a stop site in the DNA, the *terminator*, where the enzyme halts and releases both the DNA template and newly made RNA chain. The resulting RNA transcript is thus single-stranded and complementary to one of the two DNA strands.

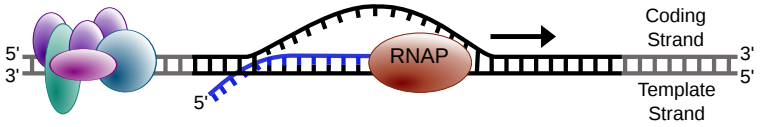



FIGURE 2.6: The process of transcription is carried out by RNA polymerase (RNAP), which uses DNA (black) as a template and produces RNA (blue). (Picture taken from [6]).

Several types of RNA are produced in cells. The majority of genes specify the amino acid sequence of proteins, and the RNA molecules that are transcribed from these genes are called **messenger RNA** (**mRNA**). Moreover, there are also non-messenger RNA: *ribosomal RNA* (*rRNA*) that forms the core of ribosomes, on which the mRNA is translated into protein, and *transfer RNA* (*tRNA*) that selects and carries amino acids to the ribosome for the protein synthesis.

### 2.2.2 Translation: from RNA to protein

The next step in gene expression is called **translation**, because it allows the conversion of the information stored into mRNA to protein. Since there are only 4 different nucleotides

in mRNA but 20 different amino acids in a protein, this translation can not be a direct one-to-one correspondence between a nucleotide in RNA and an amino acid in a protein. The rules by which the nucleotides of a gene, by means of mRNA, are translated into the amino acid sequence of a protein are known as the **genetic code**. Notably, an mRNA sequence is decoded in sets of three nucleotides, called **codons**, thus allowing $4^3 = 64$ possible combination of three nucleotides, even though only 20 amino acids are commonly found in proteins.

The translation of mRNA into protein is due to adaptor molecules that recognize and bind—through base-pairing—to a codon at one site on their surface (called anticodon) and to an amino acid at another site. These adaptors are small RNA molecules (about 80 nucleotides in length) known as **transfer RNAs** (**tRNAs**). Transfer RNAs are captured and hold in position on the mRNA strand by a large molecular machine that moves along the mRNA allowing accurate and rapid translation of the genetic code. This complex molecular machine is the **ribosome**, which is made up of more than 50 different proteins (the *ribosomal proteins*) and several RNA molecules called **ribosomal RNAs** (rRNAs).

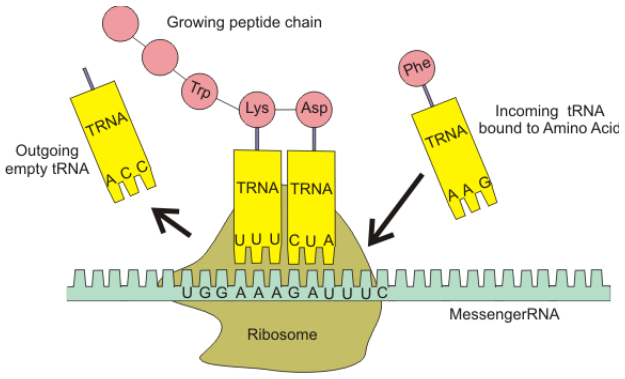



FIGURE 2.7: The process of translation is carried out by ribosome, which uses mRNA and tRNAs—charged with amino acids—to produce a protein. (Picture taken from [7])

## 2.3 Regulation of Gene expression

The regulation of gene expression is the process by which individual cell specifies which of its many thousands of genes have to be expressed. This mechanism is paramount, especially for multicellular organisms, as animals, which have to differentiate their cells in order to produce, for instance, muscle, nerve, blood cells and, eventually, all the variety of cell types seen in the adult [69]. Thus, cell **differentiation** arises because cells produce and accumulate different RNA and protein molecules [12]. But, regulation of gene expression is also widely adopted by prokaryotic/unicellular cells like bacteria. In fact, bacterial cells can change the expression of their genes in response to external signals,

for example, according to the food sources that are available in the environment [58, 133, 134].

Gene expression can be regulated at many steps in the pathway from DNA to RNA to Protein. Moreover, the stability of the final gene product, whether it is RNA or protein, also contributes to the expression level of the gene—an unstable (faster degradation) product results in a lower expression level than a stable one which degraded more slowly.

### 2.3.1 Transcriptional control

Control of transcription is mostly exerted at the initiation step. In Subsection 2.2.1 we saw that RNAP binds to the promoter of a gene to make an RNA copy of the gene. In addition to the promoter, almost all genes have **regulatory DNA sequences** that are used to activate (resp. inhibit) the gene transcription by facilitating (resp. preventing) RNAP binding to the promoter. However, these regulatory DNA sequences—to have any effect—have to be recognized by proteins called **transcription factors**, which bind to DNA. Hence, each transcription factor is able to recognize a different DNA sequence and so regulates only particular genes. Notably, a transcription factor is a **repressor** protein if, in its active form, it blocks the binding of RNAP to the promoter, thus switching genes off. But some transcription factors—called **activators**—do the opposite, that is they switch on some genes by binding nearby the promoter and helping RNAP to initiate transcription.

### 2.3.2 Post-transcriptional control

Post-transcriptional controls operate after RNAP has bound to the promoter of a gene to synthesize RNA. One of the most common ways to regulate gene expression at post-transcriptional level is to control translational initiation, so as to modulate protein synthesis.

Bacterial mRNAs, for example, have a **ribosome-binding site** (**RBS**) where translation begins. These RBS have to be recognized by a ribosome, which binds to it and starts peptide synthesis. Hence, by blocking or exposing the RBS, the bacterium can either inhibit or facilitate the translation of an mRNA.

## 2.4 Measurement Techniques

In this section we will briefly present some techniques used in molecular biology to measure gene expression. Gene expression measurement and analysis have become essential tools for medical investigations and for characterizing complex biological circumstances.

Here, without going into details—which is behind the scope of this thesis—we will list some techniques used to quantify mRNA and protein abundance.

### 2.4.1 mRNA quantification

Several techniques are available to quantify levels of mRNA in a cell, generally referred to as **DNA Microarray** [18, 111]. DNA microarray is a tool that allows the RNA of thousands of genes to be monitored at the same time, so as biologists can visualize which genes are switched on (or off) as cells grow, divide, or respond to hormones, toxins, or infections. The information contained in DNA microarrays say whether the expression of each gene has increased or decreased relative to a reference condition. It is therefore an essentially qualitative measurement.

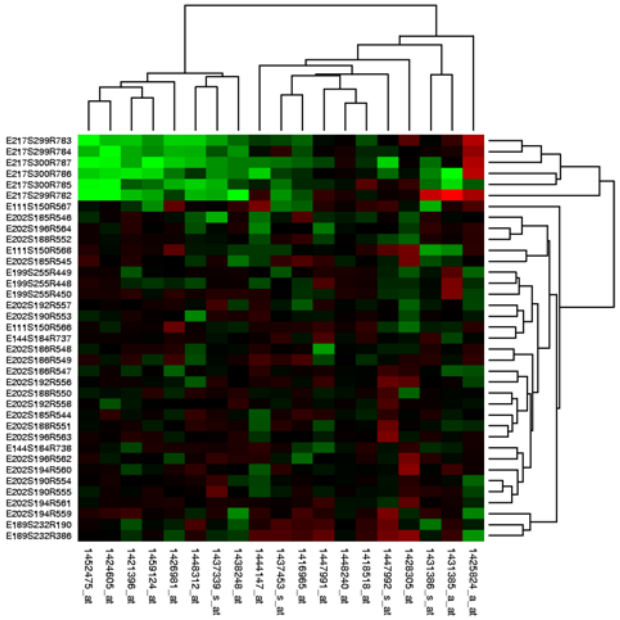



FIGURE 2.8: Heat map generated from DNA microarray data reflecting gene expression values in several conditions. (Picture taken form [8])

### 2.4.2 Protein quantification

The most common method to detect the presence of a specific protein—or a small number of them—in a sample taken from an experiment is the *Western blot* technique. The protein is extracted from the sample and together with a small number of antibodies—which recognize only specific proteins—is transferred to a nitrocellulose membrane. Different methods, for instance radioactive labelling of stains, are then used in order to produce bands, indicating the location of the protein. Finally, the intensity of the band is proportional to the amount of protein [111].

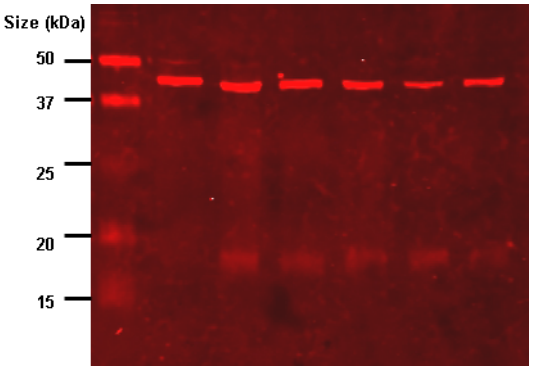



FIGURE 2.9: Western blot using a anti-lipoic acid primary antibody and an IR-dye labelled secondary antibody in Leishmania major extracts. (Picture taken from [9])

Moreover, *reporter gene* techniques can reveal when, where and how much a gene is expressed. Notably, determining the pattern and timing of a gene's expression can be obtained by fusing the regulatory region of the gene of interest to a reporter gene, which is easier to monitor. In this fashion, the amount, the timing, the cell specificity of reporter protein synthesis will reflect the function of the original gene as well as the behaviour of the regulatory sequences that belong to it. In most cases, the expression of the reporter gene is monitored by tracking the fluorescence or the luminescence of its protein product [57].

Nowadays, one of the most used reporter proteins is the **green fluorescent protein** (**GFP**), which traditionally refers to the protein first isolated from the jellyfish *Aequorea victoria*. Once GFP gene has been fused to the end of the gene that encodes a protein of interest, GFP can be monitored simply by following its fluorescence by microscopy.

Fluorescent protein methods are more and more used in combination with *flow cytometry*. Flow Cytometry devices are instruments that, at rates of up to thousands of cells per second, can count individual cells, sort them into different groups, analyse cellular characteristics such as cell size, shape, or quantity of measured fluorescence. In this fashion, it is possible, for instance, to count how many cells in a population synthesize

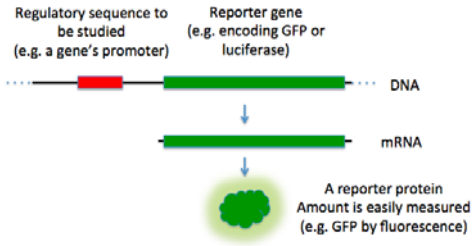



FIGURE 2.10: A diagram of a how a reporter gene is used to study a regulatory sequence. (Picture taken from [10])

a particular protein under a specific set of conditions, and to study stochasticity of gene expression by measuring fluorescence at the single cell level [169].

### 2.4.3 Measurement limitations

Even though all the measurement techniques just presented are of fundamental importance for deciphering and exploiting genetic information, all of them are intrinsically noisy because of chemical interactions in blots, production errors in arrays, or other sources of interference and artefacts. Moreover, microarray and Western blot measurements are affected by low precision because very few bits of information can be extracted from their data.

These limitations of imprecision and noise have to be taken into account in systems biology modelling because it is not always possible to tightly fit model parameters to such data. When this is the case—to overcome these difficulties—qualitative models turn out to be more suitable.

# Chapter 3

# Modelling Genetic Regulatory Network Systems

As we have seen in Chapter 2, the genome of an organism plays a fundamental role in the control of almost all cellular processes. Hence, to elucidate the functioning of organisms on the molecular level, we need to know which genes are switched on, which proteins are synthesized, when and how transcription factors interact and bind to regulatory sites of other genes.

DNA, RNA, proteins, and other small molecules are all interconnected through positive and negative feedback loops, forming the so called **genetic regulatory networks** (**GRN**). Since an intuitive understanding of the dynamics of such GRN—which could involve many components—is difficult to achieve, the help of mathematical and computer tools for modelling and simulation of genetic regulatory networks is indispensable [54, 94, 157].

This chapter reviews some formalisms that have successfully been employed in mathematical biology to model genetic regulatory systems. Notably, it will deal with gene expression modelling techniques without considering those cellular processes, such as ion channels, membrane electrical behaviour, transporters and pumps, cell cycle, etc., that might be modelled as well (see examples in [71]), but that are beyond the scope of this thesis.

Specifically, this chapter focuses on those gene expression modelling formalisms that will be used to develop the most important findings of this thesis, i.e. Boolean networks (Chapter 7), ordinary differential equations (Chapter 8), qualitative differential equations (Piecewise Linear Equations) (Chapters 4-5- 6), and stochastic equations (Chemical Master Equation and Chemical Langevin Equation) (Chapter 9).

To illustrate the advantages and disadvantages of the various formalisms, we use the *mutually inhibiting* two genes network—also called *bistable switch*—as a representative example (see Figure 3.1).

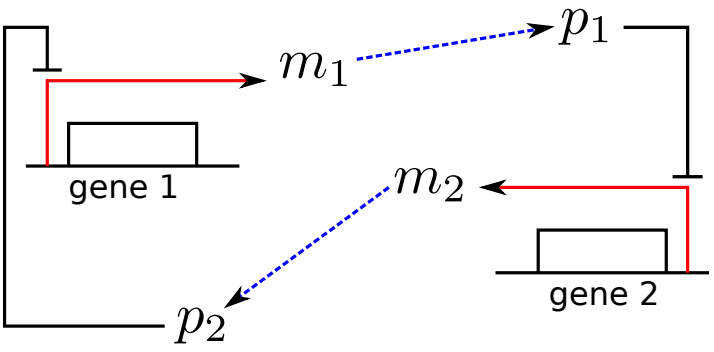



FIGURE 3.1: Example of a gene regulatory network, representing the *bistable-switch* gene circuit. It consists of four species: proteins $p_1$, $p_2$ and mRNAs $m_1$, $m_2$. Solid red arrows denote transcription processes while dashed blue arrows denote translation processes. Protein $p_1$ acts as inhibitor for gene 2, reducing synthesis of mRNA $m_2$ whereas $p_2$ acts as inhibitor for gene 1, reducing synthesis of mRNA $m_1$.

## 3.1 Boolean Models

Modelling GRNs with the Boolean framework has first been introduced by S. Kauffman and R. Thomas [85, 104, 163, 164]. Since then, relevant theoretical and applied research on the Boolean formalism has been pursued [141–143]. Moreover, the Boolean framework has turned out to be a very useful tool when dealing with large gene regulatory networks (10 or more variables) as it is successfully reported in [11, 34, 53, 72, 103, 119, 120].

As a matter of fact—at the highest level of abstraction—the state of a GRN component (gene, mRNA, protein, etc) can be represented by a Boolean variable which can only take two values: 0 or 1. In addition, these variables, or nodes, are interconnected by a set of *logical rules* (AND, OR, NOT etc.), which orchestrate their dynamical behaviour (see, for instance, [175] for a review on Boolean models in systems biology). Boolean variables are particularly suitable for modelling gene expression when experimental data are scarce and/or qualitative. In fact, in these cases, one can assume that a Boolean variable is 0 (resp., 1) when the expression of a gene or protein is weak (resp., strong).

Boolean variables will be denoted by $X = (X_1, ..., X_n)$ with the state space $\Omega = \{0, 1\}^n$, and also the time at which the system evolves is assumed to be discrete, $0 < t_1 < ... < t_{max}$, with $t_{max} \leq +\infty$. As previously said, the dynamics of a Boolean network is dictated by a set of logical rules, $\{F_i(X) : \Omega \to \{0, 1\}, i = 1, ..., n\}$ which, given the state

at $t_k$, determine the state at $t_{k+1}$: $X_i(t_{k+1}) = F_i(X(t_k))$ or, in a more compact way:

$$X_i^+ := F_i(X), \quad i = 1, ..., n.$$

Finally, to dynamically simulate a Boolean network, one must specify the mode of updating the variables of the system, that is how values of the nodes progress in time. Several updating algorithms have been developed, but the two most employed are briefly discussed next.

### 3.1.1 Synchronous and Asynchronous networks

One of the most common strategies to update the system and obtain its trajectories is that of assuming that *all variables are simultaneously updated*. This means that, at each instant $t_{k+1}$ all variables are changed to their new values:

$$\forall k > 0, \ \forall i = 1, ..., n \quad X_i(t_{k+1}) = F_i(X_i(t_k)).$$

The corresponding networks are, hence, called *synchronous Boolean networks*.

However, since the timescales of different biological processes can vary widely (transcription is generally faster than translation, and complex binding is even faster; see, for instance, some timescales in [14]), the synchronous updating is not always the more realistic choice in systems biology. To overcome this issue, a more general updating algorithm assumes that, at each time instant, only one node is updated following its logical rule, that is:

$$\forall k > 0, \ \exists! j \in 1, ..., n \quad X_j(t_{k+1}) = F_j(X_i(t_k)) \ and \ X_i(t_{k+1}) = X_i(t_k) \ i \neq j.$$

In this case, the corresponding networks are called *asynchronous Boolean networks*.

For both strategies, the trajectories of the Boolean network consist of a sequence of transitions among the $2^n$ states in $\Omega$. Notably, a transition between two states $V$, $W \in \Omega$ occurs if $V = X(t_k)$ and $W = X(t_{k+1})$, and in this case $W$ is called the *successor* of $V$. More generally, the set of all asynchronous successors of $X$ is given by:

$$\sigma(X) = \{\bar{X} \in \Omega : \ \exists j \bar{X}_j = F_j(X) \text{ and } \bar{X}_i = X_i \ i \neq j\}.$$

All possible trajectories can also be represented by a *directed graph* with $2^n$ vertexes (the cardinality of $\Omega$), with an edge connecting two vertexes whenever one state is the successor of the other. It is worth noting that there is an essential difference between

the directed graphs corresponding to synchronous or asynchronous networks. Any given sate of a synchronous network can have at most only one successor, hence giving rise to a deterministic updating, whereas the state of an asynchronous network can have up to $n$ successors, i.e. potentially leading to several different trajectories from any given state.

### 3.1.2 Graph theoretical representation

Here, we focus only on asynchronous Boolean networks since, as we have said above, they permit a more realistic interpretation of the different timescales present in biological processes. The associated directed graph will be called the *asynchronous transition graph*, and its properties can be studied using graph theory and its application [27, 176]. To this aim, a directed graph can be decomposed into **strongly connected components** (**SCCs**), which are the maximal subsets of vertexes where every pair is *mutually reachable*. Notably, two vertexes are said to be mutually reachable if there are directed paths linking one vertex to the other. SCCs can be made up of a single state, or several states, or even the whole state space for some special models. Hence, SCCs can have outgoing paths directed towards (states contained in) other SCCs. An SCC that contains no outgoing path is called a *terminal SCC* or *attractor*, since any trajectory that reaches it cannot leave. To conclude, identifying and analysing the SCCs of a Boolean network facilitates the identification of the attractors, which describe the asymptotic behavior of the system.

### 3.1.3 Example: Boolean bistable switch

The Boolean model corresponding to the bistable switch circuit depicted in Figure 3.1 reads as:

$$\begin{aligned} m_1^+ &= NOT\ p_2 \\ p_1^+ &= m_1 \\ m_2^+ &= NOT\ p_1 \\ p_2^+ &= m_2 \end{aligned} \tag{3.1}$$

To construct the asynchronous transition graph of model (3.1), one proceeds as follows:

1. for each state $X = (m_1, p_1, m_2, p_2)$, compute the possible variable changes from the synchronous Boolean table using (3.1);
2. then consider only one variable change at time, to obtain all the successors of $X$, i.e. $\sigma(X) = \{Y_1, \ldots, Y_\ell\}$ and draw an edge from $X$ to each $Y_i$.

For example:

$$X = (1, 0, 0, 1) \quad \Rightarrow \quad X^+ = (0, 1, 1, 0)$$

so the possible successors $Y_i$ are:

$$Y_1 = (0, 0, 0, 1), \quad Y_2 = (1, 1, 0, 1), \quad Y_3 = (1, 0, 1, 1), \quad Y_3 = (1, 0, 0, 0)$$

by allowing only one variable to change at each time. Hence, the corresponding asymptotic transition graph of model (3.1) is shown in Figure 3.2. In Figure 3.2 we can easily recognize the two attractors corresponding to states $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$, denoting either gene and protein 1 are ON, or gene and protein 2 are ON, respectively.

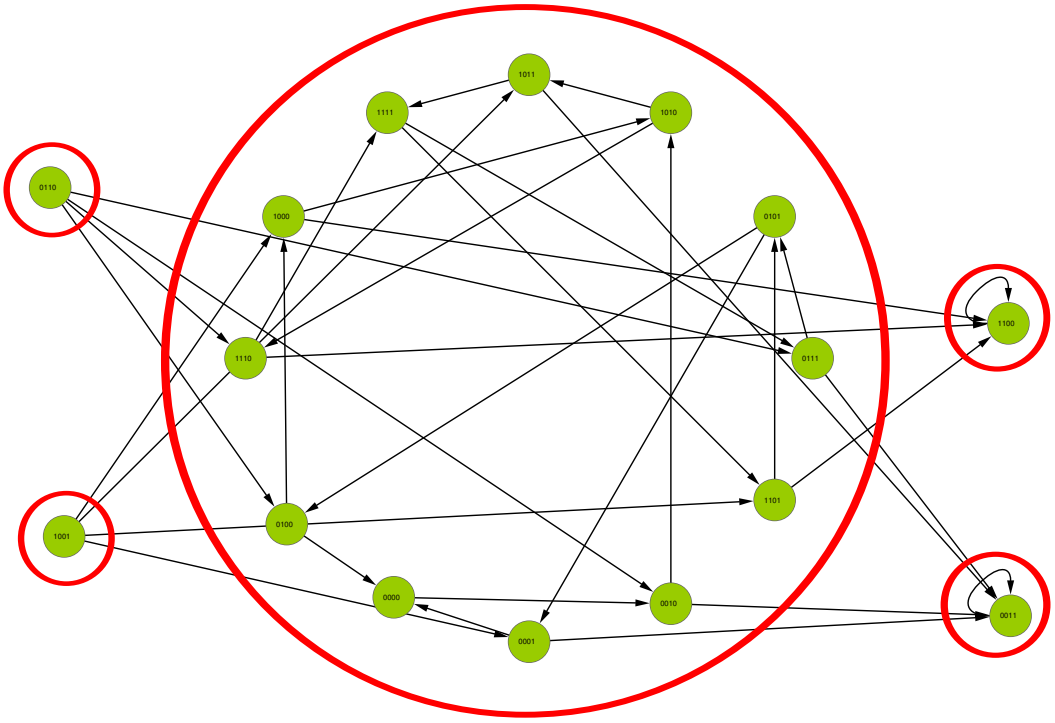



FIGURE 3.2: Asymptotic transition graph of the *bistable-switch*. There are five SCCs (states inside the red circles) and two attractors: $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$, denoting either gene and protein 1 are ON, or gene and protein 2 are ON, respectively.

## 3.2 Ordinary Differential Equation (ODE) Models

Very likely, the most used formalism for modelling gene regulatory networks is that of ordinary differential equations (ODEs). These models can generally be studied and analysed using tools developed for nonlinear systems, in order to investigate dynamics, bifurcation behaviour, system stability [65, 108]. Example of biological models involving the ODE formalism can be found in [65, 95, 105].

More specifically, when ODEs are used to model gene expression, the cellular concentration of proteins, mRNAs and other molecules are represented by *non-negative* continuous time variables. Notably, to model the typical transcription-translation process, the ODEs formalism makes use of two equations for any given gene $i$: one (3.2) to model the dynamics of the transcribed mRNA concentration and the other (3.3) for the concentration of the corresponding translated protein. Hence, given a GRN of $n$ genes, and let $m_i$, $p_i$ be the concentrations of mRNA and protein for the associated gene $i$, respectively, we have [139]:

$$\frac{dm_i}{dt} = F_i(f_i^R(p_1), f_i^R(p_2), ..., f_i^R(p_n)) - \gamma_i m_i \tag{3.2}$$

$$\frac{dp_i}{dt} = f_i^P(m_i) - \delta_i p_i \tag{3.3}$$

for $i = 1, ..., N$. Both equations (3.2), (3.3) are made up of two terms: one positive and one negative. In (3.2), the positive term $F_i(f_i^R(p_1), f_i^R(p_2), ..., f_i^R(p_n))$ denotes mRNA synthesis rate whereas the negative part $\gamma_i m_i$ stands for mRNA degradation rate. Similarly, in (3.3) $f_i^P(m_i)$ models the translation rate of mRNA $m_i$ into protein $p_i$, and $\delta_i p_i$ represents protein degradation rate. Notably, the generic function $f_i^R(p_j) : \mathbb{R} \to \mathbb{R}$ in (3.2)—usually nonlinear—describes how protein $p_j$ regulates the synthesis of $m_i$. If protein $p_j$ has no effect on $m_i$, then $f_i^R(p_j) = 0$. The function $F(\cdot)$ in (3.2) is a combination of sums and/or products of functions $f_i^R(p_j)$, describing how transcription factors $p_i$ regulate mRNA synthesis (ex. in an additive and/or multiplicative fashion).

Functions $f_i^R(p_j)$, depending on the specific transcription regulation, can follow the *law of mass action* [95, 97], the Michaelis-Menten enzymatic kinetics [153], or—as some experimental evidences suggest—can be represented by monotonic sigmoidal-shaped functions [179, 180] which increases when $p_j$ is an activator and decreases when $p_j$ is an inhibitor. A function—satisfying this property—which is widely employed in systems biology is the *Hill function* [14]. A Hill function can be *increasing*, and hence acting as an activator, or *decreasing*, if is modelling an inhibition. The *increasing Hill function*, $h^+(p_j; \theta_j, n_j) : \mathbb{R}_{\geq 0} \times \mathbb{R}^2_{>0} \to \mathbb{R}_{\geq 0}$, is an increasing function in $p_j$ with two parameters, $\theta_j$ and $n_j$, which starts from zero and approaches unity:

$$h^+(p_j; \theta_j, n_j) = \frac{p_j^{n_j}}{p_j^{n_j} + \theta_j^{n_j}} \,. \tag{3.4}$$

The parameter $\theta_j$ indicates the *threshold* of protein concentration $p_j$ needed to produce a relevant increase in mRNA synthesis (regulated by $p_j$). The parameter $n_j$ is referred to as *Hill coefficient* (or *cooperative coefficient*) and it controls the steepness of the Hill

function (the larger $n_j$, the more step-like is the Hill function). Moreover, in some cases $n_j$ could also have a biological meaning, it denotes the number of protein molecules required for binding to the DNA [177]. The *decreasing Hill function* $h^-(p_j;\theta_j,n_j)$ : $\mathbb{R}_{\geq 0}\times\mathbb{R}^2_{>0}\to\mathbb{R}_{\geq 0}$, is a decreasing function given by:

$$h^-(p_j;\theta_j,n_j)=1-h^+(p_j;\theta_j,n_j)=\frac{\theta_j^{n_j}}{p_j^{n_j}+\theta_j^{n_j}}\ . \tag{3.5}$$

Finally, the translation function $f_i^P(m_i)$ in 3.3 is often represented as a linear term proportional to the mRNA concentration $m_i$.

Because of the nonlinearity of the Hill functions, a GRN involving many genes can not generally be solved analytically, and so model reduction/approximation techniques and computer tools are necessary.

### 3.2.1 Quasi-steady-state assumption of mRNA concentration

Very often in the literature, when modelling gene expression, it is assumed that the main gene expression regulation is at the transcriptional level. This hypothesis stems from the fact that—in some GRN—the mRNA dynamics is much faster than protein dynamics, i.e. mRNA concentration reaches its equilibrium faster than that of protein. This is oftentimes due to the fact that $\gamma_i \gg \delta_i$ , that is mRNA degrades much faster than protein (typical mRNA half-lives are $2-6$ minutes, while those of proteins are on the order of hours [14]).

Mathematically, this assumption says that $\varepsilon_i\frac{dm_i}{dt}=0$, for a small positive constant $\varepsilon_i \ll 1$, in 3.2 (see *Tikhonov's theorem* for more mathematical details in [108]) and so:

$$m_i=\frac{1}{\gamma_i}F_i(f_i^R(p_1),f_i^R(p_2),...,f_i^R(p_n)). \tag{3.6}$$

Then, substituting (3.6) into (3.3) we obtain a reduced gene expression model involving only the protein concentration for each gene:

$$\frac{dp_i}{dt}=f_i^P\left(\frac{1}{\gamma_i}F(f_i^R(p_1),f_i^R(p_2),...,f_i^R(p_n))\right)-\delta_i p_i\ . \tag{3.7}$$

### 3.2.2 Example: ODE bistable switch

Now, we present the full ODE model corresponding to the bistable switch circuit depicted in Figure 3.1. The concentration of mRNA produced by *gene 1* (resp. *gene 2*) is denoted

by $m_1$ (resp. $m_2$) and the corresponding protein concentration is denoted by $p_1$ (resp. $p_2$). The inhibition of *gene 1* (resp. *gene 2*) by protein $p_2$ (resp. $p_1$) is modelled by the decreasing Hill function, $h^-(p_2;\theta_2,n_2)$ (resp. $h^-(p_1;\theta_1,n_1)$). The translation of mRNAs and the degradation of mRNAs and protein are all modelled by linear functions. Based on the above, the ordinary differential equations describing the system's dynamics read as:

$$\begin{aligned}
\dot{m_1} &= v_1 h^-(p_2;\theta_2,n_2) - \gamma_1 m_1 \\
\dot{p_1} &= \bar{k}_1 m_1 - \delta_1 p_1 \\
\dot{m_2} &= v_2 h^-(p_1;\theta_1,n_1) - \gamma_2 m_2 \\
\dot{p_2} &= \bar{k}_2 m_2 - \delta_2 p_2
\end{aligned} \tag{3.8}$$

A reduced model can be obtained by assuming that the mRNA dynamics are extremely fast when compared to the protein dynamics (realistic biological assumption) and hence reach their equilibrium instantly. Assuming quasi-steady-state mRNA concentrations (see Section 3.2.1) for the inhibition–inhibition network of Figure3.1, the dynamics can be described by just two variables, i.e. $p_1$ and $p_2$. More precisely, if we assume that $\dot{m_1} \approx 0$ and $\dot{m_2} \approx 0$ the mRNA ODEs in (3.8) yield:

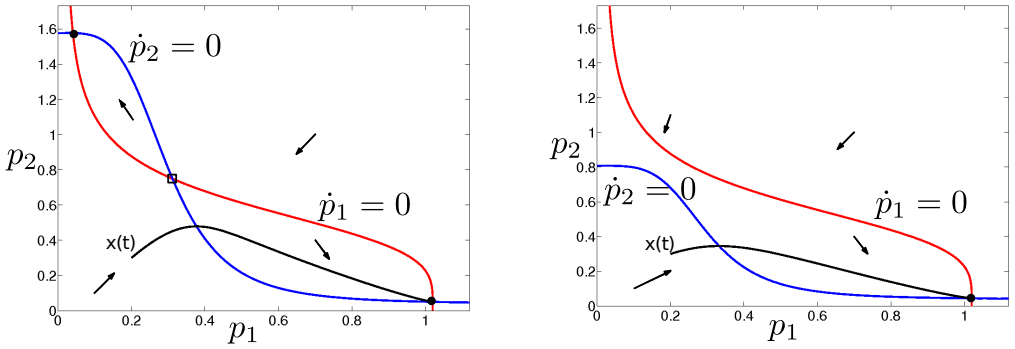
$$\begin{aligned}
m_1 &= \frac{v_1}{\gamma_1} h^-(p_2;\theta_2,n_2) \\
m_2 &= \frac{v_2}{\gamma_2} h^-(p_1;\theta_1,n_1)
\end{aligned} \tag{3.9}$$

and substituting (3.9) into $\dot{p_1}$ and $\dot{p_2}$ in (3.8) one gets:

$$\begin{aligned}
\dot{p_1} &= k_1 h^-(p_2;\theta_2,n_2) - \delta_1 p_1 \\
\dot{p_2} &= k_2 h^-(p_1;\theta_1,n_1) - \delta_2 p_2
\end{aligned} \tag{3.10}$$

where $k_1 = \frac{v_1}{\gamma_1}\bar{k}_1$ and $k_2 = \frac{v_2}{\gamma_2}\bar{k}_2$.

A dynamical analysis of this model is not the main scope of this thesis and details can be found, for instance, in [22]. Here, we would just recall that the dynamics of system (3.10) depends on the values of the parameters. Notably, system (3.10) can show two distinct dynamical scenarios: the first, in which there are two stable and one unstable steady states; the second, in which there is a unique stable steady state. The nullclines and steady states of the two scenarios mentioned above are shown in Figure 3.3.




(a) (b)

FIGURE 3.3: Phase planes for the system (3.10) for two sets of parameter values: (a) bistability, or (b) single steady. Stable steady states are marked by a black circle, and the unstable steady state by an open rectangle In each region delimited by the nullclines ($\dot{p}_1 = 0$ in red, $\dot{p}_2 = 0$ in blue), the sign of the vector field of each coordinate is unchanged. One solution, i.e. $x(t) = [p_1(t), p_2(t)]$, is shown in black. Picture taken from [22] (Chapter 2) and labels modified to match our notations.

## 3.3 Piecewise Linear (PL) models

Piecewise linear (PL) models also consist on systems of differential equations, but differently from classical ODE models, their vector fields have (finitely many) points of discontinuity. This is because the PL system state space is divided into regions (domains) in which the vector field may assume different expressions. However, these expressions must be affine or linear in each variable.

PL systems are a class of qualitative models, which can be used to facilitate the analysis of large classical ODE GRN models (see Section 3.2). In fact, intuitively, PL models can be derived from ODE models (3.2)-(3.3) with Hill functions (3.4)-(3.4) by letting the Hill coefficient tends to infinity. In this case the Hill functions $h^+$, $h^-$ turn into step functions $s^+, s^-$, respectively:

$$\begin{aligned} \lim_{n\to\infty} h^+(x;\theta,n) = s^+(x;\theta) &= \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{if } x < \theta \end{cases} \\ \lim_{n\to\infty} h^-(x;\theta,n) = s^-(x;\theta) &= \begin{cases} 1 & \text{if } x < \theta \\ 0 & \text{if } x > \theta \end{cases} \end{aligned} \tag{3.11}$$

Note that the functions $s^+(x;\theta)$ and $s^-(x;\theta)$ remain undefined at $x = \theta$, which are the points (or hyper-surfaces) of discontinuity of the vector field. Moreover, PL systems are generally used to model only protein dynamics, so—when this is the case—the quasi-steady-state assumption of mRNA concentration, discussed in Section 3.2.1, is implicitly

assumed.

Piecewise linear systems—with applications to genetic regulatory networks—have been originally introduced by [85]. From a biological point of view, the use of step functions has been motivated by the experimental observation that the activity of certain genes changes in a drastic manner at a threshold concentration of a regulatory protein ([180]). Hence, the PL model has the general form

$$\dot{x}_i = f_i(x) - d_i x_i, \quad 1 \leq i \leq n, \tag{3.12}$$

where $x = (x_1, ..., x_n)^T \in \mathbb{R}^n_{\geq 0}$ is a vector of protein concentrations. The non-negative quantities $f_i(x)$ and $d_i x_i$ represent synthesis and degradation rates for each protein $x_i$ respectively. The function $f_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ represents the expression rate of the gene $i$ depending on the whole state $x$. However, $f_i(x)$ can be detailed as:

$$f_i(x) = \sum_{l=1}^{L_i} k_{il} b_{il}(x)$$

where $k_{il} > 0$ is a rate parameter and $b_{il}(x)$ is a combination of step functions $s^+, s^-$ defined in (3.11).

Moreover, the PL system (6.1) can be written in matrix form as

$$\dot{x} = f(x) - \Gamma x, \tag{3.13}$$

where $f = (f_1, ..., f_n)$ and $\Gamma = diag(d_1, ..., d_n)$.

The dynamical properties of the PL systems are briefly summarized next.

### 3.3.1 Dynamical study of PL systems

The dynamics of PL systems can be studied in the $n$-dimensional state-space $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$, where each $\Omega_i$ is defined by $\Omega_i = \{x_i \in \mathbb{R}_{\geq 0} | 0 \leq x_i \leq max_i\}$ for some maximum concentration value $max_i$. A protein encoded by a gene will be involved in different interactions at different concentration thresholds, so for each variable $x_i$, we assume there are $p_i$ ordered thresholds $\theta_i^1, ..., \theta_i^{p_i}$ (we also define $\theta_i^0 = 0$ and $\theta_i^{p_i+1} = max_i$). The $(n-1)$-dimensional hyper-planes defined by these thresholds partition $\Omega$ into hyper-rectangular regions we call *domains*. Specifically, a domain $D \subset \Omega$ is defined

to be a set $D = D_1 \times ... \times D_n$, where $D_i$ is one of the following:

$$\begin{aligned} D_i &= \left\{x_i \in \Omega_i | 0 \leq x_i < \theta_i^1\right\} \\ D_i &= \left\{x_i \in \Omega_i | \theta_i^j < x_i < \theta_i^{j+1}\right\} \quad \text{for } j \in \{1, ..., p_i - 1\} \\ D_i &= \{x_i \in \Omega_i | \theta_i^{p_i} < x_i \leq max_i\} \\ D_i &= \left\{x_i \in \Omega_i | x_i = \theta_i^j\right\} \qquad \text{for } j \in \{1, ..., p_i\}. \end{aligned}$$

Let $\mathcal{D}$ be the set of domains in $\Omega$. A domain $D \in \mathcal{D}$ is called a *regulatory domain* if none of the variables $x_i$ has a threshold value in $D$ (it is a full open hyper-rectangle). In contrast, a domain $D \in \mathcal{D}$ is called a *switching domain* of order $k \leq n$ if exactly $k$ variables have threshold values in $D$, and the corresponding variables $x_i$ are called *switching variables* in $D$ [127]. The two sets of domains are respectively denoted by $\mathcal{D}_r$ and $\mathcal{D}_s$.

### 3.3.2 Solutions and Stability in Regular Domains

For any regulatory domain $D$, the function $f(x)$ is constant for all $x \in D$, and it follows that the PL system (3.13) can be written as

$$\dot{x} = f^D - \Gamma x, \tag{3.14}$$

where $f^D$ is constant in $D$. This implies that the equations are decoupled:

$$\dot{x}_i = f_i^D - d_i x_i, \quad i = 1, ..., n \tag{3.15}$$

and the solution $x_i^D(t)$ can be explicitly computed for all $x \in D$:

$$x_i^D(t) = (x_i(0) - \phi_i^D)e^{-d_i t} + \phi_i^D, \quad \text{where } \phi_i^D = \frac{f_i^D}{d_i}. \tag{3.16}$$

For each domain $D$, it is clear that solutions will evolve towards $\Phi^D = (\phi_1^D, ..., \phi_n^D)$, called the *focal point* of $D$.

Since $f_i^D$ is constant and the number of thresholds $\theta_i^j$ is finite (equal to $p_i$, see above), we can now be more precise and set $max_i = \max_{\forall D \in \mathcal{D}_r}\{\phi_i^D\}$. Then, the set $\Omega$ is an invariant region for the system and one may consider that $x(t) \in \Omega \ \forall \ t \geq 0$.

Each focal point $\phi_i^D$ may lie inside or outside the domain $D$ (to avoid very specific cases, it is generally assumed that focal points do not lie on switching domain, i.e. on the boundary of adjacent regular domains). Hence, if $\phi_i^D \in D$, then the domain is invariant, and the focal point turns out to be a true fixed (stable) point (see [40] for more details).

Whereas, if $\phi_i^D \notin D$, then solutions eventually leave the domain to enter another one, and the system switches to another vector field. More specifically, if the vector fields in two adjacent regular domains do not have opposite orientations, the trajectories can pass through the switching domain—which separates the two adjacent regular domains—and the overall solution is given by concatenating the solutions in the two (regular) adjacent domains. Otherwise, the vector field has to be defined as a differential inclusion along the switching surface at the boundary of the two domains, and a solution can still be constructed in the sense of Filippov ([73, 90], see also Section 3.3.3).

### 3.3.3 Solutions and Stability in Switching Domains

To provide the existence and the possibility for solutions to be continued on all domains, we have to define the right-hand side of system (3.13) at the points of discontinuity of the function $f$. To this end, we use a construction originally proposed by Filippov [73] and then applied to PL systems ([90]). The method consists of extending the system (3.13) as a differential inclusion,

$$\dot{x} \in H(x), \tag{3.17}$$

where $H$ is a set valued function (i.e. $H(x) \subseteq \mathbb{R}^n$). If $D \in \mathcal{D}_r$, then we define $H$ simply as

$$H(x) = \{f^D - \Gamma x\}, \quad \forall x \in D. \tag{3.18}$$

If $D \in \mathcal{D}_s$, we define $H$ as

$$H(x) = \overline{co}(\{f^{D'} - \Gamma x \mid D' \in R(D)\}), \quad \forall x \in D, \tag{3.19}$$

where $R(D) = \{D' \in \mathcal{D}_r \mid D \subseteq \partial D'\}$ is the set of all regulatory domains with $D$ in their boundary, and $\overline{co}(X)$ is the closed convex hull of $X$. For switching domains, $H(x)$ is typically multi-valued so solutions of the differential inclusion are defined as follows.

**Definition 3.1.** A solution of (6.12) on $[0, T]$ in the sense of Filippov is an absolutely continuous function (w.r.t. $t$) $\xi_t(x_0)$ such that $\xi_0(x_0) = x_0$ and $\dot{\xi}_t \in H(\xi_t)$, for almost all $t \in [0, T]$.

Moreover, the equilibrium points in switching domain (also called singular or Filippow equilibria) are defined in following way.

**Definition 3.2.** In a switching domain $D \in \mathcal{D}_s$, we say that a point $y \in \Omega$ is an equilibrium point for the differential inclusion if

$$0 \in H(y).$$

Now, we shall explain how to construct the set $H(x)$ at discontinuity points of $f$.

Let us consider the case where $x$ belongs to a switching domain $S$ separating two regular domains $D_1$ and $D_2$. Hence,

$$H(x) = \overline{co}(\{f^{D_1} - \Gamma x, f^{D_2} - \Gamma x\})$$

represents the segment joining the endpoints of the vectors $g_1 = f^{D_1} - \Gamma x$ and $g_2 = f^{D_2} - \Gamma x$. Trajectories can cross $S$ if the vector fields $g_1$ and $g_2$ point in a similar direction (see Figure 3.4), slide along $S$ if $g_1$ and $g_2$ point in opposite direction towards $S$ (see Figure 3.5) and be repelled from $S$ if $g_1$ and $g_2$ point in opposite direction away from $S$. The last two cases are known as *stable* and *unstable sliding motion* in the literature ([73]).

For example, for a two dimensional system let us assume that a sliding motion (stable or unstable) occurs on the switching domain $S$, which is in the hyper-plane $C_i^j = \{x \in \mathbb{R}^n_{\geq 0} \ : \ x_i = \theta_i^j\}$.

Then, for all $x \in S$ the solution may satisfy:

$$\dot{x} = \alpha f^{D_1} + (1-\alpha) f^{D_2} - \Gamma x, \quad 0 \leq \alpha \leq 1. \tag{3.20}$$

During sliding motion the state trajectories evolve on the hyper-plane $x_i = \theta_i^j$, so the parameter $\alpha$ is selected such that the velocity vector of the system (6.15) is always tangent to $C_i^j$. This, mathematically means that $\alpha$ has to satisfy the following condition:

$$\dot{x}_i = 0, \text{ for } x \in C_i^j, \iff \alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - d_i \theta_i^j = 0,$$

which, solved with respect to $\alpha$, gives:

$$\alpha = \frac{f_i^{D_2} - d_i \theta_i^j}{f_i^{D_2} - f_i^{D_1}}.$$

It is useful to define a concept analogous to the focal points defined for regulatory domains, to deal with switching domains.

**Definition 3.3.** We recall that $supp(D)$ is the $(n-k)$-dimensional hyperplane supporting $D$. Let $D$ be a switching domain of order $k$, then its focal set $\Phi(D)$ is

$$\Phi(D) = supp(D) \ \cap \ \{x : \ 0 \in H(x)\}, \tag{3.21}$$

where $H(x)$ is defined as in (6.14).

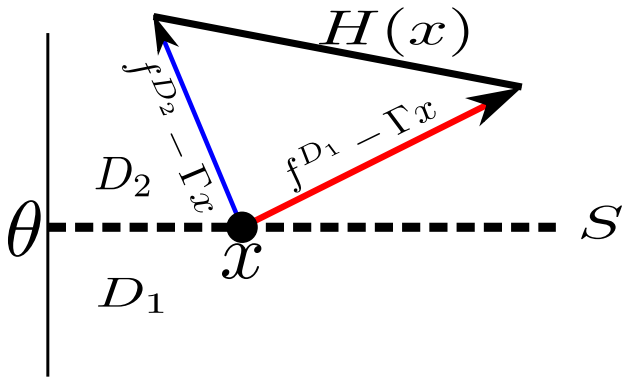



FIGURE 3.4: Example of behaviour of a bidimensional PL system at a point $x$ in a switching domain $S$, when the differential equations are generalized into differential inclusions by the method of Filippov. Here trajectories cross $S$ instantaneously, and solutions are normally continued.

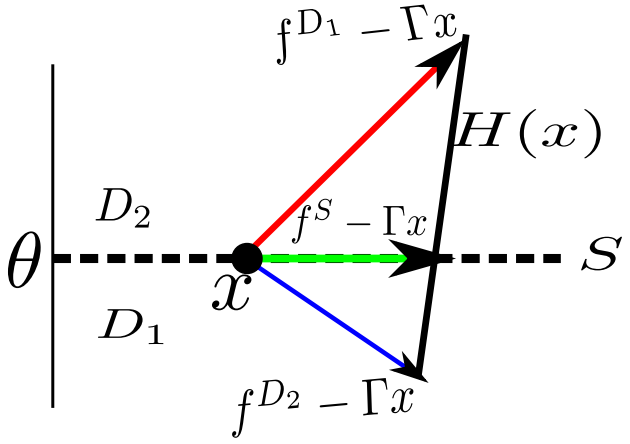



FIGURE 3.5: Example of behaviour of a bidimensional PL system at a point $x$ in a switching domain $S$, when the differential equations are generalized into differential inclusions by the method of Filippov. Here a solution can be constructed in such a way that trajectories evolve on $S$ (sliding motion).

Hence, $\Phi(D)$ for $D \in \mathcal{D}_s$ is the set containing all the equilibrium points of the differential inclusion (6.12), which lie on $supp(D)$. Thus, $\Phi(D)$ can be empty or a singleton, but more generally is a closed convex bounded set and hence is referred to as a focal set. In particular, this means that the sliding motion stops when it reaches $\Phi(D)$. If $\Phi(D)$ is empty, then the trajectory leaves $S$ and switches to another domain.

More details on dynamical analysis and applications of PL models can be found in [40, 56, 90, 144].

### 3.3.4 Example: PL bistable switch

To derive the PL system corresponding to the bistable switch example, one starts directly from the reduced ODE model (3.10) by replacing the Hill functions with the respective

step functions:

$$\begin{aligned} \dot{p_1} &= k_1 s^-(p_2;\theta_2) - \delta_1 p_1 \\ \dot{p_2} &= k_2 s^-(p_1;\theta_1) - \delta_2 p_2 \end{aligned} \tag{3.22}$$

System (3.22) is defined inside the (invariant) set $\Omega = [0, k_1/\delta_1] \times [0, k_2/\delta_2]$ which can be divided into four *regular domains*, where the vector field is uniquely defined:

$$\begin{aligned} D_1 &= \{x \in \mathbb{R}^2_{\geq 0} :\ 0 \leq p_1 < \theta_1,\ 0 \leq p_2 < \theta_2\} \\ D_2 &= \{x \in \mathbb{R}^2_{\geq 0} :\ 0 \leq p_1 < \theta_1,\ \theta_2 < p_2 \leq k_2/\delta_2\} \\ D_3 &= \{x \in \mathbb{R}^2_{\geq 0} :\ \theta_1 < p_1 \leq k_1/\delta_1,\ 0 \leq p_2 < \theta_2\} \\ D_4 &= \{x \in \mathbb{R}^2_{\geq 0} :\ \theta_1 < p_1 \leq k_1/\delta_1,\ \theta_2 < p_2 \leq k_2/\delta_2\}. \end{aligned}$$

In addition, there are also *switching domains*, where the system is defined only as a differential inclusion, corresponding to the segments where each of the variables is at a threshold ($p_i = \theta_i$ and $p_j \in [0, k_j/\delta_j]$). In each of the four regular domains, the differential system is linear, and simple to study. In $D_1$, for instance, the system is

$$\begin{aligned} \dot{p}_1 &= k_1 - \delta_1 p_1 \\ \dot{p}_2 &= k_2 - \delta_2 p_2, \end{aligned}$$

and the solution converges exponentially (see (3.16)) towards a steady state $(k_1/\delta_1, k_2/\delta_2)$.

If we suppose that $\theta_i < \frac{k_i}{\gamma_i}$, then this steady state is outside $D_1$, and the solution will switch to another system when it crosses one of the thresholds.

For the bistable switch $\theta_i < \frac{k_i}{\gamma_i}$, $i = 1, 2$, there are two classical stable steady states, $\phi_1$ and $\phi_2$, and an unstable Filippov equilibrium point, $\phi_3$, analogous to a saddle point (see Figure 3.3):

$$\phi_1 = \left(\frac{k_1}{\delta_1}, 0\right), \qquad \phi_2 = \left(0, \frac{k_2}{\delta_2}\right), \qquad \phi_3 = (\theta_1, \theta_2).$$

$\phi_1$ and $\phi_2$ belong to (the boundary of) their respective domains ($D_2$ and $D_3$), so that any trajectory entering one of these domains remains there. In contrast, trajectories starting in $D_1$ or $D_4$ will switch to another domain.

## 3.4 Stochastic Models

As seen so far, much of the mathematical modelling of GRNs represents gene expression and regulation as deterministic processes. However, considerable experimental evidence

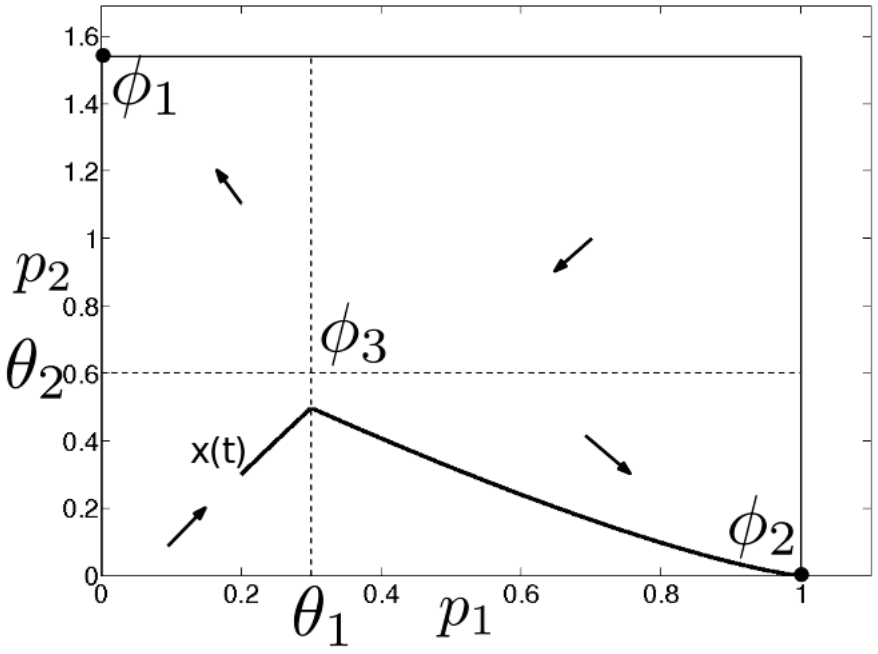



FIGURE 3.6: Phase plane for the piecewise linear system (3.22). The threshold values $\theta_1$, $\theta_2$ divide the plane into four rectangular regions, where the vector field is constant. There are two stable steady states ($\phi_1$,$\phi_2$) and an unstable Filippov equilibrium ($\phi_3$). One solution $x(t) = [p_1(t), p_2(t)]$ is shown in black, which may be compared to that shown in Figure 3.3 (a). Picture taken from [22] (Chapter 2) and labels modified to match our notations.

suggests that gene expression, both in prokaryotes and eukaryotes, is an inherently stochastic process ([68, 122, 137, 161]). This is particularly true when cellular biochemical reactions (eg. binding/unbinding of TFs to promoters, translation, etc.) are triggered by small number of molecules (typically when dealing with single cell gene expression models), a condition which makes deterministic formalisms (representing average amount of cellular components in a population of cells) poorly accurate to describe the true behaviour of the system [138, 147, 178].

The introduction of noise in gene regulation can be efficiently modelled using Chemical Master Equations (CMEs) whose realizations can be exactly simulated by Gillespie's algorithm ([80–82]). Moreover, under certain conditions discussed later on, CMEs can be approximated by stochastic differential equations (SDEs) of the Langevin type, named Langevin Chemical Equations (CLEs) [79], which are easier to handle and much less time consuming to simulate.

### 3.4.1 The Chemical Master Equation (CME)

Let us consider a well-stirred system of molecules of $N$ chemical species (mRNAs, proteins, etc) $\{S_1, ..., S_N\}$, which interact through $M$ chemical reactions (transcription, translation, degradation, etc) $\{R_1, ..., R_M\}$. It is assumed that the system is well-stirred and confined to a constant volume (cellular volume) $\Omega$.

Let $X_i(t) \in \mathbb{N}$ denote the number of molecules of species $S_i$ in the system at time $t$. The aim is to estimate the state vector $X(t) = (X_1(t), ..., X_N(t))$, given that the system was in state $X(t_0) = Z_0$ at some initial time $t_0$.

The system state $X(t)$—describing the molecular populations—will actually be a vector of *random variables*. The changes in the species populations are of course a consequence of the chemical reactions. Each reaction channel $R_j$ is characterized mathematically by two quantities:

1. **state-change vector** $\nu_j = (\nu_{1j}, ..., \nu_{Nj})$, where $\nu_{ij}$ is the change in the $S_i$ molecular population caused by one $R_j$ reaction, so if the system is in state $Z$ and one $R_j$ reaction occurs, the system immediately jumps to state $Z + \nu_j$;
2. **propensity function** $a_j$, defined so that $a_j(Z)dt =$ the probability, given $X(t) = Z$, that one $R_j$ reaction will occur somewhere inside $\Omega$ in the next infinitesimal time interval $[t, t + dt)$. The probability that more than one reaction will occur in $[t, t + dt)$ is assumed to be negligible.

Due to the probabilistic nature of the definition of $a_j$, an exact prediction of $X(t)$ is not possible, but one might hope to infer the probability

$$P(Z, t|Z_0, t_0) = \Pr\{X(t) = Z, \text{ given } X(t_0) = Z_0\},$$

which is given by the Chemical Master Equation (CME) [82]

$$\frac{\partial P(Z, t|Z_0, t_0)}{\partial t} = \sum_{j=1}^{M} \left[a_j(Z - \nu_j)P(Z - \nu_j, t|Z_0, t_0) - a_j(Z)P(Z, t|Z_0, t_0)\right]. \tag{3.23}$$

In principle, CME (3.23) completely determines $P(Z, t|z_0, t_0)$, but a close inspection reveals that (3.23) is, possibly, an infinite dimensional system of coupled ODEs, each one for every possible combination of reactant molecules. Hence, CME can be solved analytically for only a few simple cases and other approaches are needed. One fruitful approach is constructing numerical realizations of $X(t)$, which is not the same as solving the CME numerically (because that would give the probability mass function of $X(t)$),

and then histogramming or averaging the results of many realizations. Next, we present the stochastic simulation algorithm (SSA), which is a Monte Carlo procedure for numerically generating time trajectories of the molecular populations in exact accordance with the CME.

#### 3.4.1.1 Stochastic simulation algorithm (SSA)

The key to generate realizations of $X(t)$ by means of the SSA is to define a new probability function $p(\tau, j|Z, t)$ such that:

$p(\tau, j|Z, t)d\tau$ = the probability, given $X(t) = Z$, that the next reaction in the system will occur in the infinitesimal time interval $[t+\tau, t+\tau+d\tau)$, and will be an $R_j$ reaction.

Formally, $p(\tau, j|Z, t)$ is a joint probability density function of two random variables $\tau$ and $j$, representing the *time to the next reaction* and the *index of the next reaction*, respectively, given that the system at time $t$ is in the state $Z$. Taking into account the definition of $a_j$ and applying the laws of probability it can be proven that (see [82, 147] for mathematical details):

$$p(\tau, j|Z, t) = a_j(Z) \exp(-a_0(Z)\tau) \tag{3.24}$$

where:

$$a_0(Z) = \sum_{j'=1}^{M} a_{j'}(Z) \tag{3.25}$$

The stochastic simulation algorithm is mathematically based on (3.24). In fact, (3.24) implies that $\tau$ is an exponential random variable with mean (and standard deviation) $1/a_0(Z)$, while $j$ is a statistically independent random variable with probability $a_j(Z)/a_0(Z)$. There are several Monte Carlo algorithms to generate samples of $\tau$ and $j$ according the definition above [178]. Probably, the simplest approach is that which relies on the standard inversion generating method of Monte Carlo theory [145]: draw two random numbers $r_1$ and $r_2$ from the uniform distribution in the unit interval, and select $\tau$ and $j$ such that:

$$\tau = \frac{1}{a_0(Z)} \ln\left(\frac{1}{r_1}\right) \tag{3.26a}$$

$$j = \text{the smallest integer satisfying } \sum_{j'=1}^{j} a_{j'}(Z) > r_2 a_0(Z) \tag{3.26b}$$

Now, we have the following stochastic simulation algorithm (SSA), also known as the Gillespie algorithm, for generating an exact numerical realization of the stochastic process $X(t)$ satisfying CME (3.23)([84]):

1. Initialize the time $t = t_0$ and the system state $x = Z_0$.
2. With the system at state $X(t) = Z$, evaluate all $a_j(Z)$ and their sum $a_0(Z)$.
3. Generate values for $\tau$ and $j$ using (3.26a) and (3.26b).
4. Update the time and the state value by replacing $t \leftarrow t + \tau$ and $Z \leftarrow Z + \nu_j$.
5. Record $(Z, t)$ as desired. Return to Step 1, or else end the simulation.

The problem with the SSA is that it might be very slow, and this slowness is due to the factor $1/a_0(Z)$ in (3.26a), which can be very small if the population of any species is sufficiently large, and this is often the case in practice [147]. To overcome this computational issue, more efficient variations of the SSA [35, 145], or even approximated numerical methods, such as the *Tau-leaping* [83], have been implemented.

Another possible way to speed up the computation of realizations of $X(t)$ is that of directly approximating the CME by stochastic differential equations of the Langevin type, and so obtaining the so called Chemical Langevin Equation (CLE) [79].

### 3.4.2 The chemical Langevin equation (CLE)

A detailed mathematical derivation of CLE is given in [79], here we just mention the two conditions required to get a valid approximation of the CME by the CLE.

Suppose the system's state $X(t)$ of CME (3.23) at the current time $t$ is $Z$. Suppose also that, for any time $\tau > 0$, $K_j$ be the number of reactions that occur in the time interval $[t, t + \tau]$ (it is worth noting that here $\tau$ has not the same definition as given before in (3.24), since now $\tau$ assumes the meaning of a preselected time which encompasses more than one reaction event). Under the two conditions listed below:

**Condition (i):** Require $\tau$ to be *small* enough that the change in the state during $[t, t + \tau]$ will be so small that none of the propensity functions changes its value in a noticeable fashion, that is:

$$a_j(X(\bar{t})) \cong a_j(Z), \quad \forall \bar{t} \in [t, t + \tau], \quad \forall j \in [1, M]. \tag{3.27}$$

**Condition (ii):** Require $\tau$ to be *large* enough that the expected number of occurrences of each reaction $R_j$ in $[t, t+\tau]$ be much larger than 1, that is:

$$a_j(Z)\tau \gg 1, \quad \forall j \in [1, M]. \tag{3.28}$$

then, CME (3.23) can be approximated by stochastic differential equations of the Langevin type [79]:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{M} \nu_{j,i} a_j\big(x(t)\big) + \nu_{j,i}\sqrt{a_j\big(x(t)\big)}\Gamma_j(t), \tag{3.29}$$

with $i = 1, \ldots, N$, where, for $j = 1, \ldots, M$, the $\Gamma_j(t)$ are mutually uncorrelated white noise processes. Here, in (3.29) $x = [x_1, ..., x_N]$ plays the role of a continuous approximation of the molecule count $X$ in the CME.

Now, let us focus on the two conditions (i) and (ii). Condition (ii) runs counter to condition (i), and there may be cases in which both conditions can not be satisfied simultaneously. But there will be many practical circumstances in which both conditions hold. In fact, this is the case when species have large molecular population numbers, i.e. when $X_i \gg 1 \quad \forall i \in \{1, N\}$.

In principle, the CLE can be simulated using any numerical methods developed for stochastic differential equations (SDEs) (see [112] for more details on numerical simulations of SDEs). However, particular attention has to be paid to ensure non-negativity of the system's state $x$, which is not generally guaranteed for SDE simulations, differently from SSA for CME,. To this end, one approach consists in shutting down a reaction channel when the amount of any its reactants (species involved in) reaches zero [51].

### 3.4.3 Example: CME and CLE bistable switch

The bistable switch model consists of $N = 4$ species (variable): $m_1$ (mRNA of gene 1), $p_1$ (protein of gene 1), $m_2$ (mRNA of gene 2), $p_2$ (protein of gene 2), which interact via through the $M = 8$ reactions reported in Table 3.1 with the corresponding propensities $a(x) = \big[a_1(x), \ldots, a_8(x)\big]$, where $x = \big[m_1, p_1, m_2, p_2\big]$, and stoichiometric vectors $[\nu_1, \ldots, \nu_8]$:

$$[\nu_1, \ldots, \nu_8]_{4\times 8} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \tag{3.30}$$

| Synthesis | Rate $a_j$ | Degradation | Rate $a_j$ |
|---|---|---|---|
| $\varnothing \xrightarrow{a_1} m_1$ | $v_1 h^-(p_2;\theta_2,n_2)$ | $m_1 \xrightarrow{a_2} \varnothing$ | $\gamma_1 m_1$ |
| $m_1 \xrightarrow{a_3} m_1 + p_1$ | $\bar{k}_1 m_1$ | $p_1 \xrightarrow{a_4} \varnothing$ | $\delta_1 p_1$ |
| $\varnothing \xrightarrow{a_5} m_2$ | $v_2 h^-(p_1;\theta_1,n_1)$ | $m_2 \xrightarrow{a_6} \varnothing$ | $\gamma_2 m_2$ |
| $m_2 \xrightarrow{a_7} m_2 + p_2$ | $\bar{k}_2 m_2$ | $p_2 \xrightarrow{a_8} \varnothing$ | $\delta_2 p_2$ |

TABLE 3.1: Reactions of the stochastic model of the system of Figure 3.1 and corresponding propensities. An arrow from (to) symbol $\varnothing$ means synthesis (degradation).

| Parameters | Values |
|---|---|
| $v_1$ | $1\,\#\,min^{-1}$ |
| $\gamma_1$ | $0.4\ min^{-1}$ |
| $k_1$ | $1\,min^{-1}$ |
| $\delta_1$ | $0.012\,min^{-1}$ |
| $\theta_1$ | $10\#$ |
| $n_1$ | $2$ |
| $v_2$ | $1\,min^{-1}$ |
| $\gamma_2$ | $0.4\,\#\,min^{-1}$ |
| $k_2$ | $1\,min^{-1}$ |
| $\delta_2$ | $0.012\,min^{-1}$ |
| $\theta_2$ | $30\,\#$ |
| $n_2$ | $3$ |

TABLE 3.2: Parameter values for the CME and CLE (in molecule number units – symbol # denotes number of molecules).

It is worth noting that, though propensities $a_j$ reported in Table 3.1 are expressed with the same notations of deterministic synthesis and degradation rates of the ODE bistable switch model (3.8), parameter values have now different physical units. This is because CME and CLE variables denote number of molecules and not concentrations, as in the ODE example.

The CME and CLE models follow from replacing the stoichiometries $\nu_1, \ldots, \nu_8$ (3.30) and the propensities of the model of Table 3.1 into (3.23) and (3.29). In particular, the CLE can be written in the matrix form

$$\dot{x} = V\ a(x) + V\ \text{diag}\left(\sqrt{a(x)}\right)\ \Gamma \tag{3.31}$$

where $V = [\nu_1, ..., \nu_8]_{4\times 8}$, $\text{diag}\left(\sqrt{a(x)}\right)$ is the diagonal matrix having the square root of the entries of vector $a(x)$ on the diagonal, $\Gamma = [\Gamma_1, ..., \Gamma_m]^T$.

In Figure 3.7 are shown 20 realizations of protein 1 ($p_1$) and protein 2 ($p_2$) of the CME model using the Gillespie's algorithm (SSA). CLE simulations look similar and are not shown.

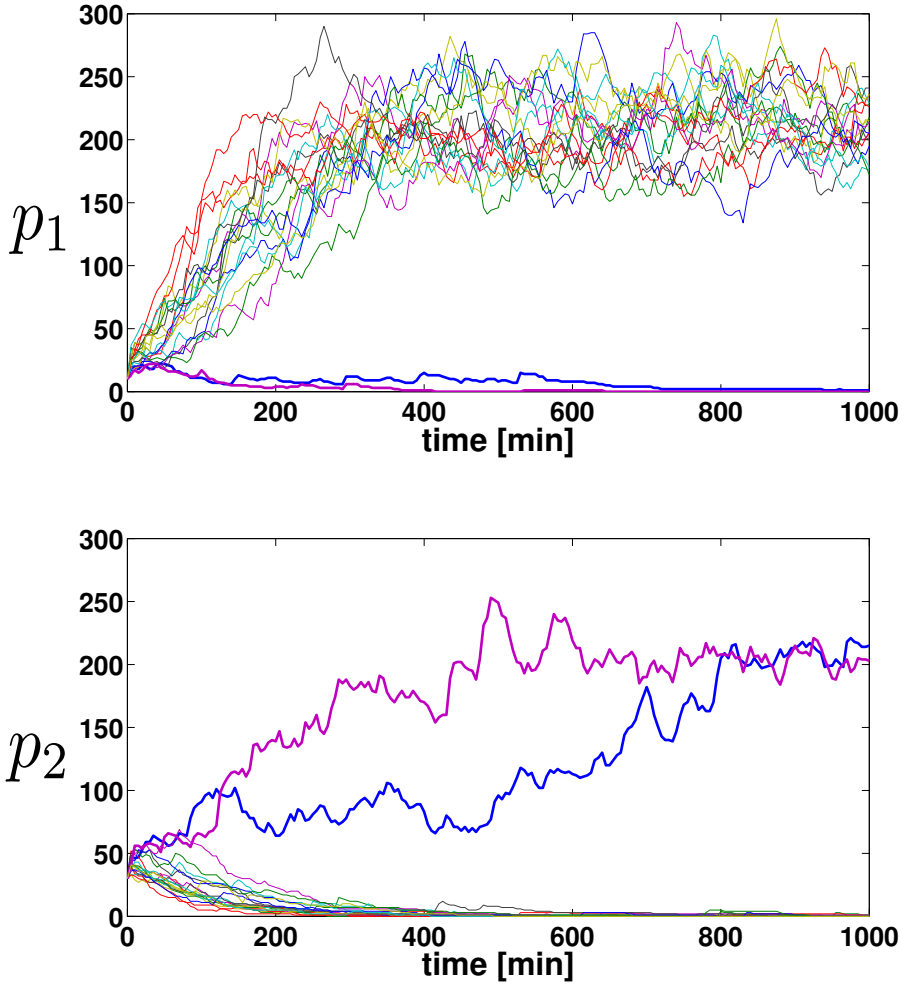



FIGURE 3.7: Plot showing 20 realizations of the $p_1$ and $p_2$ variables of the CME model of the bistable-switch system using the Gillespie's algorithm (SSA). Simulations are performed using parameter values from Table 9.3 and initial condition $x(0) = [m_1(0), p_1(0), m_2(0), p_2(0)] = [5, 10, 5, 30]$. As we can notice, for this particular choice of parameter values and initial condition, protein 1 ($p_1$) reaches its ON state (higher values) with higher frequency than protein 2 ($p_2$), which reaches its OFF state (lower values). However, due to the stochastic nature of the model, also the opposite behaviour can happen: $p_1$ is OFF and $p_2$ is ON (see thicker blue and purple lines).

## 3.5 Final comments

The mathematical formalisms discussed in this chapter allow genetic regulatory networks to be modelled in different ways, and each has some strengths and possible weakness.

### 3.5.1 Deterministic Vs stochastic models

Until recently, modelling and simulation studies of GRNs have mainly focused on deterministic models, such as Boolean models and simplified differential equation models. In

fact, when the biochemical reaction mechanisms underlying regulatory interactions are not known or are incompletely known, qualitative formalisms, as Boolean and piecewise linear methods, turn out to be the most appropriate to gain high-level informations on the network dynamics and insights on the underlying biological system.

Whereas, when dealing with regulatory networks, that have been well-characterized by experimental data (often of small size and modest complexity), quantitative ODE and stochastic models can be successfully used to simulate, analyse and validate models of biological regulatory systems.

Moreover, it is worth remembering that stochastic models, as CME and CLE, are probably the most appropriate formalisms to deal with gene expression in single cells, where the molecules' number is so small to make any deterministic approach inadequate and not realistic.

### 3.5.2 Quantitative Vs qualitative models

Another point to take into account, besides a lack of relevant biological information to build quantitative models, is that large ODE and stochastic models are also computationally very heavy to simulate, making them less attractive for real time simulation purposes. Hence, also in this case, qualitative formalisms could be successfully used both for speeding up simulations, as well as for helping modellers in identifying those subnets and/or regulation loops which determine the main biological process of interest. In this fashion, the entire system could be reduced to a smaller and, probably, lumped GRN which could be more easily addressed with quantitative ODE or stochastic methods.

However, the emergence of new experimental techniques, along with the increasing development of open biological databases of experimental data and the continuing increase of computer power, allow modellers to develop more and more quantitative and complex biological models.

# Chapter 4

# A Simple Model to Control Growth Rate of Synthetic E. coli during the Exponential Phase: Model Analysis and Parameter Estimation

In this chapter we discuss a work that has been presented at the *Computational Methods for Systems Biology* Conference in 2012 [37] and was awarded with the Best Student Paper prize.

Here, we develop and analyse a model of a minimal synthetic gene circuit, that describes part of the gene expression machinery in *Escherichia coli*, and enables the control of the growth rate of the cells during the exponential phase. This model is a piecewise non-linear system with two variables (the concentrations of two gene products) and an input (an inducer). We study the qualitative dynamics of the model and the bifurcation diagram with respect to the input. Moreover, an analytic expression of the growth rate during the exponential phase as function of the input is derived. A relevant problem is that of identifiability of the parameters of this expression supposing noisy measurements of exponential growth rate. We present such an identifiability study that we validate *in silico* with synthetic measurements.

## 4.1 Introduction

Synthetic biology has nearly emerged as a new engineering discipline. The goal of synthetic biology [15, 107, 131] is to develop and apply engineering tools to control cellular behaviour—constructing novel biological circuits in the cell—to perform new and desired functions.

Most recent synthetic designs have focused on the cell transcription machinery, which includes the genes to be expressed, their promoters, RNA polymerase and transcription factors, all serving as potential engineering components. Indeed, synthetic bio-molecular circuits are typically fabricated in *Escherichia coli* (*E. coli*), by cutting and pasting together coding regions and promoters (natural and synthetic) according to designed structures and specific purposes ([67, 76, 165]).

Along these lines, synthetic biology ultimately aims at developing synthetic bio-molecular circuitry that may help in producing bio-pharmaceuticals, bio-films, bio-fuels, novel cancer treatments and novel bio-materials (see [107] for a review on synthetic biology applications).

In the present work we focus on the gene expression machinery of the bacterium *Escherichia coli*, with the aim of controlling the growth rate of the cells. *E. coli* is a model organism that is easy to manipulate and much knowledge is available about its regulatory networks.

In the presence of a carbon source—such as glucose—*E. coli* grows in an exponential manner until it exhausts the nutrient sources, and then enters a stationary phase with practically zero growth [129]. The wild-type (namely the genetically unmodified) bacteria grow at different rates in the presence of carbon sources of different types [124]. Notably, glucose is the preferred substrate because it leads to a higher growth rate in wild type. Our control objective is to force the bacterium to significantly modify its response to glucose so as to tune the cells' growth rates. To this end, we take into account the recent applications of synthetic biology which allow us to fabricate engineered promoters which in turn can be externally controlled by inducers [102].

Notably, we will study an open loop configuration of a bi-dimensional model of a mutant *E. coli* inspired by the experiments in [160]. The two basic variables of our model, which describe the gene expression machinery that is responsible for bacterial growth are (see Fig.4.1):

1. the concentration of a *Component of the Gene Expression Machinery* (CGEM), proteins responsible for global growth (ribosomes and RNA polymerase). Without this CGEM, the bacteria cannot produce any proteins and thus cannot grow.

2. the concentration of CRP, a protein involved in the formation of the complex cAMP-CRP whose level positively correlates with less preferred carbon sources and slower growth [24].

We will assume that an engineered inducible-promoter is used to express the CGEM. Moreover it is assumed that the mutant CGEM activates its own expression. The number and location of equilibria can thus be controlled by means of an input control function of the inducer and, in particular, there can be regions of bi-stability, as observed in [160].

The type of growth rate control we present—which directly acts upon the GEM—could be useful in creating bacterial cells that divert resources used for growth towards the production of a target compound. Thus, the analysis of the simple model presented here is an attempt to help guide the construction of synthetic gene networks, which improves product yield and productivity.

This paper is structured as follow: in Section 4.2 we describe the open-loop model, providing some biological motivations for the terms forming the differential equations. Next, in Section 5.5 we qualitatively analyse the open-loop model by means of phase-plane and bifurcation diagram, showing how the steady states of the CGEM can be controlled by the external input (inducer). In Section 4.4 we derive a mathematical expression of the growth rate during the exponential phase as a function of the amount of the inducer. Finally, in Section 4.5 we present an *in silico* practical identifiability analysis of such expression.

## 4.2 The Open-loop Model

The principal modelling challenges come from incomplete knowledge of the networks, and the dearth of quantitative data for identifying kinetic parameters required for detailed mathematical models. Qualitative methods overcome both of these difficulties and are thus well-suited to the modelling and simulation of genetic networks ([55, 144]).

In this work we used a novel *piecewise non-linear* formalism—derived from piecewise linear (PL) systems (see [40, 43, 56, 90, 91] for more details)—to model gene expression affected by dilution due to growth rate.

The open-loop model depicted in Fig. 4.1—similarly to PL models of regulatory genetic networks—is built with discontinuous (step) functions. The use of step function has been

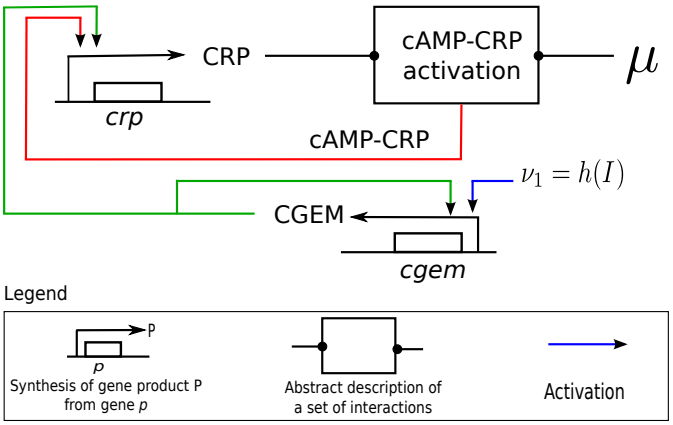



FIGURE 4.1: Regulatory network of the open-loop model in the *mutant E. coli*. The model consists of genes *crp* and synthetic-*cgem* (modified promoter of a component of the gene expression machinery (CGEM) in *E. coli*). The synthetic-*cgem* promoter is positively regulated by the inducer I—according to the input function $\nu_1 = h(I)$—and CGEM. CGEM, being responsible for the bacterial gene expression, positively regulates *crp* gene too. Moreover, *crp* transcription is induced by cAMP-CRP, a metabolite whose formation relies on CRP protein abundance and low level of bacterial growth rate $\mu$.

motivated by the experimental observation that the activity of certain genes changes in a switch-like manner at a threshold concentration of a regulatory protein [180]. The non linearity is concentrated in the removal term of differential equations, which takes into account the protein degradation and the dilution due to growth.

The open-loop model, expressed by (4.1), describes the qualitative dynamics of a CGEM responsible for bacterial growth and another protein that reflects growth, such as CRP. The CGEM is assumed to be externally controlled by an inducer I (such as IPTG (Isopropil $\beta$-D-1-tiogalattopiranoside), Tc (tetracycline) etc). This model of ODE exhibits bi-stability in CGEM expression for some parameter sets, as experimentally verified in [160]. We shall take into account this bi-stability to control the model's state to the "low" or to the "high" CGEM stable steady state. Let $x_c$, $x_p \in \mathbb{R}_{\geq 0}$ be the CRP and CGEM concentrations respectively. Thus, the open-loop model graphically depicted in Fig. 4.1, can be mathematically translated into:

$$\begin{cases} \dot{x}_c(t) = k_c^0 \; s^+(x_p, \theta_p^1) + k_c^1 \; s^+(x_p, \theta_p^2) \; s^+(x_c, \theta_c^1) \; s^-(x_p, \theta_{\bar{\mu}}) \\ \qquad\qquad - (\bar{\mu} \; x_p(t) + \gamma_c) \; x_c(t) \\ \dot{x}_p(t) = \nu_1 \; k_p^0 \; s^+(x_p, \theta_p^1) + \nu_1 \; k_p^1 \; s^+(x_p, \theta_p^2) \\ \qquad\qquad - (\bar{\mu} \; x_p(t) + \gamma_p) \; x_p(t) \end{cases} \tag{4.1}$$

where:

- $k_i^0 > 0$ $(i = c, p)$ is the basal synthesis rate constant;

- $k_i^1 > 0$ $(i = c, p)$ is the main synthesis rate constant;
- $\nu_1$ is a positive input accounting for the inducer $I$; it will be a function $\nu_1(v)$, $v$ being the concentration of $I$;
- $\gamma_i > 0$ $(i = c, p)$ is the degradation rate constant;
- $\theta_i^j > 0$ $(i = c, p;\ j = 1, 2)$ is the $x_i$ threshold concentration for activation/inhibition;
- $\theta_{\bar{\mu}} > 0$ is a growth threshold depending on which substrate is used;
- $\bar{\mu} > 0$ is a growth constant depending on which substrate is used.

and $s^+, s^-$ denote the step-like functions, defined as

$$s^+(x_i, \theta_i^j) = \begin{cases} 1 & \text{if } x_i > \theta_i^j \\ 0 & \text{if } x_i < \theta_i^j \end{cases}; \quad s^-(x_i, \theta_i^j) = 1 - s^+(x_i, \theta_i^j) \ ,$$

which are used to model the switch-like promoters' regulation carried out by the generic protein $x_i$. These $s^+, s^-$ are not defined at the threshold values so, to define solutions on the surfaces of discontinuity, i.e. $x_i = \theta_i^j$, we use the approach of Filippov [73], which extends the vector field to a differential inclusion.

In what follows, we will explain the main assumptions adopted in building the system equations (4.1), which were inspired by the models in [144, 160] and the literature on *E. coli*.

### 4.2.1 Growth rate

In bacteria, growth rate is intimately intertwined with gene expression ([114, 151]) and with the type of substrate [124]. Hence, to keep model complexity to a minimum, we assume growth rate $\mu$ to be proportional—with a constant $\bar{\mu}$ depending on the quality of medium—to the concentration of the CGEM which is responsible for bacterial growth:

$$\mu(t) = \bar{\mu}\ x_p(t) \ . \tag{4.2}$$

### 4.2.2 cAMP-CRP activation

The cAMP-CRP complex is formed from cAMP, a small metabolite, which binds the protein CRP. The cAMP concentration is higher at low growth rate and rapidly decreases

at high growth rate [24]. Thus, cAMP abundance in cells can be well captured by a negative step function of $\mu$, i.e. $s^-(\mu, \theta_\mu)$. Moreover, being cAMP association with or dissociation from CRP much faster than the synthesis and degradation of proteins [144], we have assumed that as soon as CRP reaches a certain threshold, i.e. $\theta_c$, CRP instantly binds to cAMP in a switch-like fashion . For these reasons, the positive regulation carried out by cAMP-CRP reads as:

$$b^+_{cAMP-CRP} = s^+(x_c, \theta_c)\ s^-(\mu, \theta_\mu).$$

Focusing on the negative step function $s^-(\mu, \theta_\mu)$ and taking into account the expression of $\mu$ in (4.2), we can rewrite $b^+_{cAMP-CRP}$ as:

$$b^+_{cAMP-CRP}(x_c, x_p) = s^+(x_c, \theta_c)\ s^-(x_p, \theta_{\bar{\mu}}) \tag{4.3}$$

where $\theta_{\bar{\mu}}$ is a threshold concentration of CGEM which depends on the type of carbon source.

### 4.2.3 CRP synthesis

We have assumed that a lower value of $x_p$, i.e. $\theta^1_p$, induces the basal synthesis ($k^0_c\ s^+(x_p, \theta^1_p)$) of $x_c$ while a higher value of $x_p$, i.e. $\theta^2_p$, is needed to stimulate its main expression ($k^1_c s^+(x_p, \theta^2_p)$). Moreover, the *crp* gene is regulated both positively and negatively by cAMP-CRP. However, in order to simplify, we omit the negative control of *crp*, because this mechanism only plays a role when the CRP concentration is low [144][1]. Thus, only one concentration threshold of CRP, i.e. $\theta^1_c$, is required in the model, to allow production of the cAMP-CRP complex. In conclusion, taking into account the regulation function of cAMP-CRP in (5.16), the CRP synthesis reads:

$$f_c(x) = k^0_c\ s^+(x_p, \theta^1_p) + k^1_c\ s^+(x_p, \theta^2_p)\ b^+_{cAMP-CRP}(x_c, x_p), \tag{4.4}$$

with

$$0 < \theta^1_c < max_c, \tag{4.5}$$

where $max_c$ is the maximum concentration value for CRP.

[1] We found that a model involving the negative control of *crp* by cAMP-CRP does not have any effect on the conclusion of this study.

### 4.2.4 CGEM synthesis

In this bi-dimensional model, since the CGEM is the main factor which determines growth of the cell, it is also responsible for its own synthesis. We have thus assumed that a low concentration ($\theta_p^1$) is sufficient to stimulate its basal production $k_p^0\ s^+(x_p, \theta_p^1)$ while its main production $k_p^1\ s^+(x_p, \theta_p^2)$ is stimulated only above the $\theta_p^2$ threshold. Thus, we can order the thresholds for $x_p$ as:

$$0 < \theta_p^1 < \theta_p^2 < max_p, \tag{4.6}$$

where $max_p$ is the maximum concentration value.

Moreover, the inducer effect is modelled by input $\nu_1$. For a general formulation of the activation of $x_p$ by an inducer I, we will later on assume that $\nu_1$ is a positive increasing function of $I$. Consequently, $x_p$ synthesis reads:

$$f_p(x) = \nu_1\ k_p^0\ s^+(x_p, \theta_p^1) + \nu_1\ k_p^1\ s^+(x_p, \theta_p^2). \tag{4.7}$$

### 4.2.5 Proteins removal

The negative terms in $\dot{x}_c$ and $\dot{x}_p$ of (4.1) take into account the fact that cells remove proteins by two processes: degradation and dilution due to cell growth [66]. Notably, these terms can generally be expressed as $(\mu(t)+\gamma_i)x_i$ (for $i = c, p$) where $\mu(t) = \bar{\mu}\ x_p(t)$, which is the bacterial growth rate in (4.2), is responsible for the proteins' dilution while $\gamma_i$ stands for protein's degradation.

## 4.3 Qualitative Analysis of the Open-loop Model

In this section we will qualitatively study, by means of phase-planes and bifurcation diagrams, model (4.1) in the case that cells are grown in glucose. This will elucidate how qualitative dynamics—in terms of equilibria' location and their stability—is intertwined with biological phenomena. Moreover, we shall show how—through the external input $\nu_1$—the stability of equilibria in (4.1) can be controlled, pointing out a reciprocal influence between growth rate and gene expression.

### 4.3.1 Open-loop model in glucose growth

If cells are grown in glucose, then parameters depending on the substrate become $\theta_{\bar{\mu}} = \theta_p^G$ and $\bar{\mu} = \mu_G$ in model (4.1). Moreover, in the presence of glucose or other PTS

sugars, adenylate cyclase[2] activity decreases, leading to a drop in the cellular level of cAMP [116, 134]. Thus, we have modelled this effect assuming:

$$0 < \theta_p^1 < \theta_p^G < \theta_p^2 < max_p. \tag{4.8}$$

Therefore, during growth on glucose, the state space of model (4.1) can be partitioned into eight *regular domains*, where the vector field is uniquely defined:

$$\begin{aligned}
D_1^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ 0 \leq x_p < \theta_p^1\right\} \\
D_2^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ 0 \leq x_p < \theta_p^1\right\} \\
D_3^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^1 < x_p < \theta_p^G\right\} \\
D_4^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^1 < x_p < \theta_p^G\right\} \\
D_5^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^G < x_p < \theta_p^2\right\} \\
D_6^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^G < x_p < \theta_p^2\right\} \\
D_7^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^2 < x_p \leq max_p\right\} \\
D_8^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^2 < x_p \leq max_p\right\}.
\end{aligned}$$

In addition, there are also *switching domains*, where the model is defined only as a differential inclusion [73], corresponding to the segments where each of the variables is at a threshold ($x_i = \theta_i$ and $x_j \in [0, max_j]$).

In general, for any regular domain $D$, the synthesis rates (4.4) and (4.7) are constant for all $x \in D$, and it follows that model (4.1) can be written as

$$\begin{cases}
\dot{x}_c(t) = f_c^D - (\bar{\mu}\ x_p(t) + \gamma_c)\ x_c(t) \\
\dot{x}_p(t) = f_p^D - (\bar{\mu}\ x_p(t) + \gamma_p)\ x_p(t)
\end{cases} \tag{4.9}$$

with $f_c^D, f_p^D, \bar{\mu}, \gamma_c, \gamma_p$ positive real constants. For any initial condition $x(t_0) \in D$ the unique solution of (4.9) can be found explicitly by solving first the $x_p$-equation of (4.9), which is an autonomous differential equation, and then solving the $x_c$-equation, having substituted $x_p(t)$ into it. Thus, it can be shown that $x_c(t)$ is given by:

$$x_c(t) = \frac{1}{b(t)}\left(b(t_0)x_c(t_0) + f_c^D \int_{t_0}^{t} b(s)ds\right)$$

[2]Enzyme that catalyses the conversion of ATP to cAMP and pyrophosphate.

where $b(t) = exp\left(\int_{t_0}^{t}(\bar{\mu}\ x_p(\tau) + \gamma_p)d\tau\right)$. Moreover, defining $\Phi(D) = (\bar{x}_c, \bar{x}_p)^T$ with

$$\begin{aligned} \bar{x}_c &= \frac{f_c^D}{\bar{\mu}\bar{x}_p + \gamma_c}, \\ \bar{x}_p &= \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\bar{\mu}f_p^D}}{2\bar{\mu}}, \end{aligned} \tag{4.10}$$

(it is easy to check that $\bar{x}_p$—in (6.11)—is the only positive solution of $\dot{x}_p = 0$) it turns out that either $x(t) \to \Phi(D)$ as $t \to \infty$ or $x(t)$ reaches the boundary of $D$.

**Definition 4.1.** Given a regular domain $D$, the point $\Phi(D) = (\bar{x}_c, \bar{x}_p)^T$ (defined by (6.11)) is called the *focal point* for the flow in $D$.

We will group into regions $R_j$ those domains $D_i^G$ where model (4.1)—in glucose growth—has the same dynamics and thus the same focal points. Considering Definition 6.5, we have the following focal points:

- $\forall x \in R_1 = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_1^G \cup D_2^G\right\}$

  $$x_c \to 0 \quad \wedge \quad x_p \to 0$$

  Thus, $\Phi_0^G = (0, 0)$ is the focal point of region $R_1$.

- $\forall x \in R_2 = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_3^G \cup D_4^G \cup D_5^G \cup D_6^G\right\}$

  $$\begin{aligned} x_c &\to \frac{k_c^0}{\mu_G\ \bar{x}_{p,G}^1 + \gamma_c} = \bar{x}_{c,G}^2 \\ x_p &\to \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ \nu_1\ k_p^0\ \mu_G}}{2\mu_G} = \bar{x}_{p,G}^1 \end{aligned}$$

  Thus, $\Phi_1^G = (\bar{x}_{c,G}^2, \bar{x}_{p,G}^1)$ is the focal point of region $R_2^G$.

- $\forall x \in R_3 = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_7^G \cup D_8^G\right\}$

  $$\begin{aligned} x_c &\to \frac{k_c^0}{\mu_G\ \bar{x}_{p,G}^2 + \gamma_c} = \bar{x}_{c,G}^1 \\ x_p &\to \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ \nu_1(k_p^0 + \ k_p^1)\mu_G}}{2\mu_G} = \bar{x}_{p,G}^2 \end{aligned}$$

  Thus, $\Phi_2^G = (\bar{x}_{c,G}^1, \bar{x}_{p,G}^2)$ is the focal point of region $R_3$.

The focal points $\Phi_i^G$ $(i = 1, ..., 3)$ are equilibrium points of model (4.1) provided that they belong to their respective regular domain, i.e. $\Phi(D) \in D$. The local stability of equilibrium points is given by the following theorem.

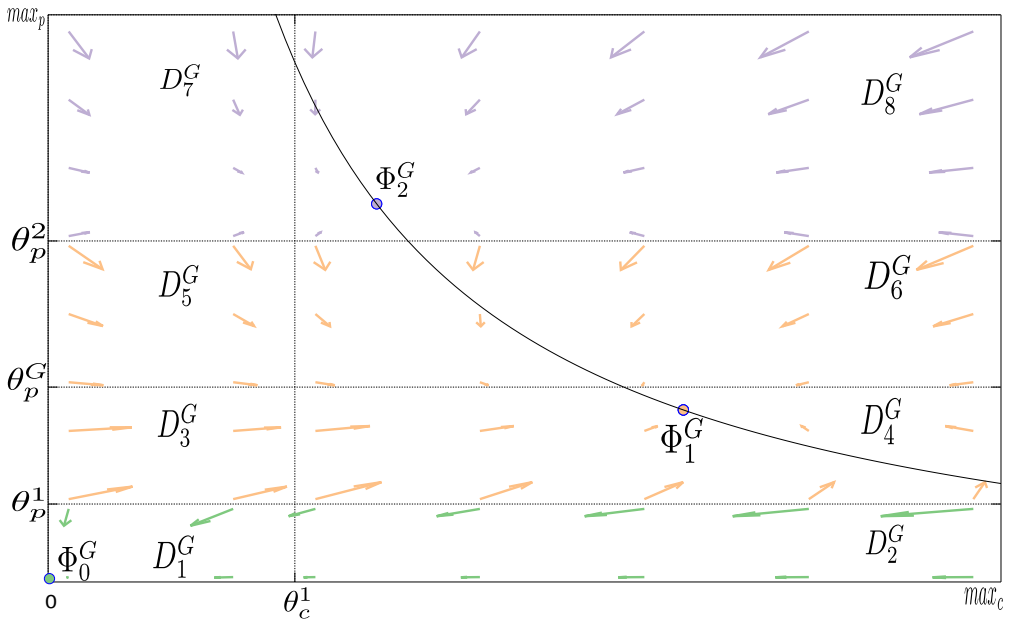



FIGURE 4.2: Phase plane of model (4.1) during growth in glucose. Parameter values used: $\theta_c^1 = 0.6$, $\theta_p^1 = 0.8$, $\theta_p^G = 2$, $\theta_p^2 = 3.5$, $k_c^0 = 7$, $k_c^1 = 10$, $k_p^0 = 40$, $k_p^1 = 50$, $\gamma_c = 1$, $\gamma_p = 1$, $\mu_G = 2$ end $\nu_1 = .5$. The black curve is the $x_c$-nullcline: $x_p = \dfrac{k_c^0}{x_c\ \mu_G} - \dfrac{\gamma_c}{\mu_G}$. Stable fixed points: $\Phi_0^G$, $\Phi_1^G$, $\Phi_2^G$.

**Theorem 4.2.** *Let $D$ be a regular domain and $\Phi(D)$ be the focal point of $D$. If $\Phi(D) \in D$, then $\Phi(D)$ is a locally stable point of model (4.1).*

*Proof.* Model (4.1) restricted to $D$ is given by (4.9). In order to assess the stability of $\Phi(D)$, we compute the Jacobian matrix of (4.9) calculated in $\Phi(D) = (\bar{x}_c, \bar{x}_p)^T$:

$$J(\bar{x}_c, \bar{x}_p) = \begin{pmatrix} -(\bar{\mu}\bar{x}_p + \gamma_c) & -\bar{\mu}\bar{x}_c \\ 0 & -(2\bar{\mu}\bar{x}_p + \gamma_p) \end{pmatrix}.$$

Since all the eigenvalues of $J$, which are the diagonal entries as $J$ is diagonal, are negative, $\Phi(D)$ turns out to be a locally stable point. □

Hence, there can be at most three locally stable steady states during growth on glucose.

Fig. 4.2 depicts the phase-plane of model (4.1). It can be seen that $\Phi_0^G$, $\Phi_1^G$, $\Phi_2^G$, (for the parameter values used) are locally stable steady states since they are within their respective regular domains (Theorem 6.3). Notably, it is easy to verify that $\Phi_0^G$ is locally stable for any set of parameters. It represents absence of growth and can happen when the initial condition $x_p(t_0)$, is too low—specifically $x_p(t_0) < \theta_p^1$—to initiate gene transcription or when the control input $\nu_1$ does not sufficiently induce CGEM expression, that is when $\bar{x}_{p,G}^1 < \theta_p^1$. We refer to $\Phi_0^G$ as the *trivial* fixed point. $\Phi_1^G$ represents CGEM

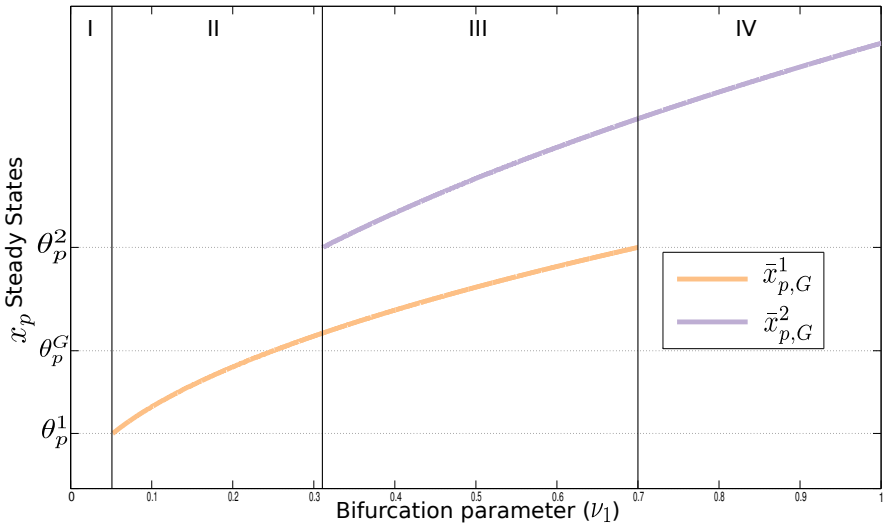



FIGURE 4.3: Bifurcation diagram for model (4.1) during growth in glucose, showing the non trivial locally stable steady states of $x_p$ as a function of the control input $\nu_1$. Other parameter values used are the same as those in Fig. 4.2. See Proposition 4.3 for more details.

basal level—leading to a low growth rate (see (4.2))— while CRP is at a high level, which is in agreement with high *crp* gene expression (by cAMP-CRP) at lower growth rate. Thus, because of the low growth rate achieved, we refer to $\Phi_1^G$ as the *low* fixed point. Conversely, at $\Phi_2^G$, CRP is at low level while CGEM , as well as $\mu$, have reached their highest stable values. Thus, $\Phi_2^G$ is named the *high* fixed point.

Since $\bar{x}_{p,G}^1(\nu_1)$ and $\bar{x}_{p,G}^2(\nu_1)$ are function of $\nu_1$, it turns out that the location of focal points $\Phi_1^G$ and $\Phi_2^G$, and thus the number of equilibria of model (4.1), depend on the control input $\nu_1$. Hence, choosing appropriate values of $\nu_1$ it is possible to control model (4.1) towards $\Phi_1^G$ or $\Phi_2^G$. To illustrate this, we have depicted in Fig. 4.3 the $x_p$-bifurcation diagram when parameter $\nu_1$ varies from 0 to 1 while the other parameter values are the same as those used in Fig. 4.2.

We notice that Fig. 4.3 is divided into four parts in which $x_p$ stability changes significantly. In part I, for those values of $\nu_1$ such that $\bar{x}_{p,G}^1 < \theta_p^1$ and $\bar{x}_{p,G}^2 < \theta_p^2$, neither $\Phi_1^G$ nor $\Phi_2^G$ are stable steady states. In this case, model (4.1) during growth on glucose converges towards the only stable point $\Phi_0^G$ (not depicted in Fig. 4.3). So, in I, the control input is too small to allow CGEM to reach a basal level, and prevents bacterial growth.

In part II, when $\bar{x}_{p,G}^1(\nu_1) > \theta_p^1$ and $\bar{x}_{p,G}^2(\nu_1) < \theta_p^2$ hold, only $\Phi_1^G$ is a stable steady state (besides the trivial one) according to Theorem 6.3. Hence, it turns out that choosing an initial condition of CGEM $x_p(t_0) > \theta_p^1$ and $\nu_1$ such that $\bar{x}_{p,G}^1(\nu_1) > \theta_p^1$ and $\bar{x}_{p,G}^2(\nu_1) < \theta_p^2$, we can control model (4.1) to the stable point $\Phi_1^G$.

In part III, characterized by $\theta_p^1 < \bar{x}_{p,G}^1(\nu_1) < \theta_p^2$ and $\bar{x}_{p,G}^2(\nu_1) > \theta_p^2$, both $\Phi_1^G$ and $\Phi_2^G$ are stable steady states: this is a region of bi-stability. Moreover, the phase plane corresponding to this configuration is depicted in Fig. 4.2, where we can also observe the presence of two separatrices $x_p = \theta_p^1$ and $x_p = \theta_p^2$. Is is clear that, depending on $x_p(t_0)$, the model can converge to $\Phi_1^G$ (if $\theta_p^1 < x_p(t_0) < \theta_p^2$) or to $\Phi_2^G$ (if $x_p(t_0) > \theta_p^2$).

In part IV, when $\bar{x}_{p,G}^1(\nu_1) > \theta_p^2$ holds, only $\Phi_2^G$ is a stable steady state and thus, whenever $x_p(t_0) > \theta_p^1$, model (4.1) converges to $\Phi_2^G$.

The open-loop control in glucose growth can be summarized as follows.

**Proposition 4.3.** *Consider model* (4.1) *with control input* $\nu_1$ *and initial condition* $x_p(t_0)$ *such that:*

- *if* $(\bar{x}_{p,G}^1(\nu_1) < \theta_p^1 \ \wedge \ \bar{x}_{p,G}^2(\nu_1) < \theta_p^2) \ \vee \ x_p(t_0) < \theta_p^1$, *then model* (4.1) *converges to the* trivial focal point $\Phi_0^G$ *(region I in Fig. 4.3);*
- *if* $\bar{x}_{p,G}^1(\nu_1) > \theta_p^1 \ \wedge \ \bar{x}_{p,G}^2(\nu_1) < \theta_p^2 \ \wedge \ x_p(t_0) > \theta_p^1$, *then model* (4.1) *converges to the* low focal point $\Phi_1^G$ *(region II in Fig. 4.3);*
- *if* $\theta_p^1 < \bar{x}_{p,G}^1(\nu_1) < \theta_p^2 \ \wedge \ \bar{x}_{p,G}^2(\nu_1) > \theta_p^2 \ \wedge \ x_p(t_0) > \theta_p^1$, *then model* (4.1) *is* bistable *(region III in Fig. 4.3) and notably:*
  - *if* $\theta_p^1 < x_p(t_0) < \theta_p^2$, *then model* (4.1) *converges to the* low focal point $\Phi_1^G$*;*
  - *if* $x_p(t_0) > \theta_p^2$, *then model* (4.1) *converges to the* high focal point $\Phi_2^G$
- *if* $\bar{x}_{p,G}^1(\nu_1) > \theta_p^2 \ \wedge \ x_p(t_0) > \theta_p^1$, *then model* (4.1) *converges to the* high focal point $\Phi_2^G$ *(region IV in Fig. 4.3).*

## 4.4 Growth rate expression for exponential phase

Here, to account for different dosage of inducer, we make an assumption to analytically characterize the function $\nu_1 = h(v)$. Notably, to describe the regulation of CGEM gene expression by the inducer, we employ a function typically used in synthetic biology [102]:

$$\nu_1(v) = \alpha + (1 - \alpha)\frac{v^n}{K_v^n + v^n} \tag{4.11}$$

where $v$ denotes inducerconcentration and $\alpha$ accounts for the basal transcriptional activity. Controlled gene expression follows Hill-type dosage-response curve with promoter-activator affinity $K_v$ and cooperative (Hill) coefficient $n$. During *exponential phase*—the period characterized by cell doubling— the bacterial culture shows a constant growth

rate [129]. This means that, according to (4.2), a stable fixed point of the CGEM has to be reached. Hence, our expression of growth rate during exponential phase reads:

$$\mu = \mu_G \bar{x}_p \tag{4.12}$$

where $\bar{x}_p$ is the CGEM concentration at steady state, which can be either $\bar{x}^1_{p,G}$ or $\bar{x}^2_{p,G}$— depending on the amount of inducerwhich determines the level of CGEM expression. Thus, our expression of growth rate during exponential phase can assume the two values below:

$$\mu(v) = \begin{cases} \mu_G \bar{x}^1_{p,G} = \dfrac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ \nu_1\ k_p^0\ \mu_G}}{2} \\ \\ \mu_G \bar{x}^2_{p,G} = \dfrac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ \nu_1\ (k_p^0 + k_p^1)\ \mu_G}}{2} \end{cases} . \tag{4.13}$$

Specifically, we assumed there is a particular value of inducer, i.e. $v^*$, such that for an appropriate choice of initial condition and for all $v \leq v^*$ the CGEM steady state is $\bar{x}^1_{p,G}$ while for all $v > v^*$ the steady state is $\bar{x}^2_{p,G}$. Thus, considering that, and substituting (4.11) into (7.9) we obtain the theoretical expression for growth rate during exponential phase:

$$\mu(v) = \begin{cases} -\dfrac{\gamma_p}{2}\left[1 - \sqrt{1 + \dfrac{4k_p^0\mu_G\alpha}{\gamma_p^2} + \dfrac{4k_p^0\mu_G(1-\alpha)}{\gamma_p^2}\dfrac{v^n}{K_v^n + v^n}}\right] & \text{if, } v \leq v^* \\ \\ -\dfrac{\gamma_p}{2}\left[1 - \sqrt{1 + \dfrac{4(k_p^0 + k_p^1)\mu_G\alpha}{\gamma_p^2} + \dfrac{4(k_p^0 + k_p^1)\mu_G(1-\alpha)}{\gamma_p^2}\dfrac{v^n}{K_v^n + v^n}}\right] & \text{if, } v > v^* \end{cases} \tag{4.14}$$

It is worthy to notice that expression (4.14) directly relates the growth rate $\mu$ during exponential phase to the amount of the inducer $v$. Hence, using (4.14) we can fine tune—by means of appropriate level of the inducer—the growth rate of the cells during the exponential phase.

## 4.5 *In silico* Identifiability Analysis of Growth Rate

Our collaborators (Jérôme Izard and Hans Geiselmann [3]) are currently performing an ongoing experiment on a synthetic *E. coli* – implementing the open-loop model depicted Fig. 4.1 – which relates the level of growth rate during the exponential phase to the amount of the inducer. In the future, these dose-response curves will be useful to calibrate and validate the growth rate expression (during exponential phase) (4.14).

[3] *Laboratoire Adaptation et Pathogénie des Microorganismes*, (CNRS UMR 5163), Université Joseph Fourier, Bâtiment Jean Roget, Faculté de Médecine-Pharmacie, La Tronche, France

Here, we used simulated data to fit the the growth rate model (4.14) and to study the identifiability of the parameters.

### 4.5.1 Problem Statement

Given a parametric non-linear model, such as (4.14), the relationship between a response variable (output) and one or more predictor variables (input) can be represented by the expression:

$$y = \eta(v, p) + \epsilon \ ,$$

where

- $y$ is an $n \times 1$ vector of observations of the response variable,
- $v$ is an $n \times m$ matrix of predictors,
- $p$ is a $q \times 1$ vector of unknown parameters to be estimated,
- $\eta$ is any function of $v$ and $p$,
- $\epsilon$ is an $n \times 1$ vector of independent, identically distributed random disturbances.

The nonlinear regression problem consists of finding a vector $\hat{p}$ minimizing a scalar cost function $J(p)$, which is generally a measurement of the agreement of experimental data with the outputs predicted by the model. The cost function that we have considered in this work is a weighted least squares criterion:

$$J(p) = \sum_{i=1}^{n} \frac{(y_i - \eta(v_i, p))^2}{y_i^2} \tag{4.15}$$

where $y_i$ denotes the $i$-th data-point of the observable $y$, measured at input-points $v_i$, and $\eta(v_i, p)$ the $i$-th observable as predicted by the parameters $p$. The parameters can be estimated numerically by:

$$\hat{p} = \arg\min \left[ J(p) \right] \ . \tag{4.16}$$

Determining the parameter vector $\hat{p}$ which minimizes $J(p)$ is only a part of the parameter estimation problem. In fact, when preparing to fit a mathematical model or expression to a set of experimental data, the prior assessment of parameter identifiability is a crucial aspect [170]. However, the structural identifiability analysis for non-linear models in systems biology is still a challenging question [47]. Whether or not parameters can be estimated uniquely depends on the model structure, the parametrization of the model and the experiment used to get the data [140].

Regarding this problem, we briefly recall two important definitions on identifiability [174]:

- the parameter $p_i$, $i = 1, ..., q$ is **structurally globally identifiable** if assuming ideal conditions (error-free model structure and unlimited noise-free observations $(v, y)$) and if for almost any $p^* \in \mathcal{P}$ (admissible parametric space $\mathcal{P}$),

$$y(p, v) = y(p^*, v), \forall v \Rightarrow p_i = p_i^*.$$

- the parameter $p_i$, $i = 1, ..., q$ is **structurally locally identifiable** if assuming ideal conditions (error-free model structure and unlimited noise-free observations $(v, y)$) and if for almost any $p^* \in \mathcal{P}$ (admissible parametric space $\mathcal{P}$), there exists a neighbourhood $V(p^*)$ such that

$$p \in V(p^*) \wedge y(p, v) = y(p^*, v), \forall v \Rightarrow p_i = p_i^*.$$

An important complement to the structural identifiability definitions is the notion of **practical identifiability**. Practical identifiability is indeed related to the quality of experimental data and their information content [61]. The question raised by this notion is the following: in the presence of observation errors and/or few data are reliable estimations of the parameters possible? Thus, once having determined the value of $\hat{p}$ minimizing the cost function $J(p)$, it is very important to find a realistic measure of how $\hat{p}$ is precise. To this end, the confidence intervals[4] of the estimated parameters have to be calculated.

It must be noted that, unlike for the linear case for which an exact theory exists, there is no exact theory for the evaluation of confidence intervals for systems which are nonlinear in the parameters. An approximate method based on a local linearisation of the output function $\eta(v, p)$ is generally used [75, 152], thus the confidence region is evaluated as a function of the parameter covariance matrix. The applicability of such approximate method requires that the response function $\eta(v, p)$ must be continuous in its arguments $(v, p)$, the first partial derivatives $\frac{\partial}{\partial p_i}\eta(v, p)$ must be continuous in its arguments $(v, p)$, and the second partial derivatives $\frac{\partial^2}{\partial p_i \partial p_j}\eta(v, p)$ must be continuous in its arguments $(v, p)$, but our model (4.14) does not satisfy these conditions because of the discontinuity in $v = v*$. Hence, in the remainder of the paper a computational method, based on *in silico* generated data, is suggested to argue the practical identifiability of non-linear discontinuous model such as (4.14).

[4] A confidence interval $[\sigma_i^-, \sigma_i^+]$ of a parameter estimate $\hat{p}_i$ to a confidence level $\alpha$ signifies that the true value $p_i^*$ is located within this interval with probability $\alpha$.

### 4.5.2 Generation of Simulated Data Sets

In order to assess the quality of parameter estimation and thus the practical identifiability of parameters in (4.14), artificial data were generated by simulation of (4.14) from a set of pre-defined parameters (to be considered as true values). The true parameter values (Tab. 4.5.2) were chosen from physiological parameters of *E.coli* cells [31, 114] and were based on similar studies of this type [160].

| $k_p^0$ $[\mu M \cdot min^{-1}]$ | $k_p^1$ $[\mu M \cdot min^{-1}]$ | $\gamma_p$ $[min^{-1}]$ | $\mu_G$ $[(\mu M \cdot min)^{-1}]$ | $\alpha$ | $K_v$ $[\mu M]$ | $n$ | $v^*$ $[\mu M]$ |
|---|---|---|---|---|---|---|---|
| 0.02 | 0.11 | 0.006 | 0.0014 | 0.1 | 30 | 2 | 50 |

TABLE 4.1: Nominal parameter values

Thus, the artificial growth rate values have been simulated considering a measurement error proportional to the nominal value of growth rate:

$$y = \mu(v) + \sigma\mu(v)\mathcal{N}(0,1) \tag{4.17}$$

where $\mathcal{N}(0,1)$ is a normally distributed random variable with zero mean and unit variance and $\sigma\mu(v)$ is the standard deviation of the observation errors. Four different types of data sets were considered to account for practical identifiability:

- data set I, with $v = [0, 5, 10, 15, ..., 295, 300, 1000]$ and $\sigma = 10^{-2}$;
- data set II, with $v = [0, 10, 20, 30, ..., 290, 300, 1000]$ and $\sigma = 10^{-2}$;
- data set III, with $v = [0, 5, 10, 15, ..., 295, 300, 1000]$ and $\sigma = 5 \cdot 10^{-2}$;
- data set IV, with $v = [0, 10, 20, 30, ..., 290, 300, 1000]$ and $\sigma = 5 \cdot 10^{-2}$;

Notably, data sets I, II, III and IV, have been generated with different number of points ($N_{exp}$) and different intensities of noise ($\sigma$) to study the practical identifiability of the parameters in four realistic experimental conditions. In particular, data sets I and III have the same number of data points, i.e. $N_{exp} = 62$, but different noise, $\sigma = 10^{-2}$ for data set I and $\sigma = 5 \cdot 10^{-2}$ for data set III. Data set II and IV have less number of points, i.e. $N_{exp} = 32$, while the level of noise considered is $\sigma = 10^{-2}$ for data set II and $\sigma = 5 \cdot 10^{-2}$ for data set VI.

### 4.5.3 Model Parametrization and Global Optimization

First, to avoid evident structural identifiability problems we will group together those parameters in (4.14) which appear as combinations of products and/or quotients between parameters. Thus, after some algebraic manipulations expression (4.14) reads as:

$$\mu(v) = \begin{cases} -\frac{\gamma_p}{2}\left[1 - \sqrt{1 + \frac{4k_p^0\mu_G\alpha}{\gamma_p^2}\left(1 + \frac{(1-\alpha)}{\alpha}\frac{v^n}{K_v^n + v^n}\right)}\right] & \text{if, } v \leq v^* \\ -\frac{\gamma_p}{2}\left[1 - \sqrt{1 + \frac{4(k_p^0 + k_p^1)\mu_G\alpha}{\gamma_p^2}\left(1 + \frac{(1-\alpha)}{\alpha}\frac{v^n}{K_v^n + v^n}\right)}\right] & \text{if, } v > v^* \end{cases} \tag{4.18}$$

Moreover, to avoid dependence on physical unit as well as to overcome possible scaling problem and to reduce the number of parameters, we decided to calculate a non-dimensional version of expression (4.18). Notably, the non-dimensional slope $\mu_N(v)$ is obtained by dividing $\mu(v)$ in (4.18) for the minimal growth rate, which is achieved at the minimum value of the inducer, i.e. at $v = v_0$, which for our data sets I, II, III, IV consists in $v_0 = 0$. Thus, considering the necessary condition $v_0 < v^*$, the non-dimensional growth rate during the exponential phase reads:

$$\mu_N(v) = \begin{cases} \dfrac{1 - \sqrt{1 + \frac{4k_p^0\mu_G\alpha}{\gamma_p^2}\left(1 + \frac{(1-\alpha)}{\alpha}\frac{v^n}{K_v^n + v^n}\right)}}{1 - \sqrt{1 + \frac{4k_p^0\mu_G\alpha}{\gamma_p^2}}} & \text{if, } v \leq v^* \\ \dfrac{1 - \sqrt{1 + \frac{4(k_p^0 + k_p^1)\mu_G\alpha}{\gamma_p^2}\left(1 + \frac{(1-\alpha)}{\alpha}\frac{v^n}{K_v^n + v^n}\right)}}{1 - \sqrt{1 + \frac{4k_p^0\mu_G\alpha}{\gamma_p^2}}} & \text{if, } v > v^* \end{cases} \tag{4.19}$$

Now, considering the following parametrization

$$p_1 = \frac{4k_p^0\mu_G\alpha}{\gamma_p^2};\ p_2 = \frac{(1-\alpha)}{\alpha};\ p_3 = K_v;\ p_4 = n;\ p_5 = \frac{4k_p^1\mu_G\alpha}{\gamma_p^2};\ p_6 = v^*$$

the expression (4.19) can be rewritten as

$$\mu_N(v, p) = \begin{cases} \dfrac{1 - \sqrt{1 + p_1\left(1 + p_2\frac{v^{p_4}}{p_3^{p_4} + v^{p_4}}\right)}}{1 - \sqrt{1 + p_1}} & \text{if, } v \leq p_6 \\ \dfrac{1 - \sqrt{1 + (p_1 + p_5)\left(1 + p_2\frac{v^{p_4}}{p_3^{p_4} + v^{p_4}}\right)}}{1 - \sqrt{1 + p_1}} & \text{if, } v > p_6 \end{cases} \tag{4.20}$$

where $p = [p_1, p_2, p_3, p_4, p_5, p_6]$ and, considering the true parameters values in Tab 4.5.2 we obtain the true vector of parameters $p^*$:

$$p^* = [0.3033, 9, 30, 2, 1.6683, 50] \ . \tag{4.21}$$

Similarly, the data sets I to IV will also be normalized to their minimal value, i.e., each output-point is divided by the minimal observation value, that is $y_{min} = \mu(v_0)$, where $v_0 = 0$.

Our approach in identifying the unknown parameters of model (4.19) consists in solving a non-linear least squares minimization problem, using a hybrid optimization approach which makes use of the functions *ga* (Genetic Algorithm [87]) and *GlobalSearch* of the *MATLAB Global Optimization Toolbox*. To start, we used the Genetic Algorithm (GA) for $10^4$ generations to get near an optimum point. The genetic algorithm does not use derivatives to detect descent in its minimization steps. Hence, it is a good choice for non-differentiable and/or discontinuous problems. Moreover, GA does not necessarily need an user supplied initial guess, which in most case leads to local sub-optimal convergence if the initial guess is far from the global optimum. The result obtained with the genetic algorithm is then used as initial point of a hybrid function, to further improve the value of the cost function $J(p)$. We decided to use the *GlobalSearch*[5] command as hybrid function since it searches many basins of attraction near the starting point given by GA, arriving faster at an even better solution.

### 4.5.4 In Silico Practical Identifiability Analysis

The practical identifiability of model (4.20) has been tested using data sets I, II, III and IV, which have different values of errors' measurement and different data points. Hence, these artificial data are suitable to mimic realistic experimental set-ups.

For each data set mentioned above, parameters' confidence intervals have been computed following a *Monte Carlo*-like approach.

Notably, $N_{simul} = 200$ runs of the previously described hybrid optimization were performed. Where, at each of the $N_{simul}$ runs, a new realization of the artificial measurements—according to the inputs and noise statistic of each data set—is considered. These $N_{simul}$

[5] *GlobalSearch* first runs *fmincon* from the start point you give. If this run converges, GlobalSearch records the start point and end point for an initial estimate on the radius of a basin of attraction. Then, *GlobalSearch* solver starts a local solver (*fmincon*) from multiple starting points and store local and global solutions found during the search process. Notably, the *GlobalSearch* solver first uses a scatter-search algorithm to randomly generate multiple starting points, then filters non-promising start points based upon objective and constraint function values and local minima already found, and finally runs a constrained nonlinear optimization solver to search for a local minimum from the remaining start points.

| | DATA SET I<br>$\sigma = 10^2$<br>$N_{exp} = 62$ | DATA SET II<br>$\sigma = 10^{-2}$<br>$N_{exp} = 32$ | DATA SET III<br>$\sigma = 5 \cdot 10^{-2}$<br>$N_{exp} = 62$ | DATA SET IV<br>$\sigma = 5 \cdot 10^{-2}$<br>$N_{exp} = 32$ |
|---|---|---|---|---|
| $CI_1$ | $0.3328 \pm 0.4939$ | $0.3738 \pm 0.5441$ | $0.2631 \pm 0.4220$ | $0.32 \pm 0.49$ |
| $CI_2$ | $9.23 \pm 3.45$ | $9.36 \pm 3.88$ | $8.63 \pm 3.06$ | $9.21 \pm 4.67$ |
| $CI_3$ | $30.16 \pm 3.55$ | $30.00 \pm 3.55$ | $29.39 \pm 5.15$ | $30.33 \pm 7.52$ |
| $CI_4$ | $2.002 \pm 0.079$ | $2.011 \pm 0.089$ | $2.006 \pm 0.232$ | $2.01 \pm 0.33$ |
| $CI_5$ | $2.053 \pm 4.192$ | $2.39 \pm 4.51$ | $1.53 \pm 3.59$ | $1.93 \pm 3.99$ |
| $CI_6$ | $53.32 \pm 4.48$ | $55.98 \pm 6.99$ | $53.06 \pm 3.58$ | $56.70 \pm 6.79$ |

TABLE 4.2: Confidence intervals of estimated parameters $\hat{p}_i$ when (4.20) is fitted to (non-dimensionalized) data sets I, II,III,IV. The confidence intervals for parameters become larger at increasing values of the measurement error and at decreasing numbers of data points, indicating possible practical identifiability problems especially for $\hat{p}_1$ and $\hat{p}_5$.

optimization yields $N_{simul}$ estimated values for each parameter $p_i$, $i = 1, \ldots, 5$. Then, for each $i$, an average value, $\hat{m}_i$, and a standard deviation, $\hat{s}_i$, were computed by fitting a Gaussian distribution $\mathcal{N}(\hat{m}_i, \hat{s}_i^2)$ to the histogram of the $N_{simul}$ values of $p_i$. Thus, the 95% confidence interval ($CI_i$) for the $p_i$ parameter is calculated as:

$$CI_i = \hat{m}_i \pm 1.96\hat{s}_i \tag{4.22}$$

This leads to the confidence intervals listed in Table 4.2.

As we can see in Table 4.2, parameters $p_i$ for $i \in \{2, 3, 4, 6\}$ do not show any practical identifiability issues, as the true value is contained in the respective CI with sufficiently precision. On the contrary, the CIs of parameters $\hat{p}_1$ and $\hat{p}_5$ tend to become very large at increasing values of the measurement's errors ($\sigma$) and at decreasing numbers of data points, indicating that in real experimental conditions (that is, limited and noisy data), the precise identification of these parameters might be impracticable. Moreover, we found that the correlation coefficient ($R$) between the two vectors of estimated parameters parameters $\hat{p}_1$ and $\hat{p}_5$ is $R = 0.99$, for all data sets. Recall that the correlation coefficient measures the interrelationship between $\hat{p}_1$ and $\hat{p}_5$ quantifying the compensation effects of changes in the parameter values on the model output. In fact, when two parameters are highly correlated, a change in the model output caused by a change in a model parameter can be balanced by a proper change in the other parameter value. Thus, instead of considering the CIs of $\hat{p}_1$ and $\hat{p}_5$ separately—which are not significant—we have computed the confidence interval of their ratio, i.e. $\hat{p}_5/\hat{p}_1$. These results are presented in Table 4.3. As we can notice in Table 4.3, the CIs of $\hat{p}_5/\hat{p}_1$ are accurate,

| | DATA SET I $\sigma = 10^{-2}$ $N_{exp} = 62$ | DATA SET II $\sigma = 10^{-2}$ $N_{exp} = 32$ | DATA SET III $\sigma = 5 \cdot 10^{-2}$ $N_{exp} = 62$ | DATA SET IV $\sigma = 5 \cdot 10^{-2}$ $N_{exp} = 32$ |
|---|---|---|---|---|
| $CI_{\hat{p}_5/\hat{p}_1}$ | $5.29 \pm 2.39$ | $5.54 \pm 2.43$ | $4.99 \pm 1.15$ | $5.2 \pm 1.3$ |

TABLE 4.3: Confidence intervals of the ratio $\hat{p}_5/\hat{p}_1$ when (4.20) is fitted to (non-dimensionalized) data sets I, II,III,IV.

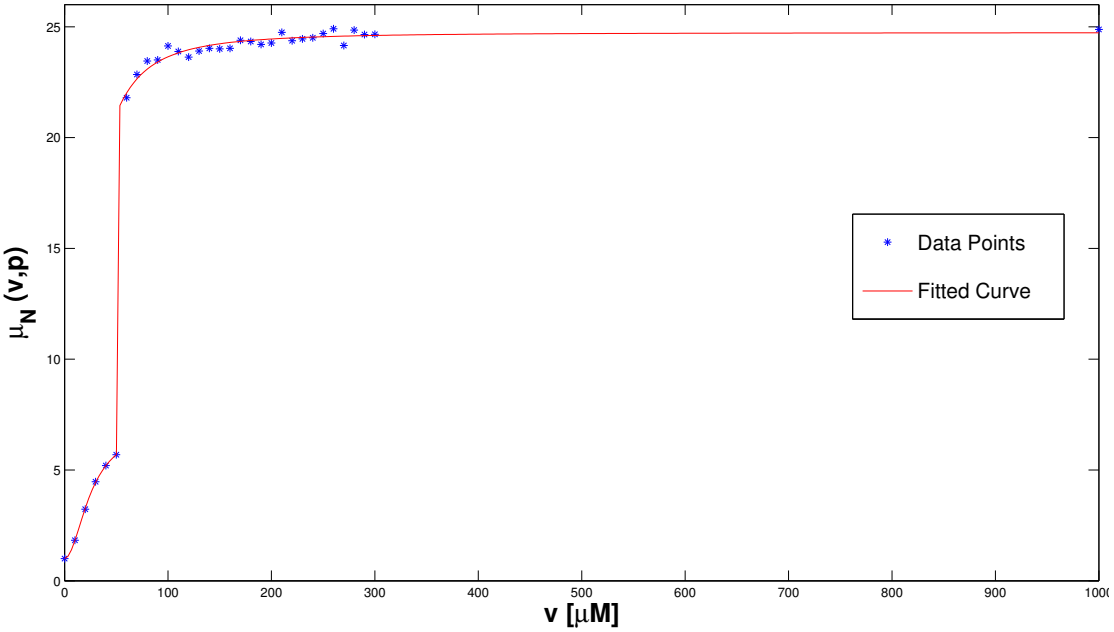



FIGURE 4.4: Fitting the growth rate function (4.20) using one realization of the non-dimensional data set II. The blue points are the normalized artificial data generated according to specification of data set II. The red curve is the function (4.20) when $\hat{p}$ is used.

since they contain the true value of the ratio $p_5^*/p_1^* = 5.5$, and more precise since their relative width is smaller than the relative width of $CI_1$ and $CI_5$.

It must be noted that a further reduced model which takes into account the correlation between $p_5$ and $p_1$ can not be achieved. This because expression (4.20) can be rewritten in terms of the ratio and either $p_5$ or $p_1$. Fig 4.4 shows the fitting of model (4.20) to one realization of data set IV.

## 4.6 Conclusions

In this paper, a minimal model consisting of two variables (the concentrations of two gene products) and an input (an inducer) was analysed and used to describe one possible mechanism to control the growth rate of *E. coli* cells during exponential phase. This

model is based on the piecewise linear formalism but a new, non-linear, term was added to account for the dilution effect during growth. The qualitative dynamics of the model can thus be studied, and the bifurcation diagram with respect to the input is obtained. Moreover, this mathematical formalism allows derivation of an analytic expression for the growth rate as function of the input. This expression has two applications:

- it can be directly fitted to experimental data to estimate a set of parameters (this is an advantage relative to the typical "indirect" parameter estimation by fitting to the numerical solutions of the differential equations);
- it provides an indication of how to control the growth rate to a desired value by adding a given quantity of inducer.

Finally, practical identifiability analysis based on numerical simulations is presented, which shows that some issues may arise with noisy measurements. In this case, our analysis suggests that the original growth rates' measurements should be adimensionalized and unknown parameters grouped into a new set of "lumped" parameters in order to obtain local identifiability. Notably, we found that only the ratio between the estimated parameters $\hat{p}_1$ and $\hat{p}_5$ can be estimated with sufficient precision in the case when only limited and noisy data are available. This study and the conclusions on identifiability will be most useful to help dealing with and solving parameter estimation problems with real data sets.

**Acknowledgements.**

This work was supported by ANR GeMCo (French national research agency).

# Chapter 5

# Controlling bacterial growth: *in silico* feedback law design to re-wire the genetic network

Here we present an unpublished work, in which we continue studying and developing the synthetic gene circuit model introduced in Chapter 4, with the aim of proposing some qualitative control strategies to tune and modify growth rate in *Escherichia coli*. First, an open loop circuit is obtained by externally controlling the gene encoding for RNA polymerase through an inducible promoter. Notably, varying the constant input we can impose a low or a high stable value to the growth rate. Finally, by appropriately rewiring and designing the gene interactions, we have formulated a feedback law, which mathematically shows the *inverse diauxie* in an appropriate range of the input values.

## 5.1 Introduction

Growth control is essential in industrial biotechnology and fundamental research of this kind could pave the way to novel types of antimicrobial strategies. Indeed, the basic effect of an antibiotic is to slow (or arrest) growth.

Here, we first introduce a novel mathematical formalism to qualitative model gene expression and dilution due to growth rate (Section 5.2). Then, we focus on the gene expression machinery of the bacterium *Escherichia coli*, with the aim of controlling the growth rate of the cells (Section 5.3). To this end, we present and analyse a bi-dimensional open-loop model in which the growth rate is controlled externally—expressing RNA polymerase from a synthetic IPTG-inducible promoter (Section 5.4,

5.5). Finally, by designing a feedback law, we obtain a closed-loop model which implements a possible rewiring of the genetic network and modifies the wild type response of the bacteria to nutrients (Section 5.6). More specifically, the model is able to theoretically generate *inverse diauxie*, notably higher growth rate on an alternative sugar than on the preferred glucose (Section 5.7). The new rewiring does allow both to control growth rate in response to a sugar source and fine-tune it in response to an (external) inducer.

## 5.2 Piecewise linear models with dilution

The regulation of gene expression plays a pivotal role in the functioning of cells. Mathematical modelling and computational techniques are fundamental to the understanding of these genetic regulatory networks ([14, 54]). The principal modelling challenges come from incomplete knowledge of the networks, and the dearth of quantitative data for identifying kinetic parameters required for detailed mathematical models. Qualitative methods overcome both of these difficulties and are thus well-suited to the modelling and simulation of genetic networks.

A class of widespread and well studied qualitative models are piecewise linear (PL) systems, originally introduced by Glass and Kauffman [85]. PL models of regulatory genetic networks are built with discontinuous (step) functions. The use of step function has been motivated by the experimental observation that the activity of certain genes changes in a switch-like manner at a threshold concentration of a regulatory protein [180]. The PL model has the general form

$$\dot{x}_i = f_i(x) - d_i x_i, \quad 1 \leq i \leq n, \tag{5.1}$$

where $x = (x_1, ..., x_n)^T$ is a non-negative vector of protein concentrations. The non-negative quantities $f_i(x)$ and $d_i x_i$ represent synthesis and degradation rates for each protein $x_i$ respectively. The function $f_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ represents the expression rate of the gene $i$ depending on the whole state $x$. However, $f_i(x)$ can be detailed as:

$$f_i(x) = \sum_{l=1}^{L_i} k_{il} b_{il}(x)$$

where $k_{il} > 0$ is a rate parameter and $b_{il}(x)$ is a combination of step functions $s^+, s^-$:

$$s^+(x_i, \theta_i^j) = \begin{cases} 1 & \text{if } x_i > \theta_i^j \\ 0 & \text{if } x_i < \theta_i^j \end{cases}; \quad s^-(x_i, \theta_i^j) = 1 - s^+(x_i, \theta_i^j).$$

More details on PL models can be found in [40, 43, 56, 90, 91].

Since our purpose is dealing with gene expression during bacterial growth, we take into account the fact that cells remove proteins by two processes: degradation and dilution due to cell growth [66]. Thus, the PL formalism can be extended setting $d_i = \mu + \gamma_i$ in (5.1), where $\mu$ is the bacterial growth rate responsible for the proteins' dilution and $\gamma_i$ is a degradation constant. Moreover, in bacteria, growth rate is intimately intertwined with gene expression ([114, 151]) and with the type of substrate [124]. Hence, keeping complexity to a minimum, we assume growth rate $\mu$ to be proportional—with a constant $\bar{\mu}$ depending on the quality of medium—to the concentration of one of $n$ proteins, which is supposed to be tightly related to $\mu$. Assuming the growth proportional protein to be $x_n$, $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ reads as:

$$\mu(t) = \bar{\mu}\ x_n(t). \tag{5.2}$$

Therefore, the PL system with dilution effect, henceforth referred to as *piecewise wise non-linear* (PWNL) system, has the general form:

$$\dot{x}_i = f_i(x) - (\bar{\mu}\ x_n + \gamma_i)x_i, \quad 1 \leq i \leq n \tag{5.3}$$

or in vector notation

$$\dot{x} = f(x) - d(x_n)x, \tag{5.4}$$

where $f = (f_1, ..., f_n)$ and $d = diag(\bar{\mu}\ x_n + \gamma_1, ..., \bar{\mu}\ x_n + \gamma_n)$, where *diag* is the diagonal matrix corresponding to the vector. The dynamics of the PWNL system (5.3) can be studied in the $n$-dimensional state-space $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$, where each $\Omega_i$ is defined by $\Omega_i = \{x_i \in \mathbb{R}_{\geq 0} | 0 \leq x_i \leq max_i\}$ for some maximum concentration value $max_i$. A protein encoded by a gene will be involved in different interactions at different concentration thresholds, so for each variable $x_i$, we assume there are $p_i$ ordered thresholds $\theta_i^1, ..., \theta_i^{p_i}$ (we also define $\theta_i^0 = 0$ and $\theta_i^{p_i+1} = max_i$ ). The $(n-1)$-dimensional hyper-planes defined by these thresholds partition $\Omega$ into hyper-rectangular regions we call *domains*. Specifically, a domain $D \subset \Omega$ is defined to be a set $D = D_1 \times ... \times D_n$, where $D_i$ is one of the following:

$$\begin{aligned}
D_i &= \left\{x_i \in \Omega_i | 0 \leq x_i < \theta_i^1\right\} \\
D_i &= \left\{x_i \in \Omega_i | \theta_i^j < x_i < \theta_i^{j+1}\right\} \quad \text{for } j \in \{1, ..., p_i - 1\} \\
D_i &= \{x_i \in \Omega_i | \theta_i^{p_i} < x_i \leq max_i\} \\
D_i &= \left\{x_i \in \Omega_i | x_i = \theta_i^j\right\} \qquad\qquad \text{for } j \in \{1, ..., p_i\}.
\end{aligned}$$

Let $\mathcal{D}$ be the set of domains in $\Omega$ . A domain $D \in \mathcal{D}$ is called a *regulatory domain* if none of the variables $x_i$ has a threshold value in $D$ (it is the full hyper-rectangle).

In contrast, a domain $D \in \mathcal{D}$ is called a *switching domain* of order $k \leq n$ if exactly $k$ variables have threshold values in $D$ [127]. The corresponding variables $x_i$ are called *switching variables* in $D$. The two sets of domains are respectively denoted by $\mathcal{D}_r$ and $\mathcal{D}_s$.

### 5.2.1 Solutions in Regular Domains

For any regulatory domain $D$, the function $f(x)$ is constant for all $x \in D$, and it follows that the PWNL system (5.4) can be written as

$$\dot{x} = f^D - d(x_n)x, \tag{5.5}$$

where $f^D$ is constant in $D$. For any $x(t_0) \in D$ the unique solution of (5.5) can be found explicitly by solving first the $n$-component of (5.5)— $\dot{x}_n = f_n^D - (\bar{\mu}\, x_n + \gamma_n)x_n$—which is an autonomous differential equation, and then solving the $i$-components ($i = 1, ..., n-1$), having substituted $x_n(t)$ into them. Thus, it can be shown that $x_i(t)$ ($i = 1, ..., n-1$) is given by:

$$x_i(t) = \frac{1}{b(t)}\left(b(t_0)x_i(t_0) + f_i^D \int_{t_0}^{t} b(s)ds\right)$$

where $b(t) = exp\left(\int_{t_0}^{t}(\bar{\mu}\, x_n(\tau) + \gamma_n)d\tau\right)$. Moreover, defining $\Phi(D) = (\bar{x}_1, ..., \bar{x}_n)^T$ with

$$\begin{aligned} \bar{x}_i &= \frac{f_i^D}{\bar{\mu}\bar{x}_n + \gamma_i}, \quad i = 1, ..., n-1 \\ \bar{x}_n &= \frac{-\gamma_n + \sqrt{\gamma_n^2 + 4\bar{\mu}f_n^D}}{2\bar{\mu}}, \end{aligned} \tag{5.6}$$

(it is easy to check that $\bar{x}_n$—in (5.6)—is the only positive solution of $\dot{x}_n = 0$) it turns out that either $x(t) \to \Phi(D)$ as $t \to \infty$ or $x(t)$ reaches the boundary of $D$.

**Definition 5.1.** Given a regulatory domain $D \in \mathcal{D}_r$, the point $\Phi(D) = (\bar{x}_1, ..., \bar{x}_n)^T \in \Omega$ (defined by (5.6)) is called the *focal point* for the flow in $D$.

Different regulatory domains will usually have different focal points. In general, all solutions in a regulatory domain $D$ flow towards the focal point $\Phi(D)$ until they either reach it or leave the domain $D$. What happens when a solution leaves a regulatory domain $D$ and enters a switching domain in the boundary of $D$? Since the step functions are not defined when a variable $x_i$ takes some threshold value $\theta_i^{q_i}$, the vector field is undefined on the switching domains. We need to precise our definition of solutions.

### 5.2.2 Solutions in Switching Domains

To provide the existence and the possibility for solutions to be continued on all domains, we have to define the right-hand side of system (5.4) at the points of discontinuity of the function $f$. To this end, we use a construction originally proposed by Filippov [73] and recently applied to PL systems ([56], [90]). The method consists of extending the system (5.4) to a differential inclusion,

$$\dot{x} \in H(x), \tag{5.7}$$

where $H$ is a set valued function (i.e. $H(x) \subseteq \mathbb{R}^n$). If $D \in \mathcal{D}_r$, then we define $H$ simply as

$$H(x) = \{f^D - d(x_n)x\}, \quad \forall x \in D. \tag{5.8}$$

If $D \in \mathcal{D}_s$, we define $H$ as

$$H(x) = \overline{co}(\{f^{D'} - d(x_n)x \mid D' \in R(D)\}), \quad \forall x \in D, \tag{5.9}$$

where $R(D) = \{D' \in \mathcal{D}_r \mid D \subseteq \partial D'\}$ is the set of all regulatory domains with $D$ in their boundary, and $\overline{co}(X)$ is the closed convex hull of $X$. For switching domains, $H(x)$ is typically multi-valued so solutions of the differential inclusion are defined as follows.

**Definition 5.2.** A solution of (5.7) on $[0, T]$ in the sense of Filippov is an absolutely continuous function (w.r.t. $t$) $\xi_t(x_0)$ such that $\xi_0(x_0) = x_0$ and $\dot{\xi}_t \in H(\xi_t)$, for almost all $t \in [0, T]$.

Now, we shall explain how to construct the set $H(x)$ at discontinuity points of $f$.

Let us consider the case where $x$ belongs to a switching domain $S$ separating two $n$-regular domains $D_1$ and $D_2$. Hence,

$$H(x) = \overline{co}(\{f^{D_1} - d(x_n)x, f^{D_2} - d(x_n)x\})$$

represents the segment joining the endpoints of the vectors $g_1 = f^{D_1} - d(x_n)x$ and $g_2 = f^{D_2} - d(x_n)x$. Trajectories can cross $S$ if the vector fields $g_1$ and $g_2$ point in a similar direction, slide along $S$ if $g_1$ and $g_2$ point in opposite direction towards $S$ and be repelled from $S$ if $g_1$ and $g_2$ point in opposite direction away from $S$. The last two cases are known as *stable* and *unstable sliding motion* in the literature [40]. Moreover, the velocity of the sliding motion (stable or unstable) on $S$ is given by

$$\dot{x} = f^S - d(x_n)x. \tag{5.10}$$

**Theorem 5.3.** *Assume that $S$ is in the hyper-plane $C_i^j = \{x \in \mathbb{R}^n_{\geq 0} \ : \ x_i = \theta_i^j\}$ and a sliding motion (stable or unstable) occurs on $S$. The vector $f^S$ in (5.10) is given by*

$$f^S = \alpha f^{D_1} + (1-\alpha) f^{D_2},$$
$$\alpha = \begin{cases} \dfrac{f_i^{D_2} - (\gamma_i + \bar{\mu} x_n)\theta_i^j}{f_i^{D_2} - f_i^{D_1}}, & \text{if } i \in \{1, ..., n-1\} \\ \dfrac{f_i^{D_2} - (\gamma_i + \bar{\mu}\theta_i^j)\theta_i^j}{f_i^{D_2} - f_i^{D_1}}, & \text{if } i = n. \end{cases} \tag{5.11}$$

*Proof.* The segment joining the endpoints of the vectors $f^{D_1} - d(x_n)x$ and $f^{D_2} - d(x_n)x$ is expressed by

$$\alpha f^{D_1} + (1-\alpha) f^{D_2} - d(x_n)x, \quad 0 \leq \alpha \leq 1.$$

Since the state trajectories during sliding motion are on the hyper-plane $x_i = \theta_i^j$, the parameter $\alpha$ in (5.11) is selected such that the state velocity vector of the system (5.10) is in this hyper-plane. Thus, depending on whether the sliding mode occurs in $x_i = \theta_i^j$ ($i \in \{1, ..., n-1\}$) or in $x_n = \theta_n^j$, we have two values of $\alpha$, which are found from the conditions

$$\alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\gamma_i + \bar{\mu} x_n)\theta_i^j = 0, \text{ if } i \in \{1, ..., n-1\}$$
$$\alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\gamma_i + \bar{\mu}\theta_i^j)\theta_i^j = 0, \text{ if } i = n.$$

□

We notice that in the case $i = n$ the value of $\alpha$ is constant $\forall x \in S$ and thus a sliding mode occurs along the entire switching domain $S$. By contrast, in the case $i \in \{1, ..., n-1\}$ the value of $\alpha$ depends on $x_n$, that means that a sliding motion occurs on $S$ as long as the vector fields point in opposite direction towards (or away) $S$ or, equivalently, as long as $0 < \alpha(x_n) < 1$ $\forall x_n \in S$. Specifically, it could happen that solutions slide for a while along $S$ and then leave it as soon as the condition $0 < \alpha(x_n) < 1$ does not hold any more.

It is useful to define a concept analogous to the focal points defined for regulatory domains, extended to deal with switching domains.

**Definition 5.4.** We recall that $supp(D)$ is the $(n-k)$-dimensional hyperplane supporting $D$. Let $D$ be a switching domain of order $k$, then its focal set $\Phi(D)$ is

$$\Phi(D) = supp(D) \ \cap \ \{x : \ 0 \in H(x)\}, \tag{5.12}$$

where $H(x)$ is defined as in (5.9).

Hence, $\Phi(D)$ for $D \in \mathcal{D}_s$ is the set containing all the equilibrium points of the differential inclusion (5.7), which lie on $supp(D)$. Thus, $\Phi(D)$ can be a singleton, but more generally is a closed convex bounded set and hence is referred to as a focal set.

To rule out some singular cases when proving results on stability, we make a technical assumption on the focal sets for our system.

***Assumption*** 1. $\forall\ D\ \in \mathcal{D}$, we assume that

$$\Phi(D) \cap supp(D') = \{\},\ \forall D^{'} \in R(D).$$

It essentially says that for every regular and switching domain $D$, the focal set $\Phi(D)$ does not intersect the supporting hyperplane of any domain $D'$ in the boundary of $D$.

### 5.2.3 Equilibria and Stability in Regular Domains

The focal points are equilibrium points of the PWNL system (5.4) provided that they belong to their respective regular domain, i.e. $\Phi(D) \in D$. If this is the case, the focal points are referred to as *regular equilibria*. The local stability of equilibrium points is given by the following theorem.

**Theorem 5.5.** *Let $D \in \mathcal{D}_r$ and $\Phi(D)$ be the focal point of $D$. If $\Phi(D) \in D$, then $\Phi(D)$ is a locally stable point of system* (5.4).

*Proof.* System (5.4) restricted to $D \in \mathcal{D}_r$ is given by (5.5). In order to assess the stability of $\Phi(D)$, we compute the Jacobian matrix of (5.5) calculated in $\Phi(D) = (\bar{x}_1, ..., \bar{x}_n)^T$:

$$J(\bar{x}_1, ..., \bar{x}_n) =$$
$$\begin{pmatrix} -(\bar{\mu}\bar{x}_n + \gamma_1) & 0 & 0 & \dots & 0 & -\bar{\mu}\bar{x}_1 \\ 0 & -(\bar{\mu}\bar{x}_n + \gamma_2) & 0 & \dots & 0 & -\bar{\mu}\bar{x}_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & -(\bar{\mu}\bar{x}_n + \gamma_{n-2}) & 0 & -\bar{\mu}\bar{x}_{n-2} \\ 0 & \dots & \dots & 0 & -(\bar{\mu}\bar{x}_n + \gamma_{n-1}) & -\bar{\mu}\bar{x}_{n-1} \\ 0 & \dots & \dots & \dots & 0 & -(2\bar{\mu}\bar{x}_n + \gamma_n) \end{pmatrix}.$$

Since all the eigenvalues of $J$, which are the diagonal entries as $J$ is diagonal, are negative, $\Phi(D)$ turns out to be a locally stable point. □

### 5.2.4 Equilibria and Stability in Switching Domains

It is possible that solutions of (5.4) reach equilibria that lie in switching domains and such equilibria are called *singular equilibria*. In general, a singular equilibrium $\bar{x}$ of system (5.4) is a point that satisfies the condition $0 \in H(\bar{x})$ and that belongs to some threshold plane. Determining whether a singular equilibrium is stable or unstable requires a detailed analysis that is beyond the scope of this paper. However, in the following theorem we present a procedure to assess the stability of singular equilibria that can occur on $x_n$-hyperplane.

**Theorem 5.6.** *Assume that a sliding motion occurs on a switching domain $S$, which lies in the hyper-plane $C_n^j = \{x \in \mathbb{R}^n_{\geq 0} \ : \ x_n = \theta_n^j\}$, separating two regular domains $D_1$ and $D_2$. Let $\bar{x} = (\bar{x}_1, ..., \bar{x}_n)$ be the singular equilibrium point of the sliding motion. If $\bar{x} \in S$ and if the sliding motion is stable (resp. unstable), then $\bar{x}$ is locally stable (resp. unstable).*

*Proof.* Assuming the presence of a such stable sliding motion in $S$ and $\bar{x} \in S$, this implies that there exists a neighbourhood of $\bar{x}$ where the $n$-component of trajectories are approaching $\bar{x}_n = \theta_n^j$. Notably, the velocity of motion of the other $i$-components $(i = 1, ..., n-1)$ is given by

$$\dot{x}_i = \alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\bar{\mu}\theta_n^j + \gamma_i) x_i \tag{5.13}$$

with $\alpha$ equal to the second value in (5.11). Hence, the stability of $\bar{x}$ follows by the fact that (5.13) is of the PL form. If the $\bar{x} \in S$, but the sliding motion is unstable, the instability of $\bar{x}$ follows from the instability of the sliding motion. □

## 5.3 Introduction to the control problem

In this paper we focus on the gene expression machinery of the bacterium *Escherichia coli*, with the aim of controlling the growth rate of the cells. *E. coli* is a model organism that is easy to manipulate and much knowledge is available about its regulatory networks.

In the presence of a carbon source—such as the preferred glucose, or alternatively maltose or other sugars—*E. coli* grows in an exponential manner until it exhausts the nutrient sources, and then enters a stationary phase with practically zero growth [129]. The wild-type (namely the genetically unmodified) bacteria grow at different rates in the presence of carbon sources of different types [124]. Notably, glucose is the preferred

substrate because it leads to a higher growth rate in wild type. In the presence of two carbon sources (one is glucose), two successive growth phases are observed: a faster growth rate until glucose is exhausted and then a slower growth rate corresponding to consumption of the alternative sugar. This phenomenon is called *diauxie* [133].

Our control objective is to force the bacterium to significantly modify its response to glucose and an alternative carbon source. Using the PWNL formalism introduced in Section 5.2 and taking into account the recent applications of synthetic biology, we propose a bi-dimensional model of a mutant *E. coli*. The two basic variables, which describe the gene expression machinery that is responsible for bacterial growth are (see Figure5.1):

1. the concentration of RNA polymerase (RNAP), a special enzyme which enables the transcription of any gene (in other words if there is no RNAP, the bacteria can not produce proteins and thus can not grow);
2. the concentration of CRP, a protein involved in the formation of the complex cAMP-CRP whose level positively correlates with less preferred carbon sources and slower growth [24].

For this model we will study an open loop (see Figure5.1) configuration where the number and location of equilibria can be controlled by means of an input control function, which directly acts on the synthetic promoter of RNAP, tuning its gene expression. This choice can be justified by the experiment in [160] where the synthetic promoter of RNAP is controlled by IPTG as in Figure5.1.

Then, starting from the open-loop model, we extend it by adding a feedback control of the system's variables. The objective of the closed loop system is to explore, in silico, the possibility of constructing a mutant *E. coli* by re-wiring the gene expression machinery. In particular, the model suggests a re-wired network (see Figure5.6) that has the property of making the bacteria grow faster on an alternative sugar than on glucose (with appropriate set of parameters) leading de facto to *inverse diauxie* as desired.

## 5.4 Open-loop model

The open-loop model, expressed by (5.14), describes the qualitative dynamics of RNAP and CRP concentrations when the *rpoBC* gene—encoding for RNAP—is controlled externally by an IPTG-inducible promoter ($\nu_1 = h(IPTG)$). This RNAP control induces bi-stability in RNAP expression for some parameter sets, as experimentally verified

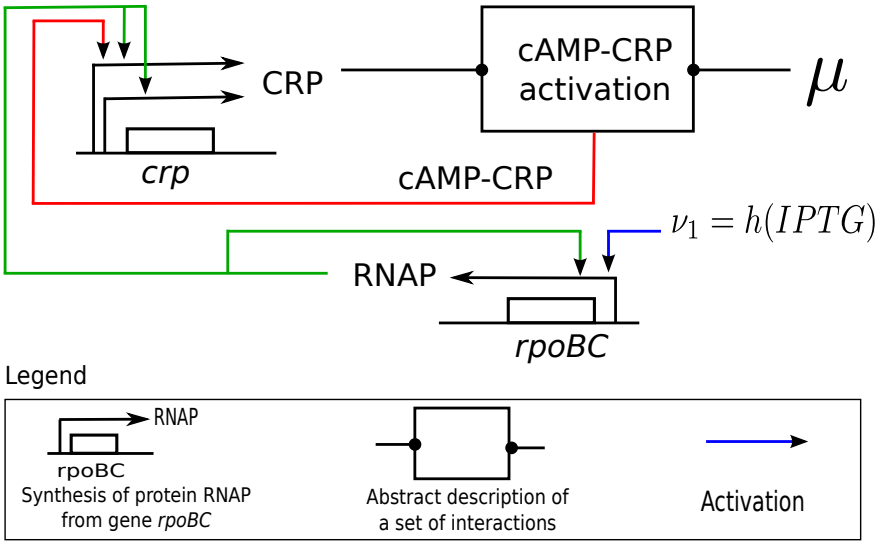



FIGURE 5.1: Regulatory network of the open-loop model in the *mutant E. coli*. The system consists of genes *crp* and synthetic-*rpoBC*. The synthetic-*rpoBC* promoter is positively regulated by IPTG and RNA-Polymerase (RNAP), which initiates the transcription of all genes (crp gene too). Moreover, *crp* transcription is induced by cAMP-CRP, a metabolite whose formation relies on CRP protein abundance and low level of bacterial growth rate $\mu$.

in [160]. We shall take into account this bi-stability to control the system to the "low" or to the "high" RNAP stable steady state. Let $x_c$, $x_p \in \mathbb{R}_{\geq 0}$ be the CRP and RNAP concentrations respectively. Thus, the open-loop model graphically depicted in Figure 5.1, can be translated into a PWNL system as (5.3) considering $n = 2, x_1 = x_c$ and $x_2 = x_p$:

$$\begin{cases} \dot{x}_c(t) = k_c^0\ s^+(x_p, \theta_p^1) + k_c^1\ s^+(x_p, \theta_p^2)\ s^+(x_c, \theta_c^1)\ s^-(x_p, \theta_{\bar{\mu}}) \\ \qquad\qquad - (\bar{\mu}\ x_p(t) + \gamma_c)\ x_c(t) \\ \dot{x}_p(t) = \nu_1\ k_p^0\ s^+(x_p, \theta_p^1) + \nu_1\ k_p^1\ s^+(x_p, \theta_p^2) \\ \qquad\qquad - (\bar{\mu}\ x_p(t) + \gamma_p)\ x_p(t) \end{cases} \tag{5.14}$$

where:

- $k_i^0 > 0$ $(i = c, p)$ are the basal synthesis rate constants;
- $k_i^1 > 0$ $(i = c, p)$ are the main synthesis rate constants;
- $\nu_1$ is a positive input accounting for IPTG induction of the synthetic *rpoBC* promoter;
- $\gamma_i > 0$ $(i = c, p)$ is the degradation rate constant;
- $\theta_i^j > 0$ $(i = c, p;\ j = 1, 2)$ is the $x_i$ threshold concentration for activation/inhibition;

- $\theta_{\bar{\mu}} > 0$ is a growth threshold depending on which substrate is used;
- $\bar{\mu} > 0$ is a growth constant depending on which substrate is used.

In what follows, we will explain the main assumptions adopted in building the system (5.14).

### 5.4.1 Growth rate

In bacteria, growth rate is highly correlated with RNAP [159], which plays a key role in gene transcription [77], and with the type of substrate [124]. Hence, according to (7.9), we modelled growth rate $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as:

$$\mu(t) = \bar{\mu}\ x_p(t). \tag{5.15}$$

### 5.4.2 cAMP-CRP activation

The cAMP-CRP activation box (Figure 5.1) summarizes the formation of the cAMP-CRP complex from cAMP, a small metabolite, which binds the protein CRP. The cAMP concentration is higher at low growth rate and rapidly decreases at high growth rate [24]. Thus, cAMP abundance in cells can be well captured by a negative step function of $\mu$, i.e. $s^-(\mu, \theta_\mu)$. Moreover, since cAMP association with or dissociation from CRP is much faster than the synthesis and degradation of proteins [144], we have assumed that as soon as CRP reaches a certain threshold, i.e. $\theta_c$, CRP instantly binds to cAMP in a switch-like fashion . For these reasons, the positive regulation carried out by cAMP-CRP reads as:

$$b^+_{cAMP-CRP} = s^+(x_c, \theta_c)\ s^-(\mu, \theta_\mu).$$

Focusing on the decreasing step function $s^-(\mu, \theta_\mu)$ and taking into account the expression of $\mu$ in (5.15), we can rewrite $b^+_{cAMP-CRP}$ as:

$$b^+_{cAMP-CRP}(x_c, x_p) = s^+(x_c, \theta_c)\ s^-(x_p, \theta_{\bar{\mu}}) \tag{5.16}$$

where $\theta_{\bar{\mu}}$ is a RNAP threshold concentration, which depends on the type of carbon source.

### 5.4.3 CRP synthesis

With *CRP synthesis* we indicate the production of CRP, namely the positive term of $\dot{x}_c$ in (5.14). CRP—as every other protein—needs, to be expressed, that RNAP binds to

the relative promoter in order to initiate transcription [77]. Thus, we have assumed that a lower value of $x_p$, i.e. $\theta_p^1$, induces the basal synthesis ($k_c^0\ s^+(x_p, \theta_p^1)$) of $x_c$ while a higher value of $x_p$, i.e. $\theta_p^2$, is needed to stimulate its main expression ($k_c^1 s^+(x_p, \theta_p^2)$). Moreover, the *crp* gene is regulated both positively and negatively by cAMP-CRP. However, in order to simplify, we omit the negative control of *crp*, because this mechanism only plays a role when the CRP concentration is low [144][1]. Thus, only one concentration threshold of CRP, i.e. $\theta_c^1$, is required in the model, to allow production of the cAMP-CRP complex. In conclusion, taking into account the regulation function of cAMP-CRP in (5.16), the CRP synthesis reads:

$$f_c(x) = k_c^0\ s^+(x_p, \theta_p^1) + k_c^1\ s^+(x_p, \theta_p^2)\ b^+_{cAMP-CRP}(x_c, x_p),$$

with

$$0 < \theta_c^1 < max_c, \tag{5.17}$$

where $max_c$ is the maximum concentration value for CRP.

### 5.4.4 RNAP synthesis

Similarly, RNAP synthesis denotes RNAP yield, expressed by the positive term of $\dot{x}_p$ in (5.14). We have assumed that a low concentration of RNAP ($\theta_p^1$) is sufficient to stimulate its basal production $k_p^0\ s^+(x_p, \theta_p^1)$ while only above the $\theta_p^2$ threshold RNAP is able to stimulate its main production $k_p^1\ s^+(x_p, \theta_p^2)$. Thus, we can order RNAP's thresholds as:

$$0 < \theta_p^1 < \theta_p^2 < max_p, \tag{5.18}$$

where $max_p$ is the maximum concentration value for RNAP. Moreover, the synthetic *rpoBC* promoter can be induced by IPTG whose effect is modelled by input $v_1$. Since we do not known whether the effect of IPTG is linear, we will later on assume, more generally, that $v_1$ is a positive increasing function of IPTG, to be denoted by $v_1 = h(IPTG)$. Thus, RNAP synthesis reads:

$$f_p(x) = \nu_1\ k_p^0\ s^+(x_p, \theta_p^1) + \nu_1\ k_p^1\ s^+(x_p, \theta_p^2).$$

### 5.4.5 CRP and RNAP removal

The negative terms of $\dot{x}_c$ and $\dot{x}_p$ in (5.14) take into account proteins' dilution and degradation as discussed in Section 5.2 and generally expressed in (5.3).

[1] We found that a model involving the negative control of *crp* by cAMP-CRP does not have any effect on the conclusion of this study.

## 5.5 Qualitative analysis of the open-loop system

In this section we will qualitatively study, by means of phase-plane analysis and bifurcation diagrams, system (5.14) in two cases: 1) cells grown in glucose and 2) cells grown in a medium with a different sugar, which we will assume is maltose, for simplicity. This will elucidate how different growth rates, induced by different substrates, influence gene expression and vice versa. Moreover, we shall show how—through the external input $\nu_1$—the stability of equilibria in (5.14) can be controlled.

### 5.5.1 Open-loop system in glucose growth

If cells are grown in glucose, then parameters depending on the substrate become $\theta_{\bar{\mu}} = \theta_p^G$ and $\bar{\mu} = \mu_G$ in system (5.14). Moreover, in the presence of glucose or other PTS sugars, adenylate cyclase[2] activity decreases, leading to a drop in the cellular level of cAMP [116] [134]. Thus, we have modelled this effect assuming:

$$0 < \theta_p^1 < \theta_p^G < \theta_p^2 < max_p. \tag{5.19}$$

Therefore, during growth on glucose, the state space of system (5.14) can be partitioned into eight *regular domains*, where the vector filed is uniquely defined:

$$\begin{aligned}
D_1^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ 0 \leq x_p < \theta_p^1\right\} \\
D_2^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ 0 \leq x_p < \theta_p^1\right\} \\
D_3^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^1 < x_p < \theta_p^G\right\} \\
D_4^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^1 < x_p < \theta_p^G\right\} \\
D_5^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^G < x_p < \theta_p^2\right\} \\
D_6^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^G < x_p < \theta_p^2\right\} \\
D_7^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_c < \theta_c^1,\ \theta_p^2 < x_p \leq max_p\right\} \\
D_8^G &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_c^1 < x_c \leq max_c,\ \theta_p^2 < x_p \leq max_p\right\}.
\end{aligned}$$

The *switching domains* are not listed here, but they are as defined in Section 5.2. We will group into regions $R_j^G$ those domains where system (5.14)—in glucose growth— has the same dynamics and thus the same focal points. Considering Definition 5.1, we have the following focal points:

[2]Enzyme that catalyses the conversion of ATP to cAMP and pyrophosphate.

- $\forall x \in R_1^G = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_1^G \cup D_2^G\right\}$

$$x_c \to 0$$
$$x_p \to 0$$

Thus, $\Phi_0^G = (0,0)$ is the focal point of region $R_1^G$.

- $\forall x \in R_2^G = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_3^G \cup D_4^G \cup D_5^G \cup D_6^G\right\}$

$$x_c \to \frac{k_c^0}{\mu_G\ \bar{x}^1_{p,G} + \gamma_c} = \bar{x}^2_{c,G}$$
$$x_p \to \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ v_1\ k_p^0\ \mu_G}}{2\mu_G} = \bar{x}^1_{p,G}$$

Thus, $\Phi_1^G = (\bar{x}^2_{c,G}, \bar{x}^1_{p,G})$ is the focal point of region $R_2^G$.

- $\forall x \in R_3^G = \left\{x \in \mathbb{R}^2_{\geq 0} : x \in D_7^G \cup D_8^G\right\}$

$$x_c \to \frac{k_c^0}{\mu_G\ \bar{x}^2_{p,G} + \gamma_c} = \bar{x}^1_{c,G}$$
$$x_p \to \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ v_1(k_p^0 +\ k_p^1)\mu_G}}{2\mu_G} = \bar{x}^2_{p,G}$$

Thus, $\Phi_2^G = (\bar{x}^1_{c,G}, \bar{x}^2_{p,G})$ is the focal point of region $R_3^G$.

Hence, there can be at most three locally stable steady states during growth on glucose.

Figure 5.2 depicts the phase-plane of system (5.14). It can be seen that $\Phi_0^G$, $\Phi_1^G$, $\Phi_2^G$, (for the parameter values used) are locally stable steady states since they are within their respective regular domains (Theorem 5.5). Notably, it is easy to verify that $\Phi_0^G$ is locally stable for any set of parameters. It represents absence of growth and can happen when RNAP initial condition, i.e. $x_p(t_0)$, is too low—specifically $x_p(t_0) < \theta_p^1$—to initiate gene transcription or when the control input $\nu_1$ does not sufficiently induce RNAP expression, that is when $\bar{x}^1_{p,G} < \theta_p^1$. We refer to $\Phi_0^G$ as the *trivial* fixed point. $\Phi_1^G$ represents RNAP basal level—leading to a low growth rate (see (5.15))— while CRP is at a high level, which is in agreement with high *crp* gene expression (by cAMP-CRP) at lower growth rate. Thus, because of the low growth rate achieved, we refer to $\Phi_1^G$ as the *low* fixed point. Conversely, at $\Phi_2^G$, CRP is at low level while RNAP, as well as $\mu$, have reached their highest stable values. Thus, $\Phi_2^G$ is named the *high* fixed point.

Since $\bar{x}^1_{p,G}(\nu_1)$ and $\bar{x}^2_{p,G}(\nu_1)$, it turns out that the location of focal points $\Phi_1^G$ and $\Phi_2^G$, and thus the number of equilibria of system (5.14), depend on the control input $\nu_1$.

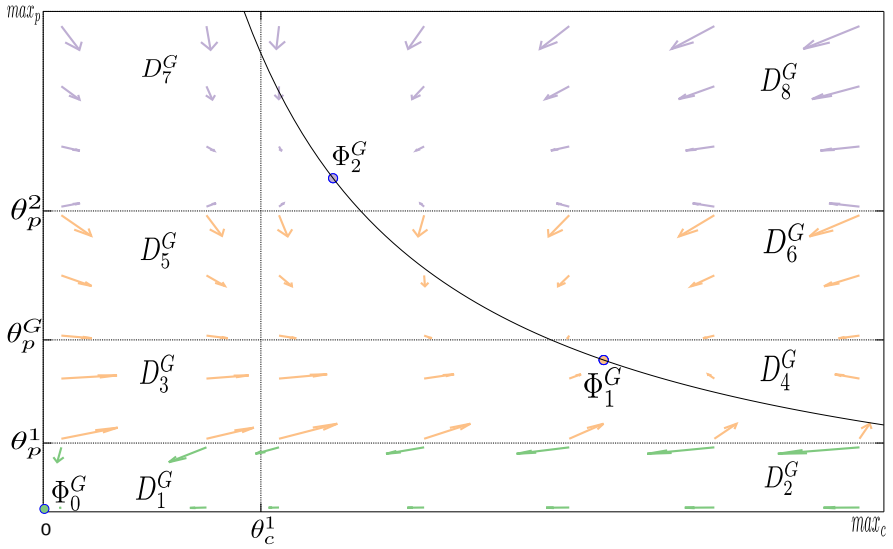



FIGURE 5.2: Phase plane of system (5.14) during growth in glucose. Parameter values used: $\theta_c^1 = 0.6$, $\theta_p^1 = 0.8$, $\theta_p^G = 2$, $\theta_p^2 = 3.5$, $k_c^0 = 7$, $k_c^1 = 10$, $k_p^0 = 8$, $k_p^1 = 26$, $\gamma_c = 1$, $\gamma_p = 1$, $\mu_G = 2$ e $v_1 = 1$. The black curve is the $x_c$-nullcline: $x_p = \dfrac{k_c^0}{x_c\ \mu_G} - \dfrac{\gamma_c}{\mu_G}$. Stable fixed points: $\Phi_0^G$, $\Phi_1^G$, $\Phi_2^G$.

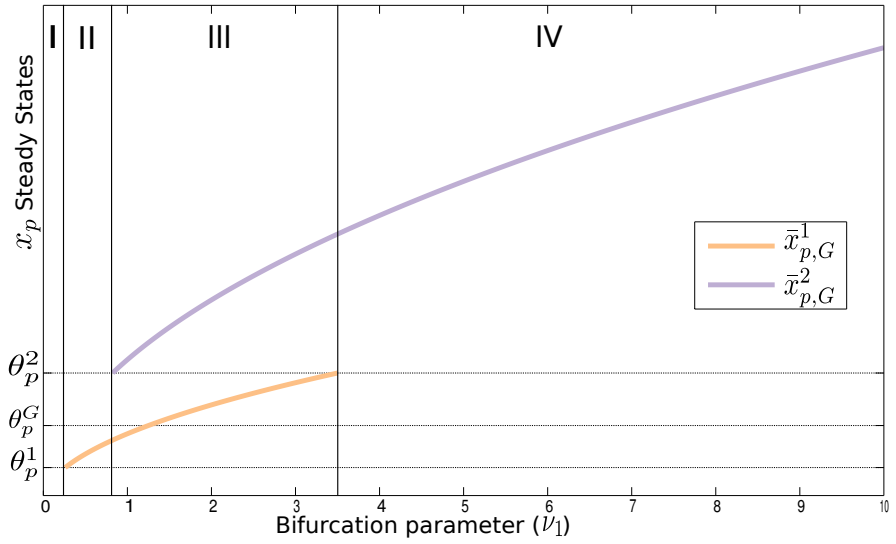



FIGURE 5.3: Bifurcation diagram for system (5.14) during growth in glucose, showing the non trivial locally stable steady states of $x_p$ as a function of the control input $v_1$. Other parameter values used are the same of those in Figure 5.2. See Proposition 5.7 for more details.

Hence, choosing appropriate values of $\nu_1$ it is possible to control system (5.14) towards $\Phi_1^G$ or $\Phi_2^G$. To illustrate this, we have depicted in Figure 5.3 the $x_p$-bifurcation diagram when parameter $v_1$ varies from 0 to 10 while the other parameter values are the same of those used in Figure 5.2.

We notice that Figure 5.3 is divided into four parts in which $x_p$ stability changes significantly. In part I, for those values of $\nu_1$ such that $\bar{x}^1_{p,G} < \theta^1_p$ and $\bar{x}^2_{p,G} < \theta^2_p$, neither $\Phi_1^G$ nor $\Phi_2^G$ are stable steady states. In this case, system (5.14) during growth on glucose converges towards the only stable point $\Phi_0^G$ (not depicted in Figure 5.3). So, in I, the control input is too small to allow RNAP to reach a basal level, and prevents bacterial growth.

In part II, when $\bar{x}^1_{p,G}(\nu_1) > \theta^1_p$ and $\bar{x}^2_{p,G}(\nu_1) < \theta^2_p$ hold, only $\Phi_1^G$ is a stable steady state (besides the trivial one) according to Theorem 5.5. Hence, it turns out that choosing an initial condition of RNAP $x_p(t_0) > \theta^1_p$ and $\nu_1$ such that $\bar{x}^1_{p,G}(\nu_1) > \theta^1_p$ and $\bar{x}^2_{p,G}(\nu_1) < \theta^2_p$, we can control system (5.14) to the stable point $\Phi_1^G$.

In part III, characterized by $\theta^1_p < \bar{x}^1_{p,G}(\nu_1) < \theta^2_p$ and $\bar{x}^2_{p,G}(\nu_1) > \theta^2_p$, both $\Phi_1^G$ and $\Phi_2^G$ are stable steady states: this is a region of bi-stability. Moreover, the phase plane corresponding to this configuration is depicted in Figure 5.2, where we can also observe the presence of two separatrices $x_p = \theta^1_p$ and $x_p = \theta^2_p$. Is is clear that, depending on $x_p(t_0)$, the system can converge to $\Phi_1^G$ (if $\theta^1_p < x_p(t_0) < \theta^2_p$) or to $\Phi_2^G$ (if $x_p(t_0) > \theta^2_p$).

In part IV, when $\bar{x}^1_{p,G}(\nu_1) > \theta^2_p$ holds, only $\Phi_2^G$ is a stable steady state and thus, whenever $x_p(t_0) > \theta^1_p$, system (5.14) converges to $\Phi_2^G$.

The open-loop control in glucose growth can be summarized as follows.

**Proposition 5.7.** *Consider system* (5.14) *with control input* $\nu_1$ *and initial condition* $x_p(t_0)$ *such that:*

- *if* $(\bar{x}^1_{p,G}(\nu_1) < \theta^1_p \ \wedge \ \bar{x}^2_{p,G}(\nu_1) < \theta^2_p) \ \vee \ x_p(t_0) < \theta^1_p$, *then system* (5.14) *converges to the* trivial focal point $\Phi_0^G$ *(region I in Figure 5.3);*
- *if* $\bar{x}^1_{p,G}(\nu_1) > \theta^1_p \ \wedge \ \bar{x}^2_{p,G}(\nu_1) < \theta^2_p \ \wedge \ x_p(t_0) > \theta^1_p$, *then system* (5.14) *converges to the* low focal point $\Phi_1^G$ *(region II in Figure 5.3);*
- *if* $\theta^1_p < \bar{x}^1_{p,G}(\nu_1) < \theta^2_p \ \wedge \ \bar{x}^2_{p,G}(\nu_1) > \theta^2_p \ \wedge \ x_p(t_0) > \theta^1_p$, *then system* (5.14) *is* bistable *(region III in Figure 5.3) and notably:*
  - *if* $\theta^1_p < x_p(t_0) < \theta^2_p$, *then system* (5.14) *converges to the* low focal point $\Phi_1^G$*;*
  - *if* $x_p(t_0) > \theta^2_p$, *then system* (5.14) *converges to the* high focal point $\Phi_2^G$

- *if* $\bar{x}^1_{p,G}(\nu_1) > \theta^2_p \ \wedge \ x_p(t_0) > \theta^1_p$, *then system* (5.14) *converges to the* high focal point $\Phi^G_2$ *(region IV in Figure 5.3).*

### 5.5.2 Open-loop system under an alternative carbon source

We consider now system (5.14) in the case that *E. coli* is grown in a medium with an alternative non PTS sugar, such as maltose. It follows that $\theta_{\bar{\mu}} = \theta^M_p$ and $\bar{\mu} = \mu_M$ in system (5.14). To take into account the fact that—in the absence of PTS sugar (as glucose)—cAMP formation is not strongly inhibited [134], we have assumed:

$$0 < \theta^1_p < \theta^2_p < \theta^M_p < max_p \tag{5.20}$$

(compare to (5.19)).

Figure 5.4 depicts the phase-plane of system (5.14) during maltose growth with the location of regular domains $D^M_j$. It turns out that there are new four (distinct focal points) possible steady states. The focal points' coordinates and their expressions are presented below:

$$\Phi^M_0 = (0,0) \qquad \Phi^M_1 = (\bar{x}^2_{c,M}, \bar{x}^1_{p,M})$$
$$\Phi^M_2 = (\bar{x}^1_{c,M}, \bar{x}^2_{p,M}) \qquad \Phi^M_3 = (\bar{x}^3_{c,M}, \bar{x}^2_{p,M})$$

where:

$$\bar{x}^1_{c,M} = \frac{k^0_c}{\mu_M \ \bar{x}^2_{p,M} + \gamma_c}, \qquad \bar{x}^1_{p,M} = \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4 \ v_1 \ k^0_p \ \mu_M}}{2\mu_M},$$
$$\bar{x}^2_{c,M} = \frac{k^0_c}{\mu_M \ \bar{x}^1_{p,M} + \gamma_c}, \qquad \bar{x}^2_{p,M} = \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4 \ v_1 (k^0_p + \ k^1_p) \mu_M}}{2\mu_M},$$
$$\bar{x}^3_{c,M} = \frac{k^0_c + k^1_c}{\mu_M \ \bar{x}^2_{p,M} + \gamma_c}.$$

We have that $\Phi^M_0$ is the focal point of domains $D^M_1$ and $D^M_2$; $\Phi^M_1$ is the focal point of domains $D^M_3$ and $D^M_4$; $\Phi^M_2$ is the focal point of domains $D^M_5$, $D^M_7$ and $D^M_8$; and finally, $\Phi^M_3$ is the focal point of domain $D^M_6$. Notably, Figure 5.4 shows the case where $\Phi^M_0$, $\Phi^M_1$, $\Phi^M_3$ are stable steady states since they are within their respective regular domains (Theorem 5.5) while $\Phi^M_2$ is not a stable point (but only a focal point) because it does not belong to any of its generating domains.

The maltose and glucose induced bacterial growth differ in the existence of one more focal point, i.e. $\Phi^M_3$, in the maltose case. This arises from the fact that maltose has

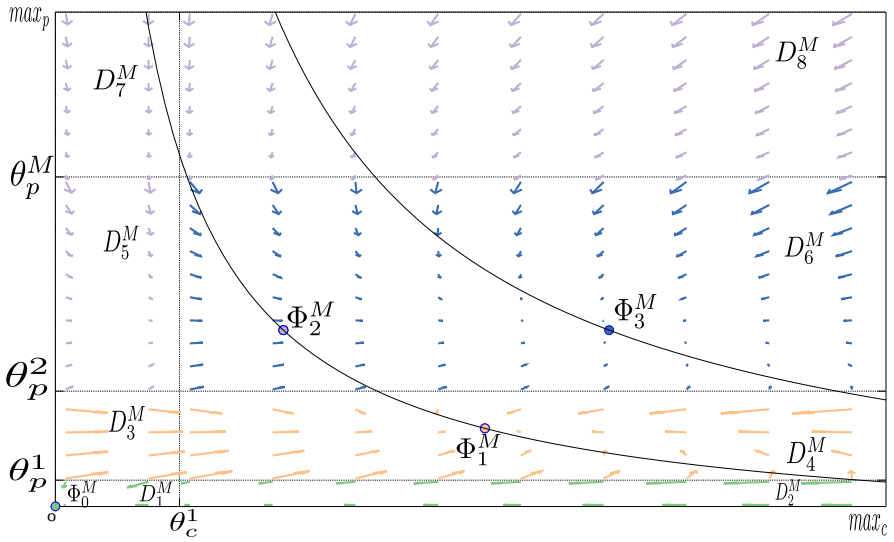



FIGURE 5.4: Phase plane of system (5.14) during growth in maltose. Parameter values used: $\theta_c^1 = 0.6$, $\theta_c^2 = 2.7$, $\theta_p^1 = 0.8$, $\theta_p^2 = 3.5$, $\theta_p^M = 10$, $k_c^0 = 5$, $k_c^1 = 10$, $k_p^0 = 8$, $k_p^1 = 26$, $\gamma_c = 1$, $\gamma_p = 1$, $\mu_M = 1$ e $v_1 = 1$. The black curves are the $x_c$-nullclines: $x_p = \dfrac{k_c^0}{x_c\ \mu_G} - \dfrac{\gamma_c}{\mu_G}$ and $x_p = \dfrac{k_c^0 + k_c^1}{x_c\ \mu_G} - \dfrac{\gamma_c}{\mu_G}$. Stable fixed points: $\Phi_0^M$, $\Phi_1^M$, $\Phi_3^M$.

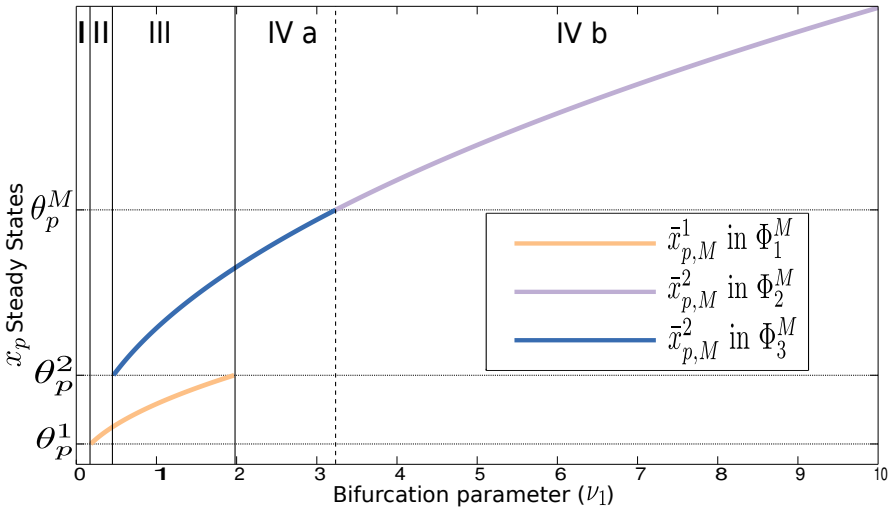



FIGURE 5.5: Bifurcation diagram for system (5.14) during growth in maltose, showing the non trivial locally stable steady states of $x_p$ as a function of the control input $v_1$. Other parameter values used are the same of those in Figure 5.4. See Proposition 5.8 for more details.

no direct effect on cAMP formation—in contrast to glucose which is observed to inhibit it [134]. Moreover, when $\Phi_3^M$ is a stable point (if $\Phi_3^M \in D_6^M$, Theorem 5.5) it ensures a higher level of CRP expression, which reflects the higher level of cAMP-CRP as well.

Again in the maltose growth case, it is possible to control system (5.14)—by means of the input $\nu_1$— to a "low" or a "high" stable point. Indeed, Figure 5.5 depicts how the stability of $x_p$ changes as a function of $\nu_1$. Notably—always without considering the trivial

fixed point $\Phi_0^M$—we observe that choosing $\nu_1$ such that $\bar{x}^1_{p,M}(\nu_1) > \theta_p^1$ and $\bar{x}^2_{p,M}(\nu_1) < \theta_p^2$, system (5.14) converges to the "low" stable point $\Phi_1^M$ for any $x_p(t_0) > \theta_p^1$.

The convergence to the "high" stable points, i.e. $\Phi_2^M$ or $\Phi_3^M$, is guaranteed for those $\nu_1$ such that $\bar{x}^1_{p,M}(\nu_1) > \theta_p^2$. In fact, in this case system (5.14) with $x_p(t_0) > \theta_p^1$ converges to $\Phi_3^M$, if $\theta_p^2 < \bar{x}^2_{p,M}(\nu_1) < \theta_p^M$, or to $\Phi_2^M$, if $\bar{x}^2_{p,M}(\nu_1) > \theta_p^M$, both of them having the "high" level of RNAP, i.e. $\bar{x}^2_{p,M}$.

When $\theta_p^1 < \bar{x}^1_{p,M}(\nu_1) < \theta_p^2$ and $\bar{x}^2_{p,M}(\nu_1) > \theta_p^2$ hold, we are in the bi-stability case depicted in Figure 5.4, where both $\bar{x}^1_{p,M}$ and $\bar{x}^2_{p,M}$ are stable steady states of $x_p$ differential equation. The convergence to one or to the other stable steady state depends on the initial condition with respect to the separatrices $\theta_p^1$ and $\theta_p^2$. Notably, if $\theta_p^1 < x_p(t_0) < \theta_p^2$, $x_p(t)$ converges to $\bar{x}^1_{p,M}$ while, if $x_p(t_0) > \theta_p^2$, $x_p(t)$ converges to $\bar{x}^2_{p,M}$.

The open-loop control in maltose growth is summarized below.

**Proposition 5.8.** *Consider system* (5.14) *(during growth on maltose) with control input* $\nu_1$ *and initial condition* $x_p(t_0)$ *such that:*

- *if* $(\bar{x}^1_{p,M}(\nu_1) < \theta_p^1 \ \wedge \ \bar{x}^2_{p,M}(\nu_1) < \theta_p^2) \ \vee \ x_p(t_0) < \theta_p^1$, *then system* (5.14) *converges to the* trivial *focal point* $\Phi_0^M$ *(region I in Figure 5.5);*
- *if* $\bar{x}^1_{p,M}(\nu_1) > \theta_p^1 \ \wedge \ \bar{x}^2_{p,M}(\nu_1) < \theta_p^2 \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.14) *converges to the* low *focal point* $\Phi_1^M$ *(region II in Figure 5.5);*
- *if* $\theta_p^1 < \bar{x}^1_{p,M}(\nu_1) < \theta_p^2 \wedge \bar{x}^2_{p,M}(\nu_1) > \theta_p^2$, *then system* (5.14) *is* non-trivially bistable *(region III in Figure 5.5), and notably:*
  - *if* $\theta_p^1 < x_p(t_0) < \theta_p^2$, *then system* (5.14) *converges to the* low *focal point* $\Phi_1^M$ *;*
  - *if* $x_p(t_0) > \theta_p^2$, *then system* (5.14) *converges to the* high *focal point* $\Phi_3^M$*;*
- *if* $\bar{x}^1_{p,M}(\nu_1) > \theta_p^2 \ \wedge \ \bar{x}^2_{p,M}(\nu_1) < \theta_p^M \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.14) *converges to the* high *focal point* $\Phi_3^M$ *(region IVa in Figure 5.5).*
- *if* $\bar{x}^1_{p,M}(\nu_1) > \theta_p^2 \ \wedge \ \bar{x}^2_{p,M}(\nu_1) > \theta_p^M \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.14) *converges to the* high *focal point* $\Phi_2^M$ *(region IVb in Figure 5.5).*

To conclude, we remark that the most relevant difference in comparison with growth on glucose, is the higher level of CRP, and thus the higher level of cAMP-CRP, which can be achieved in bacteria grown in maltose. We shall see how this information shall be useful in constructing a feedback control law later on.

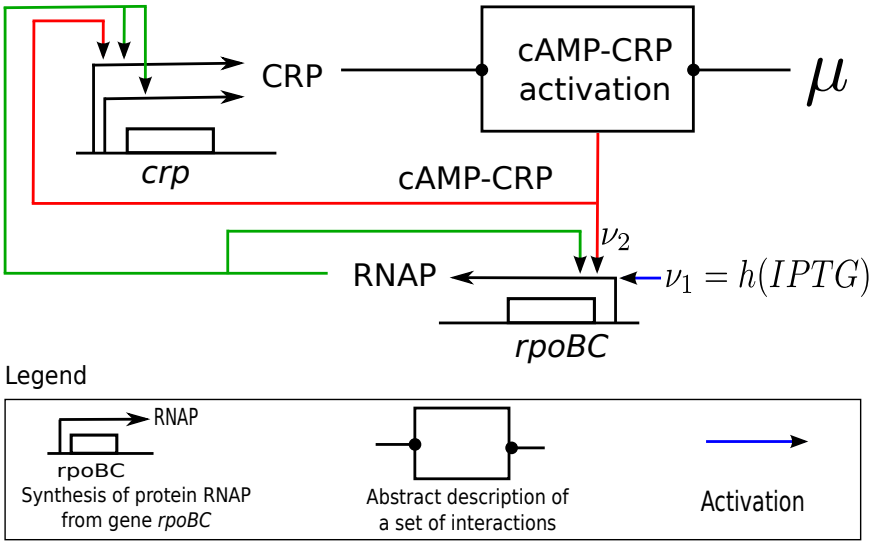



FIGURE 5.6: Closed loop-model of gene expression machinery in mutant *E. coli*. The the term $\nu_2$ indicates the feedback control law implemented by re-wiring the open loop model (Figure 5.1), that is putting RNAP transcription under positive regulation of cAMP-CRP.

## 5.6 Closed-loop model

As motivated in Section 5.3, the aim of the closed loop model is to suggest a way for building a synthetic strain of *E. coli* able to generate *inverse diauxie*. In the wild type case, glucose is the "preferred" substrate because it supports the highest growth rate [133] while in a modified bacterium, using the closed-loop model, we aim at inverting this mechanism to get higher growth on maltose than on glucose.

In [160], the RNAP promoter was modified in such a way that it can be controlled externally through IPTG. We will assume that further modifications can be introduced such that RNAP will eventually be controlled by some other variable. A new control system involving two inputs is thus considered: assuming a new synthetic *rpoBC* promoter is used, in which both $k_p^0$ and $k_p^1$ are controlled by $\nu_1$ while only $k_p^1$ is controlled by $\nu_2$, we end up with the following PWNL system:

$$\begin{cases} \dot{x}_c(t) = k_c^0\ s^+(x_p, \theta_p^1) + k_c^1\ s^+(x_p, \theta_p^2)\ s^+(x_c, \theta_c^1)\ s^-(x_p, \theta_{\bar{\mu}}) \\ \qquad\qquad - (\bar{\mu}\ x_p(t) + \gamma_c)\ x_c(t) \\ \dot{x}_p(t) = \nu_1\ k_p^0\ s^+(x_p, \theta_p^1) + \nu_1\ \nu_2\ k_p^1\ s^+(x_p, \theta_p^2) \\ \qquad\qquad - (\bar{\mu}\ x_p(t) + \gamma_p)\ x_p(t). \end{cases} \tag{5.21}$$

During *exponential phase*—the period characterized by cell doubling—the bacterial culture shows a constant growth rate [129]. This means that, according to (5.15), a stable

fixed point of RNAP has to be reached. Thus, constructing a closed-loop control to generate *inverse diauxie*, mathematically means finding a function of the system's variables $\nu_2 = g(x_c, x_p)$ that regulates the transcription of the gene *rpoBC* in such a way that:

$$\mu_M \ \bar{x}_{p,M} > \mu_G \ \bar{x}_{p,G} \tag{5.22}$$

where $\bar{x}_{p,M}$ and $\bar{x}_{p,G}$ are two possible RNAP steady states in maltose and glucose media, respectively.

In the following Theorem we state sufficient conditions and a feedback function for model (5.21) to exhibit a higher growth rate on maltose then on glucose.

**Theorem 5.9.** *Assume*

$$g(x_c, x_p) = s^+(x_c, \theta_c^*) \ s^-(x_p, \theta_{\bar{\mu}}), \tag{5.23}$$

$$\theta_c^1 < \theta_c^* < \frac{k_c^0 + k_c^1}{\mu_M \ \theta_p^M + \gamma_c}, \tag{5.24}$$

$$\frac{\mu_G}{\mu_M} < 1 + \frac{k_p^1}{k_p^0}, \tag{5.25}$$

$$x_p(t_0) > \theta_p^1, \tag{5.26}$$

*then setting* $\nu_2 = g(x_c, x_p)$*, there exists an interval* $I_1 = (\nu_1^{min}, \nu_1^{max})$ *such that the inverse diauxie condition in* (5.22) *is satisfied by* (5.21)*, for all* $\nu_1 \in I_1$*.*

Briefly, function (5.23) asserts that the regulation of RNAP transcription is carried out by cAMP-CRP, which positively regulates *rpoBC* main synthesis rate. The mathematical formulation of such feedback control function arises from the modelling and analysis of the open-loop system (Section 5.4, 5.5). Condition (5.24) gives a range for the cAMP-CRP affinity, i.e. $\theta_c^*$, of the synthetic *rpoBC* promoter. Condition (5.25) indicates how the basal ($k_p^0$) and the main ($k_p^1$) synthesis rates of RNAP have to be related with $\mu_G$ and $\mu_M$. Finally, condition (5.26) establishes which is the minimum RNAP initial value that guarantees the inverse diauxie.

The proof of Theorem 5.9 will be given in the remainder of the article, by analysing system (5.21) in glucose and maltose medium.

## 5.7 Qualitative analysis of the closed-loop system

Now, we will study—both in glucose and in maltose cultures—the closed-loop system that results from (5.21) when input $v_2$ is given by the feedback law (5.23). Moreover, we

will assume that conditions (5.24), (5.25) hold in the following analysis. Thus, when we refer to system (5.21), we implicitly assume that all the conditions stated in Theorem 5.9 hold, to ensure the *inverse diauxie*.

### 5.7.1 Closed-loop system in glucose growth

When *E. coli* is grown in glucose we have $\theta_{\bar{\mu}} = \theta_p^G$ and $\bar{\mu} = \mu_G$ in system (5.21). Since glucose inhibits cAMP formation as previously motivated and expressed in (5.19), it turns out that the product $s^+(x_p, \theta_p^2)s^-(x_p, \theta_G)$, in the second equation in (5.21), is always zero. Thus, system (5.21) during glucose growth can be simplified to:

$$\begin{cases} \dot{x}_c\,(t) = k_c^0\ s^+(x_p, \theta_p^1) - (\mu_G\ x_p(t) + \gamma_c)\ x_c(t) \\ \qquad 0 < \theta_c^1 < max_c \\ \dot{x}_p\,(t) = v_1\ k_p^0\ s^+(x_p, \theta_p^1) - (\mu_G\ x_p(t) + \gamma_p)\ x_p(t) \\ \qquad 0 < \theta_p^1 < \theta_p^G < \theta_p^2 < max_p \end{cases} \tag{5.27}$$

System (5.27) clearly shows that the value of $\theta_c^*$ does not play a role during growth on glucose. The focal points of system (5.27) are:

$$\Psi_0^G = (0, 0), \quad \Psi_1^G = (\bar{x}_{c,G}^1, \bar{x}_{p,G}^1)$$

where:

$$\bar{x}_{c,G}^1 = \frac{k_c^0}{\mu_G\ \bar{x}_p^1 + \gamma_c}; \quad \bar{x}_{p,G}^1 = \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ v_1\ k_p^0\ \mu_G}}{2\mu_G}.$$

Figure 5.7 depicts the phase-plane of system (5.27) in the case when both $\Psi_0^G$ and $\Psi_1^G$ are stable fixed points (Theorem 5.5). We notice that, the main difference with the open-loop model is that system (5.27) no longer exhibits the high focal point ($\Phi_2^G$) with $\bar{x}_{p,G}^2$ as RNAP equilibrium (see Figure 5.2 for comparison). This is due to cAMP-CRP which through the feedback law $g(x_c, x_p)$, inhibits the *rpoBC* promoter and shuts off RNAP main expression. Biologically, this means that the growth rate of *E. coli*—at steady state—is limited by the low level of RNAP expression at $\bar{x}_{p,G}^1$.

For reasons of space we omit the $x_p$ bifurcation diagram as function of $\nu_1$ for system (5.27), but the main results on equilibria stability are stated in Proposition 5.10.

**Proposition 5.10.** *Consider system* (5.27) *with control input* $\nu_1$ *and initial condition* $x_p(t_0)$ *such that:*

- *if* $\bar{x}_{p,G}^1(\nu_1) < \theta_p^1 \ \vee \ x_p(t_0) < \theta_p^1$, *then system* (5.27) *converges to the* trivial *focal point* $\Psi_0^G$;

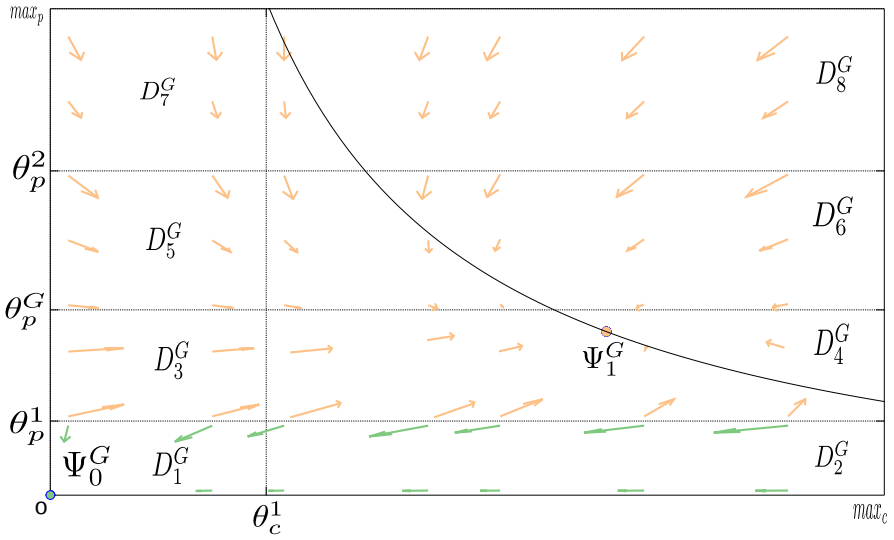



FIGURE 5.7: Phase plane of system (5.27) during growth in glucose. Parameter values used: $\theta_c^1 = 0.6$, $\theta_p^1 = 0.8$, $\theta_p^G = 2$, $\theta_p^2 = 3.5$, $k_c^0 = 5$, $k_c^1 = 10$, $k_p^0 = 8$, $k_p^1 = 26$, $\gamma_c = 1$, $\gamma_p = 1$, $\mu_G = 1$ and $v_1 = 1$. The black curve is the $x_c$-nullcline: $x_p = \dfrac{k_c^0}{x_c\ \mu_G} - \dfrac{\gamma_c}{\mu_G}$. Stable fixed points: $\Psi_0^G$, $\Psi_1^G$.

- *if* $\bar{x}^1_{p,G}(\nu_1) > \theta_p^1 \ \wedge\ x_p(t_0) > \theta_p^1$, *then system* (5.27) *converges to the* low *focal point* $\Psi_1^G$.

### 5.7.2 Closed-loop system in maltose growth

We consider now system (5.21) in the case when *E. coli* is grown in maltose, thus we have $\theta_{\bar{\mu}} = \theta_p^M$, $\bar{\mu} = \mu_M$ and (5.20) holds.

It follows that the focal points of system (5.21) during maltose growth are:

$$\Psi_0^M = (0,0) \qquad \Psi_1^M = (\bar{x}^1_{c,M}, \bar{x}^1_{p,M})$$
$$\Psi_2^M = (\bar{x}^2_{c,M}, \bar{x}^2_{p,M}) \qquad \Psi_3^M = (\bar{x}^3_{c,M}, \bar{x}^1_{p,M})$$

where:

$$\bar{x}^1_{c,M} = \frac{k_c^0}{\mu_M\ \bar{x}^1_{p,M} + \gamma_c}, \quad \bar{x}^1_{p,M} = \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ v_1\ k_p^0\ \mu_M}}{2\mu_M},$$
$$\bar{x}^2_{c,M} = \frac{k_c^0 + k_c^1}{\mu_M\ \bar{x}^2_{p,M} + \gamma_c}, \quad \bar{x}^2_{p,M} = \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4\ v_1(k_p^0 + k_p^1)\mu_M}}{2\mu_M},$$
$$\bar{x}^3_{c,M} = \frac{k_c^0 + k_c^1}{\mu_M\ \bar{x}^1_{p,M} + \gamma_c}.$$

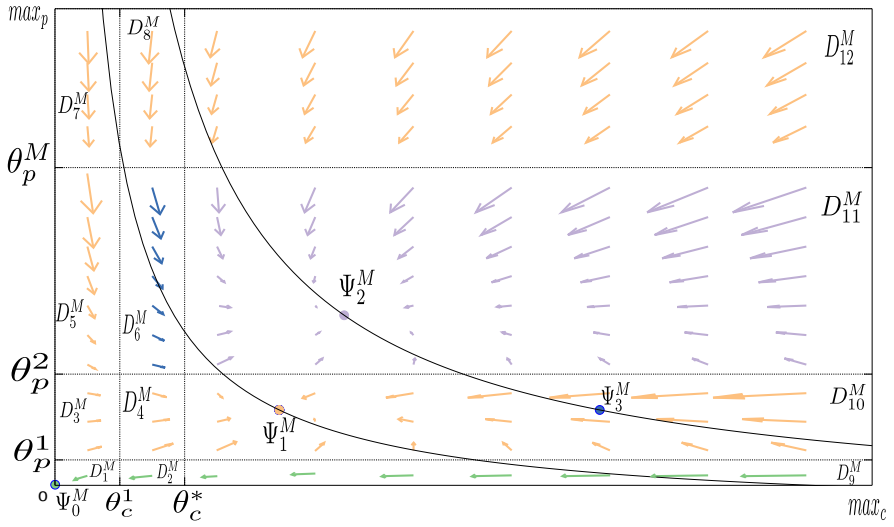



FIGURE 5.8: Phase plane of system (5.21) during growth in maltose. Parameter values used: $\theta_c^1 = 0.6$, $\theta_c^* = 1.2$, $\theta_p^1 = 0.8$, $\theta_p^2 = 3.5$, $\theta_p^M = 10$, $k_c^0 = 5$, $k_c^1 = 10$, $k_p^0 = 8$, $k_p^1 = 26$, $\gamma_c = 1$, $\gamma_p = 1$, $\mu_M = 1$ and $v_1 = 1$. The black curves are the $x_c$-nullclines: $x_p = \dfrac{k_c^0}{x_c\ \mu_M} - \dfrac{\gamma_c}{\mu_M}$ and $x_p = \dfrac{k_c^0 + k_c^1}{x_c\ \mu_M} - \dfrac{\gamma_c}{\mu_M}$. Stable fixed points: $\Psi_0^M$, $\Psi_1^M$ and $\Psi_2^M$.

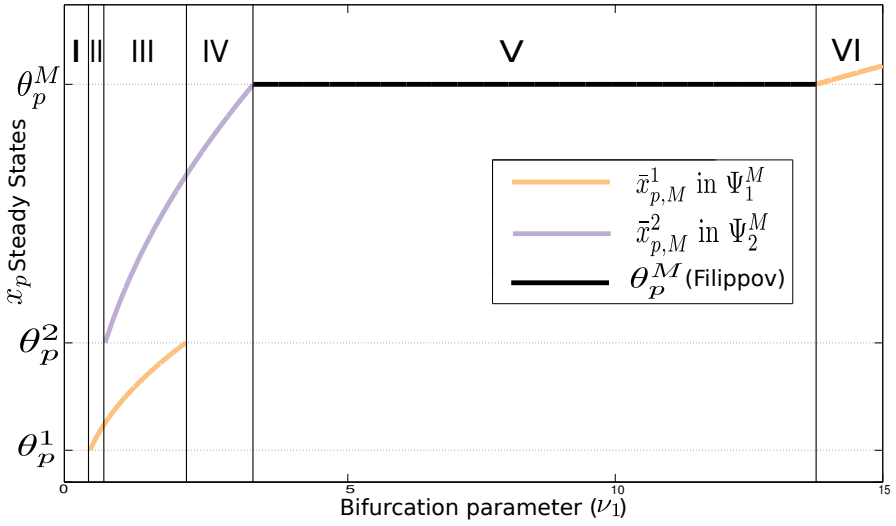



FIGURE 5.9: Bifurcation diagram for system (5.21) during growth in maltose, showing the non trivial locally stable steady states of $x_p$ as a function of the control input $v_1$. Other parameter values used are the same of those in Figure 5.8. See Proposition 5.11 for more details.

Figure 5.8 depicts the phase plane of system (5.21) when maltose is the substrate. Unlike in the glucose case, the closed loop model for maltose preserves the high focal point, allowing RNAP to be expressed at a high rate and hence to reach its maximal equilibrium in $\bar{x}_{p,M}^2$. Figure 5.9 shows the $x_p$ stable steady states as a function of the control input $v_1$. The main differences with the $x_p$-bifurcation diagram of the open-loop model in Figure 5.5 are the regions V and VI, which are due to the feedback control $\nu_2 = g(x_c, x_p)$. Notably, region V of Figure 5.9 shows that, according to Theorem 5.6,

a Filippov stable fixed point is generated on the switching domain $x_p = \theta_p^M$ when $\theta_p^2 < \bar{x}_{p,M}^1 < \theta_p^M$ and $\bar{x}_{p,M}^2 > \theta_p^M$. Besides, because of the inhibition of RNAP by cAMP-CRP (in VI) $\bar{x}_{p,M}^1$ becomes again a stable steady state when $\bar{x}_{p,M}^1(\nu_1) > \theta_p^M$.

Finally, Proposition 5.11 summarizes the stable equilibria of system (5.21) with maltose as medium.

**Proposition 5.11.** *Consider system* (5.21) *during growth in maltose with control input* $\nu_1$ *and initial condition* $x_p(t_0)$ *such that:*

- *if* $(\bar{x}_{p,M}^1(\nu_1) < \theta_p^1 \ \wedge \ \bar{x}_{p,M}^2(\nu_1) < \theta_p^2) \ \vee \ x_p(t_0) < \theta_p^1$, *then system* (5.21) *converges to the* trivial *focal point* $\Psi_0^M$ *(region I in Figure 5.9);*
- *if* $\bar{x}_{p,M}^1(\nu_1) > \theta_p^1 \ \wedge \ \bar{x}_{p,M}^2(\nu_1) < \theta_p^2 \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.21) *converges to the* low *focal point* $\Psi_1^M$ *(region II in Figure 5.9);*
- *if* $\theta_p^1 < \bar{x}_{p,M}^1(\nu_1) < \theta_p^2 \wedge \bar{x}_{p,M}^2(\nu_1) > \theta_p^2$, *then system* (5.21) *is* non-trivially bistable *(region III in Figure 5.9), and notably:*
  - *if* $\theta_p^1 < x_p(t_0) < \theta_p^2$, *then system* (5.21) *converges to the* low *focal point* $\Psi_1^M$ *;*
  - *if* $x_p(t_0) > \theta_p^2$, *then system* (5.21) *converges to the* high *focal point* $\Psi_2^M$*;*
- *if* $\bar{x}_{p,M}^1(\nu_1) > \theta_p^2 \ \wedge \ \bar{x}_{p,M}(\nu_1)^2 < \theta_p^M \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.21) *converges to the* high *focal point* $\Psi_2^M$ *(region IV in Figure 5.9).*
- *if* $\theta_p^2 < \bar{x}_{p,M}^1(\nu_1) < \theta_p^M \ \wedge \ \bar{x}_{p,M}^2(\nu_1) > \theta_p^M \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.21) *converges to a* Filippov *focal point, which arises on the* $x_p = \theta_p^M$ *switching domain (region V in Figure 5.9).*
- *if* $\bar{x}_{p,M}^1(\nu_1) > \theta_p^M \ \wedge \ x_p(t_0) > \theta_p^1$, *then system* (5.21) *converges to the focal point* $\Psi_1^M$ *(region VI in Figure 5.9).*

## 5.8 Inverse Diauxie

We conclude by showing how the feedback control law $\nu_2 = g(x_c, c_p)$ and the sufficient conditions in Theorem 5.9 lead to the *inverse diauxie*.

In a glucose medium, under conditions (5.23), (5.24), (5.25) and (5.26), only a low level of stable RNAP can be reached, i.e. $\bar{x}_{p,G}^1$, while in a maltose medium, also a stable high level of RNAP can be achieved, i.e. $\bar{x}_{p,M}^2$. These facts are fundamental for ensuring

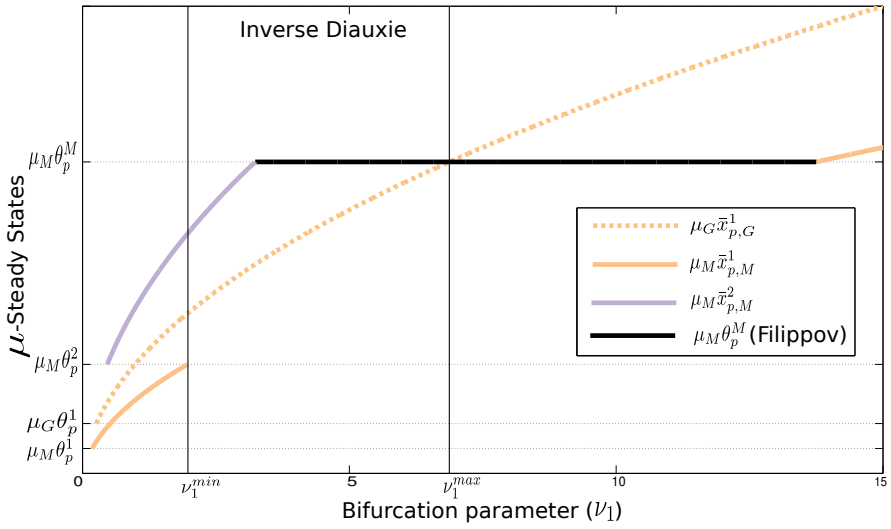



FIGURE 5.10: Bifurcation diagram of closed-loop bacterial growth rate, showing $\mu$ stable steady states as a function of $v_1$. System's parameters are the same of those in Figure 5.7 and Figure 5.8. $(\nu_1^{min}, \nu_1^{max})$ is the range in which the closed loop system (5.21) with (5.23) exhibits the *inverse diauxie* phenomenon.

the *inverse diauxie* phenomenon (5.22) and proving Theorem 5.9. In fact, considering $\bar{x}_{p,G} = \bar{x}_{p,G}^1$ (as in $\Psi_1^G$) and $\bar{x}_{p,M} = \bar{x}_{p,M}^2$ (as in $\Psi_2^M$) in (5.22), we end up with

$$\mu_M \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4v_1(k_p^0 + k_p^1)\mu_M}}{2\mu_M} > \mu_G \frac{-\gamma_p + \sqrt{\gamma_p^2 + 4v_1 k_p^0 \mu_G}}{2\mu_G}$$

which is satisfied if and only if condition (5.25) holds. Moreover, condition (5.24) is essential to guarantee that $\Psi_2^M$ stays within its domain $D_{11}^M$, and hence the stability of $\bar{x}_{p,M}^2$ (Theorem 5.5) for all $\nu_1$ such that $\theta_p^2 < \bar{x}_{p,M}^2 < \theta_p^M$. Finally, condition (5.26) is needed to avoid the convergence of system (5.21) with the feedback law (5.23) to the trivial steady state. Figure 5.10 graphically illustrates Theorem 5.5 and range $(\nu_1^{min}, \nu_1^{max})$ in which the stable growth rate on maltose is higher than that one on glucose. Notably, $\nu_1^{min}$ is the minimum $\nu_1$ such that $\bar{x}_{p,M}^1(\nu_1) > \theta_p^2$ while $\nu_1^{max}$ is the maximum $\nu_1$ such that $\bar{x}_{p,G}^1(\nu_1) < \theta_p^M$.

To conclude, Figure 5.11 depicts the growth rate (5.15) of the open-loop system (5.14) and the closed-loop system (5.21) in the case when cells are grown in glucose and in maltose. Notably, the growth curves in Figure 5.11 are calculated considering the same parameter values of Figures 5.2, 5.4, 5.7, 5.8, the same initial condition and the same control input $\nu_1$, which satisfies $\nu_1^{min} < \nu_1 < \nu_1^{max}$. Observing the final part of the growth curves, which corresponds to stable steady states, we clearly notice the *inverse diauxie* phenomenon (higher growth rate on maltose than that of glucose) achieved by the closed-loop model and not by the open-loop model.

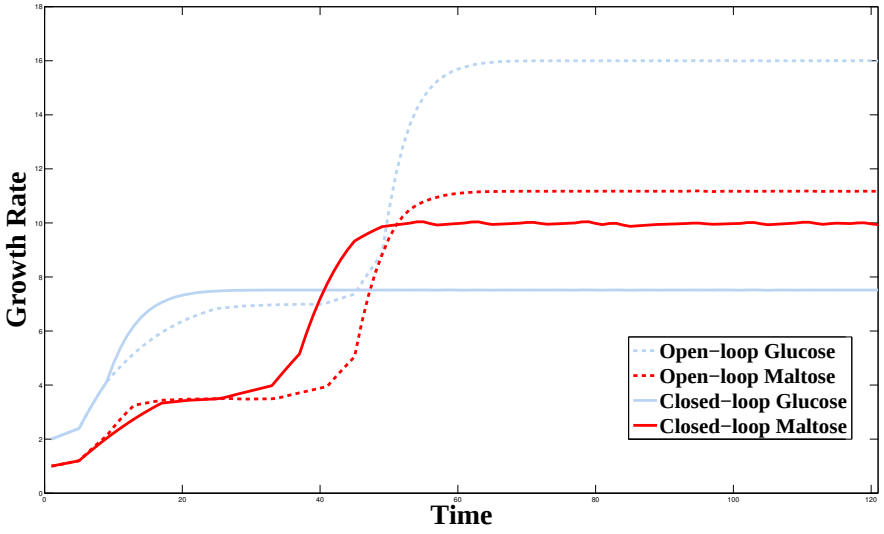



FIGURE 5.11: Growth rate as function of time (5.15) for the open-loop system (5.14) and the closed-loop system (5.21) when cells are grown in glucose and in maltose. $\nu_1 = 4$ satisfies $\nu_1^{min} < \nu_1 < \nu_1^{max}$ as in Theorem 5.9, $x_p(t_0) = 1$ satisfies condition (5.26). The other parameter values used are the same of those in Figures 5.2, 5.4, 5.7, 5.8.

## 5.9 Conclusions

A new qualitative formalism to model gene expression and growth rate in bacteria has been proposed. Within this new modelling framework, named *piecewise non linear* (PWNL) systems, we presented a bi-dimensional open loop model of *E. coli* gene expression in which the synthetic promoter encoding for RNAP is controlled externally by an input (IPTG). Since *E. coli* growth rate is related to the type of carbon source as well as the amount of RNAP, it was shown out that the bacterial exponential growth rate can be controlled—by means of appropriate values of the input (IPTG)—towards a low or a high level. Then, starting from the analysis of the open-loop system in the cases that cells are grown in glucose and in maltose, we developed a closed-loop system, obtained by rewiring the open-loop's gene interactions, which is able to mathematically show the *inverse diauxie*, namely reaching an higher growth rate on maltose then on glucose.

Our results combine the control based approaches of engineering with biology, with the aim of implementing new functions in the cells (synthetic biology), which may have some practical and relevant importance in applications such as production of biopharmaceuticals, novel bio-materials and developing new bio-technologies.

To conclude, the work presented here is an attempt to help guide the construction of synthetic gene networks, by allowing biologists to select experimental conditions most likely to yield successful results.

# Chapter 6

# Switched piecewise quadratic models of biological networks: application to control of bacterial growth

The content of this chapter has been submitted to the AUTOMATICA journal and it consists of an extended version of a previously paper presented at the Nolcos 2013 Conference [38].

Here, a novel qualitative formalism to model gene expression dynamics dependent on dilution due to cell growth rate is proposed and explored for qualitatively controlling the bacterial growth rate. Notably, we extend the piecewise linear (PL) systems by keeping the use of step functions to model the interactions between the elements and adding a growth rate expression to model the dilution effect. We model the growth rate as the weighted minimum of two limiting gene products responsible for bacterial growth. The resulting system is a switched system with two piecewise quadratic (PQ) modes. We study the stability of such switched piecewise quadratic (SPQ) system starting from the stability analysis of the (PQ) modes. Then, we propose and analyse a bidimensional SPQ open-loop model—describing the bacterial gene expression machinery—in which the growth rate of cells can be controlled by an external input which tunes the synthesis of one of the variables. The stability of the SPQ open-loop model is thus studied by means of bifurcation diagram with respect to the input and the phase plane of a relevant scenario—showing bi-mode bistability—is presented.

## 6.1 Introduction

The overall aim of synthetic biology is to develop and apply engineering tools to control cellular behaviour— constructing novel biological circuits in the cell—to achieve desired functions [131].

Most recent synthetic designs have focused on the cell transcription machinery, which includes the genes to be expressed, their promoters, RNA polymerase, Ribosomes and transcription factors, all serving as potential engineering components. Indeed, synthetic bio-molecular circuits are typically constructed in *Escherichia coli* (*E. coli*), by "cutting and pasting"together coding regions and promoters (natural and synthetic) according to designed structures and specific purposes ([67],[76], [165]).

Along these lines, control-based approaches are entering more and more into the core of synthetic biology ([16], [117], [126], [181]). Control theory is equally applicable, but with some limitations due to biological constrains ([44], [158]), for instance, to implement controllers of particular cellular processes that ultimately may result in great advances which would not be possible when using other approaches. Indeed, control theory approaches may help in producing bio-pharmaceuticals, bio-films, bio-fuels, novel cancer treatments and novel bio-materials (see [107] for a review on synthetic biology applications).

Here, we first introduce a novel mathematical formalism to qualitatively model gene expression and dilution due to cell growth. In fact, one of the aims of systems and synthetic biology is to link molecular-level mechanisms (e.g. gene expression) to cell-level behaviour (e.g. growth rate) ([14, 110]). In the last years much work has focused on the impact of molecular and gene networks on cellular physiology, but less is known about how cellular physiology can influence the machinery of transcription and translation ([114]). To this aim, we present a bacterial growth rate model—where growth is limited by different factors—which ultimately leads to a *switched piecewise quadratic* (SPQ) formalism–derived from *piecewise linear* (PL) systems ([90]).

Then, we focus on the gene expression machinery of the bacterium *Escherichia coli*, with the aim of designing laws to control the growth rate of the cells. Growth control is essential in industrial biotechnology and fundamental research of this kind could pave the way to novel types of antimicrobial strategies. Indeed, the basic effect of an antibiotic is to slow (or arrest) growth. To this end, we present and analyse a bi-dimensional open-loop model of bacterial gene expression in which the growth rate is controlled externally by tuning the synthesis of a component of the gene expression machinery (RNA Polymerase). This type of control can be easily implemented, for instance, by means of inducers that activate synthetic inducible promoters [102].

The paper is organized as follows. In Section 6.2 we briefly recall PL systems. In Section 6.3 we define the cell growth rate model used in this paper. In Section 6.4 we present the SPQ system, which extends the PL system by considering the dilution due to cell growth. In Section 6.5 we state and prove some theorems on the stability of PQ subsystems that will be useful in Section 6.6 to study the stability of the entire SPQ system. Then, in Section 7.3 we present and analyse a bi-dimensional SPQ open-loop control system of bacterial gene expression. Finally, in Section 6.8 we state our conclusion and perspectives.

## 6.2 Piecewise Linear systems overview

The regulation of gene expression plays a pivotal role in the functioning of cells. Mathematical modelling and computational techniques are fundamental to the understanding of these genetic regulatory networks ([14, 54]). The principal modelling challenges come from incomplete knowledge of the networks, and the dearth of quantitative data for identifying kinetic parameters required for detailed mathematical models. Qualitative methods overcome both of these difficulties and are thus well-suited to the modelling and simulation of genetic networks ([144]).

A class of widespread and well studied qualitative models are piecewise linear (PL) systems, originally introduced by [85]. PL models of regulatory genetic networks are built with discontinuous (step) functions. The use of step function has been motivated by the experimental observation that the activity of certain genes changes in a drastic manner at a threshold concentration of a regulatory protein ([180]). The PL model has the general form

$$\dot{x}_i = f_i(x) - d_i x_i, \quad 1 \leq i \leq n, \tag{6.1}$$

where $x = (x_1, ..., x_n)^T \in \mathbb{R}^n_{\geq 0}$ is a vector of protein concentrations. The non-negative quantities $f_i(x)$ and $d_i x_i$ represent synthesis and degradation rates for each protein $x_i$ respectively. The function $f_i : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ represents the expression rate of the gene $i$ depending on the whole state $x$. However, $f_i(x)$ can be detailed as:

$$f_i(x) = \sum_{l=1}^{L_i} k_{il} b_{il}(x)$$

where $k_{il} > 0$ is a rate parameter and $b_{il}(x)$ is a combination of step functions $s^+, s^-$:

$$s^+(x_i, \theta_i^j) = \begin{cases} 1 & \text{if } x_i > \theta_i^j \\ 0 & \text{if } x_i < \theta_i^j \end{cases}; \quad s^-(x_i, \theta_i^j) = 1 - s^+(x_i, \theta_i^j).$$

More details on dynamical analysis and applications of PL models can be found in [40, 56, 90].

## 6.3 The growth rate model

In our model proteins' concentration decreases by two processes: *degradation* and *dilution* due to cell growth ([66]). However, PL systems do not take into account protein dilution, but they only consider protein degradation. To overcome this issue, the PL formalism (6.1) can be extended setting:

$$d_i = \mu(\cdot) + \gamma_i \tag{6.2}$$

where $\mu(\cdot)$ is the bacterial growth rate accounting for proteins' dilution and $\gamma_i$ is a degradation constant.

Several studies have reported that bacterial growth rate is intimately intertwined with gene expression ([31, 114, 151, 159]) and with the type of substrate ([124, 129]).

Here, to keep the discussion as general as possible, we assume that growth rate is limited by the amount of two generic cellular components which are necessary to sustain the gene expression machinery of cell. Notably, let $x_a$, $x_b \in \mathbb{R}_{\geq 0}$ be the concentrations of cellular components $a$ and $b$, respectively we model the bacterial growth rate $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as:

$$\mu(t) = \min(\mu_a\ x_a(t), \mu_b\ x_b(t)) \tag{6.3}$$

where $\mu_a$ and $\mu_b$ are proportion factors depending for instance on the carbon source used. We note that expressions of the type (6.3) are widely used in ecology to model the specific growth rate of species, determined by the resource that is most limiting according to Liebig's "law of the minimum" ([98]), and recently, a similar expression to (6.3) has also been applied in a model of ribosomal regulation in *E. coli* ([155]).

## 6.4 The Switched Piecewise Quadratic (SPQ) system

Since our purpose is dealing with gene expression dependent on bacterial growth, we substitute expression (6.2) with $\mu$ of the form (6.3) into the PL systems (6.1).

Without loss of generality, we choose for $x_a$ and $x_b$ the two last components of $x$. That is, considering $x_a = x_{n-1}$ and $x_b = x_n$ in (6.3), we obtain a new system, which takes

into account the dilution effect, whose general form is:

$$\dot{x}_i = f_i(x) - [\min(\mu_{n-1}\ x_{n-1}, \mu_n\ x_n) + \gamma_i]\, x_i, \quad 1 \leq i \leq n\ . \tag{6.4}$$

We note that system (6.4), according to the evaluation of the function min in $\mu$, can be split into two subsystems (or modes):

$$\begin{aligned} &\text{I: if } \ x_{n-1} < \frac{\mu_n}{\mu_{n-1}}\ x_n \\ &\dot{x}_i = f_i(x) - [\mu_{n-1}\ x_{n-1} + \gamma_i]\, x_i, \quad 1 \leq i \leq n \\ &\text{II: if } \ x_n < \frac{\mu_{n-1}}{\mu_n}\ x_{n-1} \\ &\dot{x}_i = f_i(x) - [\mu_n\ x_n + \gamma_i]\, x_i, \quad 1 \leq i \leq n \end{aligned} \tag{6.5}$$

which share the same structure and properties. Thus, system (6.4) belongs to the class of *switched systems* ([121]) in which the growth rate $\mu$ acts as a rule that orchestrates the switching between the two subsystems in (6.5). Furthermore, noting that subsystems in (6.5) are piecewise quadratic, we named system (6.4) *switched piecewise quadratic* (SPQ).

Moreover, each piecewise quadratic (PQ) system (or mode) in (6.5) can be written in matrix form as

$$\dot{x} = f(x) - d(x_q)x, \tag{6.6}$$

where $f = (f_1, ..., f_n)$ and $d(x_q) = diag(\mu_q\ x_q + \gamma_1, ..., \mu_q\ x_q + \gamma_n)$, where *diag* is the diagonal matrix corresponding to the vector and $q = n - 1$ or $q = n$ depending on whether we refer to mode I or mode II in (6.5), respectively.

To study the dynamics of the SPQ system (6.4) we need first to characterize the dynamics of its PQ modes (6.6), and then investigate the properties arising from the switching condition. To this end, in the next section we present a dynamical study of the PQ subsystem.

## 6.5 The PQ subsystem: dynamical study

For simplicity, we provide a dynamical study only for mode II in (6.5), that is when $x_q = x_n$ in (6.6), but equivalent results can be derived for mode I considering $x_q = x_{n-1}$. The dynamics of the PQ subsystem can be studied in the $n$-dimensional state-space $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$, where each $\Omega_i$ is defined by $\Omega_i = \{x_i \in \mathbb{R}_{\geq 0} | 0 \leq x_i \leq max_i\}$ for some maximum concentration value $max_i$. A protein encoded by a gene will be involved in different interactions at different concentration thresholds, so for each variable $x_i$, we

assume there are $p_i$ ordered thresholds $\theta_i^1, ..., \theta_i^{p_i}$ (we also define $\theta_i^0 = 0$ and $\theta_i^{p_i+1} = max_i$ ). The $(n-1)$-dimensional hyper-planes defined by these thresholds partition $\Omega$ into hyper-rectangular regions we call *domains*. Specifically, a domain $D \subset \Omega$ is defined to be a set $D = D_1 \times ... \times D_n$, where $D_i$ is one of the following:

$$D_i = \left\{ x_i \in \Omega_i | \theta_i^j < x_i < \theta_i^{j+1} \right\} \qquad \text{for } j \in \{0, ..., p_i\}$$
$$D_i = \left\{ x_i \in \Omega_i | x_i = \theta_i^j \right\} \qquad \text{for } j \in \{0, ..., p_i\}.$$

Let $\mathcal{D}$ be the set of domains in $\Omega$. A domain $D \in \mathcal{D}$ is called a *regulatory domain* if none of the variables $x_i$ has a threshold value in $D$ (it is the full hyper-rectangle). In contrast, a domain $D \in \mathcal{D}$ is called a *threshold domain* of order $k \leq n$ if exactly $k$ variables have threshold values in $D$ (in [127] threshold domains are called switching domains, but we avoid this definition to prevent misunderstandings with switched system). The corresponding variables $x_i$ are called *threshold variables* in $D$. The two sets of domains are respectively denoted by $\mathcal{D}_r$ and$\mathcal{D}_t$.

### 6.5.1 Solutions and Stability in Regular Domains

For any regulatory domain $D$, the function $f(x)$ is constant for all $x \in D$, and it follows that the PQ system (6.6) (for $x_q = x_n$) can be written as

$$\dot{x} = f^D - d(x_n)x, \tag{6.7}$$

where $f^D$ is constant in $D$. We note that (6.7) is a hierarchical system, since the differential equation governing $x_n(t)$ depends only on $x_n(t)$ while $\dot{x}_i(t)$ depends only on $x_i(t)$ and $x_n(t)$, but not on $x_j(t)$ for $n > j > i$. Thus, for any $x(t_0) \in D$ the unique solution of (6.7) can be found explicitly by solving first the $n$-component of (6.7)—$\dot{x}_n = f_n^D - (\mu_n\ x_n + \gamma_n)x_n$—which is an autonomous differential equation, and then solving the $i$-components $(i = 1, ..., n-1)$, having substituted $x_n(t)$ into them.

**Lemma 6.1.** *Equation*

$$\dot{x}_n = f_n^D - (\mu_n\ x_n + \gamma_n)x_n \tag{6.8}$$

*admits only one positive equilibrium, which is globally asymptotically stable in* $\mathbb{R}_{\geq 0}$.

*Proof.* Given the initial condition $x_n(t_0)$, it is easy to check that the solution $x_n(t)$ of (6.8) is given by:

$$x_n(t) = x_n^+ + \frac{(x_n^+ - x_n^-)\exp\left[-\mu_n(x_n^+ - x_n^-)(t - t_0)\right]}{\beta - \exp\left[-\mu_n(x_n^+ - x_n^-)(t - t_0)\right]} \tag{6.9}$$

where

$$\begin{aligned} x_n^+ &= \frac{-\gamma_n + \sqrt{\gamma_n^2 + 4\mu_n f_n^D}}{2\mu_n} \\ x_n^- &= \frac{-\gamma_n - \sqrt{\gamma_n^2 + 4\mu_n f_n^D}}{2\mu_n} \end{aligned} \tag{6.10}$$

are, respectively, the unique roots of $f_n^D - (\mu_n\ x_n + \gamma_n)x_n = 0$ and $\beta = \frac{x_n(t_0)-x_n^-}{x_n(t_0)-x_n^+}$. Moreover, since all parameters in (6.8) are positive, it turns out that $x_n^+ > 0$ and $x_n^- < 0$. Finally, from (6.9) follows that

$$\lim_{t\to+\infty} x_n(t) = x_n^+.$$

Hence, $x_n^+$ is the only globally asymptotically stable equilibrium of (6.8) in $\mathbb{R}^n_{\geq 0}$. □

Given $x_n(t)$ from (6.9), $x_i(t)$ $(i = 1, ..., n-1)$ follows by:

$$x_i(t) = \frac{1}{b(t)}\left(b(t_0)x_i(t_0) + f_i^D \int_{t_0}^{t} b(s)ds\right)$$

where $b(t) = \exp\left(\int_{t_0}^{t}(\mu_n\ x_n(\tau) + \gamma_n)d\tau\right)$.

Moreover, regarding the stability of system (6.7) we can state the following two theorems.

**Theorem 6.2.** *Assuming that $D = \mathbb{R}^n_{\geq 0}$, then point $\Phi(D) = (\bar{x}_1, ..., \bar{x}_n)^T$ defined as*

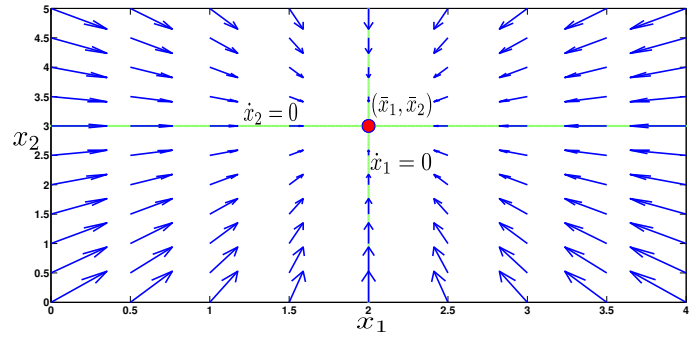
$$\begin{aligned} \bar{x}_i &= \eta(\bar{x}_n, f_i^D, \mu_n, \gamma_i) = \frac{f_i^D}{\mu_n \bar{x}_n + \gamma_i}, \quad i = 1, ..., n-1 \\ \bar{x}_n &= \varphi(f_n^D, \mu_n, \gamma_n) = \frac{-\gamma_n + \sqrt{\gamma_n^2 + 4\mu_n f_n^D}}{2\mu_n}, \end{aligned} \tag{6.11}$$

*is a globally asymptotically stable equilibrium of the system* (6.7)

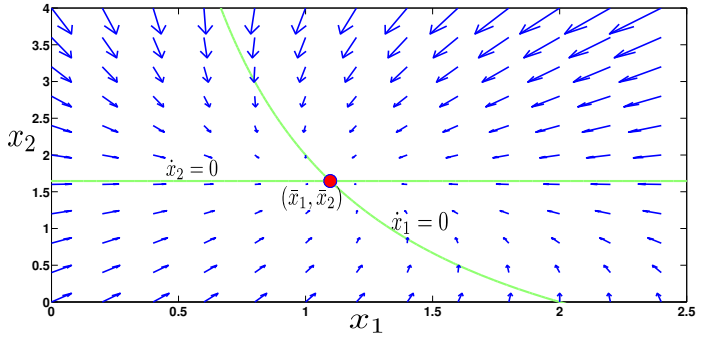
*Proof.* Since $\bar{x}_n = x_n^+$ (see (6.10)), from Lemma 6.1 it turns out that $\bar{x}_n$ is a globally asymptotically stable equilibrium of (6.8). Moreover, $\bar{x}_i$ is a globally asymptotically stable equilibrium for $\dot{x}_i = f_n^D - (\mu_n \bar{x}_n + \gamma_n)x_i$ (which is of the form of PL systems). Finally, for the state $x$ of (6.7), there exists a positively invariant compact set $D' \subset D$ containing $\Phi(D)$, such that $D'$ attracts all the solution of system (6.7) (see [56]). Hence, the proof follows by Theorem 1 in [172]. □

What happens to the stability of system (6.7) if we relax the hypothesis $D = \mathbb{R}^n_{\geq 0}$ and consider $D \in \mathcal{D}_r$?

**Theorem 6.3.** *Let $D \in \mathcal{D}_r$. If $\Phi(D) \in D$, then $\Phi(D)$ is a locally asymptotically stable point of system* (6.7).




(a)




(b)

FIGURE 6.1: Examples of vector-field in a 2D regular domain for the PL system (a) and PQ system (b). In both cases, the focal point (red circle) is inside the domain. The green lines represent the nullclines of the dynamical system. We note how the PL and PQ systems differ in nullclines shape and vector field orientation.
(a): $\dot{x}_1 = k_1 - \gamma_1 x_1$, $\dot{x}_2 = k_2 - \gamma_2 x_2$;
(b): $\dot{x}_1 = k_1 - (\mu_2 x_2 + \gamma_1) x_1$, $\dot{x}_2 = k_2 - (\mu_2 + \gamma_2) x_2$.
Parameter values: $k_1 = 2$, $k_2 = 3$, $\gamma_1 = 1$, $\gamma_2 = 1$, $\mu_2 = 0.5$.

*Proof.* Any $D \in \mathcal{D}_r$, is not in general a positively invariant compact set for $x$, so there could exist solutions $x$ of (6.7) which exit the domain $D$ without converging towards $\Phi(D)$. Hence, $\Phi(D)$ is only locally asymptotically stable with respect to $D$. □

*Remark* 6.4. We note that Theorem 6.3 states a novel behaviour of the PQ systems with respect to PL systems, that is the convergence towards the equilibrium point is not assured from every point within the domain containing the equilibrium as, conversely, it is for PL systems. In fact, if in Figure 6.1(b) there was a threshold domain $x_1 = \theta$ such that $\bar{x}_1 < \theta < \dfrac{k_1}{\mu_2 x_2 + \gamma_1}$, then convergence towards the equilibrium point $(\bar{x}_1, \bar{x}_2)$ would not be assured from any $(x_1, x_2) \in D$, as—for instance—if we consider $\bar{x}_1 < x_1 < \dfrac{k_1}{\mu_2 x_2 + \gamma_1}$ and $x_2 < \bar{x}_2$ (see vector filed orientation).

**Definition 6.5.** Given a regulatory domain $D \in \mathcal{D}_r$, the point $\Phi(D) = (\bar{x}_1, ..., \bar{x}_n)^T \in \Omega$ (defined by (6.11)) is called the *focal point* for the flow in $D$.

The focal points are equilibrium points of the PQ system (6.6) provided that they belong to their respective regular domain, i.e. $\Phi(D) \in D$. If this is the case, the focal points are referred to as *regular equilibria*.

Different regulatory domains will usually have different focal points. In general, all solutions in a regulatory domain $D$ flow towards the focal point $\Phi(D)$ until they either reach it or leave the domain $D$. What happens when a solution leaves a regulatory domain $D$ and enters a threshold domain in the boundary of $D$? Since the step functions are not defined when a variable $x_i$ takes some threshold value $\theta_i^{p_i}$, the vector field is undefined on the threshold domains. We need to precise our definition of solutions.

### 6.5.2 Solutions and Stability in Threshold Domains

To provide the existence and the possibility for solutions to be continued on all domains, we have to define the right-hand side of system (6.6) at the points of discontinuity of the function $f$. To this end, we use a construction originally proposed by Filippov [73] and then applied to PL systems ([90]). The method consists of extending the system (6.6) to a differential inclusion,

$$\dot{x} \in H(x), \tag{6.12}$$

where $H$ is a set valued function (i.e. $H(x) \subseteq \mathbb{R}^n$). If $D \in \mathcal{D}_r$, then we define $H$ simply as

$$H(x) = \{f^D - d(x_n)x\}, \quad \forall x \in D. \tag{6.13}$$

If $D \in \mathcal{D}_t$, we define $H$ as

$$H(x) = \overline{co}(\{f^{D'} - d(x_n)x \mid D' \in R(D)\}), \quad \forall x \in D, \tag{6.14}$$

where $R(D) = \{D' \in \mathcal{D}_r \mid D \subseteq \partial D'\}$ is the set of all regulatory domains with $D$ in their boundary, and $\overline{co}(X)$ is the closed convex hull of $X$. For threshold domains, $H(x)$ is typically multi-valued so solutions of the differential inclusion are defined as follows.

**Definition 6.6.** A solution of (6.12) on $[0, T]$ in the sense of Filippov is an absolutely continuous function (w.r.t. $t$) $\xi_t(x_0)$ such that $\xi_0(x_0) = x_0$ and $\dot{\xi}_t \in H(\xi_t)$, for almost all $t \in [0, T]$.

Now, we shall explain how to construct the set $H(x)$ at discontinuity points of $f$.

Let us consider the case where $x$ belongs to a threshold domain $S$ separating two $n$-regular domains $D_1$ and $D_2$. Hence,

$$H(x) = \overline{co}(\{f^{D_1} - d(x_n)x, f^{D_2} - d(x_n)x\})$$

represents the segment joining the endpoints of the vectors $g_1 = f^{D_1} - d(x_n)x$ and $g_2 = f^{D_2} - d(x_n)x$. Trajectories can cross $S$ if the vector fields $g_1$ and $g_2$ point in a similar direction, slide along $S$ if $g_1$ and $g_2$ point in opposite directions towards $S$ and be repelled from $S$ if $g_1$ and $g_2$ point in opposite directions away from $S$. The last two cases are known as *stable* and *unstable sliding motion* in the literature ([73]). Moreover, the velocity of the sliding motion (stable or unstable) on $S$ is given by

$$\dot{x} = f^S - d(x_n)x. \tag{6.15}$$

**Theorem 6.7.** *Assume that $S$ is in the hyper-plane $C_i^j = \{x \in \mathbb{R}^n_{\geq 0} \ : \ x_i = \theta_i^j\}$ and a sliding motion (stable or unstable) occurs on $S$. The vector $f^S$ in (6.15) is given by*

$$\begin{aligned} f^S &= \alpha f^{D_1} + (1-\alpha) f^{D_2}, \\ \alpha &= \begin{cases} \dfrac{f_i^{D_2} - (\gamma_i + \mu_n x_n)\theta_i^j}{f_i^{D_2} - f_i^{D_1}}, & \text{if } i \in \{1, ..., n-1\} \\ \dfrac{f_i^{D_2} - (\gamma_i + \mu_n \theta_i^j)\theta_i^j}{f_i^{D_2} - f_i^{D_1}}, & \text{if } i = n. \end{cases} \end{aligned} \tag{6.16}$$

*Proof.* The segment joining the endpoints of the vectors $f^{D_1} - d(x_n)x$ and $f^{D_2} - d(x_n)x$ is expressed by

$$\alpha f^{D_1} + (1-\alpha) f^{D_2} - d(x_n)x, \quad 0 \leq \alpha \leq 1.$$

Since the state trajectories during sliding motion are on the hyper-plane $x_i = \theta_i^j$, the parameter $\alpha$ in (6.16) is selected such that the state velocity vector of the system (6.15) is in this hyper-plane. Thus, depending on whether the sliding motion occurs in $x_i = \theta_i^j$ ($i \in \{1, ..., n-1\}$) or in $x_n = \theta_n^j$, we have two values of $\alpha$, which are found from the conditions

$$\begin{aligned} &\alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\gamma_i + \mu_n x_n)\theta_i^j = 0, \text{ if } i \in \{1, ..., n-1\} \\ &\alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\gamma_i + \mu_n \theta_i^j)\theta_i^j = 0, \text{ if } i = n. \end{aligned}$$

□

*Remark* 6.8. In the case $i = n$ the value of $\alpha$ is constant $\forall x \in S$ and thus a sliding motion occurs along the entire threshold domain $S$. By contrast, in the case $i \in \{1, ..., n-1\}$ the value of $\alpha$ depends on $x_n$, this means that a sliding motion occurs on $S$ as long as the vector fields point in opposite direction towards (or away) $S$ or, equivalently, as long as $0 < \alpha(x_n) < 1$ $\forall x_n \in S$. Specifically, it could happen—in contrast to PL systems—that solutions slide for a while along $S$ and then leave it as soon as the condition $0 < \alpha(x_n) < 1$ does not hold any more.

It is useful to define a concept analogous to the focal points defined for regulatory domains, extended to deal with threshold domains.

**Definition 6.9.** We recall that $supp(D)$ is the $(n-k)$-dimensional hyperplane supporting $D$. Let $D$ be a threshold domain of order $k$, then its focal set $\Phi(D)$ is

$$\Phi(D) = supp(D) \ \cap \ \{x : \ 0 \in H(x)\}, \tag{6.17}$$

where $H(x)$ is defined as in (6.14).

Hence, $\Phi(D)$ for $D \in \mathcal{D}_t$ is the set containing all the equilibrium points of the differential inclusion (6.12), which lie on $supp(D)$. Thus, $\Phi(D)$ can be a singleton, but more generally is a closed convex bounded set and hence is referred to as a focal set.

From now on—to rule out some singular cases when proving results on stability—we make the following technical assumption on the focal sets for our system:

***Assumption*** 2. $\forall \ D \ \in \mathcal{D}$, we assume that

$$\Phi(D) \cap supp(D') = \{\}, \ \forall D^{'} \in \partial(D).$$

It essentially says that for every regular and threshold domain $D$, the focal set $\Phi(D)$ does not intersect the supporting hyperplane of any domain $D'$ in the boundary of $D$.

It is possible that solutions of (6.6) reach equilibria that lie in threshold domains and such equilibria are called *singular equilibria*. In general, a singular equilibrium $\bar{x}$ of system (6.6) is a point that satisfies the condition $0 \in H(\bar{x})$ and that belongs to some threshold plane. Determining in the most general case whether a singular equilibrium is stable or unstable requires a detailed analysis that for the sake of space is not mentioned in this paper. However, in the following theorem we present a procedure to assess the stability of singular equilibria that can occur on $x_n$-hyperplane.

**Theorem 6.10.** *Assume that a sliding motion occurs on a threshold domain $S$, which lies in the hyper-plane $C_n^j = \{x \in \mathbb{R}^n_{\geq 0} \ : \ x_n = \theta_n^j\}$, separating two $n$-domains $D_1$ and $D_2$. Let $\bar{x} = (\bar{x}_1, ..., \bar{x}_n)$ be the singular equilibrium point of the sliding motion. If $\bar{x} \in S$ and if the sliding motion is stable (resp. unstable), then $\bar{x}$ is locally stable (resp. unstable).*

*Proof.* Assuming the presence of a such stable sliding motion in $S$ and $\bar{x} \in S$, this implies that there exists a neighbourhood of $\bar{x}$ where the $n$-component of trajectories are approaching $\bar{x}_n = \theta_n^j$. Notably, the velocity of motion of the other $i$-components

$(i = 1, ..., n-1)$ is given by

$$\dot{x}_i = \alpha f_i^{D_1} + (1-\alpha) f_i^{D_2} - (\mu_n \theta_n^j + \gamma_i) x_i \tag{6.18}$$

with $\alpha$ equal to the second value in (6.16). Hence, the stability of $\bar{x}$ follows by the fact that (6.18) is of the PL form. If the $\bar{x} \in S$, but the sliding motion is unstable, the instability of $\bar{x}$ follows from the instability of the sliding motion. □

After having studied the stability of PQ modes, we are able–in the next section—to present the stability analysis of the entire SPQ system.

## 6.6 Stability Analysis of the SPQ system

Within the hybrid systems literature, much has been written on the stability of switching systems ([28]). Usually, Multiple Lyapunov Functions are used to prove Lyapunov stability for switched systems ([28, 121]). However, many results using this approach are not directly applicable to systems with sliding motions and/or cases when the domains do not have a common focal point. Moreover, the structure of the SPQ system (6.4) is particular and the problem we consider quite specific, which allows us to take a different approach.

More specifically, we can define two regions $\chi_I$ and $\chi_{II} \subset \mathbb{R}^n$, in which system (6.4) is active following the I-mode and the II-mode, respectively:

$$\begin{aligned} \chi_I &= \{[x_1, ..., x_n]^T \in \mathbb{R}^n : \mu_{n-1} x_{n-1} - \mu_n x_n < 0\} \\ \chi_{II} &= \{[x_1, ..., x_n]^T \in \mathbb{R}^n : \mu_{n-1} x_{n-1} - \mu_n x_n > 0\} \ . \end{aligned} \tag{6.19}$$

In addition, a switching surface between the I and II modes, i.e. a boundary between $\chi_I$ and $\chi_{II}$, is given by:

$$S_{I,II} = \{[x_1, ..., x_n]^T \in \mathbb{R}^n : \mu_{n-1} x_{n-1} - \mu_n x_n = 0\} \ . \tag{6.20}$$

We will now state two definitions and an hypothesis useful to enunciate a theorem for the stability of system (6.4).

**Definition 6.11.** Let $\Psi_m$ $(m = I, II.)$ be the set containing all the locally stable points of the $m$-mode.

**Definition 6.12.** Let $\Lambda_m$ $(m = I, II.)$ be the set containing all the unstable points of the $m$-mode.

We recall that the procedures to determine $\Psi_m$ and $\Lambda_m$, that is the stable and unstable points of the two modes, have been presented in Section 6.5.

***Assumption*** 3. Assume that:

$$\Psi_m \cap S_{I,II} = \{\} \ (\forall m = I, II), \quad \Lambda_m \cap S_{I,II} = \{\} \ (\forall m = I, II).$$

Assumption 3 states that equilibria of the I and II modes do not lay on switching surface $S_{I,II}$. This rules out some very special cases when proving results on stability of the switched system (6.4).

**Theorem 6.13.** *Assume that hypothesis 3 holds, then the set of locally stable points of (6.4), i.e. $\Psi$, and set of unstable points of (6.4), i.e. $\Lambda$, are given by:*

$$\begin{aligned} \Psi &= (\Psi_I \cap \chi_I) \cup (\Psi_{II} \cap \chi_{II}) \\ \Lambda &= (\Lambda_I \cap \chi_I) \cup (\Lambda_{II} \cap \chi_{II}) \,. \end{aligned} \tag{6.21}$$

*Proof.* The proof follows by observing that a stable (resp. unstable) point of the m-mode, is also a stable (resp. unstable) point of the switched system (6.4) only if it is within the space region in which the m-mode is active, i.e. $\chi_m$. □

## 6.7 Open loop control of the RNAP-ribosomes system

From here on we focus on the gene expression machinery of the bacterium *Escherichia coli*, with the aim of controlling the growth rate of the cells. The model and the control we shall introduce have been developed in collaboration with our biologist research partners (Ibis team Inria and UJF Grenoble).

In the presence of a carbon source *E. coli* grows in an exponential manner until it exhausts the nutrient sources, and then enters a stationary phase with practically zero growth [129]. Our control objective is to force the bacterium to significantly modify its response to the carbon source so as to tune the growth rate during the exponential phase. To this end, we take into account the recent applications of synthetic biology which allow us to engineer promoters which in turn can be externally controlled by inducers [102].

Notably, we will study an open loop configuration of a bi-dimensional SPQ model of a mutant *E. coli* inspired by the experiments in [160]. The two basic variables of our model, which describe the gene expression machinery that is responsible for bacterial growth are (see Fig. 6.2):

1. the concentration of RNA Polymerase (RNAP), protein that catalyses the transcription of all genes. Without RNAP, the bacteria cannot produce any RNAs.

2. the concentration of Ribosomes (RIB), a complex molecular machine serving for the production of all proteins (translation).

Moreover, we will assume that an engineered inducible-promoter is used to externally control the expression of RNAP (similarly to [160]). We note that we could theoretically have decided to control the synthesis of Ribosomes, but this seems to be much more complicated to implement in practice since Ribosomes are more complex entities.

Finally, as in [38] we assume that growth rate is intimately related to the capacity of cells to produce *bulk proteins*, which represent cell building and maintenance proteins essential for bacterial growth. Bulk proteins, as any other protein, are produced in a two-step process (gene expression) in which RNAP and ribosomes play a pivotal role ([77]). The first step, i.e. transcription, is catalysed by RNAP which allows the synthesis of mRNA from DNA. During the second step, i.e. translation, the mRNA is translated into proteins by ribosomes. Taking this into account, the cell's growth rate–considered as a sort of production rate of bulk proteins–is thus limited by two potential limiting factors: RNAP and ribosomes. Thus, let $x_p$, $x_r \in \mathbb{R}_{\geq 0}$ be the concentrations of RNAP and ribosomes, respectively we modeled the bacterial growth rate $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ as:

$$\mu(t, u) = \min(\mu_p \; x_p(t, u), \mu_r \; x_r(t)) \tag{6.22}$$

where $\mu_p$ and $\mu_r$ are proportion factors depending on the carbon source used, whereas $u$ is the control input that allows us to modulate the expression of RNAP. It is evident from (6.22) that controlling the expression of RNAP will eventually lead to a growth rate tuning.

During exponential phase—the period characterized by cell doubling— the bacterial culture shows a constant growth rate [129]. This means that, according to (6.22), stable fixed points of RNAP and RIB have to be reached. Hence, our expression of growth rate during exponential phase reads:

$$\bar{\mu}(u) = \min(\mu_p \; \bar{x}_p(u), \mu_r \; \bar{x}_r) \tag{6.23}$$

where $\bar{x}_p$ and $\bar{x}_r$ are concentrations of RNAP and RIB at steady state, respectively. Hence, controlling bacterial growth rate will result in controlling the location of $x_p$, $x_r$ stable points by means of the control input $u$ so as to have a desired constant $\bar{\mu}$ at exponential phase.

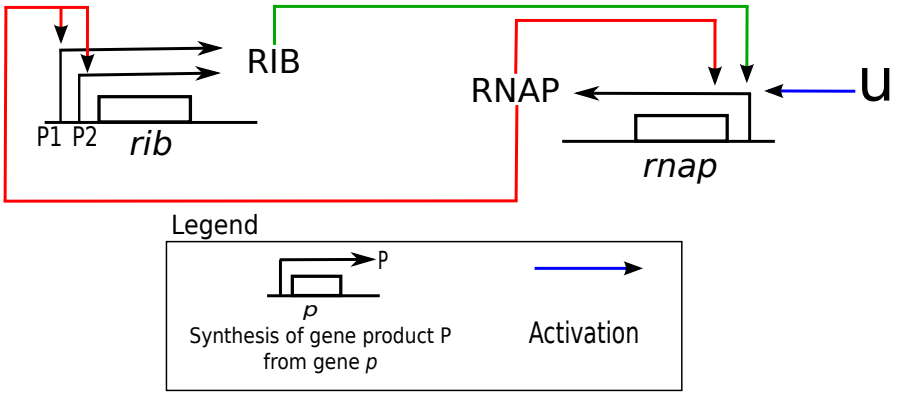



FIGURE 6.2: Regulatory network of the SPQ model (6.24). The system consists of RNAP and ribosomes (RIB), encoded by their respective proxy genes *rnap* and *rib* (it is worth saying that RNAP and RIB are made up of different subunits encoded by different genes, but here, for the sake of simplicity, we consider that they are originated by lumped (proxy) gene classes (*rnap* and *rib*)). Notably, the synthesis of RNAP is positively regulated by RIB, RNAP and the control input $u$ while the synthesis of RIB is positively regulated by RNAP from the promoters P1 and P2.

### 6.7.1 SPQ model of the open-loop control system

In order to show how we can control the bacterial growth rate acting on the RNAP expression, we propose a bidimensional SPQ open-loop model (depicted in Figure 6.2) describing the concentrations' dynamics of RNAP and ribosomes. Let $x_p$, $x_r$ be the concentration of RNAP and ribosomes respectively, the SPQ model is given by:

$$\begin{aligned}\dot{x}_r &= k_r^1 s^+(x_p,\theta_p^1) + k_r^2 s^+(x_p,\theta_p^2) - (\min(\mu_p x_p, \mu_r x_r) + \gamma_r)x_r \\ \dot{x}_p &= u\left[k_p^0 s^+(x_p,\theta_p^1)s^+(x_r,\theta_r^1) + k_p^1 s^+(x_p,\theta_p^2)s^+(x_r,\theta_r^2)\right] \\ &\quad - (\min(\mu_p x_p, \mu_r x_r) + \gamma_p)x_p\end{aligned} \tag{6.24}$$

We have considered that the synthesis of ribosomes is limited by the production rate of stable-RNAs, which in turn are essentially regulated at the level of transcription by RNAP ([89]). More specifically, stable-RNAs are produced from two promoters, P1 and P2. We assumed that a lower concentration of RNAP, i.e. $\theta_p^1$, activates the house-keeping promoter P2 while a higher RNAP concentration, i.e. $\theta_p^2$, is needed to stimulates the promoter P1, whose activity increases with growth rate.

Regarding RNAP, we note that the control input $u$ acts on the RNAP synthesis rate modulating its level of expression. Notably, we assumed that $u \in [0,1]$ in order to allow $u$ to mimic the effect of typical regulation functions (e.g. Hill type functions) employed in synthetic biology for modelling engineered promoters controlled by inducers. Moreover, for the RNAP synthesis we have considered both transcription and translation regulation. Hence, a lower concentration of RNAP (accounting for transcription), i.e.

$\theta_p^1$, and a lower concentration of ribosomes (according for translation), i.e. $\theta_r^1$, are required for the basal synthesis ($u\ k_p^0$) of RNAP whereas, for the main synthesis of RNAP ($u[k_p^0 + k_p^1]$), higher concentrations of RNAP ($\theta_p^2$) and ribosomes ($\theta_r^2$) are needed. From the considerations above, it follows that:

$$0 \leq \theta_r^1 \leq \theta_r^2 \leq max_r; \quad 0 \leq \theta_p^1 \leq \theta_p^2 \leq max_p\ .$$

Therefore, the state space of each of two modes of system (6.24) can be partitioned into nine *regular domains*, where the vector field is uniquely defined:

$$\begin{aligned}
D_1 &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_r < \theta_r^1,\ 0 \leq x_p < \theta_p^1\right\}\\
D_2 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^1 < x_r < \theta_r^2,\ 0 \leq x_p < \theta_p^1\right\}\\
D_3 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^2 < x_r \leq max_r,\ 0 \leq x_p < \theta_p^1\right\}\\
D_4 &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_r < \theta_r^1,\ \theta_p^1 < x_p < \theta_p^2\right\}\\
D_5 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^1 < x_r < \theta_r^2,\ \theta_p^1 < x_p < \theta_p^2\right\}\\
D_6 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^2 < x_r \leq max_r, \theta_p^1 < x_p < \theta_p^2\right\}\\
D_7 &= \left\{x \in \mathbb{R}^2_{\geq 0} : 0 \leq x_r < \theta_r^1,\ \theta_p^2 < x_p \leq max_p\right\}\\
D_8 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^1 < x_r < \theta_r^2,\ \theta_p^2 < x_p \leq max_p\right\}\\
D_9 &= \left\{x \in \mathbb{R}^2_{\geq 0} : \theta_r^2 < x_r \leq max_r, \theta_p^2 < x_p \leq max_p\right\}.
\end{aligned}$$

The *threshold domains* are not listed here, but they are as defined in Section 6.5.

Let mode-I be active when $\min(\mu_p x_p, \mu_r x_r) = \mu_p x_p$ and mode-II be active when $\min(\mu_p x_p, \mu_r x_r) = \mu_r x_r$ in (6.24). Hence, according to (6.11) we can calculate the focal points of modes I and II for each regular domains $D_j$ $(j = 1, ..., 9)$:

| $D_j$ | I-mode | II-mode |
|---|---|---|
| $D_1$ | $\bar{x}_p = \varphi(0, \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(0, \mu_r, \gamma_r)$ |
| $D_2$ | $\bar{x}_r = \eta(\bar{x}_p, 0, \mu_p, \gamma_r)$ | $\bar{x}_p = \eta(\bar{x}_r, 0, \mu_r, \gamma_p)$ |
| $D_3$ | | |
| $D_4$ | $\bar{x}_p = \varphi(0, \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(k_r^1, \mu_r, \gamma_r)$ |
| | $\bar{x}_r = \eta(\bar{x}_p, k_r^1, \mu_p, \gamma_p)$ | $\bar{x}_p = \eta(\bar{x}_r, 0, \mu_r, \gamma_p)$ |
| $D_5$ | $\bar{x}_p = \varphi(uk_p^0, \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(k_r^1, \mu_r, \gamma_r)$ |
| $D_6$ | $\bar{x}_r = \eta(\bar{x}_p, k_r^1, \mu_p, \gamma_r)$ | $\bar{x}_p = \eta(\bar{x}_r, uk_p^0, \mu_r, \gamma_p)$ |
| $D_7$ | $\bar{x}_p = \varphi(0, \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$ |
| | $\bar{x}_r = \eta(\bar{x}_p, k_r^1 + k_r^2, \mu_p, \gamma_r)$ | $\bar{x}_p = \eta(\bar{x}_r, 0, \mu_r, \gamma_p)$ |
| $D_8$ | $\bar{x}_p = \varphi(uk_p^0, \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$ |
| | $\bar{x}_r = \eta(\bar{x}_p, k_r^1 + k_r^2, \mu_p, \gamma_r)$ | $\bar{x}_p = \eta(\bar{x}_r, uk_p^0, \mu_r, \gamma_p)$ |
| $D_9$ | $\bar{x}_p = \varphi(u[k_p^0 + k_p^1], \mu_p, \gamma_p)$ | $\bar{x}_r = \varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$ |
| | $\bar{x}_r = \eta(\bar{x}_p, k_r^1 + k_r^2, \mu_p, \gamma_r)$ | $\bar{x}_p = \eta(\bar{x}_r, u[k_p^0 + k_p^1], \mu_r, \gamma_p)$ |

(6.25)

It turns out from (6.25) that focal points' location of modes I and II of system (6.24) depends on the control input $u$. Therefore, the number of equilibria of system (6.24) depends on $u$ too. This means that model (6.24) can be controlled towards different equilibrium points by choosing appropriate values of $u$ and this, eventually, will result in controlling the bacterial growth rate during the exponential phase. In fact, exponential growth rate is related to the equilibrium reached by the system (6.24) through expression (6.23).

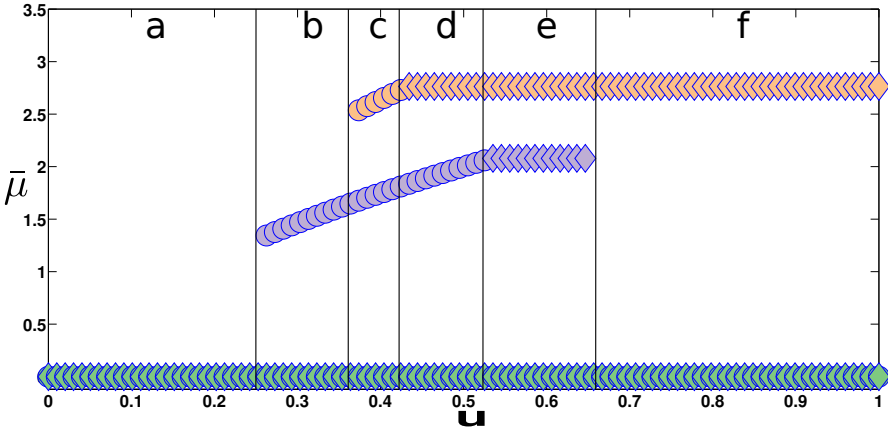



FIGURE 6.3: Growth rate bifurcation diagram as function of the control input $u$. In circles are plotted growth rate values when mode-I is active while in diamonds those when mode-II is active. Notably, green circles denote $\bar{\mu} = \mu_p\varphi(0, \mu_p, \gamma_p)$, green diamonds denote $\bar{\mu} = \mu_r\varphi(0, \mu_r, \gamma_r)$, purple circles denote $\bar{\mu} = \mu_p\varphi(uk_p^0, \mu_p, \gamma_p)$, orange circles denote $\bar{\mu} = \mu_p\varphi(u[k_p^0 + k_p^1], \mu_p, \gamma_p)$, purple diamonds denote $\bar{\mu} = \mu_r\varphi(k_r^1, \mu_r, \gamma_r)$, orange diamonds denote $\bar{\mu} = \mu_r\varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$. Parameter values used: $\theta_r^1 = 1.1$, $\theta_p^1 = 1.1$, $\theta_p^2 = 2.1$, $\theta_r^2 = 2.1$, $k_r^1 = 8$, $k_r^2 = 5$, $k_p^0 = 10$, $k_p^1 = 10$, $\gamma_r = 1$, $\gamma_p = 1$, $\mu_r = 0.8$ end $\mu_p = 0.5$.

To illustrate this, we have depicted in Fig. 6.3 the $\mu$-bifurcation diagram when the control input $u$ varies from 0 to 1 while the other parameter values stay constant and set to values reported in Fig. 6.3. Notably, to draw Fig. 6.3 we first calculated the equilibria of the SPQ system (6.24) as function of $u$ applying Theorem 6.13 and then computed the growth rate expression at steady state using (6.22).

As shown in Fig. 6.3, for the parameter values used, we have six different scenarios which correspond to different equilibria of system (6.24) and so to different growth rates. Moreover, to point out which mode contributes to the system's equilibria we have plotted growth rate equilibria resulting from mode-I ($\mu_p\bar{x}_p$) in circles and growth rate equilibria resulting from mode-II ($\mu_r\bar{x}_r$) in diamonds in Figure 6.3.

**Proposition 6.14** (Trivial Equilibrium)**.** Assume the control input $u$ is such that $\bar{x}_p = \varphi(uk_p^0, \mu_p, \gamma_p) < \theta_p^1$, then the only equilibrium of the system is the origin, which is an equilibrium for both modes (scenario-$a$). Moreover, system (6.24) always converges

to the origin—independently from $u$—when the initial condition $x_p(t_0)$ satisfies $x_p(t_0) < \theta_p^1$(all scenarios in Figure 6.3).

From a biological point of view, Proposition 6.14 says that the origin is an equilibrium of system (6.24) either when $u$ does not sufficiently induce RNAP expression (scenario-$a$) or when RNAP initial condition is is too low to initiate gene transcription (all scenarios in Figure 6.3). The origin represents absence of growth and we refer to it as the *trivial* equilibrium, for this reason we will not mention it any more in the discussion which follows.

**Proposition 6.15** (Monostability)**.** Assume $u$ is such that the only equilibrium of the system is given by the focal point of $D_5$, $D_6$ of mode-I (see (6.25)), then system (6.24) is monostable and the growth rate reads as $\bar{\mu} = \mu_p\varphi(uk_p^0, \mu_p, \gamma_p)$ (scenario-$b$). Moreover, assume $u$ is such that the only equilibrium of the system is given by the focal point of $D_9$ of mode-II (see (6.25)), then system (6.24) is still monostable and the growth rate reads as $\bar{\mu} = \mu_r\varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$ (scenario-$f$).

When $u$ is such that system (6.24) has two (non trivial) equilibria which belong to the same mode we say that system (6.24) is *mono-mode bistable*. Notably, we can have that the two equilibria belong to mode-I (scenario-$c$) or to mode-II (scenario-$e$).

**Proposition 6.16** (Mono-mode Bistability)**.** Assume $u$ is such that system (6.24) has two equilibria in the same mode, then we can have:

- scenario-$c$: equilibria belong to mode-I and they are the focal point of $D_5$, $D_6$ (for the lower one) and the focal point of $D_9$ (for the higher one). Hence, the two growth rate expressions are: $\bar{\mu} = \mu_p\varphi(uk_p^0, \mu_p, \gamma_p)$ and $\bar{\mu} = \mu_p\varphi(u[k_p^0 + k_p^1], \mu_p, \gamma_p)$.
- scenario-$e$: equilibria belong to mode-II and they are given by the focal point of $D_5$, $D_6$ (for the lower one) and by the focal point of $D_9$ (for the higher one). Thus, now the two growth rate expressions are: $\bar{\mu} = \mu_r\varphi(k_r^1, \mu_r, \gamma_r)$ and $\bar{\mu} = \mu_r\varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$.

When $u$ is such that system (6.24) has two (non trivial) equilibria shared by the two modes, we say that system (6.24) is *bi-mode bistable*. This is the case of scenario-$d$ in Figure 6.3.

**Proposition 6.17** (Bi-mode Bistability)**.** Assume that the equilibria are given by the focal point of $D_5$, $D_6$ of mode-I and by the focal point of $D_9$ of mode-II (see (6.25)), then the two growth rate expressions are $\bar{\mu} = \mu_p\varphi(uk_p^0, \mu_p, \gamma_p)$ and $\bar{\mu} = \mu_r\varphi(k_r^1 + k_r^2, \mu_r, \gamma_r)$.

We think that scenario-$d$ is the most intriguing case, because it shows how both modes can contribute to system stability. For this reason, the phase plane of system (6.24) representing scenario-$d$ (bi-mode bistability) is shown in Figure 6.4.

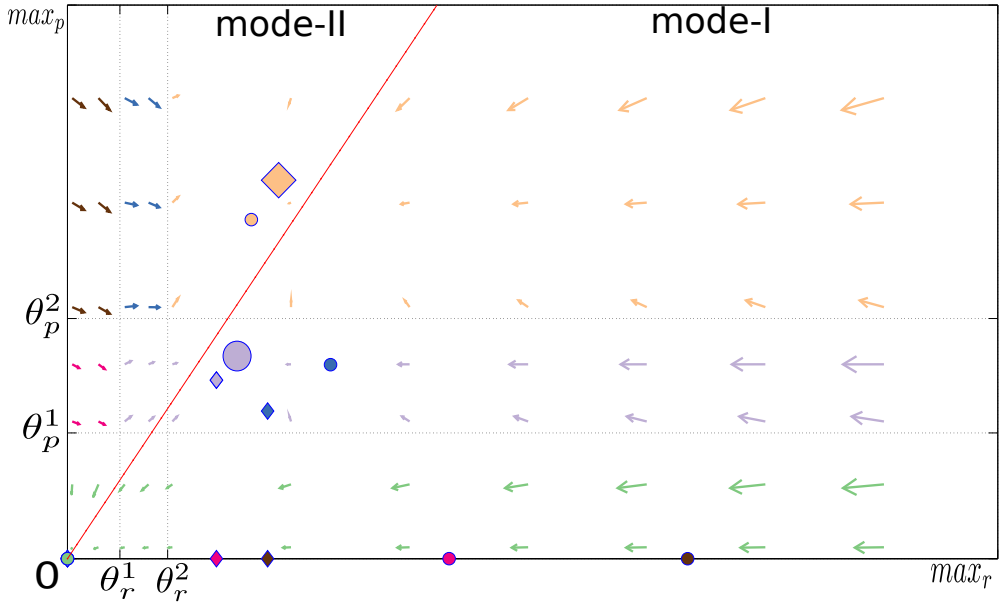



FIGURE 6.4: Phase planes of the SPQ system (6.24) of scenario-$d$ depicted in Figure 6.3 (*bi-mode bistable*). For each regular domain are drawn the focal points of I-mode (circles) and II-mode (diamonds). The color of the focal points is the same of the arrows of the domain whence they are originated. The red line represents the switching surface. The equilibria of the two modes are drawn in larger size. Notably, the stable points of the SPQ system are $(0,0)$, $(\eta(\bar{x}_p, k_r^1, \mu_p, \gamma_r), \varphi(uk_p^0, \mu_p, \gamma_p))$, $(\varphi(k_r^1 + k_r^2, \mu_r, \gamma_r), \eta(\bar{x}_r, u[k_p^0 + k_p^1], \mu_r, \gamma_p))$.

## 6.8 Conclusion

In this paper we proposed a new mathematical formalism to model bacterial gene expression dependent on dilution due to growth rate. This novel modelling approach can be considered as an extension of piecewise linear (PL) systems, which have been modified by introduction of an expression for the growth rate to model the dilution effect. The derived system is a switched system whose two modes are piecewise quadratic (PQ), hence we named this new formalism as switched piecewise quadratic (SPQ). We have first focused on the characterization of equilibria of the PQ subsystems, both for equilibria within regular domains and equilibria that lie on surfaces of discontinuity (threshold domains) due to the use of step functions (as in PL models). Then, we took into account the switching behaviour of the SPQ system to formulate a criterium assessing the stability of its equilibria.

Finally, a minimal SPQ model consisting of two variables (RNA Polymerase and Ribosomes) and an input was analysed and used to describe one possible mechanism to control the growth rate of *E. coli* cells during exponential phase. The qualitative dynamics of the model can thus be studied, and the growth rate bifurcation diagram with respect to the input is obtained. Notably, the growth rate bifurcation diagram has shown out that the bacterial exponential growth rate can be controlled—by means of appropriate values of the input—towards different levels which correspond to different equilibria of the SPQ system. Notably, we point out the interesting bi-mode bistability scenario where both PQ modes contributes to the equilibria of the SPQ system.

Our results combine the control based approaches of engineering with biology, with the aim of implementing new functions in the cells (synthetic biology), which may have some practical and relevant importance in applications such as production of biopharmaceuticals, novel bio-materials and developing new bio-technologies.

One possible extension for the SPQ systems—to get even more realistic growth rate values in some biological conditions—would be that of considering a cut-off value of the growth rate, that is a maximal ($\mu_{max}$) growth rate. In this case, expression (6.3) would be modified in:

$$\mu(t) = \min(\mu_a \; x_a(t), \mu_b \; x_b(t), \mu_{max}) \tag{6.26}$$

With a growth rate model as (6.26), the SPQ system will have a third mode, which will be active when $\mu(t) = \mu_{max}$. Moreover, it is easy to check that this third mode is a PL system and so no other specific studies are required to handle this extension.

To conclude, we believe that the SPQ formalism is a promising approach for qualitative modelling gene expression dynamics dependent on dilution and a valid starting point to help guide the construction of synthetic gene networks, by allowing biologists to select experimental conditions most likely to yield successful results.

# Chapter 7

# Attractor computation using interconnected Boolean networks: testing growth models in *E. Coli*

This chapter is taken from a recent accepted paper that Madalena Chaves and I wrote for the *Theoretical Computer Science* journal [42].

A recently developed method has shown that the attractors, or asymptotic behaviour, of an asynchronous Boolean network can be computed at a much lower cost if the network is written as an interconnection of two smaller modules. We have applied this methodology to study the interconnection of two Boolean models to explore bacterial growth and its interactions with the cellular gene expression machinery, with a focus on growth dynamics as a function of ribosomes, RNA polymerase and other "bulk" proteins inside the cell. Notably, some of the growth rate models tested here are Boolean versions of those growth rate functions introduced in Chapters 4-5-6. The discrete framework permits easier testing of different combinations of biochemical interactions, leading to hypotheses elimination and model discrimination, and thus providing useful insights for the construction of a more detailed dynamical growth model. In this work, I did not participate in the development of the Boolean methods, whereas my contribution was in the design of the different tested Boolean models and in the interpretation of results.

## 7.1 Introduction

Large networks with complex interactions are hard to analyse in detail, but logical and discrete models can facilitate this task. Based essentially on the structure and

topology of the network interactions, logical models provide qualitative information on the dynamical properties of system [85, 163], which can be used for model discrimination and guidance in model improvement. There are many recent examples of applications of discrete models including *Drosophila* embryo pattern formation [11, 148], yeast cell cycle [154], T-cell response [146], or an apoptosis network [34].

Boolean networks are a class of logical models whose variables are described in terms of only two levels (1 or 0; presence or absence; "on" or "off"), which have been useful for biochemical systems [175]. The dynamics of a Boolean model is determined by specifying an updating mode, most commonly synchronous (all nodes updated simultaneously) or asynchronous (only one node updated at any given instant). Since the state space is finite, the dynamics can be represented in terms of a transition graph, which can be studied using some classical algorithms from graph theory [49]. Other, more specific tools are available for an exact and rigorous analysis of the transition graph [123], computation of attractors (or asymptotic behaviour) [45], and other properties [164]. In addition, a wide range of computer tools are available for simulation and analysis of discrete models [88], model reduction [132], or model checking [70].

It is clear that discrete models are not appropriate to finely describe the behaviour of a system, since they cannot represent continuous effects (such as indicating whether an oscillation is sustained or damped), but they are useful to verify whether a given network of interactions is feasible and compatible with known properties of the system. This is a first step towards the construction of a more detailed and informed model.

As an application, we will analyse a network of interactions involved in determining bacterial growth of *Escherichia Coli*, which varies non-linearly with different factors, such as availability of nutrients or the concentration of the necessary enzymes and proteins needed for cell division [24, 155]. Mathematical models have been developed to describe and reproduce several regulatory modules and their response to nutrient availability [92, 144]. One of the least understood aspects in these studies remains the actual modelling of bacterial growth: while it is clear that growth depends on the general availability of "bulk" proteins, ribosomes, and RNA polymerase, it is difficult to find a reasonable mathematical model that reproduces all these effects [86]. In many cases, growth is considered to be a given constant and the model is designed to reproduce a single phase of bacterial growth.

Here, we propose to test and study a dynamical function for bacterial growth in terms of the major components involved in bacterial cell division, that is, gene transcription (RNA polymerase) and translation (ribosomes). To test the feasibility of mathematical growth functions, we will focus on a qualitative model of the network involved in the

carbon starvation response [144] and its interconnection with a basic model describing the dynamics of ribosomes and RNA polymerase (see Section 7.3).

We will use two methods for analysis of qualitative systems (see Section 7.2): first, a method that transforms piecewise affine (PWA) systems into discrete and then Boolean models [46, 99]; and, second, a recently developed method to compute the attractors of an interconnection of two Boolean modules [45, 166]. Our analysis generates a general view of the dynamical properties of a model which is a first step towards verifying the feasibility of the model's structure –by comparing to experimental observations– and facilitates hypotheses testing. The results indicate that at least two (positive) qualitative levels for growth rate (such as "high" and "intermediate" rates) are needed in order to reproduce both the stationary and exponential growth phases (see Section 7.4).

## 7.2 Methodology

In this section, we briefly recall two mathematical methods which are very useful for analysis of qualitative systems and, in particular, interconnections of Boolean models.

### 7.2.1 From discrete to Boolean models

Although Boolean variables can only take the values 0 or 1, it is nevertheless possible to construct Boolean models that describe variables with a discrete number of values [46, 171]. Consider a discrete model $\Sigma_{disc} = (\Omega_d, F_d)$, with variables $V = (V_1, \ldots, V_n)'$, state space $\Omega_d = \Pi_{i=1}^{n}\{0, 1, 2, \ldots, d_i\}$, where $d_i \in \mathbb{N}$ is the number of levels of variable $V_i$ $(i = 1, \ldots, n)$, and a state transition table $F_d : \Omega_d \to \Omega_d$. The state of the system at the next instant $k+1$ is given as a function of the state of the system at the current instant $k$, according to the rules $F_d$, using the notation:

$$V^+ = \tilde{F}_d(V).$$

Throughout this paper, the function $\tilde{F}_d$ is obtained from $F_d$ by assuming an *asynchronous dynamic updating rule*, that is, exactly one variable is updated at any given time:

$$V^+ \in \{W \in \Omega_d : \ \exists k \text{ s.t. } W_k = (F_d)_k(V) \neq V_k \text{ and } W_j = V_j, \ \forall j \neq k \ \}. \tag{7.1}$$

Furthermore, for a more realistic model, we consider that each variable $V_i$ can only switch from its current level to an immediately adjacent level [164], that is:

$$V_i^+ \in \{V_i - 1, V_i, V_i + 1\}, \quad \forall i. \tag{7.2}$$

The idea is to create an extended Boolean model $\Sigma_{bool} = (\Omega_b, F_b)$ where each discrete variable $V_i$ is represented by $d_i$ Boolean variables, for instance, $\{X_{i,1}, \ldots, X_{i,d_i}\}$, so that the state space of the model is $\Omega_b = \{0,1\}^{d_1+\cdots+d_n}$. There are several possible ways to convert the discrete into the Boolean variables, but here we chose to use the same criterion as in [46] which stipulates that

$$V_i = k \quad \Leftrightarrow \quad (X_{i,1} = \cdots = X_{i,k} = 1,\ X_{i,k+1} = \cdots = X_{i,d_i} = 0), \tag{7.3}$$

meaning that a variable $i$ is at a state $k$ if and only if all the first $k$ Boolean variables are ON. In particular, note that this criterion implies the partition of the state space of the extended Boolean model into *permissible and forbidden regions*:

$$\Omega_p = \{X \in \Omega_b :\ k < l \quad \Leftrightarrow \quad X_{i,k} \geq X_{i,l}\}, \qquad \Omega_f = \{0,1\}^{d_1+\cdots+d_n} \setminus \Omega_p.$$

Thus, to generate the Boolean transition table $F_b$ we need to guarantee that no transitions from a permissible to a forbidden state take place. The method described in [46] deals with this problem in a natural way, and guarantees that no transitions from permissible to forbidden states take place.

### 7.2.2 Dynamics of Boolean models

This section contains a brief summary of some useful objects that characterize the dynamics of a Boolean model. There are several possible ways of defining the dynamical updating rules [175] of a Boolean network $\Sigma = (\Omega, F_b)$, but here we will assume asynchronous updates, so the definitions and rules (7.1) stated for discrete systems also apply, with $d_i = 1$ for all $i$. Note that (7.2) is immediately satisfied for Boolean models.

The *asynchronous transition graph*, $G = (\Omega, E)$, of system $\Sigma$ is a directed graph whose vertices (or nodes) are the elements of $\Omega$, and the edges are given by $E$. There are thus $2^n$ nodes in $G$. Given any two elements $a, \tilde{a} \in \Omega$ the edge "$a \to \tilde{a}$" is in $E$ iff:

$$\tilde{a} \in \{w \in \Omega :\ \exists k \text{ s.t. } w_k = (F_b)_k(a) \neq a_k \text{ and } w_j = a_j,\ \forall j \neq k\ \}.$$

A *path* $a_1 \rightsquigarrow a_2$ in $G$ is a sequence of edges linking $a_1$ to $a_2$.

A *strongly connected component (SCC)* of $G$ is a maximal subset $C \subset \Omega$, that contains a path joining any pair of its elements. In general, a SCC may have both incoming and outgoing edges. An SCC with no outgoing edges is called *terminal*.

An *attractor* $\mathcal{A}$ of $G$ is a terminal strongly connected component, that is, once a trajectory enters $\mathcal{A}$ it cannot leave again. Therefore, the attractors can be said to characterize

the asymptotic behaviour of the network. An asynchronous transition graph always has at least one, but can have multiple, attractors. An attractor can be formed of a single state (we will call it a *singleton*) or of a subset of $\Omega$.

### 7.2.3 Interconnection of Boolean models

To study the interconnection of the two systems, we will use a method based on control theory concepts recently developed by one of the authors [45, 166]. This method analyses the asymptotic behaviour of the interconnection of two systems directly from the behaviour of the two subsystems, without having to construct or analyse the full interconnected system. The advantage is a much reduced computational cost, while still obtaining exact results: indeed, for large (e.g., $n \geq 15$) Boolean models, the computation of the asynchronous transition graph and its attractors is infeasible, as it involves the analysis of a $2^n \times 2^n$ matrix. The idea is to first study each individual system for each set of inputs, obtain the corresponding attractors, and then construct a new object, the *asymptotic graph*. This new graph is much smaller than the state transition graph of the full model, but it contains all the information on its asymptotic dynamics, namely all the attractors of the full model correspond to attractors in the asymptotic graph. Some notation is next introduced.

Consider two asynchronous Boolean models, $\Sigma_A$ and $\Sigma_B$, with a set of inputs ($U_i$) and a set of outputs ($H_i$):

$$\begin{aligned}\Sigma_A &= (\Omega_A, U_A, H_A, F_A): \quad \Omega_A = \{0,1\}^{n_A}, \quad U_A = \{0,1\}^{p_A}, \quad H_A = \{0,1\}^{q_A},\\ \Sigma_B &= (\Omega_B, U_B, H_B, F_B): \quad \Omega_B = \{0,1\}^{n_B}, \quad U_B = \{0,1\}^{p_B}, \quad H_B = \{0,1\}^{q_B}.\end{aligned}$$

The following notation will be used: states will be denoted $a \in \Sigma_A$ and $b \in \Sigma_B$, inputs $u \in U_A$ and $v \in U_B$, and the output corresponding to state $a$ will be denoted $h_A(a) \in H_A$ (resp., $h_B(b) \in H_B$ for state $b$). The synchronous rules are written:

$$a^+ = F_A(a;u), \qquad \text{and} \qquad b^+ = F_B(b;v).$$

For each fixed $u \in U_A$, there is a set of attractors of system $\Sigma_A$, which will be denoted $A_u^i$, $i \in \mathbb{N}$. Similarly for system $\Sigma_B$, $B_v^j$, $j \in \mathbb{N}$.

The interconnection of these two systems is formed by letting the input of each system be the output of the other

$$v = h_A(a) \in U_B \qquad u = h_B(b) \in U_A,$$

where it is assumed without loss of generality that $q_A = p_B$ and $q_B = p_A$. The new system will be denoted:

$$\Sigma = (\Omega, F_{bool}): \quad \Omega = \{0,1\}^{n_A+n_B}, \quad F_{bool} : \Omega \to \Omega$$

with the Boolean rules $F_{bool}$ given by the appropriate combination of $F_A$, $F_B$:

$$F_{bool}(a,b) = (F_A(a; h_B(b)), F_B(b; h_A(a))).$$

Note that $F_A$, $F_B$, and $F_{bool}$ contain the synchronous table of state transitions. Here, we will consider that the dynamics is asynchronous, so that only one variable is updated at a given time. The asynchronous transition graphs of the two modules (one for each fixed input) and that of the full interconnected system will be called, respectively, $G^{A,u}$, $G^{B,v}$, and $G$.

**Transition graphs and semi-attractors** The first step of the method is to compute all the transition graphs $G^{A,u}$ and $G^{B,v}$, compute their attractors, and then divide each of these into subsets corresponding to a fixed output. These will be called *semi-attractors* of the individual system and are defined as follows:

$A^i_{u\alpha}$ = the $i$-th semi-attractor of system $\Sigma_A$, corresponding to input $u$, with output $\alpha$
$B^j_{v\beta}$ = the $j$-th semi-attractor of system $\Sigma_B$, corresponding to input $v$, with output $\beta$.

Note that the standard attractor is the union of all corresponding "semi-attractors":

$$A^i_u = \cup_{\text{all } \alpha} A^i_{u\alpha}.$$

**The asymptotic graph** The second step of the method is to construct the asymptotic graph $G^{as}$ whose nodes are the cross-products of semi-attractors:

$$A^i_{u\alpha} \times B^j_{v\beta}.$$

There is an edge between two of the nodes

$$A^i_{u\alpha} \times B^j_{v\beta} \to A^i_{u\alpha} \times B^{\tilde{j}}_{\alpha\tilde{\beta}}$$

if there is a path in the graph $G^{B,\alpha}$ that leads from some state in $B^j_{v\beta}$ to some state in $B^{\tilde{j}}_{\alpha\tilde{\beta}}$. Similarly for an edge $A^i_{u\alpha} \times B^j_{v\beta} \to A^{\tilde{i}}_{\beta\tilde{\alpha}} \times B^j_{v\beta}$. In order to satisfy an asynchronous updating scheme, only one set of variables is allowed to change for each edge. The computational cost can be further reduced by observing that all nodes with $u \neq \beta$ and

$v \neq \alpha$ are transient (shown in [166]); hence, to compute the attractors of the asymptotic graph we only need to include the edges between nodes satisfying either $u = \beta$ or $v = \alpha$.

### 7.2.4 Attractors of an interconnection

The third step of the method is to compute all the attractors of $G^{as}$ which contain, in fact, a representative of each of the attractors of $G$. This is theoretically proven in [45, 166]:

**Theorem 7.1.** *[45] If $Q$ is an attractor of $G$, then there exists at least one corresponding attractor in $G^{as}$, $Q^{as} = Q^{as}(Q)$. Moreover, if $Q_1 \neq Q_2$ are two distinct attractors of $G$, then $Q^{as}(Q_1) \neq Q^{as}(Q_2)$.*

In other words, we recover *all the attractors of the interconnection, without explicitly constructing the interconnected system.* In broad terms, Theorem 7.1 says that any attractor of $G$ generates an attractor in $G^{as}$, but the converse is not necessarily true and $G^{as}$ may have more attractors than $G$.

To better illustrate Theorem 7.1, and show its advantages as well as limitations, a purely theoretical example is next given. For convenience, in the following examples, the attractors are labelled using the decimal representation for the Boolean inputs and outputs, that is:

$$000 \Leftrightarrow u = 1, \quad 001 \Leftrightarrow u = 2, \ \ldots, \ 111 \Leftrightarrow u = 8, \quad \textit{etc.}$$

**Example I.** Consider the following bi-dimensional systems $A$ and $B$, with $n_A = n_B = 2$ and $p_A = p_B = 1$:

$$\begin{aligned} a_1^+ &= u \ \texttt{and} \ (a_1 \ \texttt{and not} \ a_2), \\ a_2^+ &= [u \ \texttt{and} \ (\texttt{not} \ a_1 \ \texttt{or} \ a_2)] \ \texttt{or} \ [\texttt{not} \ u \ \texttt{and} \ a_1], \\ h_A(a) &= a_2, \\ \\ b_1^+ &= [v \ \texttt{and not} \ b_2)] \ \texttt{or} \ [\texttt{not} \ v \ \texttt{and} \ (b_1 \ \texttt{xor} \ b_2)], \\ b_2^+ &= [v \ \texttt{and} \ b_1 \ \texttt{and} \ b_2] \ \texttt{or} \ [\texttt{not} \ v \ \texttt{and} \ (b_1 \ \texttt{or} \ b_2)], \\ h_B(b) &= b_2, \end{aligned} \tag{7.4}$$

whose asynchronous transition graphs $G^{A,u}$ and $G^{B,v}$ are shown in Fig. 7.1, for convenience. Note that the attractors in all graphs are singletons except for $B^2_{2,2} = \{01, 11\}$. However, since the two states have the same output ($h_B(01) = h_B(11) = 1$), in this example the semi-attractors are in fact the actual attractors.

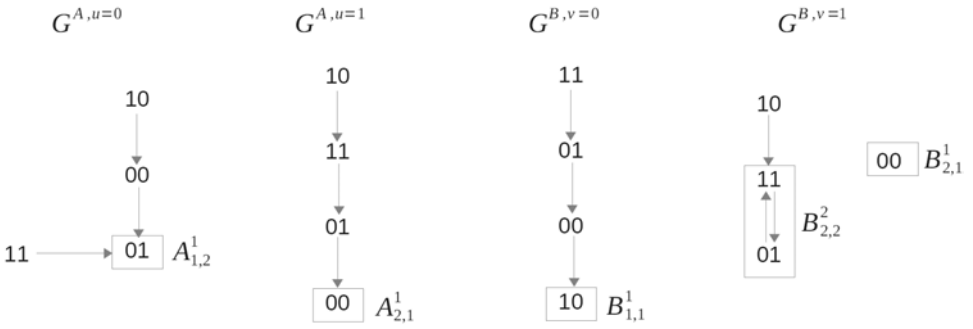



FIGURE 7.1: Example I: the asynchronous transition graphs that define the dynamics of the two systems (7.4).

The corresponding asymptotic graph is shown in Fig. 7.2. To illustrate the computation of an edge, consider the product $A^i_{u\alpha} \times B^j_{v\beta} = A^1_{21} \times B^1_{11}$: since $\alpha = 1 = v$, the system $A$ does not induce any change in the variables $b$; in contrast, the fact that $\beta = 1$ will induce a trajectory between a state in $A^1_{21}$ and an attractor in the graph $G^{A,2}$ (corresponds to Boolean input u=1). In the graph $G^{A,2}$, the state 00 is in the basin of attraction of $\{01\} = A^1_{12}$. Therefore, there is an edge $A^1_{21} \times B^1_{11} \to A^1_{12} \times B^1_{11}$. All other edges are similarly computed.

Note that the full interconnected system has four variables and hence its dynamics is given by an asynchronous transition graph $G$ with $2^4 = 16$ states. To compute the attractors of $G$ we needed to compute a transition graph with only $2 \times 3 = 6$ states (2 attractors from system $A$ and 3 from system $B$). Furthermore, as remarked above, the size of $G^{as}$ can be further reduced by excluding the cross-product states known to be transient. In this example only $A^1_{21} \times B^1_{21}$ satisfies the condition $u \neq \beta$ and $v \neq \alpha$, and can be excluded. For higher order systems, such a size reduction can represent very significant savings in computational cost.

The $G^{as}$ of Example I has two attractors: $Q_1 = \{A^1_{12} \times B^1_{21}\}$ and $Q_2 = \{A^1_{12} \times B^1_{11}, A^1_{12} \times B^2_{22}, A^1_{21} \times B^1_{11}, A^1_{21} \times B^2_{22}\}$. For this 4-dimensional example, it is easy to check that $Q_1$ is a true attractor of the full interconnected system (see also Prop. 7.2), while $Q_2$ is a "spurious" attractor, that is, not a real attractor of $G$. To see this, it suffices to note that there is a pathway that leads from a state within $Q_2$ to $Q_1$, and which is not "covered" by $G^{as}$:

$$Q_2 \ni A^1_{21} \times B^2_{22} \ni (00, 01) \xrightarrow{G^{B,1}} (00, 00) \xrightarrow{G^{A,1}} (01, 00) \in Q_1$$

This Example shows that even very simple (and deterministic) individual asynchronous dynamics can lead to asymptotic graphs that exhibit spurious attractors. However,

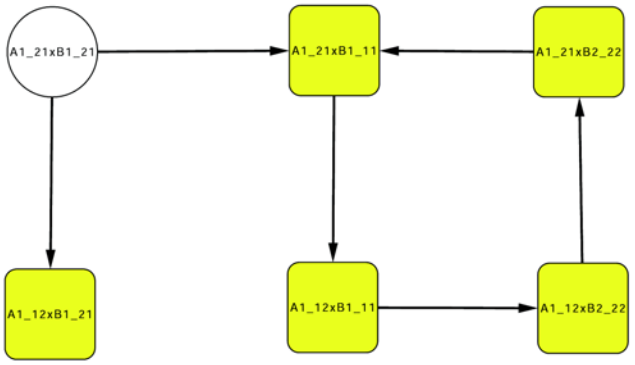



FIGURE 7.2: Example I: the asymptotic graph of the interconnection of the two systems (7.4). The cross-products inside shaded squares belong to an attractor. The cross-product inside a white circle represents a transient state that can be excluded from the computation.

note that this example was specifically contrived to illustrate the generation of spurious attractors; its Boolean rules are not necessarily biologically plausible.

In view of Example I, it would be useful to complement Theorem 7.1 by conditions permitting to decide whether an attractor of $G^{as}$ is also an attractor of $G$. An exact result was also proved in [166] –i.e., recovering exactly all the attractors in $G$ from the cross-products of semi-attractors, with no spurious generation,– by computing the *cross graph* which is similar to $G^{as}$ but involves *cross-products of all (semi-)SCCs* (as opposed to considering only semi-attractors). However, depending on the number of SCCs, the cross graph can often be more costly to compute than the full graph $G$, hence the usefulness of establishing sufficient conditions for deciding whether an attractor of $G^{as}$ is a "true" attractor.

Some preliminary results were presented in Proposition 1 of [166], which are improved below in Prop. 7.2. To state this, we need to introduce projection functions, for $V = A^i_{u\alpha} \times B^j_{v\beta}$, and $R = \{V_1, \ldots, V_r\}$:

$$\begin{aligned} \pi(V) &= \{(a,b) \in \Omega : \ a \in A^i_{u\alpha}, \ b \in B^j_{v\beta}\}, \\ \pi(R) &= \cup_{V \in R} \, \pi(V), \\ \pi^A(R) &= \{a \in \Omega_A : \ \exists b \text{ such that } \ (a,b) \in \pi(R)\}. \end{aligned}$$

The *A-output* of $R$ is the set:

$$A\text{-output } = \{h_A(a) : \ a \in \pi^A(R)\} \ \subset \ H_A$$

Similar definitions apply for the projection $\pi^B(R)$ and the *B-output* of $R$.

Recall that we are assuming $q_A = p_B$ and $q_B = p_A$, hence $H_A \equiv U_B$ and $H_B \equiv U_A$ and the $A$-output (resp., $B$-output) of $R$ is also contained in $U_B$ (resp., $U_A$). The new result of Prop. 7.2 is in parts (ii), (iii), which previously stated "for all $u \in U_A$" or "for all $v \in U_B$". The new conditions are much less restrictive, although the proof is similar. If Proposition 7.2 is not applicable, then one may still verify *a posteriori* whether $R$ represents an attractor of $G$ by simulating all trajectories starting from all states in $\pi(R)$ and checking whether any of them leaves $R$ (however, this "direct force" procedure may also involve some computational costs).

**Proposition 7.2.** *Let $R$ be a terminal SCC of $G^{as}$. If either one of the following conditions is satisfied:*

*i) $R$ is a singleton (i.e., contains a single product $V$);*

*ii) the $A$-output of $R$ is a singleton and the set $\pi^A(R)$ is an attractor of $G^{A,u}$ for all $u$ in the $B$-output of $R$;*

*iii) the $B$-output of $R$ is a singleton and the set $\pi^B(R)\}$ is an attractor of $G^{B,v}$ for all $v$ in the $A$-output of $R$;*

*then $R^{as} = \pi(R)$ is an attractor of $G$.*

*Proof:* We will use the notation $(a, b) \rightsquigarrow_G (a', b')$ to denote a path connecting the two elements in the transition graph $G$ and $(a, b) \rightarrow_G (a, b')$ to denote a one-step transition.

Part (i) is unchanged from [166]. Parts (ii) and (iii) are very similar, so we will only prove part (iii). If the $B$-output of $R$ is a singleton, say $\{\alpha\}$, then any $V \in R$ must be of the form

$$A^j_{\alpha(\cdot)} \times B^{(\cdot)}_{v\alpha}, \quad \text{for some } v \text{ in the } A\text{-output of } R.$$

In particular, (see definition of semi-SCC) all $A^j_{\alpha(\cdot)}$ belong to the same attractor $A^j_\alpha$ of $G^{A,\alpha}$.

Suppose now that the set $\pi^B(R)$ is an attractor for all $v$ in the $A$-output of $R$. Then, to show that $\pi(R)$ is an attractor of $G^{as}$, it suffices to show that: (1) $\pi(R)$ is a strongly connected set, and (2) $\pi(R)$ contains all its successors. If (1) and (2) hold, then $\pi(R)$ is indeed a terminal SCC.

To show (1), let $(a, b)$ and $(a', b')$ be any two elements of $\pi(R)$. Then

$$(a, b) \rightsquigarrow_G (a, b'), \quad \text{since } \pi^B(R) \text{ is an attractor of } G^{B,h_A(a)} (v = h_A(a) \in A\text{-output})$$
$$(a, b') \rightsquigarrow_G (a', b'), \quad \text{since } a, a' \text{ belong to the same attractor } A^j_\alpha \text{ of } G^{A,\alpha}.$$

To show (2), observe that there are two forms of successors: either $(a,b) \to_G (a',b)$ or $(a,b) \to_G (a,b')$. We want to prove that both $(a',b)$ and $(a,b')$ are in $\pi(R)$. In the first case, since $a, a'$ belong to the same attractor $A^j_\alpha$, it is immediate to see that $(a',b) \in \pi(R)$. In the second case, since $b' \in \pi^B(R)$ and $\pi^B(R)$ is an attractor of $G^{B,h_A(a)}$, by definition of $\pi^B(R)$ there some exists $a'$ such that $(a',b') \in \pi(R)$. Recall that the $B$-output is a singleton so $h_B(b') = \alpha$. This implies

$$(a,b') \rightsquigarrow_G (a',b') \rightsquigarrow_G (a,b'), \quad \text{since } a, a' \text{ belong to the same attractor } A^j_\alpha \text{ of } G^{A,\alpha}.$$

Therefore, $(a,b') \in \pi(R)$ as wanted. □

**Remark.** The generalization of points (ii) and (iii) of Proposition 7.2 to multiple $A$-outputs and $B$-outputs is not clear, due to Example I where the spurious attractor $Q_2$ satisfies $A$-output=$B$-output=$\{1,2\}$. Other examples exist where an attractor of $G^{as}$ of the same form as $Q_2$ is indeed and attractor of $G$ (see Example 2 in [45]).

If Proposition 7.2 cannot be applied, there may be other methods to decide whether an attractor of $G^{as}$ is a true attractor, such as identifying invariant sets of the system that contain the given attractor: examples of this are given below in Propositions 7.4 and 7.5.

**Example II.** To illustrate the relevance of Prop. 7.2, another theoretical example is now given. The two systems $A$ and $B$ are more conveniently represented by their asynchronous transitions graphs, one for each fixed input (Fig. 7.3). The dimensions are $n_a = 2$, $n_B = 3$, $p_A = 1$, $p_B = 2$ and their outputs are as follows:

$$h_A(a) = (a_1, a_2)', \qquad h_B(b) = b_1.$$

Note that attractor $A^1_2$ splits into two semi-attractors, $A^1_{21}$ and $A^2_{23}$, and the attractor $B^1_2$ splits into $B^1_{21}$ and $B^2_{22}$. The full interconnected system has five variables and hence its dynamics is given by an asynchronous transition graph $G$ with $2^5 = 32$ states. To compute the attractors of $G$ we needed to compute a transition graph with $4 \times 7 - 8 = 20$ states: 4 semi-attractors from system $A$ and 7 from system $B$, and 8 transient cross-products (see also Fig. 7.4).

The $G^{as}$ of Example II (Fig. 7.4) has two attractors: $Q_1 = \{A^1_{11} \times B^1_{11}\}$ and $Q_2 = \{A^1_{21} \times B^1_{32}, A^1_{21} \times B^2_{12}, A^2_{23} \times B^1_{12}, A^2_{23} \times B^1_{32}\}$. It is easy to check that $Q_1 = \{00000\}$ is an attractor of $G$, by Prop. 7.2(i). Likewise

$$Q_2 = \{10111, 10101, 10100, 00111, 00101, 00100\}$$

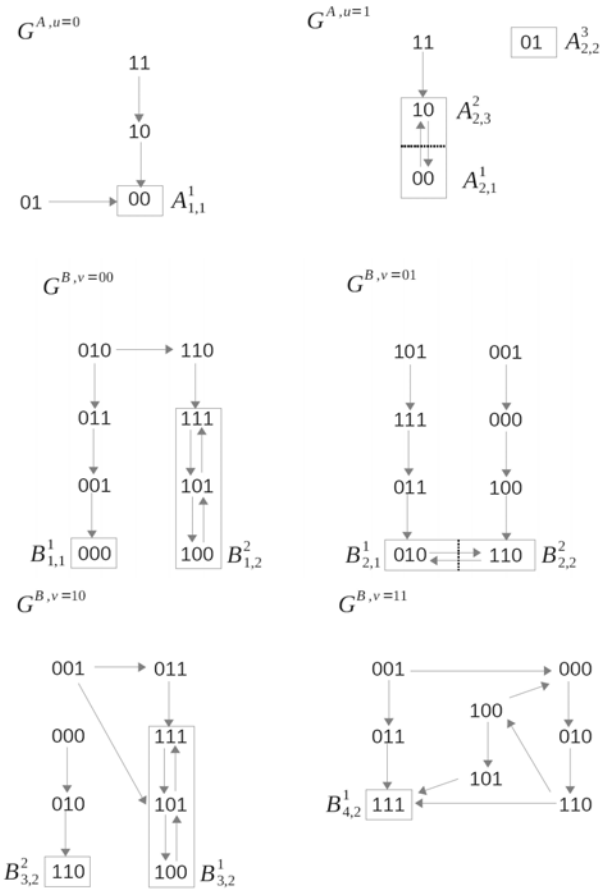



FIGURE 7.3: Example II: the asynchronous transition graphs of systems $A$ and $B$, for each fixed input.

is also an attractor of $G$, by Prop. 7.2(iii): the $B$-output is a singleton since $\{h_B(b) : b \in \pi^B(Q_2)\} = \{1\}$; the $A$-output of $Q_2$ is $\{h_A(a) : a \in \pi^A(Q_2)\} = \{10, 00\}$; and, finally, the set $\pi^B(Q_2) = \{111, 101, 100\}$ is indeed an attractor of both $G^{B,v=10}$ and $G^{B,v=00}$.

## 7.3 Application: a model for *E. Coli* growth mechanism

The bacteria *Escherichia Coli* are unicellular micro-organisms (present in the human gut, for instance) which grow and divide in the presence of a carbon source, such as glucose or other sugars. In typical experiments, in a carbon rich medium, the bacteria are observed to grow at a constant growth rate, which is referred to as the *exponential phase* [124]. In the absence of carbon, the bacteria enter a *stationary phase*, with no cellular growth or division. *E. Coli* use a network of genes and proteins to detect the presence or absence of carbon sources and respond accordingly, by adjusting their gene expression levels.

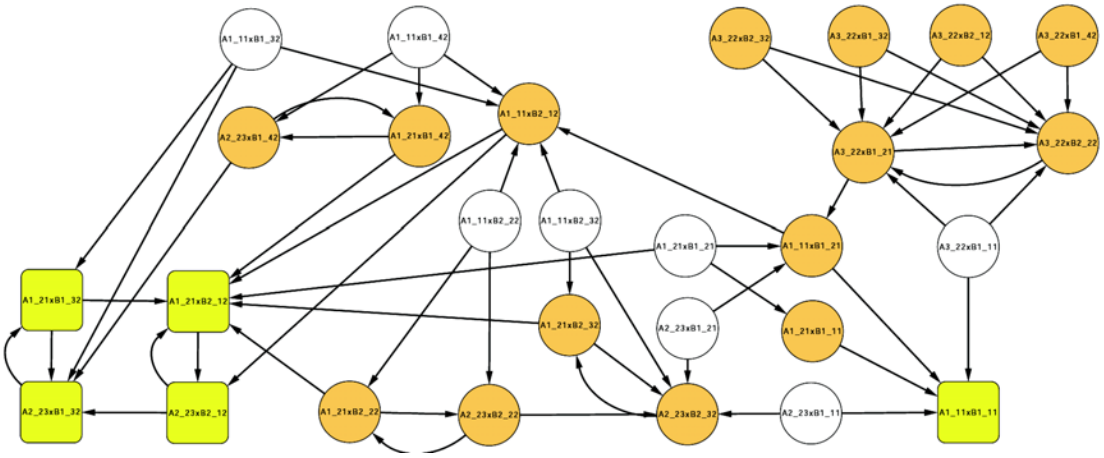



FIGURE 7.4: Example II: the asymptotic graph of the interconnection of systems $A$ and $B$ defined in Fig. 7.3. States inside light shaded squares belong to some attractor; there are two attractors in this graph. States inside white circles represent known transient state, which can be discarded from the computation. States inside light shaded circles represent all other states.

The major players in this nutritional response network are well characterized (see, for instance, [24, 92, 144] and references therein) but, in contrast, it has been difficult to find an appropriate dynamical expression for modelling the growth rate of *E. Coli* [124]. In other words, if one wishes to add a model variable to describe growth rate, what should its mathematical rule be? To overcome this problem, models often focus on either the exponential or the stationary phases, thereby considering growth rate to be either constant or zero, respectively [86]. However, such models are not able to describe the transition from one phase to the other, thus failing to provide intuition on a crucial cellular mechanism.

Growth should depend on the capacity of the bacteria to produce all the "bulk" proteins necessary to its development and cellular division. In its turn, the synthesis of "bulk" proteins depends on the transcription and translation steps, which are limited, respectively, by the concentrations of RNA polymerase and ribosomes. Some models have thus tried to include these effects to obtain a more accurate expression for growth rate. For instance, one may have a dependence on one step:

$$\text{Growth rate} \sim \text{RNA polymerase} \tag{7.5}$$

as tested previously in [37], or in two (or more) steps, each of them separately limiting growth rate, hence the use of the *minimum* function:

$$\text{Growth rate} \sim \min\{\text{ ribosomal proteins, bulk proteins }\} \tag{7.6}$$

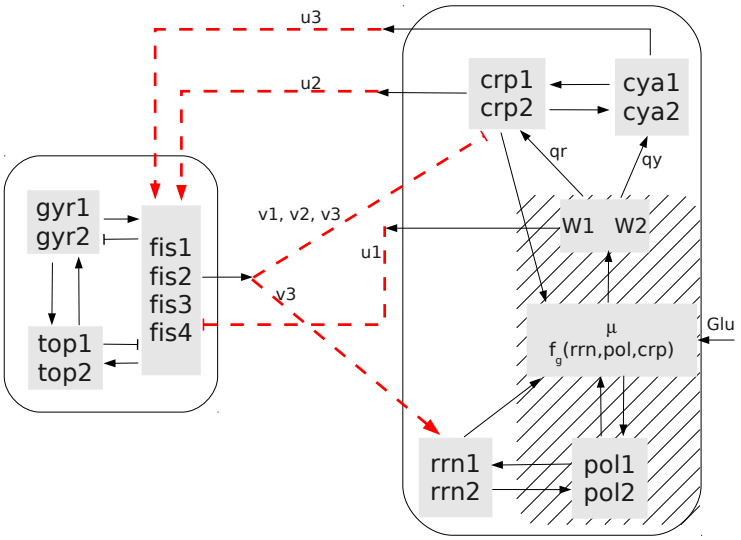



FIGURE 7.5: The interconnection of the *fis* global regulatory module (left rectangle) and a basic cellular growth model (right rectangle). Each module has three inputs and outputs: $u = (u_1, u_2, u_3)$, $v = (v_1, v_2, v_3)$. The dashed lines represent the interconnection: i.e., the output of one system becomes the input of the other. Bacterial growth rate is internally computed as a function of the external nutrient sources (Glu), ribosomes (here represented by $rrn_i$), RNA polymerase ($pol_i$) or "bulk" proteins (which will be basically represented by *crp*). Growth rate is first translated into two qualitative levels, $W_1$ and $W_2$, which signal downstream. The region under hatching represents the new variables and interactions added to the original model in [144]. Several different forms for $f_g$ will be tested (see text).

as considered in [155], or

$$\text{Growth rate } \sim \min\{ \text{ ribosomal proteins, RNA polymerase } \} \tag{7.7}$$

as we considered in [38]. In this Section, our goal is to test these expressions, by interconnecting a well known nutritional response module with a basic transcription/translation model, using the Boolean interconnection method described in Section 7.2.3.

### 7.3.1 *E. Coli* nutritional stress response module

The nutritional stress response network developed in [144] involves three groups of variables, each representing a different regulatory effect: DNA super-coiling (determined by the enzymes GyrAB and TopA), carbon response (involving the proteins Crp, Cya), and a global regulator (protein Fis) that sends the carbon availability signal down to the stable RNAs (*rrn*). The latter are limiting factors in ribosome production, and are thus a measure of the growth of the bacteria.

A very brief description of the main biological steps in response to nutritional stress is as follows (see [144] and references therein): in answer to carbon depletion, the bacteria increase their cyclic AMP concentration (cAMP); this small molecule will bind to Crp (cAMP receptor protein) to form a complex that controls the expression of different genes, some involved in the synthesis of enzymes that allow the bacteria to make use of other carbon sources, others involved in morphological changes and motility. The complex cAMP-Crp also activates the enzyme Cya (adenylate cyclase), which contributes to produce cAMP from ATP, and represses the global regulator Fis, a protein which is available at high concentration during the exponential phase, and is responsible for the control of many other genes. The protein Fis also represses the complex cAMP-Crp and, among others, it controls two enzymes involved in DNA topology regulation: Gyrase AB (GyrAB) which induces negative super-coils in the DNA and Topoisomerase A which restores super-coiling to "normal" levels. Finally, Fis also stimulates the transcription of stable RNAs, a necessary condition for the production of ribosomes and hence necessary for bacterial growth.

The model developed in [144] includes a constant external input named "Signal" that represents nutritional stress, that is, the presence ("Signal" = 0) or absence ("Signal" = 1) of carbon sources, while the variable *rrn* is simply an output, as it does not influence the other variables. Growth rate was summarized into the effect of the complex cAMP-Crp on the other variables, namely Fis, Crp, abd Cya. Depending on the value of "Signal", the network reproduced two steady states corresponding to the stationary or exponential phases of *E. coli*, characterized in Table 7.1. The two states predicted by this model are consistent with experimental observations: in the exponential phase, Fis is present at high levels, as well as stable RNAs, and the cAMP receptor protein is not strongly present. The opposite happens in the stationary phase.

In our model, the interactions are reorganized in order to include the explicit effect of growth rate. It is known that the complex cAMP-Crp is growth dependent [23], so we replaced this complex by an equivalent expression that depends on Crp, Cya, and growth, now represented by the arrows $u_1$, $q_r$ and $q_y$ in Fig. 7.5. The components inside the hatched region in Fig. 7.5 were not present in model [144], and the *rrn* variable did not influence the system. The objective in this paper is thus to refine the effect of growth in the system, as described below in Section 7.3.2.

The model [144] consists of a piecewise affine system on six variables, it was further studied in [48, 91, 167] and has been written as an extended Boolean model in [46], using the procedure briefly described in Section 7.2. The first Boolean module is formed by the 8 variables corresponding to genes *fis*, *gyr*, and *top*, since *fis* is described by 4

Boolean variables and *gyr*, *top* by 2 each (see Fig. 7.5). The rules for this module are given in the Appendix.

Since each variable may have several discrete values, the Boolean models will use $var_i$, $i \in \{1, 2, \ldots, d\}$ to denote the corresponding $d$ Boolean variables (see Section 7.2.1) (similarly for the other variables). The discrete variable can be recovered simply by adding the Boolean variables:

$$var = \sum_{i=1}^{d} var_i. \tag{7.8}$$

TABLE 7.1: The two *E. coli* modes reproduced by the model [144]. If a variable has more than one value, this means that the asymptotic solution is oscillatory among those values.

| "Signal" | *fis* | *gyr* | *top* | *crp* | *cya* | *rrn* | Phase |
|---|---|---|---|---|---|---|---|
| 0 | 1,2,3,4 | 1,2 | 0 | 1 | 2 | 0,1 | Exponential |
| 1 | 0 | 2 | 0 | 2 | 2 | 0 | Stationary |

### 7.3.2 The cellular growth module

To test the dependence of growth rate on some of the major model components, we will study a "closed-loop system": that is, use the state of the system to construct a mathematical expression for bacterial growth rate and then feed it back to the system, by letting proteins Cya and Crp depend on it. Thus, cellular growth rate (represented by $\mu$) now appears explicitly in the model, as an internal variable that depends dynamically on the state of the system at each instant (see Fig. 7.5). In agreement with the variables of the system, growth rate will have two positive discrete levels (translated to $W_1$ and $W_2$, see equation (7.9) below). This also implies that the effect of growth on *fis*, *crp* and *cya* has to be updated relative to the original model [144]. In Fig. 7.5, there are thus three links (respectively, $u_1$, $q_r$ and $q_y$) which are not fixed for now, but for which several possible combinations will be tested, with a view to better understand growth signalling (see Section 7.4). The motivation for building this closed-loop system is to test the dynamical dependence of bacterial growth rate on the system's variables, a question which is still not well understood. Thus, for this example, *an expression for growth rate will be considered valid if the refined system in Fig. 7.5 is able to reproduce the same results as the (more schematic) model [144].*

As indicated in Fig. 7.5, the second Boolean module will describe the expression of the genes encoding for *crp*, *cya*, *rrn*, and will further include *pol*, to represent the expression

of RNA polymerase, the enzyme responsible for gene transcription (2 Boolean variables each). The presence of carbon sources will be represented by the external input Glu. Previously [37], we have studied a mathematical expression for bacterial growth rate that is dependent only on RNA polymerase, for a simple 2-dimensional model. However, experimental data [155] suggests that ribosomes play a major role, hence we wish to improve our results by analysing models that consider different combinations of limiting factors, and checking their compatibility with known results.

The growth variable, $\mu$, and its downstream signals will be given by:

$$\begin{aligned} \mu &= \text{Glu} \quad \texttt{and} \quad f_g(rrn_1, rrn_2, pol_1, pol_2, crp_1, crp_2); \\ W &= 2 - \mu; \\ W_1 &= \text{sign}(W); \\ W_2 &= \max(0, W - 1); \end{aligned} \tag{7.9}$$

where $\text{sign}(W) = 1$ if $W > 0$ and $\text{sign}(W) = 0$ if $W = 0$ (by construction, $\text{sign}(W)$ is never negative). The variables $W_1$ and $W_2$ correspond, respectively, to:

$$W_1 = 1 \Leftrightarrow \mu \leq 1, \quad W_2 = 1 \Leftrightarrow \mu = 0.$$

and satisfy $W_1 \geq W_2$. Following (7.5)-(7.7) and (7.8), different expressions for the function $f_g$ will be tested, namely:

$$\begin{aligned} f_g^r &= rrn_1 + rrn_2, \\ f_g^p &= pol_1 + pol_2, \\ f_g^b &= crp_1 + crp_2, \\ f_g^{rp} &= \min(rrn_1 + rrn_2, pol_1 + pol_2) \\ f_g^{rb} &= \min(rrn_1 + rrn_2, crp_1 + crp_2), \end{aligned} \tag{7.10}$$

where the protein Crp is used as a surrogate for the level of expression of "bulk" proteins. In addition, to describe how the growth rate affects the genetic machinery, two functions need to be chosen: these correspond to the arrows labelled $q_y$ and $q_r$ (see below), which will also be a function of $W_1$ and $W_2$. Several possible combinations will be tested and the final results compared to the original model.

### 7.3.3 System interconnection

The full discrete system will thus have 7 variables,

$$V = (fis, gyr, top, crp, cya, rrn, pol)',$$

with discrete levels $d_1 = 4$, $d_j = 2$ for $j = 2, \ldots, 7$ and state space:

$$\Omega_d = \{0, 1, \ldots, 4\} \times \{0, 1, 2\} \times \ldots \times \{0, 1, 2\}.$$

The extended Boolean model will have 16 variables. As described in Section 7.2.3, the interconnection of two input/output asynchronous Boolean networks such as systems (7.13) and (7.14), is obtained by setting $u = h_B(b)$ and $v = h_A(a)$. Most of the input/output functions are already fixed by model [144]. There is a new interaction between the two modules, due to the effect of the growth rate in *fis*, which is represented by $u_1$ in Fig. 7.5:

$$\begin{aligned} u_1 &\in \{W_1, W_2\}, & v_1 &= fis_1, \\ u_2 &= crp_1 \ \texttt{or} \ crp_2, & v_2 &= fis_2 \ \texttt{or} \ fis_4, \\ u_3 &= cya_1 \ \texttt{or} \ cya_2, & v_3 &= fis_3. \end{aligned}$$

The goal in this paper is the discrimination between different variants of the model in Fig. 7.5, in order to choose the mechanism that better represents bacterial response. The variants cover:

- models for growth rate: $f_g^r$, $f_g^b$, $f_g^{rb}$, and $f_g^{rp}$;
- interactions between growth signals and the genetic machinery response: $q_r$, $q_y$, and $u_1$.

As remarked above (Section 7.3.1), the interactions $q_r$, $q_y$, and $u_1$ in some sense replace the effect of the complex cAMP-Crp on the system, by including an explicit dependence on growth rate. To evaluate the new rules we will consider that there are two signalling stages, corresponding to the response of Cya/cAMP (the initial steps in the case of nutritional stress) and of Fis (global regulator). The response of *crp* will be timed with one or the other:

$$q_r = u_1 \quad \text{or} \quad q_r = q_y.$$

The following distinct combinations for $q_y$, $q_r$ and $u_1$ will be tested:

$$\begin{aligned}
(I) \quad & q_y = W_1, \quad q_r = W_1, \quad u_1 = W_1 \\
(II) \quad & q_y = W_1, \quad q_r = W_1, \quad u_1 = W_2 \\
(III) \quad & q_y = W_2, \quad q_r = W_2, \quad u_1 = W_1 \\
(IV) \quad & q_y = W_1, \quad q_r = W_2, \quad u_1 = W_2 \\
(V) \quad & q_y = W_2, \quad q_r = W_2, \quad u_1 = W_2 \\
(VI) \quad & q_y = W_2, \quad q_r = W_1, \quad u_1 = W_1
\end{aligned} \tag{7.11}$$

## 7.4 Results

As discussed above (cf. Section 7.3), the goal is to recover the behaviour of the system as described in Ropers et al [144] (Table 7.1) but now with growth rate "actually computed" by the bacteria, for the system in closed loop form which uses the state of the system. Various combinations of interactions and growth rate functions were tested, with the results summarized in Table 7.2 and discussed below. As an indication of the

TABLE 7.2: The attractors for each combination of growth rate function and interactions $u_1$, $q_r$, $q_y$. Attractors $\sigma_i$ satisfy $rrn = pol = 0$, while attractors $\alpha_j$, $j \in \{2, 4, 24, 48, 52, 72\}$, satisfy $rrn \geq 1$ and $pol \geq 1$ (full characterizations are given in Sections 7.4.2 and 7.4.3). The indexes $i$, $j$ denote the number of distinct states contained in the attractor. All the attractors of $G^{as}$ are also attractors of $G$: either they satisfy Prop. 7.2 and/or other methods, as indicated. The highlighted row $(***)$ represents the model variants which better reproduce Table 7.1 results (see Section 7.4.4).

| Growth rate function | Interactions (7.11) | Attractors, $G^{as}$ (Glu=1) Stationary | Attractors, $G^{as}$ (Glu=1) Exponential | Attractors, $G^{as}$ (Glu=0) Stationary |
|---|---|---|---|---|
| $f_g^r, f_g^p, f_g^{rp}, f_g^{rb}$ | I | $\sigma_4$ [Prop. 7.2(i)] | $\alpha_4$ [Prop. 7.2(i)] | $\sigma_4$ [Prop. 7.3] |
| | II | $\sigma_4$ [Prop. 7.2(i)] | $\alpha_{72}$ [Prop. 7.4] | $\sigma_4$ [Prop. 7.3] |
| | III,VI | $\sigma_4$ [Prop. 7.2(i)] | $\alpha_2$ [Prop. 7.2(i)] | $\sigma_4$ [Prop. 7.3] |
| $***$ | IV,V | $\sigma_4$ [Prop. 7.2(i)] | $\alpha_{24}$ [Prop. 7.2(iii)] | $\sigma_4$ [Prop. 7.3] |
| $f_g^b$ | I,III,VI | $\sigma_{52}$ [Prop. 7.5(i)] | $\alpha_{52}$ [Prop. 7.5(ii)] | $\sigma_4$ [Prop. 7.3] |
| | II | $\sigma_{48}$ [Prop. 7.5(i)] | $\alpha_{48}$ [Prop. 7.5(ii)] | $\sigma_4$ [Prop. 7.3] |
| | IV,V | $\sigma_{24}$ [Prop. 7.2(iii)] | $\alpha_{24}$ [Prop. 7.2(iii)] | $\sigma_4$ [Prop. 7.3] |

computational costs, application of the method presented in Section 7.2.3 to compute the attractors for model $f_g^{rp}$, case IV, gave the following results:

- there are eight constant-input asynchronous transition graphs for each system ($G^{A,u}$, or $G^{B,v}$);
- on these graphs there are a total of 22 semi-attractors for system $\Sigma_A$ and 20 for $\Sigma_B$;
- the total number of vertices in the asymptotic graph is thus $20 \times 22 = 440$;

- as remarked in Section 7.2.3 (and [166]), the number of vertices in $G^{as}$ can be further reduced by eliminating those which are known to have no incoming arrow. This leads to only 90 vertices;
- the computational cost of finding the attractors of the interconnected system $\Sigma$ has therefore been reduced from analysis of a size $2^{16} = 65536$ to a size 90 matrix;
- one should nevertheless consider the cost of computing this size 90 matrix, which involves reachability calculations in the $2 \times 8$ individual asynchronous transition graphs (the full process was very fast here, taking between 30-60 seconds for each model variant).

### 7.4.1 General properties

Some immediate observations from the results are:

- a common point to all model variants is that, in the presence of nutrient (Glu=1), $G^{as}$ always has two attractors which are both attractors of $G$, by application of Prop. 7.2(i) or (iii), or other methods (see Prop. 7.4, 7.5).
- for all model variants, the first attractor ($\sigma_i$, $i \in \{4, 24, 48, 52\}$) has $rrn = pol = 0$ and the second attractor ($\alpha_j$, $j \in \{2, 4, 24, 48, 52, 72\}$) $rrn = pol = 1$. The first may be said to represent stationary phase, while the second stands for exponential phase.
- also for all model variants, in the absence of nutrient (Glu=0), there is only one attractor, $\sigma_4$; this can be verified directly (see Prop. 7.3 below). The *stationary phase* attractor $\sigma_4$ has four states and is characterized by:

  $$\sigma_4 : \quad fis = 0,\ gyr \in \{1, 2\},\ top = 0,\ crp = 2,\ cya \in \{1, 2\},\ rrn = 0,\ pol = 0 \quad (7.12)$$

  coinciding with the stationary attractor of model [144] (see Table 7.1) with the exception of *gyr* and *cya* which oscillate between 1 and 2 (instead of being fixed at 2).
- all models involving the ribosomes or RNA polymerase as a growth rate limiting factor exhibit the same stationary phase and a similar exponential phase attractors, depending only on the choice of feedback interactions.

As an example, for model variant $f_g^{rp}$, case IV, the basins of attraction for $\sigma_4$ and $\alpha_{24}$ are disconnected. The stationary phase attractor is formed of a single vertex, while the

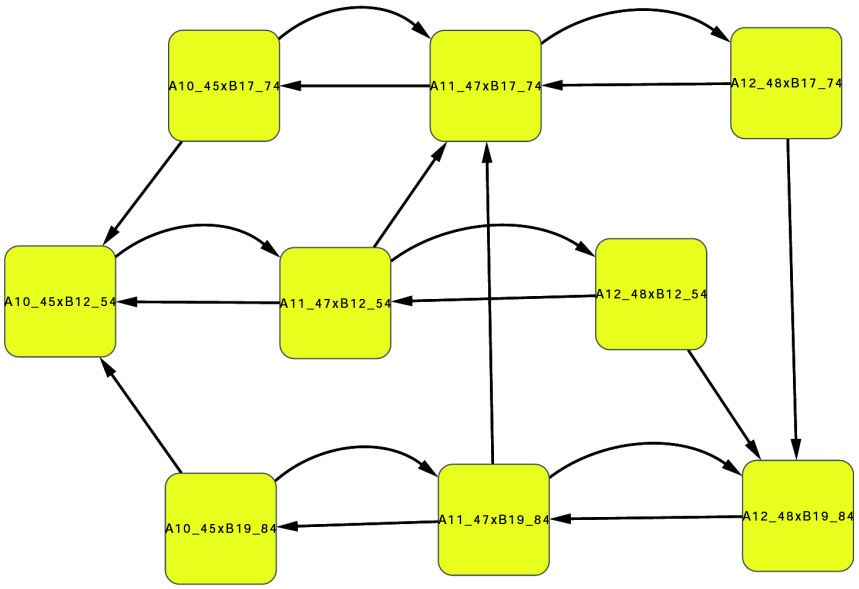



FIGURE 7.6: The exponential growth phase attractor (model $f_g^{rp}$, case IV). Since the $B^j_{\alpha 4}$ components are all equal, this attractor can be reduced to (either) one of the horizontal rows, with three components only (see text for more details).

exponential phase attractor is composed of 9 vertices, as shown in Fig. 7.6. However, all system $B$ semi-attractors coincide (Prop. 7.2(iii) is satisfied):

$$B^{12}_{54} = B^{17}_{74} = B^{19}_{84} = \{10111010\},$$

while the $A$ system semi-attractors are characterized by the levels of *fis*, with either $fis = 1$,$fis = 2$ or $fis \geq 3$:

$$\begin{aligned} A^{10}_{45} &= \{10000000, 10000010, 10001000, 10001010, 10001100, 10001110\}, \\ A^{11}_{47} &= \{11000000, 11000010, 11001100, 11001010, 10001100, 11001110\}, \\ A^{12}_{48} &= \{11100000, 11100010, 11101100, 11101010, 11101100, 11101110, \\ &\quad 11110000, 11110010, 11111100, 11111010, 11111100, 11111110\}. \end{aligned}$$

In practice, the attractor in Fig. 7.6 can be reduced to (either) one of the horizontal rows, with three components only. All concentrations are fixed, except for *fis*, *gyr*, and *top* which are allowed to oscillate in any given increasing or decreasing order, provided that $fis \geq 1$ and $top \leq 1$.

In the case where no carbon sources are present, it can be shown that all model variants become the same, and hence exhibit the same stationary phase attractor. This is essentially due to the direct effect of growth rate on the synthesis of RNA polymerase.

**Proposition 7.3.** *Assume that Glu=0. Then, the asymptotic graph for all model variants exhibits only one attractor, $\sigma_4$.*

*Proof*: In the case Glu=0, we immediately have the steady state values for *rrn* and *pol*:

$$\mu = 0 \;\Rightarrow\; pol_1 = pol_2 = 0 \;\Rightarrow\; rrn_1 = rrn_2 = 0$$

For the interactions $W$, $q_y$, $q_r$, and $u_1$ it also follows that:

$$W_1 = W_2 = 1 \;\Rightarrow\; q_y = q_r = u_1 = 1.$$

Together with the rules in Appendix, this leads to:

$$cya_1 = crp_1 = 1 \;\Rightarrow\; u_2 = u_3 = 1 \;\Rightarrow\; fis_i = hf_i$$

which simplifies to

$$fis_1^+ = fis_2, \quad fis_2^+ = fis_3, \quad fis_3^+ = fis_4, \quad fis_4^+ = fis_3 \texttt{ and not } fis_4.$$

Thus, at steady state, the values for *fis* satisfy $fis_i = 0$, for all $i$, which in turn imply that all the outputs of system $A$ are zero: $v_i = 0$ for all $i$. The remaining concentrations can now be easily established from the Boolean rules, so it follows that there is only one attractor and that it is $\sigma_4$ (7.12). □

### 7.4.2 Growth Rate limited by ribosomes or RNA polymerase

For the model variants using $f_g^r$, $f_g^p$,$f_g^{rp}$, or $f_g^{rb}$, the *stationary phase* attractor $\sigma_4$ is always the same (as described in Section 7.4.1). The *exponential phase* attractor, $\alpha_j$, $j \in \{2, 4, 24, 72\}$, depends on the wiring and has $j$ states characterized by :

$$\alpha_2 : \; fis = 0, \; gyr \in \{1,2\}, \; top = 0, \; crp = 2, \; cya = 2, \; rrn = pol = 1,$$

$$\alpha_4 : \; fis = 0, \; gyr \in \{1,2\}, \; top = 0, \; crp = 2, \; cya \in \{1,2\}, \; rrn = pol = 1,$$

$$\alpha_{24} : \; fis \in \{1,2,3,4\} \; gyr \in \{1,2\}, \; top = 0, \; crp = 1, \; cya = 2, \; rrn = pol = 1,$$

$$\alpha_{72} : \; fis \in \{1,2,3,4\}, \; gyr \in \{0,1,2\}, \; top = 0, \; crp \in \{1,2\}, \; cya \in \{1,2\}, \; rrn = pol = 1.$$

Note that cases IV and V ($\alpha_{24}$) are similar to the exponential phase attractor of [144] (see Table 7.1) (the only difference is *rrn* now fixed at 1, which seems reasonable for

the exponential phase). Cases I,III,VI ($\alpha_2$,$\alpha_4$) fail to reproduce the levels of *fis* during exponential phase (here they are fixed at zero). Case II ($\alpha_{72}$) also exhibits oscilations in *crp* and *cya*, which are not observed in Table 7.1. This attractor does not fit into Proposition 7.2, but an alternative way to show that it is not a spurious attractor, is to note that the set of states with ribosomes, RNA polymerase and Fis all at discrete level 1 is invariant, so trajectories cannot leave this set; therefore, an attractor with such properties must exist, with the only possible candidate being $\alpha_{72}$.

**Proposition 7.4.** *Assume that* $f_g \in \{f_g^r, f_g^p, f_g^{rp}, f_g^{rb}\}$. *The set* $Q = \{x \in \Omega_d : \ rrn = 1,\ pol = 1,\ fis \geq 1\}$ *is invariant.*

*Proof*: From the Boolean rules (see Appendix), it suffices to note that:

$$rrn = pol = 1 \ \Rightarrow \ \mu = 1 \ \Rightarrow \ W_1 = 1,\ W_2 = 0 \ \Rightarrow \ q_y = q_r = 1,\ u_1 = 0$$

and also

$$rrn = pol = 1, \quad \mu \geq 1 \ \Rightarrow \ rrn = pol = 1.$$

And then:

$$u_1 = 0 \ \Rightarrow \ fis_1 = h01 \equiv 1 \ \Rightarrow \ fis \geq 1.$$

Therefore, the set $Q$ is invariant. □

### 7.4.3 Growth Rate limited by bulk proteins

For the model variants using $f_g^b$, the exponential phase attractors are characterized as follows:

$$\begin{aligned}
\alpha_{24}: &\ fis \in \{1,2,3,4\}\ \ gyr \in \{0,1,2\},\ top = \{0,1\},\ crp = 1,\ cya = 2,\ rrn = pol = 1,\\
\alpha_{48}: &\ fis \in \{1,2,3,4\}\ \ gyr \in \{0,1,2\},\ top = \{0,1\},\ crp \in \{1,2\},\ cya = 2,\ rrn = pol = 1,\\
\alpha_{52}: &\ fis \in \{0,1,2,3,4\},\ gyr \in \{0,1,2\},\ top = \{0,1\},\ crp \in \{1,2\},\ cya = 2,\ rrn = pol = 1.
\end{aligned}$$

The stationary phase attractors are similar in all variables except that $rrn = pol = 0$. Comparison with Table 7.1 shows many differences with respect to model [144]. These attractors are also true attractors of $G$, as shown by application of the following result.

**Proposition 7.5.** *Define the sets* $P_0$ *and* $P_1$*:*

$$P_0 = \{x \in \Omega_d : \ rrn = 0,\ pol = 0\}, \qquad P_1 = \{x : \ rrn = 1,\ pol = 1\}.$$

*Then:*

*(i) The set $P_0$ is invariant independently of the function $f_g$;*

*(ii) The set $P_1$ is invariant if $f_g = f_g^b$.*

*Proof*: Invariance of $P_0$ follows directly from the Boolean rules for $rrn_i$ and $pol_i$. For $P_1$, it suffices to note that the Boolean rules imply (see Appendix): $cya \geq 1$ and $crp \geq 1$ which imply $\mu \geq 1$. Then, $rrn = pol = 1$ and $\mu \geq 1$ imply $rrn = pol = 1$. □

### 7.4.4 Model discrimination

Based on the observations above and comparison of Tables 7.1 and 7.2, it seems clear that the growth rate function should depend on the ribosomes and/or RNA polymerase. With this model, at steady state there is no difference between a dependence on ribosomes or RNA polymerase, although the transient dynamics do depend differently on these two species (simulations in Section 7.4.5). (This may be due to a very simplified model for the transcription/translation steps which is, however, not our aim to study here.) The model variants corresponding to I,III,VI do not satisfy the properties of the exponential phase attractor and can thus be eliminated. The interconnection of type II has most of the correct properties, but it allows the concentration of *crp* and *cya* to oscillate, in contrast to the original model. The cases that better fit the original model are IV and V, whose asymptotic behaviour is indistinguishable. This is consistent with the observations (summarized in Ropers et al. [144]) that: immediately upon carbon starvation, or absence of carbon source, transcription of the gene *cya* is activated, which leads to production of the protein Cya. Phosphorylation of Cya leads to synthesis of cyclic AMP, which in turn will bind to Crp and produce a complex [cAMP-Crp]. This complex will then control a variety of genes which are directly involved in the adaptive response of *E. coli* to a deprivation of carbon. Among others, it activates *crp*, and inactivates *cya* and the global regulator *fis*.

Furthermore, to establish the transition to exponential phase, and guarantee the presence of global regulator *fis*, the wiring interactions should be as in cases II, IV, and V, which all satisfy $u_1 = W_2$: in other words, since $W_2 = 1$ corresponds to $\mu = 0$, *fis* is inhibited only at low growth rate, as also observed in [144]. The levels of *crp* and *cya* are in agreement with those of Table 7.1 if *crp* is activated at high growth rate level and the inhibition effect on *cya* is not later than the activation of *crp*, i.e., $q_r = W_2$, $q_y \geq q_r$ (recall that $W_1 \geq W_2$).

In conclusion, to develop a more detailed continuous model, bacterial growth rate should depend on the ribosomes. For simplicity, one may even consider the ribosomes to be the

only variable influencing growth rate (besides the external input), because no differences were observed between models with $f_g^r$, $f_g^{rp}$, or $f_g^{rb}$.

### 7.4.5 Dynamical behaviour

By virtue of Theorem 7.1 and Propositions 7.2 to 7.5, we know that any trajectory of the interconnected model will eventually reach either the exponential or stationary phase attractors, depending on the initial condition and path in the graph $G$. To illustrate possible dynamical behaviours, we can generate trajectories in $G$, by randomly choosing the variable to be updated at the next instant, according to the rules. Note that this simulation does not need the graph $G$ to be constructed; but, on the other hand, such simulations cannot characterize the full behaviour of the network. Hence the usefulness of the asymptotic graph $G^{as}$, which can now be completed with some statistical results on initial conditions and attractors reached.

For the statistical analysis we choose interconnection model IV, and growth rate models $f_g^{rp}$, $f_g^r$, and $f_g^p$. It must be noted that the asymptotic graph "looses" some trajectories of the full interconnected system as, to construct $G^{as}$, the system is assumed to evolve in one of the constant-input/constant-output graphs $\{a\} \times G^{B,\alpha}$ or $G^{A,\beta} \times \{b\}$ until reaching an attractor. In simulations, however, the system is allowed to switch before reaching an attractor, meaning that the basins of attraction are not really disconnected as might be suggested by the asymptotic graph.

Monte Carlo simulations of the full model ($10^4$ randomly generated trajectories) assume that all transitions in $G$ are equally probable and show that the two attractors are reached with similar frequencies: for $f_g^{rp}$, a fraction of 0.59 (0.57 for $f_g^r$, or 0.58 for $f_g^p$) trajectories converge to the exponential phase attractor.

Note that the invariance results in Propositions 7.4 and 7.5 already provide an idea of the basins of attractions, since they imply notably that initial conditions of the form $rrn = pol = 0$ (resp., $rrn = pol = 1$) lead immediately to the stationary (resp., exponential) phase attractor. To obtain more information on the distribution of the basins of attraction, we have further analysed the probability that the system converges to either attractor given an initial condition with variable $var_i = \ell$ (where $var_i$ runs over the sixteen Boolean variables of the system and $\ell \in \{0, 1\}$). We found that the convergence to either attractor depends essentially on the initial concentrations of RNA polymerase and ribosomes, while all other concentrations play minor roles (in agreement with Propositions 7.4 and 7.5). An interesting observation is that, for all variables except the polymerase, and for any initial condition, the probability of converging to exponential phase is higher than to stationary phase. It is also evident that the absence

of polymerase immediately prevents convergence to exponential phase. In addition, we observe that all trajectories converging to exponential phase need to start with an intermediate (or higher) level of RNA polymerase ($pol \geq 1$). Table 7.3 summarizes the statistics obtained from the Monte Carlo simulations. Our studies lead to the conclusion

TABLE 7.3: Initial conditions and attractor reached, for some model variants, with interconnection of type IV.

| Initial conditions | Attractor reached $f_g^{rp}$, $f_g^r$, $f_g^p$ |
|---|---|
| $pol = 0$ | stationary |
| $pol \geq 1$ and $rrn = 0$ | either |
| $pol \geq 1$ and $rrn \geq 1$ | exponential |

that RNA polymerase and ribosomes are both crucial for bacterial growth, but exert their roles at different times: initially, the presence of RNA polymerase is necessary to grow and reach the exponential phase (otherwise, if RNA polymerase is absent at time zero, the bacteria enter the stationary phase even in the presence of carbon sources), while ribosomes can be absent; at later times, the presence of ribosomes is essential to guarantee the entry into exponential phase.

## 7.5 Conclusions

Several dynamic model variants for bacterial growth rate that consider limitation by availability of the proteins needed for cell division (RNA polymerase for transcription, ribosomes for translation, or other "bulk" proteins) were tested and compared to a well established model. The main goal was to analyse (qualitative) feasibility of the wiring network, as well as the logical coherence of each model variant. This was accomplished by using a Boolean version of the model for nutritional stress response in [144], coupled with a basic cellular growth module.

We can conclude that Boolean models provide a useful framework for analysis of a system's dynamical behaviour, convenient for hypotheses testing and model discrimination. This framework presents several advantages from a computational point of view, as many tools and algorithms are available for the study and rigorous analysis of the networks. In particular, using the interconnection of two Boolean modules, it is possible to compute the attractors of a large network at a much lower cost than with classical graph theoretical tools. However, the drawbacks of this methodology include problems related to identifying the two (or more) Boolean modules as well as the corresponding inputs and outputs, which are not always obvious (see also [166]). As the number of

modules and inputs increases, also the computational cost will increase and a balance must be found. This is a topic that should be further developed in future work.

A number of interesting points arise from our qualitative analysis. First, it was clear that limitation of growth rate by the ribosomes is needed in order to correctly reproduce the asymptotic modes, as well as transient dynamics, of the original model [144]. Second, in the presence of nutrient, our closed-loop model –where bacteria internally compute their growth rate, rather than responding to an already fixed signal– has the capacity for bistability (i.e., two asymptotic modes, representing exponential and stationary phases). Thus the closed-loop model also recovers the correct response to initial conditions: if both ribosomes and RNA polymerase concentration is very low, then the bacteria cannot grow even in the presence of nutrient. In the absence of nutrient, only the stationary phase attractor remains, as should be expected. Finally, by comparison to [144], we were able to discard most of the model variants and retain several properties necessary to reproduce the original model's attractors.

Since our main goal was essentially theoretical, we have not fully explored the directions for model improvement suggested by our analysis. For instance, a more detailed module for transcription/translation including other components besides ribosomes and RNA polymerase, or the modelling of the "bulk" proteins in a more precise way. To conclude, although discrete models are, of course, not appropriate for a detailed description of a system or to answer more specific questions, this analysis constitutes a very useful preliminary study of growth rate models. It provides many indications and clues for future work on constructing a more detailed, continuous model of the system.

## Acknowledgements

We are especially grateful to Laurent Tournier for many discussions and for providing part of the Matlab codes used here to analyse the asynchronous transitions graphs (specifically, the decomposition into strongly connected components and subsequent hierarchical organization). We also thank Jean-Luc Gouzé and our reviewers for many useful suggestions that helped improve the paper.

This work was supported in part by projects GeMCo (ANR 2010 BLAN0201-01) and ColAge (Inria-INSERM large scale initiative action)./

# Appendix

## Boolean rules of the two *E. coli* modules

The Boolean model for the Fis module is defined by a set of rules which use some auxiliary expressions of the form $h-$ given below:

$$
\begin{aligned}
fis_1^+ &= (\texttt{not } u_1 \texttt{ and } h01) \texttt{ or } (u_1 \texttt{ and } h11);\\
fis_2^+ &= (\texttt{not } u_1 \texttt{ and } h02) \texttt{ or } (u_1 \texttt{ and } h12);\\
fis_3^+ &= (\texttt{not } u_1 \texttt{ and } h03) \texttt{ or } (u_1 \texttt{ and } h13);\\
fis_4^+ &= (\texttt{not } u_1 \texttt{ and } h04) \texttt{ or } (u_1 \texttt{ and } h14);\\
gyr_1^+ &= (\texttt{not } fis_3 \texttt{ and not } fis_4) \texttt{ or } (gyr_2 \texttt{ and } hf_3);\\
gyr_2^+ &= \texttt{not } fis_3 \texttt{ and not } fis_4 \texttt{ and } gyr_1 \texttt{ and } (\texttt{not } gyr_2 \texttt{ or } top_1 \texttt{ or } top_2);\\
top_1^+ &= (\texttt{not } fis_3 \texttt{ and not } fis_4 \texttt{ and } top_2) \texttt{ or}\\
&\quad (hf_3 \texttt{ and } ((\texttt{not } gyr_2 \texttt{ and } top_2) \texttt{ or } (gyr_2 \texttt{ and } (\texttt{not } top_1 \texttt{ or } top_2))));\\
top_2^+ &= 0.
\end{aligned}
\tag{7.13}
$$

with the auxiliary expressions:

$$
\begin{aligned}
hf_2 &= fis_1 \texttt{ and } fis_2;\\
hf_3 &= fis_1 \texttt{ and } fis_2 \texttt{ and } fis_3;\\
hf_4 &= fis_1 \texttt{ and } fis_2 \texttt{ and } fis_3 \texttt{ and } fis_4;\\
hf_{4n} &= fis_1 \texttt{ and } fis_2 \texttt{ and } fis_3 \texttt{ and not } fis_4;\\
h01 &= 1;\\
h02 &= (fis_1 \texttt{ and } gyr_1 \texttt{ and not } top_2) \texttt{ or } hf_3;\\
h03 &= (hf_2 \texttt{ and } gyr_1 \texttt{ and not } top_2) \texttt{ or } hf_4;\\
h04 &= hf_{4n} \texttt{ and } gyr_1 \texttt{ and not } top_2;\\
h11 &= ((u_2 \texttt{ or } u_3) \texttt{ and } hf_2) \texttt{ or } ((\texttt{not } u_2 \texttt{ or not } u_3) \texttt{ and } h01);\\
h12 &= ((u_2 \texttt{ or } u_3) \texttt{ and } hf_3) \texttt{ or } ((\texttt{not } u_2 \texttt{ or not } u_3) \texttt{ and } h02);\\
h13 &= ((u_2 \texttt{ or } u_3) \texttt{ and } hf_4) \texttt{ or } ((\texttt{not } u_2 \texttt{ or not } u_3) \texttt{ and } h03);\\
h14 &= (\texttt{not } u_2 \texttt{ or not } u_3) \texttt{ and } h04;
\end{aligned}
$$

The rules for the cellular growth module can be written as follows:

$$
\begin{aligned}
crp_1^+ &= 1; \\
crp_2^+ &= (\texttt{not } q_r \texttt{ and } crp_1 \texttt{ and not } v_1) \texttt{ or } (q_r \texttt{ and } crp_1 \texttt{ and not } (v_2 \texttt{ or } v_3)); \\
cya_1^+ &= 1; \\
cya_2^+ &= (\texttt{not } q_y \texttt{ and } cya_1) \texttt{ or } (q_y \texttt{ and } (hy1 \texttt{ or } hy2)); \\
rrn_1^+ &= pol_1 \texttt{ or } rrn_2; \\
rrn_2^+ &= pol_2 \texttt{ and } rrn_1 \texttt{ and } v_3; \\
pol_1^+ &= (\text{sign}(\mu) \texttt{ and } rrn_1 \texttt{ and } pol_1) \texttt{ or } pol_2; \\
pol_2^+ &= \text{sign}(\mu) \texttt{ and } rrn_2 \texttt{ and } pol_2;
\end{aligned}
\tag{7.14}
$$

where the auxiliary expressions are

$$
\begin{aligned}
hy1 &= cya_1 \texttt{ and } (\texttt{not } crp_1 \texttt{ or not } crp_2); \\
hy2 &= cya_1 \texttt{ and not } cya_2 \texttt{ and } crp_1 \texttt{ and } crp_2;
\end{aligned}
$$

# Chapter 8

# A coarse-grained dynamical model of *E. coli* gene expression machinery at varying growth rates

This chapter is written in the form of a technical report (deliverable) for the ANR GeMCo project [41].

The mathematical model of *E. coli* gene expression machinery presented here can be seen as a reduced version of a higher dimensional and more detailed kinetic model (unpublished work) developed by our collaborators in Grenoble (D. Ropers and E. Grac, Inria Ibis team). The aim of developing a reduced model was mainly that of facilitating the identification of the system's parameters. In fact, a low-dimensional, but still quantitative model, for which the parameters are known, is a crucial starting point to design and study—*in silico*—possible growth rate control strategies. Along these lines, this reduced, but quantitative model, could be easily extended considering the network motifs and dynamical growth rate expressions of Chapters 4-5-6, which describe, qualitatively, possible synthetic gene modifications to control the growth rate of the cells. In this way, one could use qualitative and quantitative modelling formalisms together, in an iterative process, in order to develop more accurate reduced models and test novel control laws.

## 8.1 Introduction

In bacteria the rate of cell proliferation (growth rate) and the gene expression machinery (GEM) are tightly coupled. In fact, bacterial gene expression not only depends on specific regulation of particular molecules, but also on cell growth rate because important components of GEM (ribosomes, RNA Polymerase, etc) are all growth-rate dependent [114]. Moreover, early works on bacterial physiology [31, 124, 151, 173] have shown that many parameters of the cells, such as their macro-molecular composition and cell size, depend only on the growth rate and not on the nutrient used to achieve that growth rate.

In this study we focus on the GEM of *E. coli* with the intent of developing a minimal, but quantitative, dynamical model of the bacterial transcription-translation apparatus dependent on growth rate. We sought to formulate a model of *E. coli* that could predict macro-molecular cell composition (concentrations of RNAP, ribosomes, proteins etc) from growth rate, DNA related parameters (concentrations of genes of interest) and known key physical process in the cell, while keeping the complexity of the model to a minimum.

Similarly to [159], we lumped the *E. coli* genome into a small set of *gene classes* (*rnn*, *rpoBC*, protein-encoding (*bulk*) genes) accounting for all transcription and translation within the cell for a given growth rate. Then, according to the formalism in [115] we developed an ODE model whose variables account for the macromolecule content of the cell (ribosomes, RNAP, proteins, mRNAs etc). This coarse-grained model of *E.coli* GEM has then been calibrated to data from [31] showing that the model is able to reproduce the macromolecular composition of the cell at several growth rate values.

This type of model is interesting *per se* because it may help elucidate how the *E. coli* GEM works on a global scale, pointing out the most important and fundamental regulations at the gene expression level. Moreover, this GEM model could also be used as host-cell model where more complex/specific gene networks are embedded into it. This, for example, would allow one to study host-synthetic circuit interactions and help biologists in engineering biological circuits that require the knowledge of host factors that compete or interface with designed function [36].

## 8.2 *E. coli* GEM network: biological description

In Figure 8.1 is depicted the coarse-grained model of *E. coli* GEM. This picture describes the expression of the three gene classes *rnn*, *rpoBC* and *bulk* representing, respectively,

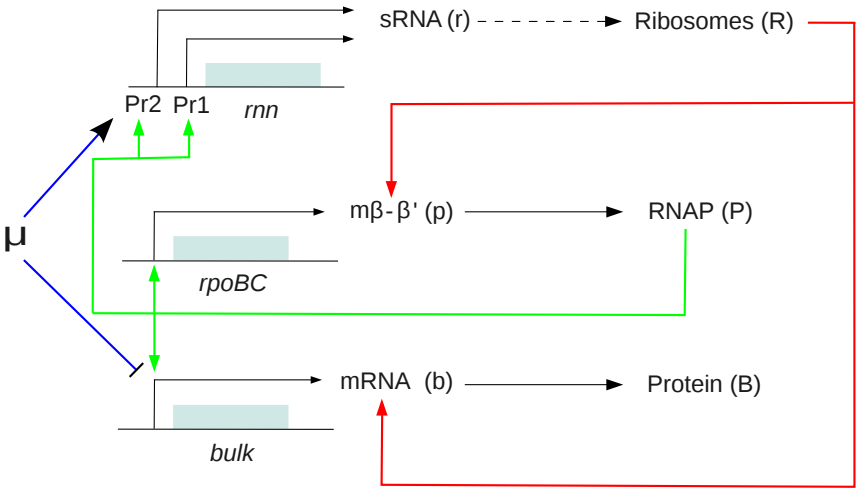



FIGURE 8.1: **The coarse-grained model of *E. coli* GEM.**

the rRNA operon determining the ribosome synthesis [31], the operon encoding for $\beta - \beta'$ subunits [130] which are the limiting factors in RNAP assembling and the proxy gene class which is a consolidation of all identified gene-encoding protein. Moreover, Figure 8.1 shows the various feedback regulations between the components of the GEM. Notably, transcription of all gene classes is catalysed by RNAP while translation of $\beta - \beta'$ mRNA ($m\beta - \beta'$) and *bulk* mRNA (mRNA) is operated by ribosomes [77]. In addition to these mechanistic regulations, we considered also the known effect that the physiological state of the cell—represented by the growth rate $\mu$—has on the regulation of stable-RNA (sRNA) and mRNA [31, 89]. That is, at increasing growth rate values more stable-RNAs are produced whereas mRNA synthesis is reduced. In what follows we will discuss the expression of the three gene classes in more detail.

### 8.2.1 Ribosomes synthesis and function

Ribosomes are composed of stable-RNAs (sRNA) and proteins (r-protein). Since the synthesis of r-proteins is regulated so as to match that of sRNA [59], it turns out that ribosome regulation centers on the transcriptional control of *rnn* genes. Hence, we will not take into account r-proteins synthesis and their assembling with sRNA to produce ribosomes, but we will consider that ribosome synthesis is essentially limited by the rate of sRNA synthesis. This assumption is graphically represented by the dashed arrow in Figure 8.1.

The *rnn* genes, from which sRNA is produced, are organized in 7 operons (*rnnA, rnnB, rnnC, rnnD, rnnE, rnnG, rnnH*) and are under the control of two promoters: the Pr1 constitutive promoter which is essentially regulated by free RNAP and the Pr2 growth rate dependent promoter whose activation strongly correlates with increasing

growth rate values [89]. In rapidly growing cells, sRNA is strongly synthesized and commandeers the majority of RNAP in the cell, whereas sRNA synthesis is suppressed at lower growth (e.g. nutrient starvation, amino acids depletion) [100]. At molecular level, Pr2 promoter is inhibited by ppGpp, an alarmone molecule, which rapidly accumulates at lower growth rates following the stringent response caused by amino acid and/or nutrient starvation [59]. Since ppGpp and growth rate regulations of Pr2 are strongly interconnected and superposed [100], for the sake of simplicity we will only take into account the growth rate positive effect on sRNA synthesis.

Ribosome role in the GEM is that of synthesizing proteins (translation) linking amino acids together in the order specified by messenger-RNA molecules. The red arrows in Figure 8.1, starting from ribosomes and ending in the mRNA species ($m\beta - \beta'$, mRNA), are indicating the positive regulation effect carried out by ribosomes in the translation process.

### 8.2.2 RNAP synthesis and function

In *E. coli* cells the transcription is catalysed by RNA Polymerase (RNAP) core enzyme (green arrows in Figure 8.1), which is made up of five subunits ($\alpha_2\beta\beta'\sigma$). The amount of $\beta\beta'$ turns out to be the limiting factor in RNAP synthesis [31]. Hence—in our model—we will omit the other subunits and focus only on $\beta\beta'$ formation, assuming that $\beta\beta'$ subunits' amount equals RNAP core enzyme amount.

The $\beta\beta'$ subunits are the product of *rpoBC* genes expression. The *rpoBC* genes are regulated both at transcriptional and translational level [64], but some of these feedback mechanisms are still unclear. In our model, we will only consider the positive transcriptional regulation of RNAP and the positive translational regulation of ribosomes which are known and well understood regulations.

RNAP initiates transcription at specific DNA sequences known as promoters. It then moves along DNA producing an RNA chain (mRNA sRNA), which is complementary to the template DNA strand. The process of adding nucleotides to the RNA strand is known as elongation. Finally, RNAP will preferentially release its RNA transcript at specific DNA sequences encoded at the end of genes, which are known as terminators.

### 8.2.3 Proteins synthesis and function

In our model, the *bulk* gene class represents a consolidation of all 4288 identified gene-encoded protein [25] in *E. coli*. Hence, mRNA and protein in Figure 8.1 represent proxy products of all messenger-RNAs and proteins expressed in *E. coli*. In bacteria, mRNA

synthesis is regulated at transcription level by RNAP and by growth rate dependent mechanisms (probably due to ppGpp) [32] that—at higher growth rates—reduces mRNA synthesis to divert resources towards sRNA synthesis. This is because—to achieve higher growth rate values—cells need more ribosomes to be engaged in translating proteins and so resources are diverted from producing mRNA towards stable-RNA molecules, which substantially determine ribosome amount.

## 8.3 Mathematical background

To model the dynamics of gene products depicted in Figure 8.1 we used a mathematical formalism developed by A. Kremling in [115]. This formalism uses ordinary differential equations coupled with algebraic mass conservation equations (to derive free RNAP and ribosomes) to model—in a quite detailed fashion—the dynamics of those molecular mechanisms involved in the transcription and translation of a single gene. Here, we basically report, as an example, the derivation of transcription and translation model equations of a generic gene. These equations will be then used to model the entire GEM depicted in Figure 8.1.

### 8.3.1 Transcription

Let consider a single gene $g$ of length $L$, being $L$ the number of nucleotides of gene $g$. Free RNAP, denoted by $FP$, binds in a reversible way to the free promoter $D$ of gene $g$ and forms the complex $PD$:

$$FP + D \underset{k_1}{\overset{k_2}{\leftrightarrows}} PD \tag{8.1}$$

After binding, RNAP clears the promoter and is ready to start moving (and reading nucleotides) along the gene $g$:

$$PD \overset{k_g}{\rightarrow} Y + D \tag{8.2}$$

where $Y$ denotes RNAP just after having cleared the promoter, but still bound to the DNA. Then, RNAP moves along the gene $g$, reads its information and enlarges the chain of the nascent mRNA $Y_i$ with nucleotides $Nu$. Eventually, when RNAP has finished to

elongate the RNA chain (RNA), it unbinds to the DNA and returns to its free state:

$$\begin{aligned} Y + Nu &\xrightarrow{c_g} Y_1 \\ Y_1 + Nu &\xrightarrow{c_g} Y_2 \\ &\vdots \\ Y_{L-1} + Nu &\xrightarrow{c_g} FP + RNA \end{aligned} \tag{8.3}$$

Then, the completed RNA molecule is subject to degradation $\gamma_g$ and dilution due to cell growth rate $\mu$:

$$RNA \xrightarrow{\gamma_g+\mu} \varnothing \tag{8.4}$$

In this model, reactions (8.3) are supposed not dependent on nucleotides, since they are assumed to be in excess. Let $P_f$, $d_f$, $P_d$, $w$,$w_i$, $x$ be the molar concentration of species $FP$, $D$, $PD$, $Y$, $Y_i$ and $RNA$, respectively. Hence, assuming classical mass-action kinetics, one can derive the following ODEs from reactions above (8.1)-(8.2)-(8.3)-(8.4) as in [19, 20]:

$$\begin{aligned} \dot{P}_d &= k_1 P_f d_f - k_2 P_d - k_g P_d \\ \dot{d}_f &= -k_1 P_f d_f + k_2 P_d + k_g P_d \\ \dot{P}_f &= -k_1 P_f d_f + k_2 P_d + c_g w_{L-1} \\ \dot{w} &= k_g P_d - c_g w \\ \dot{w}_1 &= c_g w - c_g w_1 \\ &\vdots \\ \dot{w}_{L-1} &= c_g w_{L-2} - c_g w_{L-1} \\ \dot{x} &= c_g w_{L-1} - (\gamma_g + \mu) x \end{aligned} \tag{8.5}$$

Along with system (8.5), there are also two mass conservation equations regarding the total concentration of promoter, i.e. $d$, and the total concentration of RNAP, i.e. $P$:

$$d = d_f + P_d \tag{8.6a}$$

$$P = P_f + P_d + w + w_1 + \ldots + w_{L-1}. \tag{8.6b}$$

Hence, solving the promoter conservation equation (8.6a) with respect to the free promoter concentration $d_f$, system (8.5) can be reduced in:

$$\begin{aligned}
\dot{P}_d &= k_1 P_f (d - P_d) - k_2 P_d - k_g P_d \\
\dot{P}_f &= -k_1 P_f (d - P_d) + k_2 P_d + c_g w_{L-1} \\
\dot{w} &= k_g P_d - c_g w \\
\dot{w}_1 &= c_g w - c_g w_1 \\
&\vdots \\
\dot{w}_{L-1} &= c_g w_{L-2} - c_g w_{L-1} \\
\dot{x} &= c_g w_{L-1} - (\gamma_g + \mu) x
\end{aligned} \tag{8.7}$$

Then, assuming that all RNAP complexes $P_d$, $w$, $w_1$,...,$w_{L-1}$ are at steady state (the mostly accepted biological hypothesis beyond this assumption is that biochemical reactions involved in complexes' formation are generally much faster than the entire transcription process), one gets:

$$P_d = \frac{P_f}{P_f + \theta_P} d \tag{8.8}$$

$$w_{L-1} = w_{L-2} = \ldots = w_1 = w = \frac{k_g}{c_g} P_d = \frac{k_g}{c_g} \frac{P_f}{P_f + \theta_P} d \tag{8.9}$$

where $\theta_P = \dfrac{k_2 + k_g}{k_1}$. Moreover, substituting (8.8) and (8.9) into (8.6b), the total RNAP conservation equation can be rewritten as:

$$P = P_f + \left( \frac{P_f}{P_f + \theta_P} \right) d \left( 1 + L \frac{k_g}{c_g} \right) \tag{8.10}$$

Finally, the ODE modelling the dynamics of the transcription of gene $g$, which leads to the formation of the completed RNA—expressed in molar concentration by $x$, reads as:

$$\dot{x} = k_g \frac{P_f}{P_f + \theta_P} d - (\gamma_g + \mu) x. \tag{8.11}$$

Biologically speaking, equation (8.10) makes sense only if the biological system under investigation was made up of only gene $g$ (RNAP would either be free or bound to gene $g$), but in reality organisms consist of thousands of genes. Hence, considering an organism with a total number of $n$ genes, a more realistic conservation equation for RNAP would be:

$$P = P_f + \sum_{l=1}^{n} \left( \frac{P_f}{P_f + \theta_P^l} \right) d_l \left( 1 + L_l \frac{k_l}{c_l} \right) \tag{8.12}$$

As explained in [115], this model does not consider stearic effects of RNAP, that is RNAP when bound to the DNA occupies only one nucleotide.

### 8.3.2 Translation

Modelling the translation process is a little bit more elaborated than transcription, since one has to take into account both nascent mRNAs, i.e. uncompleted mRNAs that stem from RNAPs which are still in the process of transcribing, and completed mRNA.

#### 8.3.2.1 Translation of nascent mRNA

Every complex $Y_i$ (the moving RNAP on gene $g$) represents a starting point for translation, since from each $Y_i$ can originate a new, but uncompleted, mRNA where free ribosomes can bind and start translation. Hence, free ribosome $FR$ binds to the free ribosome binding site $\overline{Y}_i$ of $Y_i$ in a reversible fashion forming the complex $R\overline{Y}_i$:

$$FR + \overline{Y}_i \underset{k_3}{\overset{k_4}{\leftrightarrows}} R\overline{Y}_i \tag{8.13}$$

After binding, ribosome clears the ribosome binding site and is ready to start moving (and reading nucleotides) along the nascent mRNA:

$$R\overline{Y}_i \overset{k_G}{\rightarrow} Z_i + \overline{Y}_i \tag{8.14}$$

where $Z_i$ denotes ribosome just after the clearance, bound to the nascent mRNA and waiting for reading nucleotides.

Then, the moving ribosome $Z_i^j$ reads the information stored in the nascent mRNA and enlarges the chain of the corresponding protein with amino acids $AA$. Here we do not model amino acid formation, which are considered not limiting, nor the charging and uncharging of $AA$ on transfer-RNA, which are considered not limiting as well. The protein chain is considered to grow at the maximal length $s$ amino acids. This maximal length of $s$ is related to the length $i$—in nucleotides—of the nascent mRNA $Y_i$ by the relation $s = i/3$, since every triplet of nucleotides (codon) determines one amino acid.

$$\begin{aligned} Z_i + AA &\xrightarrow{c_G} Z_i^1 \\ Z_i^1 + AA &\xrightarrow{c_G} Z_i^2 \\ &\vdots \\ Z_i^{s-1} + AA &\xrightarrow{c_G} Z_i^s \\ Z_i^s + AA &\xrightarrow{c_G} \dots \end{aligned} \tag{8.15}$$

Degradation of nascent mRNA is not considered.

Let $R_f$, $\overline{w}_i^f$, $\overline{w}_i$, $z_i$, $z_i^j$ be the molar concentration of species $FR$, $\overline{Y}_i$, $R\overline{Y}_i$, $Z_i$, $Z_i^j$, respectively. Hence, assuming classical mass-action kinetics, one can derive the following ODEs from reactions above (8.13)-(8.14)-(8.15):

$$\begin{aligned} \dot{\overline{w}}_i &= k_3 R_f \overline{w}_i^f - k_4 \overline{w}_i - k_G \overline{w}_i \\ \dot{\overline{w}}_i^f &= -k_3 R_f \overline{w}_i^f + k_4 \overline{w}_i + k_G \overline{w}_i \\ \dot{R}_f &= -k_3 R_f \overline{w}_i^f + k_4 \overline{w}_i \\ \dot{z}_i &= k_G \overline{w}_i - c_G z_i \\ \dot{z}_i^1 &= c_G z_i - c_G z_i^1 \\ &\vdots \\ \dot{z}_i^{s-1} &= c_G z_i^{s-2} - c_G z_i^{s-1} \\ \dot{z}_i^s &= c_G z_i^{s-1} - c_G z_i^s \end{aligned} \tag{8.16}$$

Since for every nascent mRNA $Y_i$ the following conservation equation is valid:

$$w_i = \overline{w}_i^f + \overline{w}_i \tag{8.17}$$

considering $\overline{w}_i^f = w_i - \overline{w}_i$, we can reduce system (8.24) to:

$$\begin{aligned} \dot{\overline{w}}_i &= k_3 R_f (w_i - \overline{w}_i) - k_4 \overline{w}_i - k_G \overline{w}_i \\ \dot{R}_f &= -k_3 R_f (w_i - \overline{w}_i) + k_4 \overline{w}_i \\ \dot{z}_i &= k_G \overline{w}_i - c_G z_i \\ \dot{z}_i^1 &= c_G z_i - c_G z_i^1 \\ &\vdots \\ \dot{z}_i^{s-1} &= c_G z_i^{s-2} - c_G z_i^{s-1} \\ \dot{z}_i^s &= c_G z_i^{s-1} - c_G z_i^s \end{aligned} \tag{8.18}$$

Now, considering steady state assumption for all ribosome complexes $R\overline{Y}_i$, $Z_i$ and $Z_i^j$ one gets:

$$\overline{w}_i = \frac{R_f}{R_f + \theta_R} w_i \tag{8.19}$$

$$\overline{z}_i^s = \overline{z}_i^{s-1} = \ldots = \overline{z}_i^1 = \overline{z}_i = \frac{k_G}{c_G}\overline{w}_i = \frac{k_G}{c_G}\frac{R_f}{R_f + \theta_R} w_i \tag{8.20}$$

where $\theta_R = \dfrac{k_4 + k_G}{k_3}$.

#### 8.3.2.2 Translation of completed mRNA

Now we focus on the translation of completed mRNAs. Let $RNA$ be the completed mRNA, free from RNAP and the gene whereby it has been generated.

The process of ribosome binding to the completed mRNA is similar to that for the nascent mRNA presented above. Hence, free ribosome $FR$ binds to the free ribosome binding site $\overline{RNA}_f$ of $mRNA$ in a reversible way forming the complex $RRNA$:

$$FR + \overline{RNA}_f \underset{k_3}{\overset{k_4}{\leftrightarrows}} RRNA \tag{8.21}$$

After binding, ribosome clears the ribosome binding site and is ready to start moving (and reading nucleotides) along the mRNA:

$$RRNA \overset{k_G}{\rightarrow} Z + \overline{RNA}_f \tag{8.22}$$

where $Z$ denotes ribosome just after the clearance, bound to the completed mRNA and waiting for reading nucleotides.

Then, the moving ribosome $Z^j$ reads all the information stored into the completed mRNA, and elongates the corresponding protein $G$ with amino acids $AA$ (encoded by nucleotides in mRNA) until $G$ has reached its length $M = L/3$. As before, we do not model neither amino acid formation nor the charging and uncharging of $AA$ on transfer-RNA. Moreover, protein $G$ is subject to degradation $\gamma_G$ and dilution due to cell growth

rate $\mu$.

$$\begin{aligned} Z + AA &\overset{c_G}{\rightarrow} Z^1 \\ Z^1 + AA &\overset{c_G}{\rightarrow} Z^2 \\ &\vdots \\ Z^{M-1} + AA &\overset{c_G}{\rightarrow} G + R_f \\ G &\overset{\gamma_G+\mu}{\rightarrow} \varnothing \end{aligned} \tag{8.23}$$

Let $R_f$, $\overline{w}^f$, $\overline{w}$, $z$, $z^j$,$X$ be the molar concentration of species $FR$, $\overline{RNA}$, $RRNA$, $Z$, $Z^j$, $G$, respectively. Hence, assuming classical mass-action kinetics, one can derive the following ODEs from reactions above (8.21)-(8.22)-(8.23):

$$\begin{aligned} \dot{\overline{w}} &= k_3 R_f \overline{w}^f - k_4 \overline{w} - k_G \overline{w} \\ \dot{\overline{w}}^f &= -k_3 R_f \overline{w}^f + k_4 \overline{w} + k_G \overline{w} \\ \dot{R}_f &= -k_3 R_f \overline{w}^f + k_4 \overline{w} \\ \dot{z} &= k_G \overline{w} - c_G z \\ \dot{z}^1 &= c_G z - c_G z^1 \\ &\vdots \\ \dot{z}^{M-1} &= c_G z^{M-2} - c_G z^{M-1} \\ \dot{X} &= c_G z^{M-1} - (\gamma_G + \mu) X \end{aligned} \tag{8.24}$$

Since for the completed mRNA $RNA$ the following conservation equation is valid:

$$x = \overline{w}^f + \overline{w} \tag{8.25}$$

we can reduce system (8.24), considering $\overline{w}^f = x - \overline{w}$, to:

$$\begin{aligned} \dot{\overline{w}} &= k_3 R_f (x - \overline{w}) - k_4 \overline{w} - k_G \overline{w} \\ \dot{R}_f &= -k_3 R_f (x - \overline{w}) + k_4 \overline{w} \\ \dot{z} &= k_G \overline{w} - c_G z \\ \dot{z}^1 &= c_G z - c_G z^1 \\ &\vdots \\ \dot{z}^{M-1} &= c_G z^{M-2} - c_G w^{M-1} \\ \dot{X} &= c_G z^{M-1} - (\gamma_G + \mu) X \end{aligned} \tag{8.26}$$

Now, considering steady state assumption for all ribosome complexes $RRNA$, $Z$ and $Z^j$ one gets:

$$\overline{w} = \frac{R_f}{R_f + \theta_R} x \tag{8.27}$$

$$\overline{z}^{M-1} = \overline{z}^{M-2} = \ldots = \overline{z}^1 = \overline{z} = \frac{k_G}{c_G}\overline{w} = \frac{k_G}{c_G}\frac{R_f}{R_f + \theta_R} x \tag{8.28}$$

where $\theta_R = \dfrac{k_4 + k_G}{k_3}$.

Finally, the ODE for the protein $G$ reads as:

$$\dot{X} = k_G \frac{R_f}{R_f + \theta_R} x - (\gamma_G + \mu) X \tag{8.29}$$

#### 8.3.2.3 Comments on ribosome engaged in translation

To calculate the total concentration of ribosomes engaged in the translation process, one has to consider two terms: one for those translating nascent mRNAs and the other one for those translating the completed mRNA.

Let us first consider the total concentration of ribosomes that are involved in translating one nascent mRNA which stems from the moving RNAP $Y_i$. Since from each moving RNAP complex $Y_i$ only one nascent mRNA originates from it, from now on $Y_i$ will also indicate the corresponding nascent mRNA which is being transcribed from the moving RNAP ($Y_i$). For every $Y_i$ the ribosomes that are bound to it are $RRNA_i$, $Z_i$, $Z_i^1$, $Z_i^2$,...,$Z_i^s$, denoting the ribosome bound to the ribosome binding site, the ribosome after the clearance, and all moving ribosomes. Hence, let $R_i^n$ be the concentration of ribosomes engaged in the translation of the nascent mRNA $Y_i$, we have:

$$R_i^n = \overline{w}_i + z_i + \sum_{j=1}^{s=iM/L} z_i^j \tag{8.30}$$

and knowing that all the nascent mRNA on gene $g$ are $Y_i$ for $i = 1, 2, \ldots, L-1$, it turns out that the total concentration of ribosomes engaged in translating all nascent mRNA, i.e. $R^n$, is given by:

$$R^n = \sum_{i=1}^{L-1} \left( \overline{w}_i + z_i + \sum_{j=1}^{iM/L} z_i^j \right). \tag{8.31}$$

Substituting expressions (8.19)-(8.20) into (8.31), one gets:

$$\begin{aligned} R^n &= \sum_{i=1}^{L-1}\left(\frac{R_f}{R_f+\theta_R}w_i + \frac{k_G}{c_G}\frac{R_f}{R_f+\theta_R}w_i + \sum_{j=1}^{iM/L}\frac{k_G}{c_G}\frac{R_f}{R_f+\theta_R}w_i\right) \\ &= \frac{R_f}{R_f+\theta_R}\sum_{i=1}^{L-1}\left(1+\frac{k_G}{c_G}+i\frac{M}{L}\frac{k_G}{c_G}\right)w_i \end{aligned} \tag{8.32}$$

then, since $w_i = \frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d$ as stated in (8.9), we have:

$$\begin{aligned} R^n &= \frac{R_f}{R_f+\theta_R}\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d\sum_{i=1}^{L-1}\left(1+\frac{k_G}{c_G}+i\frac{M}{L}\frac{k_G}{c_G}\right) \\ &= \frac{R_f}{R_f+\theta_R}\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d\left[(L-1)+\frac{k_G}{c_G}(L-1)+\frac{M}{L}\frac{k_G}{c_G}\sum_{i=1}^{L-1}i\right] \\ &= \frac{R_f}{R_f+\theta_R}\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d\left[(L-1)+\frac{k_G}{c_G}(L-1)+\frac{M(L-1)}{2}\frac{k_G}{c_G}\right] \\ &= \frac{R_f}{R_f+\theta_R}\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d(L-1)\left[1+\frac{k_G}{c_G}\left(1+\frac{M}{2}\right)\right] \end{aligned} \tag{8.33}$$

Now we will focus on the ribosomes engaged in translating the completed mRNA. Similarly as above, for the completed mRNA the ribosomes that are bound to it are $RRNA$, $Z$, $Z^1$, $Z^2$,...,$Z^{M-1}$, denoting the ribosome bound to the ribosome binding site, the ribosome after the clearance, and all moving ribosomes. Hence, let $R^c$ be the concentration of ribosomes engaged in the translation of the completed mRNA, we have:

$$R^c = \overline{w} + z + \sum_{j=1}^{M-1} z^j \tag{8.34}$$

Substituting expressions (8.27)-(8.28) into (8.34), one gets:

$$\begin{aligned} R^c &= \left(\frac{R_f}{R_f+\theta_R}x + \frac{k_G}{c_G}\frac{R_f}{R_f+\theta_R}x + \sum_{j=1}^{M-1}\frac{k_G}{c_G}\frac{R_f}{R_f+\theta_R}x\right) \\ &= \frac{R_f}{R_f+\theta_R}x\left(1+\frac{k_G}{c_G}M\right) \end{aligned} \tag{8.35}$$

Finally, the total concentration of ribosomes engaged in the translation process for the single gene $g$, denoted by $R^g$, are:

$$\begin{aligned} R^g &= R^n + R^c \\ &= \frac{R_f}{R_f+\theta_R}\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d(L-1)\left[1+\frac{k_G}{c_G}\left(1+\frac{M}{2}\right)\right] + \frac{R_f}{R_f+\theta_R}x\left(1+\frac{k_G}{c_G}M\right) \\ &= \frac{R_f}{R_f+\theta_R}\left\{(L-1)\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d\left[1+\frac{k_G}{c_G}\left(1+\frac{M}{2}\right)\right] + \left(1+\frac{k_G}{c_G}M\right)x\right\} \end{aligned} \tag{8.36}$$

### 8.3.3 Final conclusions

Here we summarize the most important modelling equations of the transcription-translation model just presented above:

$$\dot{x} = k_g\frac{P_f}{P_f+\theta_P}d - (\gamma_g+\mu)x \tag{8.37a}$$

$$\dot{X} = k_G\frac{R_f}{R_f+\theta_R}x - (\gamma_G+\mu)X \tag{8.37b}$$

$$P^g = \left(\frac{P_f}{P_f+\theta_P}\right)d\left(1+L\frac{k_g}{c_g}\right) \tag{8.37c}$$

$$R^g = \frac{R_f}{R_f+\theta_R}\left\{(L-1)\frac{k_g}{c_g}\frac{P_f}{P_f+\theta_P}d\left[1+\frac{k_G}{c_G}\left(1+\frac{M}{2}\right)\right] + \left(1+\frac{k_G}{c_G}M\right)x\right\} \tag{8.37d}$$

Notably, (8.37a) represents the dynamics of mRNA concentration $x$, produced during the transcription of gene $g$; (8.37b) models the dynamics of protein concentration $X$, produced during the translation of the mRNA of gene $g$; (8.37c) denotes the concentration of RNAP engaged in the transcription of gene $g$; (8.37d) counts for the total concentration of ribosomes engaged in the translation of both nascent and completed mRNA of gene $g$.

It is worth saying that equations (8.37) describe the transcription-translation case of a *constitutive* gene $g$, i.e. a gene which is not specifically regulated, neither at the transcriptional level by transcription factors and/or external inducers, nor at the translational level. One possible way to extend model (8.37), to account for these specific regulations of the gene expression, would be that of considering two generic functions $F_1(\cdot)$ and $F_2(\cdot)$ describing all the specific regulations at the transcriptional and at the translational level, respectively. Moreover, also the growth rate $\mu$ might be modeled as a function $\mu(\cdot)$ of the system variables. With all these considerations above, model (8.37)

can be generalized as follows:

$$\dot{x} = k_g \frac{P_f}{P_f + \theta_P} F_1(\cdot) d - (\gamma_g + \mu(\cdot)) x \tag{8.38a}$$

$$\dot{X} = k_G \frac{R_f}{R_f + \theta_R} F_2(\cdot) x - (\gamma_G + \mu(\cdot)) X \tag{8.38b}$$

$$P^g = \left( \frac{P_f}{P_f + \theta_P} \right) F_1(\cdot) d \left( 1 + L \frac{k_g}{c_g} \right) \tag{8.38c}$$

$$R^g = \frac{R_f}{R_f + \theta_R} F_2(\cdot) \left\{ (L-1) \frac{k_g}{c_g} \frac{P_f}{P_f + \theta_P} F_1(\cdot) d \left[ 1 + \frac{k_G}{c_G} \left( 1 + \frac{M}{2} \right) \right] + \left( 1 + \frac{k_G}{c_G} M \right) x \right\} \tag{8.38d}$$

To conclude, we note that all parameters used in model (8.38) have a biological meaning and each can be expressed with physical units as reported in Table 8.1.

| Descriptions | Parameters | Physical Units |
|---|---|---|
| Gene length | $L$ | *nucleotides* |
| Promoter concentration | $d$ | *Molar concentration* |
| Max transcription initiation rate | $k_g$ | $1/time$ |
| Michaelis threshold transcription | $\theta_P$ | *Molar concentration* |
| RNA elongation rate | $c_g$ | *nucleotides/time* |
| RNA degradation rate | $\gamma_g$ | $1/time$ |
| Protein length | $M$ | *amino acids* |
| Max translation initiation rate | $k_G$ | $1/time$ |
| Michaelis threshold translation | $\theta_R$ | *Molar concentration* |
| Protein degradation rate | $\gamma_G$ | $1/time$ |
| Protein elongation rate | $c_G$ | *amino acids/time* |

TABLE 8.1: Model Parameters.

## 8.4 *E. coli* GEM dynamical model

Here, using the mathematical formalism presented in Section 8.3 and summarized in (8.38), we aim at representing the transcription and translation machinery in *E. coli* depicted in Figure 8.1. Notably, we will apply model (8.38) to each of the gene classes, i.e. *rnn*, *rpoBC*, *bulk*, to derive a dynamical model whose system's variables are $R$, $p$, $P$, $b$, $B$, denoting molar concentration of ribosomes, RNAP mRNA, RNAP protein, total mRNA, total protein, respectively.

### 8.4.1 *rnn* gene expression model

As stated in Section 8.2.1, *rnn* genes encode for the stable-RNAs (sRNA), which are the limiting factors in ribosome formation, hence here we assume that sRNA basically determine the amount of ribosomes. This means that only the transcription model equation (8.38a) will be used to model the dynamics of ribosome concentration. Moreover, equation (8.38a) has to be slightly be modified to account for the fact that *rnn* genes have two promoters: *Pr1*, which is a growth rate dependent promoter (presumably regulated by ppGpp) and *Pr2*, which is a constitutive one. Therefore, the dynamics of ribosomes concentration $R$ reads as:

$$\dot{R} = k_r \left( \frac{P_f}{P_f + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f}{P_f + \theta_P^{r2}} \right) d_r - (\mu + \gamma_r) R \tag{8.39}$$

where $h^+(\mu, \theta_\mu^r, n_r)$ denotes an increasing Hill function of the growth rate $\mu$, modelling the specific growth rate dependent regulation of promoter *Pr1*. Whereas the second Michaelis-Menten term of free RNAP ($P_f$) denotes the constitutive regulation of promoter $Pr2$. The concentration of RNAP engaged in translating the *rnn* genes $P^r$ can be easily derived from (8.38c):

$$P^r = \left( \frac{P_f}{P_f + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f}{P_f + \theta_P^{r2}} \right) d_r \left( 1 + L_r \frac{k_r}{c_r} \right) \tag{8.40}$$

Explanations of parameters in (8.39) and in (8.40) are given in Table 8.2.

| Descriptions | Parameters | Physical Units |
|---|---|---|
| *rnn* gene length | $L_r$ | *nucleotides* |
| *rnn* gene concentration | $d_r$ | *Molar concentration* |
| *rnn* Max transcription initiation rate | $k_r$ | $1/time$ |
| Michaelis threshold for RNAP regulation of *Pr1* | $\theta_P^{r1}$ | *Molar concentration* |
| Hill threshold for $\mu$ regulation of *Pr1* | $\theta_\mu^r$ | $1/time$ |
| Hill coefficient for $\mu$ regulation of *Pr1* | $n_r$ | dimensionless |
| Michaelis threshold for RNAP regulation of *Pr2* | $\theta_P^{r2}$ | *Molar concentration* |
| sRNA elongation rate | $c_r$ | *nucleotides/time* |
| Ribosome degradation rate | $\gamma_R$ | $1/time$ |

TABLE 8.2: *rnn* gene expression model model parameters.

### 8.4.2 *rpoBC* gene expression model

As reported in Section 8.2.2, *rpoBC* genes encode for the $\beta - \beta'$ subunits, which are the limiting factors for RNAP formation. Hence, here we assume that concentrations

of $\beta - \beta'$ mRNA, i.e. $p$, and $\beta - \beta'$ protein, i.e. $P$, equal those of RNAP mRNA and RNAP protein, respectively. Even though *rpoBC* gene expression seems to be specifically regulated both at transcriptional and translational level (but mechanisms have still to be well elucidated), here—for the sake of simplicity—we assume that *rpoBC* genes have no specific regulations, that is $F_1(\cdot) = F_2(\cdot) = 1$ in (8.38). Hence, *rpoBC* gene expression model reads as:

$$\dot{p} = k_p \left( \frac{P_f}{P_f + \theta_P^p} \right) d_p - (\mu + \gamma_p) p \tag{8.41a}$$

$$\dot{P} = k_P \left( \frac{R_f}{R_f + \theta_R^P} \right) p - (\mu + \gamma_P) P \tag{8.41b}$$

$$P^p = \left( \frac{P_f}{P_f + \theta_P^p} \right) d_p \left( 1 + L_p \frac{k_p}{c_p} \right) \tag{8.41c}$$

$$R^p = \frac{R_f}{R_f + \theta_R^P} \left\{ (L_p - 1) \frac{k_p}{c_p} \left( \frac{P_f}{P_f + \theta_P^p} \right) d_p \left[ 1 + \frac{k_P}{c_P} \left( 1 + \frac{M_P}{2} \right) \right] + \left( 1 + M_P \frac{k_P}{c_P} \right) p \right\} \tag{8.41d}$$

where $P^p$ and $R^p$ denote RNAP engaged in the transcription of *rpoBC* genes and ribosomes engaged in translation of $\beta - \beta'$ mRNA, respectively.

Explanations of parameters in (8.41) are given in Table 8.3.

| Descriptions | Parameters | Physical Units |
|---|---|---|
| *rpoBC* gene length | $L_p$ | *nucleotides* |
| *rpoBC* promoter concentration | $d_p$ | *Molar concentration* |
| *rpoBC* Max transcription initiation rate | $k_p$ | $1/time$ |
| *rpoBC* Michaelis threshold transcription | $\theta_P^p$ | *Molar concentration* |
| *rpoBC* mRNA elongation rate | $c_p$ | *nucleotides/time* |
| *rpoBC* mRNA degradation rate | $\gamma_p$ | $1/time$ |
| *rpoBC* protein length | $M_P$ | *amino acids* |
| *rpoBC* max translation initiation rate | $k_P$ | $1/time$ |
| *rpoBC* Michaelis threshold translation | $\theta_R^P$ | *Molar concentration* |
| *rpoBC* protein degradation rate | $\gamma_P$ | $1/time$ |
| *rpoBC* protein elongation rate | $c_P$ | *amino acids/time* |

TABLE 8.3: *rpoBC* gene expression model model parameters.

### 8.4.3 *bulk* gene expression model

As reported in Section 8.2.3, *bulk* genes are a proxy gene class representing the consolidation of all identified gene-encoding protein in *E. coli*, i.e. 4288 genes [25]. Hence, the transcription of *bulk* genes will model the total mRNA production in *E. coli*. The synthesis of mRNA concentration, i.e. $b$, is determined by free RNAP availability, but

also by the global and specific growth-rate regulation that at higher growth rate diverts part of the resources (free RNAP etc.) from mRNA transcription towards stable-RNA transcription to sustain the translation apparatus (as discussed in Section 8.2.3).

Finally, total protein concentration $B$ is simply determined by the translation of the proxy total mRNA $b$, without any specific translational regulation.

Hence, *rpoBC* gene expression model reads as:

$$\dot{b} = k_b \left( \frac{P_f}{P_f + \theta_P^b} \right) h^-(\mu; \theta_\mu^b, n_b) d_b - (\mu + \gamma_b) b \tag{8.42a}$$

$$\dot{B} = k_B \left( \frac{R_f}{R_f + \theta_R^B} \right) b - (\mu + \gamma_B) B \tag{8.42b}$$

$$P^b = \left( \frac{P_f}{P_f + \theta_P^b} \right) h^-(\mu; \theta_\mu^b, n_b) d_b \left( 1 + L_b \frac{k_b}{c_b} \right) \tag{8.42c}$$

$$R^b = \frac{R_f}{R_f + \theta_R^B} \left\{ (L_b - 1) \frac{k_b}{c_b} \left( \frac{P_f}{P_f + \theta_P^b} \right) h^-(\mu; \theta_\mu^b, n_b) d_b \left[ 1 + \frac{k_B}{c_B} \left( 1 + \frac{M_B}{2} \right) \right] + \left( 1 + M_B \frac{k_B}{c_B} \right) b \right\} \tag{8.42d}$$

where $h^-(\mu; \theta_\mu^b, n_b)$ is a decreasing Hill function of the growth rate $\mu$, modelling the negative (specific) regulation effect that $\mu$ exerts on the mRNA transcription at higher growth rates. Whereas, $P^b$ and $R^b$ denote RNAP engaged in the transcription of *bulk* genes and ribosomes engaged in translation of mRNA, respectively.

Explanations of parameters in (8.42) are given in Table 8.4.

| Descriptions | Parameters | Physical Units |
|---|---|---|
| *bulk* gene length | $L_b$ | *nucleotides* |
| *bulk* promoter concentration | $d_b$ | *Molar concentration* |
| mRNA max transcription initiation rate | $k_b$ | $1/time$ |
| *bulk* Michaelis threshold transcription | $\theta_P^b$ | *Molar concentration* |
| Hill threshold for $\mu$ regulation of *bulk* | $\theta_\mu^b$ | $1/time$ |
| Hill coefficient for $\mu$ regulation of *bulk* | $n_b$ | dimensionless |
| *bulk* mRNA elongation rate | $c_b$ | *nucleotides/time* |
| *bulk* mRNA degradation rate | $\gamma_b$ | $1/time$ |
| *bulk* protein length | $M_B$ | *amino acids* |
| *bulk* max translation initiation rate | $k_B$ | $1/time$ |
| *rpoBC* Michaelis threshold translation | $\theta_R^B$ | *Molar concentration* |
| *bulk* protein degradation rate | $\gamma_B$ | $1/time$ |
| *bulk* protein elongation rate | $c_B$ | *amino acids/time* |

TABLE 8.4: *bulk* gene expression model parameters.

### 8.4.4 Complete dynamical model of *E. coli* GEM

Here, we will group together all the gene expression dynamical models of each gene class to have a comprehensive dynamical model, which describes the concentrations of the GEM products depicted in Figure 8.1. Notably, considering the ODE equations in (8.39), (8.41), (8.42), we have

$$\begin{aligned}
\dot{R} &= k_r \left( \frac{P_f}{P_f + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f}{P_f + \theta_P^{r2}} \right) d_r - (\mu + \gamma_r) R \\
\dot{p} &= k_p \left( \frac{P_f}{P_f + \theta_P^{p}} \right) d_p - (\mu + \gamma_p) p \\
\dot{P} &= k_P \left( \frac{R_f}{R_f + \theta_R^{P}} \right) p - (\mu + \gamma_P) P \\
\dot{b} &= k_b \left( \frac{P_f}{P_f + \theta_P^{b}} \right) h^-(\mu; \theta_\mu^b, n_b) d_b - (\mu + \gamma_b) b \\
\dot{B} &= k_B \left( \frac{R_f}{R_f + \theta_R^{B}} \right) b - (\mu + \gamma_B) B
\end{aligned} \tag{8.43}$$

where $P_f$ and $R_f$, denoting free RNAP and free ribosomes respectively, have to fulfil the following conservation equations:

$$\begin{aligned}
P &= P_f + P^r + P^p + P^b + P^{ns} \\
R &= R_f + R^p + R^b
\end{aligned} \tag{8.44}$$

where $P^r$, $P^p$, $R^p$, $P^b$, $R^b$, have been defined in (8.40), (8.41), (8.42); whereas $P^{ns} = \left( \frac{P_f}{P_f + \theta_P^{ns}} \right) d_{ns}$ models RNAP concentration that binds to non-specific binding sites present on the DNA [113], reducing in this way the available RNAP for the transcription process.

Hence, replacing expressions (8.40), (8.41), (8.42), into (8.44), we have:

$$
\begin{aligned}
P = P_f &+ \left(\frac{P_f}{P_f+\theta_P^{r1}}\cdot h^+(\mu,\theta_\mu^r,n_r) + \frac{P_f}{P_f+\theta_P^{r2}}\right) d_r\left(1+L_r\frac{k_r}{c_r}\right) \\
&+ \left(\frac{P_f}{P_f+\theta_P^{p}}\right) d_p\left(1+L_p\frac{k_p}{c_p}\right) \\
&+ \left(\frac{P_f}{P_f+\theta_P^{b}}\right) h^-(\mu;\theta_\mu^b,n_b) d_b\left(1+L_b\frac{k_b}{c_b}\right) \\
&+ \left(\frac{P_f}{P_f+\theta_P^{ns}}\right) d_{ns}
\end{aligned}
$$

$$
\begin{aligned}
R = R_f &\\
&+ \frac{R_f}{R_f+\theta_R^P}\left\{(L_p-1)\frac{k_p}{c_p}\left(\frac{P_f}{P_f+\theta_P^p}\right) d_p\left[1+\frac{k_P}{c_P}\left(1+\frac{M_P}{2}\right)\right] + \left(1+M_P\frac{k_P}{c_P}\right)p\right\} \\
&+ \frac{R_f}{R_f+\theta_R^B}\left\{(L_b-1)\frac{k_b}{c_b}\left(\frac{P_f}{P_f+\theta_P^b}\right) h^-(\mu;\theta_\mu^b,n_b) d_b\left[1+\frac{k_B}{c_B}\left(1+\frac{M_B}{2}\right)\right] + \left(1+M_B\frac{k_B}{c_B}\right)b\right\}
\end{aligned}
\tag{8.45}
$$

## 8.5 Model calibration

In this section we will discuss how the parameter values of model (8.43) with (8.45) have been identified. Notably, some parameters have been fixed to values given in the literature (Section 8.5.2), some others have been calculated using biological experimental formulas (Section 8.5.3), and the remaining ones have been identified using steady state experimental data from the literature (Section 8.5.4).

### 8.5.1 Experimental data

The experimental data we used to calibrate the GEM model (8.43) with (8.45) were taken from [31], where macromolecular composition of exponentially growing *E. coli B/r* were measured at 37 ℃for five growth rate values. Here we recall that when bacteria are in the exponential phase, growth rate $\mu$ is constant, and macromolecular composition of the cells is supposed to be at steady state.

For the sake of clarity, in Table 8.5 are reported the experimental data as they appear in [31].

We note that data in Table 8.5 are not expressed in consistent physical units for direct calibration of model (8.43) with (8.45), since they have first to be converted in molar concentration [M]. To do that, we need to know the cell volume ($V_c$) at each $\mu$ reported

| | | | Growth rate $\mu$ [1/h] | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Symbol | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 |
| RNAP molecules/cell | $N_P$ | $\#/cell$ | 1500 | 2800 | 5000 | 8000 | 11400 |
| ribosomes/cell | $N_R$ | $\#/cell$ | 6800 | 13500 | 26300 | 45100 | 72000 |
| protein/cell | $N_B$ | $10^8 aa/cell$ | 5.6 | 8.7 | 13 | 18.9 | 25 |
| mRNA synth. rate/cell | $r_m$ | $10^5 Nu/min/cell$ | 4.3 | 9.2 | 13.7 | 18.7 | 23.4 |
| sRNA synth. rate/cell | $r_s$ | $10^5 Nu/min/cell$ | 3 | 9.9 | 29 | 66.4 | 132.5 |

TABLE 8.5: Data taken from [31]. We note that in [31] growth rate $\mu$ in expressed in [doublings/h] while here it has been converted in [1/h] considering the conversion $[1/h] = \ln(2) \cdot [doublings/h]$. Symbols $\#$, $aa$, $Nu$ denote number of molecules, amino acid, nucleotide, respectively.

in Table 8.5, the protein length ($L_B$) in amino-acids (aa), the mRNA length ($L_b$) in nucleotides (Nu) and stable-RNA length ($L_r$) in nucleotides (Nu).

To calculate the cell volume, which is a growth-rate-dependent parameter, we interpolated a quadratic function to *E. coli* cell volume data from [168], measured at five other different growth rate values, so as to derive *E. coli* cell volumes at growth rates of interest indicated in Table 8.5. A quadratic interpolation has been adopted since it was the one giving smaller errors. The interpolation of cell volume data from [168] is depicted in Figure 8.2, whereas interpolated cell volume values—at growth rate of interest—are reported in Table 8.6.

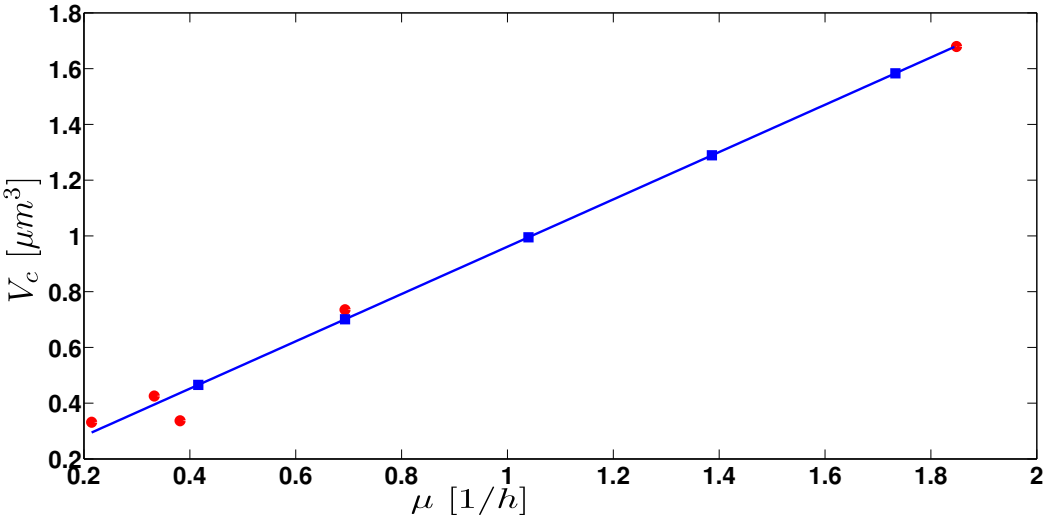



FIGURE 8.2: Cell volume plotted against the growth rate. Red circles represent data from [168]; blue curve is the quadratic interpolated function: $V_c = 0.0002\,\mu^2 + 0.8482\,\mu + 0.1129$; blue squares are $V_v$ at $\mu$ of interest in Table 8.5.

Regarding stable-RNA length, mRNA length and protein length we assumed $L_r = 6623\ Nu$, $L_b = 1000\ Nu$ and $L_B = 333\ aa$ as reported in [159]. In particular it is

worth noting that $L_b = 1000\ Nu$ and $L_B = 333\ aa$ represent average *E. coli* mRNA and protein length respectively.

Hence we can now convert data in Table 8.5 in $[\mu M]$ units using the following expressions:

$$\begin{aligned}
[N_P] &= \frac{N_P}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot N_a} \cdot 10^6 \\
[N_R] &= \frac{N_R}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot N_a} \cdot 10^6 \\
[N_B] &= \frac{N_B}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot L_B \cdot N_a} \cdot 10^6 \\
[r_m] &= \frac{r_m}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot L_b \cdot N_a} \cdot 60\ [min/h] \cdot 10^6 \\
[r_s] &= \frac{r_s}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot L_r \cdot N_a} \cdot 60\ [min/h] \cdot 10^6
\end{aligned} \tag{8.46}$$

where $N_a = 6.022 \cdot 10^{23}\ [molecules/mol]$ is the Avogadro constant.

In Table 8.6 are reported the conversion of data in Table 8.5 using formulas (8.46).

| | | | Growth rate $\mu$ [1/h] | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Symbol | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 |
| Cell Volume | $V_c$ | $\mu m^3$ | 0.47 | 0.70 | 0.99 | 1.29 | 1.58 |
| RNAP concentration | $[N_P]$ | $\mu M$ | 5.35 | 6.63 | 8.34 | 10.31 | 11.96 |
| ribosomes concentration | $[N_R]$ | $\mu M$ | 24.25 | 31.98 | 43.89 | 58.10 | 75.52 |
| protein concentration | $[N_B]$ | $\mu M$ | 5996.5 | 6189.6 | 6515.3 | 7311.1 | 7874.1 |
| mRNA synthesis rate | $[r_m]$ | $\mu M/h$ | 92.00 | 130.77 | 137.18 | 144.53 | 147.25 |
| stable RNA synthesis rate | $[r_s]$ | $\mu M/h$ | 9.69 | 21.25 | 43.85 | 77.49 | 125.90 |

TABLE 8.6: Data in Table 8.5 after conversion in $\mu M$ using formulas (8.46).

Data in Table 8.6 will then be used (Section 8.5.4) to estimate unknown parameters of model 8.43-8.45, but let us first see how most of the model parameters have either been fixed to literature values, or calculated using genetic informations and experimental biological formulas.

### 8.5.2 Parameters taken from literature

In the last decades much research has focused on the mutual influence between bacterial gene expression and global regulation of cell growth rate [31, 113, 124, 151, 159]. These research works allowed us to derive some of the parameter values used for the GEM model (8.43)-(8.45).

Notably, some of the model parameter values taken from the literature are growth-rate-independent, i.e. their values do not change when $\mu$ changes, while others are growth-rate-dependent, i.e. they assume different values for each of the $\mu$ values reported in Table 8.6.

Hence, in Table 8.7 are reported growth-rate-independent parameter values taken from literature with their references, whereas in Table 8.8 are presented the growth-rate-dependent parameter values at $\mu$ of interest with references whereby they have been deduced.

| Parameters | Units | Values | Notes |
|---|---|---|---|
| $k_r$ | $1/h$ | 6600 | from [159] |
| $k_b$ | $1/h$ | 600 | from [113] |
| $k_p$ | $1/h$ | 600 | assumed to be equal to $k_b$ |
| $k_B$ | $1/h$ | 4800 | from [159] |
| $k_P$ | $1/h$ | 4800 | assumed to be equal to $k_B$ |
| $c_r$ | $Nu/h$ | 306000 | from [31] |
| $L_r$ | $Nu$ | 6623 | from [159] |
| $L_b$ | $Nu$ | 1000 | from [159] |
| $L_p$ | $Nu$ | 8253 | from EcoCyc [106] |
| $L_B$ | $aa$ | 333 | from [159] |
| $L_P$ | $aa$ | 2749 | from EcoCyc [106] |

TABLE 8.7: Growth-rate-independent parameter values based on literature.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $c_b$ | $Nu/h$ | 140400 | 162000 | 180000 | 187200 | 198000 | from [31] |
| $c_p$ | $Nu/h$ | 140400 | 162000 | 180000 | 187200 | 198000 | $c_p = c_b$ (assumed) |
| $c_B$ | $aa/h$ | 43200 | 57600 | 64800 | 72000 | 75600 | from [31] |
| $c_P$ | $aa/h$ | 43200 | 57600 | 64800 | 72000 | 75600 | $c_P = c_B$ (assumed) |

TABLE 8.8: Growth-rate-dependent parameter values taken from literature.

### 8.5.3 Calculated growth-rate-dependent parameters

Other growth-rate-dependent parameters are the promoter concentrations denoted by $d_r$, $d_p$, $d_b$, $d_{ns}$ in (8.43)-(8.45). To calculate these quantities we have first to introduce the concepts of *average DNA per cell* and *individual gene copy number per cell*.

**Average DNA per cell:** In an exponential culture, the average number of DNA per cell or the genome equivalent per cell is given by the *Helmstetter and Cooper* equation [31]:

$$G_c = \frac{\tau}{C \cdot \ln(2)} \times \left(2^{\frac{C+D}{\tau}} - 2^{\frac{D}{\tau}}\right) \tag{8.47}$$

where $C$, $D$ denote the time required to replicate chromosome (in [min]) and the interval time (in [min]) between chromosomal replication and cell division, respectively. Both $C$, $D$ are measured growth-dependent parameters [31] whose values for $\mu$ of interest are reported in Table 8.9. Whereas $\tau$ denote time required for cell division (in [min]) and it is related to growth rate $\mu$ (in [1/h]) by the following formula:

$$\tau = \frac{\ln(2)}{\mu} \cdot 60 \ [min/h]. \tag{8.48}$$

| | | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|---|
| Descriptions | Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| Division time | $\tau$ | $min$ | 100 | 60 | 40 | 30 | 24 | from [31] |
| Replication period | $C$ | $min$ | 67 | 50 | 45 | 43 | 42 | from [31] |
| Period from termination of replication to cell division | $D$ | $min$ | 30 | 27 | 25 | 24 | 23 | from [31] |

TABLE 8.9: Growth-rate-dependent cell cycle parameter values taken from [31].

**Individual gene copy number per cell:** In an exponential culture the number of copies of a gene $x$ is calculated by the following equation [31]:

$$N_x = 2^{[C(1-m')+D]/\tau} \tag{8.49}$$

where $m'$ is the map location of gene $x$ relative to location or replication origin:

$$m' = \begin{cases} (m+16)/50 & \text{if } 0 \leq m < 36 \\ (84–m)/50 & \text{if } 36 \leq m < 84 \\ (m–84)/50 & \text{if } 84 \leq m < 100 \end{cases} \tag{8.50}$$

while $m$ denotes the absolute map location of gen $x$.

Having presented the concepts of *average DNA per cell* and *individual gene copy number per cell* with their respective formulas, we can now calculate the promoter concentrations of each gene class, i.e. $d_r$, $d_p$, $d_b$, and the concentration of not specific binding sites $d_{ns}$.

#### 8.5.3.1 Promoter concentration of *rnn* operon

The *rnn* operon is made up of seven genes and their $m$ and $m'$ values taken from [31] are reported in Table 8.10 .

| *rnn* operon | | |
|---|---|---|
| Gene $x$ | $m$ | $m'$ |
| *rnnA* | 87 | 0.06 |
| *rnnB* | 89.5 | 0.11 |
| *rnnC* | 85 | 0.02 |
| *rnnD* | 72 | 0.24 |
| *rnnE* | 90.5 | 0.13 |
| *rnnG* | 57 | 0.54 |
| *rnnH* | 5 | 0.42 |

TABLE 8.10: Absolute map location ($m$) and relative map location($m'$) of the seven *rnn* genes. Values taken from [31].

Using formula (8.49) with values in Tables 8.9-8.10 we can calculate the gene copy number per cell of operon *rnn* ($N_{rnn}$) at growth rates of interest. Eventually, the promoter concentration $d_r$ (in $[\mu M]$) of operon *rnn* is given by:

$$d_r = \frac{N_{rnn}}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot N_a} \cdot 10^6 \tag{8.51}$$

$N_{rnn}$ and $d_r$ values at growth rates of interest are reported in Table 8.11.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $N_{rnn}$ | $genes/cell$ | 12.44 | 15.11 | 20.07 | 26.94 | 35.96 | calculated |
| $d_r$ | $\mu M$ | 0.0444 | 0.0358 | 0.0335 | 0.0347 | 0.0377 | calculated |

TABLE 8.11: Growth-rate-dependent *rnn* copy number and concentration.

#### 8.5.3.2 Promoter concentration of *rpoBC* genes

In Table 8.12 are reported the $m$ and $m'$ values of *rpoBC* genes. The *rpoBC* $m$ value has been taken from [130].

Using formula (8.49) with values in Tables 8.9-8.12 we can calculate the gene copy number per cell of *rpoBC* ($N_{rpoBC}$) at growth rates of interest. Eventually the promoter concentration $d_p$ (in $[\mu M]$) of *rpoBC* is given by:

$$d_p = \frac{N_{rpoBC}}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot N_a} \cdot 10^6 \tag{8.52}$$

| *rpoBC* genes | | |
|---|---|---|
| Gene $x$ | $m$ | $m'$ |
| *rpoBC* | 90 | 0.12 |

TABLE 8.12: Absolute map location ($m$) and relative map location($m'$) of the *rpoBC* genes. Values taken from [130].

$N_{rpoBC}$ and $d_p$ values at growth rates of interest are reported in Table 8.13.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $N_{rpoBC}$ | *genes/cell* | 1.85 | 2.27 | 3.06 | 4.17 | 5.65 | calculated |
| $d_p$ | $\mu M$ | 0.0066 | 0.0054 | 0.0051 | 0.0054 | 0.0059 | calculated |

TABLE 8.13: Growth-rate-dependent *rpoBC* copy number and concentration.

#### 8.5.3.3 Promoter concentration of *bulk* genes

The *bulk* gene class is a consolidation of the 4288 identified protein encoding-genes in *E. coli* [25]. Here, we assume that these 4288 genes are uniformly distributed across the chromosome. This implies that:

$$m'_i = i/4288, \quad \text{for } i = 0, 1, 2, ..., 4287$$

and the gene copy number per cell is given by:

$$N_b = \sum_{i=0}^{4287} 2^{[C(1-m'_i)+D]/\tau} = 2^{(C+D)/\tau} \frac{1 - 2^{-C/\tau}}{1 - 2^{-C/(4288 \cdot \tau)}} \tag{8.53}$$

Using formula (8.53) with values in Tables 8.9 we can calculate the gene copy number per cell of *bulk* genes ($N_b$) at growth rates of interest. Eventually, the promoter concentration $d_p$ (in $[\mu M]$) of *bulk* genes is given by:

$$d_b = \frac{N_b}{V_c \cdot 10^{-15} [L/\mu m^3] \cdot N_a} \cdot 10^6 \tag{8.54}$$

where $N_b$ and $d_b$ values at growth rates of interest are reported in Table 8.14.

#### 8.5.3.4 Promoter concentration of non-specific binding sites

RNAP could also bind to non-specific binding sites on the DNA, because the number of sites for non-specific binding greatly exceeds the number of promoters [113]. In theory,

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $N_b$ | $genes/cell$ | 6719.36 | 7928.62 | 10016.49 | 12781.57 | 16237.20 | calculated |
| $d_b$ | $\mu M$ | 23.96 | 18.78 | 16.72 | 16.46 | 17.03 | calculated |

TABLE 8.14: Growth-rate-dependent values of *bulk* copy number and concentration.

each nucleotide of the DNA may be a (non-specific) binding site of RNAP, and when RNAP is bound to these non-specific binding sites not only transcription does not start, but also the pool of free RNAP is reduced.

The number of non-specific binding sites per cell $N_{ns}$ is given by [113]:

$$N_{ns} = g\ G_c$$

where $g = 4.6 \cdot 10^6$ is the genome size ([113]) and $G_c$ is genome equivalent per cell expressed in 8.47. Finally the non-specific promoter concentration $d_{ns}$ (in $[\mu M]$) is given by:

$$d_{ns} = \frac{N_{ns}}{V_c \cdot 10^{-15}[L/\mu m^3] \cdot N_a} \cdot 10^6. \tag{8.55}$$

In Table 8.15 are reported the calculated values of $G_c$, $N_{ns}$ and $d_{ns}$ at growth rates of interest.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $G_c$ | $DNA/cell$ | 1.6 | 1.8 | 2.3 | 3.0 | 3.8 | calculated |
| $N_{ns}$ | $10^6\ bs/cell$ | 7.3 | 8.28 | 10.58 | 13.8 | 17.48 | calculated |
| $d_{ns}$ | $\mu M$ | 26244.18 | 19616.19 | 17657.09 | 17776.45 | 18333.50 | calculated |

TABLE 8.15: Growth-rate-dependent values of non-specific promoter concentration.

Finally, to recap, all calculated growth rate parameter values are reported in Table 8.16.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $d_r$ | $\mu M$ | 0.0444 | 0.0358 | 0.0335 | 0.0347 | 0.0377 | calculated |
| $d_p$ | $\mu M$ | 0.0066 | 0.0054 | 0.0051 | 0.0054 | 0.0059 | calculated |
| $d_b$ | $\mu M$ | 23.96 | 18.78 | 16.72 | 16.46 | 17.03 | calculated |
| $d_{ns}$ | $\mu M$ | 26244.18 | 19616.19 | 17657.09 | 17776.45 | 18333.50 | calculated |

TABLE 8.16: Calculated growth-rate-dependent parameter values of promoter concentrations.

### 8.5.4 Estimated parameters

This section deals with estimating the remaining model parameters from experimental data reported in Table 8.6. Notably, the remaining parameters to be estimated are grouped in vector $\phi$, which reads as:

$$\phi = [\theta_P^{r1},\ \theta_\mu^r,\ n_r,\ \theta_P^{r2},\ \gamma_r,\ \theta_P^p,\ \gamma_p,\ \theta_R^P,\ \gamma_P,\ \theta_P^b,\ \theta_\mu^b,\ n_b,\ \theta_R^B,\ \gamma_B,\ \theta_P^{ns}] \tag{8.56}$$

where parameters in (8.56) are assumed to be growth rate independent.

In general, the problem of estimating the unknown parameters of a model can be formulated as a minimization problem, where the theoretical model predictions are compared to the experimental data, and the parameter values are adjusted in order to minimize the distance between the two. Hence, with this in mind and since experimental data are steady state measurements (at five given growth rate values) of ribosome, RNAP, protein concentrations and mRNA and sRNA synthesis rates (see Table 8.6), we have to derive the same theoretical quantities from model (8.43)-(8.45). This means that we need to calculate steady state variables of model (8.43)-(8.45), and the theoretical mRNA synthesis rate $\overline{r}_m$ and sRNA synthesis rate $\overline{r}_s$. However, let us first introduce the vector of steady state growth rate $\overline{\mu} = [\mu_1, ..., \mu_5]$, where $\mu_i$ $(i = 1, ..., 5)$ is the $i-th$ growth rate value at which measurements in Table 8.6 are taken. Hence, for each

$i = 1, ..., 5$, the steady state model predictions read as:

$$
\begin{aligned}
\overline{R}_i &= \frac{k_r}{\mu + \gamma_r}\left(\frac{P_f^i}{P_f^i + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f^i}{P_f^i + \theta_P^{r2}}\right) d_r^i \\
\overline{p}_i &= \frac{k_p}{\mu + \gamma_p}\left(\frac{P_f^i}{P_f^i + \theta_P^{p}}\right) d_p^i \\
\overline{P}_i &= \frac{k_P}{\mu + \gamma_P}\left(\frac{R_f^i}{R_f^i + \theta_R^{P}}\right) \overline{p}_i \\
\overline{b}_i &= \frac{k_b}{\mu + \gamma_b}\left(\frac{P_f^i}{P_f^i + \theta_P^{b}}\right) h^-(\mu; \theta_\mu^b, n_b) d_b^i \\
\overline{B}_i &= \frac{k_B}{\mu + \gamma_B}\left(\frac{R_f^i}{R_f^i + \theta_R^{B}}\right) \overline{b}_i
\end{aligned}
$$

$$
\begin{aligned}
\overline{P}_i = P_f^i &+ \left(\frac{P_f^i}{P_f^i + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f^i}{P_f^i + \theta_P^{r2}}\right) d_r^i \left(1 + L_r \frac{k_r}{c_r}\right) \\
&+ \left(\frac{P_f^i}{P_f^i + \theta_P^{p}}\right) d_p^i \left(1 + L_p \frac{k_p}{c_p^i}\right) \\
&+ \left(\frac{P_f^i}{P_f^i + \theta_P^{b}}\right) h^-(\mu; \theta_\mu^b, n_b) d_b^i \left(1 + L_b \frac{k_b}{c_b^i}\right) \\
&+ \left(\frac{P_f^i}{P_f^i + \theta_P^{ns}}\right) d_{ns}^i
\end{aligned}
$$

$$
\begin{aligned}
\overline{R}_i = R_f^i &\\
&+ \frac{R_f^i}{R_f^i + \theta_R^{P}} \left\{ (L_p - 1) \frac{k_p}{c_p^i} \left(\frac{P_f^i}{P_f^i + \theta_P^{p}}\right) d_p^i \left[1 + \frac{k_P}{c_P^i}\left(1 + \frac{M_P}{2}\right)\right] + \left(1 + M_P \frac{k_P}{c_P^i}\right) \overline{p}_i \right\} \\
&+ \frac{R_f^i}{R_f^i + \theta_R^{B}} \left\{ (L_b - 1) \frac{k_b}{c_b^i} \left(\frac{P_f^i}{P_f^i + \theta_P^{b}}\right) h^-(\mu; \theta_\mu^b, n_b) d_b^i \left[1 + \frac{k_B}{c_B^i}\left(1 + \frac{M_B}{2}\right)\right] + \left(1 + M_B \frac{k_B}{c_B^i}\right) \overline{b}_i \right\}
\end{aligned}
$$

$$
\begin{aligned}
\overline{r}_s^i &= k_r \left(\frac{P_f^i}{P_f^i + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f^i}{P_f^i + \theta_P^{r2}}\right) d_r^i \\
\overline{r}_m^i &= k_b \left(\frac{P_f^i}{P_f^i + \theta_P^{b}}\right) h^-(\mu; \theta_\mu^b, n_b) d_b^i
\end{aligned}
\tag{8.57}
$$

We note that index $i$ indicates that model predictions (8.57) are calculated considering growth rate dependent parameters at their $i - th$ growth rate values (see Tables 8.8 and 8.16).

Now, to quantify the distance between the model predictions in (8.57) and experimental data in Table 8.6, calculate the cost function $J(\phi)$:

$$J(\phi) = \sum_i^n \left( \frac{\overline{R}_i(\phi) - [N_R]_i}{0.1 \cdot [N_R]_i} \right)^2 + \sum_i^n \left( \frac{\overline{P}_i(\phi) - [N_P]_i}{0.1 \cdot [N_P]_i} \right)^2 + \sum_i^n \left( \frac{\overline{B}_i(\phi) - [N_B]_i}{0.1 \cdot [N_B]_i} \right)^2 + \sum_i^n \left( \frac{\bar{r}_s^i(\phi) - [r_s]_i}{0.1 \cdot [r_s]_i} \right)^2 + \sum_i^n \left( \frac{\bar{r}_m^i(\phi) - [r_m]_i}{0.1 \cdot [r_m]_i} \right)^2 \tag{8.58}$$

where $n = 5$ is the number of experimental data points, one for each growth rate values in Table 8.6.

To find the parameters in $\phi$ that minimize $J(\phi)$ we used an optimization approach written in $MATLAB^{®}$ which first uses the *GlobalSearch* function and then, to refine the estimation, the *pattern search* function.

Hence, the estimated parameter values are reported in Table 8.17.

| Parameters | Units | Values |
|---|---|---|
| $\theta_P^{r1}$ | $\mu M$ | 16.3938 |
| $\theta_\mu^r$ | $1/h$ | 1.2106 |
| $n_r$ | – | 7.4251 |
| $\theta_P^{r2}$ | $\mu M$ | 45.3137 |
| $\gamma_r$ | $1/h$ | 0.0344 |
| $\theta_P^p$ | $\mu M$ | 5.9876 |
| $\gamma_p$ | $1/h$ | 12.6818 |
| $\theta_R^P$ | $\mu M$ | 124.3815 |
| $\gamma_P$ | $1/h$ | 0.2179 |
| $\theta_P^b$ | $\mu M$ | 107.1302 |
| $\theta_\mu^b$ | $1/h$ | 0.2291 |
| $n_b$ | – | 0.7424 |
| $\gamma_b$ | $1/h$ | 20.7052 |
| $\theta_R^B$ | $\mu M$ | 29.8494 |
| $\gamma_B$ | $1/h$ | 0.0330 |
| $\theta_P^{ns}$ | $\mu M$ | 7739.7769 |

TABLE 8.17: Estimated parameter values.

In Figure 8.3 are shown the fitting results of steady state model (8.43) and conservation equations (8.45) using parameter values in Table 8.18. As we can notice, all model predictions are within the 95% confidence intervals of measurements. Measurement 95% confidence intervals are calculated considering 10% measurement errors as stated in [31].

To recap, in Table 8.18 are reported all parameter values of model (8.43) and conservation equations (8.45).

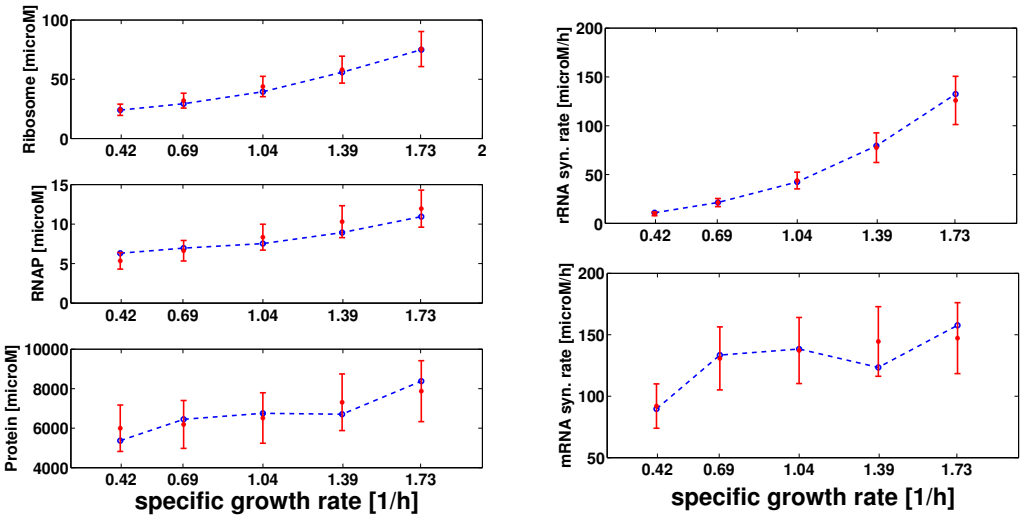



FIGURE 8.3: Model fitting. Red circles denote experimental data [31] after molar conversion (see Table 8.6); blue open circles are model predictions with parameter values from Table 8.18. Measurement 95% confidence intervals are calculated considering 10% measurement errors as stated in [31].

| $R$ | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $L_r$ | $Nu$ | 6623 | 6623 | 6623 | 6623 | 6623 | from [159] |
| $c_r$ | $Nu/h$ | 306000 | 306000 | 306000 | 306000 | 306000 | from [31] |
| $d_r$ | $\mu M$ | 0.0082 | 0.0083 | 0.0094 | 0.0113 | 0.0142 | calculated |
| $k_r$ | $1/h$ | 6600 | 6600 | 6600 | 6600 | 6600 | from [159] |
| $\theta_P^{r1}$ | $\mu M$ | 16.3938 | 16.3938 | 16.3938 | 16.3938 | 16.3938 | estimated |
| $\theta_\mu^r$ | $1/h$ | 1.2106 | 1.2106 | 1.2106 | 1.2106 | 1.2106 | estimated |
| $n_r$ | – | 7.4251 | 7.4251 | 7.4251 | 7.4251 | 7.4251 | estimated |
| $\theta_P^{r2}$ | $\mu M$ | 45.3137 | 45.3137 | 45.3137 | 45.3137 | 45.3137 | estimated |
| $\gamma_r$ | $1/h$ | 0.0344 | 0.0344 | 0.0344 | 0.0344 | 0.0344 | estimated |
| $p$ | | Growth rate $\mu$ [1/h] | | | | | |
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $L_p$ | $Nu$ | 8253 | 8253 | 8253 | 8253 | 8253 | from EcoCyc [106] |
| $c_p$ | $Nu/h$ | 140400 | 162000 | 180000 | 187200 | 198000 | $c_p = c_b$ (assumed) |
| $d_p$ | $\mu M$ | 0.0066 | 0.0054 | 0.0051 | 0.0054 | 0.0059 | calculated |
| $k_p$ | $1/h$ | 600 | 600 | 600 | 600 | 600 | $k_p = k_b$ (assumed) |
| $\theta_P^p$ | $\mu M$ | 5.9876 | 5.9876 | 5.9876 | 5.9876 | 5.9876 | estimated |
| $\gamma_p$ | $1/h$ | 12.6818 | 12.6818 | 12.6818 | 12.6818 | 12.6818 | estimated |
| $P$ | | Growth rate $\mu$ [1/h] | | | | | |

| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
|---|---|---|---|---|---|---|---|
| $L_P$ | $aa$ | 2749 | 2749 | 2749 | 2749 | 2749 | from EcoCyc [106] |
| $c_P$ | $aa/h$ | 43200 | 57600 | 64800 | 72000 | 75600 | $c_P = c_B$ (assumed) |
| $k_P$ | $1/h$ | 4800 | 4800 | 4800 | 4800 | 4800 | $k_P = k_B$ (assumed) |
| $\theta_R^P$ | $\mu M$ | 124.3815 | 124.3815 | 124.3815 | 124.3815 | 124.3815 | estimated |
| $\gamma_P$ | $1/h$ | 0.2179 | 0.2179 | 0.2179 | 0.2179 | 0.2179 | estimated |
| $\theta_P^{ns}$ | $\mu M$ | 7739.7769 | 7739.7769 | 7739.7769 | 7739.7769 | 7739.7769 | estimated |
| $d_{ns}$ | $\mu M$ | 26244.18 | 19616.19 | 17657.09 | 17776.45 | 18333.50 | calculated |
| $b$ | | | | Growth rate $\mu$ [1/h] | | | |
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $L_b$ | $Nu$ | 1000 | 1000 | 1000 | 1000 | 1000 | from [159] |
| $c_b$ | $Nu/h$ | 140400 | 162000 | 180000 | 187200 | 198000 | from [31] |
| $d_b$ | $\mu M$ | 23.96 | 18.78 | 16.72 | 16.46 | 17.03 | calculated |
| $k_b$ | $1/h$ | 600 | 600 | 600 | 600 | 600 | from [113] |
| $\theta_P^b$ | $\mu M$ | 107.1302 | 107.1302 | 107.1302 | 107.1302 | 107.1302 | estimated |
| $\theta_\mu^b$ | $1/h$ | 0.2291 | 0.2291 | 0.2291 | 0.2291 | 0.2291 | estimated |
| $n_b$ | – | 0.7424 | 0.7424 | 0.7424 | 0.7424 | 0.7424 | estimated |
| $\gamma_b$ | $1/h$ | 20.7052 | 20.7052 | 20.7052 | 20.7052 | 20.7052 | estimated |
| $B$ | | | | Growth rate $\mu$ [1/h] | | | |
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $L_B$ | $aa$ | 333 | 333 | 333 | 333 | 333 | from [159] |
| $c_B$ | $aa/h$ | 43200 | 57600 | 64800 | 72000 | 75600 | from [31] |
| $k_B$ | $1/h$ | 4800 | 4800 | 4800 | 4800 | 4800 | from [159] |
| $\theta_R^B$ | $\mu M$ | 29.8494 | 29.8494 | 29.8494 | 29.8494 | 29.8494 | estimated |
| $\gamma_B$ | $1/h$ | 0.0330 | 0.0330 | 0.0330 | 0.0330 | 0.0330 | estimated |

TABLE 8.18: All parameter values of model (8.43) and conservation equations (8.45).

## 8.6 Free RNAP and Free ribosomes

We can use model (8.43) at steady state with (8.45) and the parameter values in Table 8.18 as a tool to calculate free concentrations of RNAP and ribosomes which are difficult to measure in practice.

In Table 8.19 are reported the free concentrations of RNAP and ribosomes at growth rate of interest.

| | | Growth rate $\mu$ [1/h] | | | | | |
|---|---|---|---|---|---|---|---|
| Parameters | Units | 0.42 | 0.69 | 1.04 | 1.39 | 1.73 | Notes |
| $P_f$ | $\mu M$ | 1.7362 | 4.3230 | 6.3803 | 6.8421 | 9.9190 | prediction |
| $R_f$ | $\mu M$ | 4.0206 | 5.5573 | 9.3218 | 16.5172 | 23.4799 | prediction |

TABLE 8.19: Growth-rate-dependent Free RNAP and ribosomes predictions using model (8.43) with (8.45) at steady state and the parameter values from Table 8.18.

## 8.7 Model reduction

Model (8.43) can be reduced by considering a quasi-steady-state approximation for the mRNA variables $p$ and $b$. In fact, since $g_p \gg g_P$ and $g_b \gg g_B$ we can assume that $\dot{p} \approx 0$ and $\dot{b} \approx 0$. Hence, the reduced model reads as:

$$\begin{aligned}
\dot{R} &= k_r \left( \frac{P_f}{P_f + \theta_P^{r1}} \cdot h^+(\mu, \theta_\mu^r, n_r) + \frac{P_f}{P_f + \theta_P^{r2}} \right) d_r - (\mu + \gamma_r) R \\
p &= \frac{k_p}{\mu + \gamma_p} \left( \frac{P_f}{P_f + \theta_P^p} \right) d_p \\
\dot{P} &= k_P \left( \frac{R_f}{R_f + \theta_R^P} \right) p - (\mu + \gamma_P) P \\
b &= \frac{k_b}{\mu + \gamma_b} \left( \frac{P_f}{P_f + \theta_P^b} \right) h^-(\mu; \theta_\mu^b, n_b) d_b \\
\dot{B} &= k_B \left( \frac{R_f}{R_f + \theta_R^B} \right) b - (\mu + \gamma_B) B
\end{aligned} \tag{8.59}$$

where $P_f$ and $R_f$, denoting free RNAP and free ribosomes respectively, have still to fulfil the conservation equations in (8.45).

## 8.8 Conclusions

In this Chapter, we presented a core dynamical model of the gene expression machinery of the bacterium *E. coli*. In particular, the entire *E.coli* genome has been divided into three main gene classes: *rnn* genes, *rpoBC* genes and the proxy *bulk* genes. As a matter of fact, *rnn* genes account for ribosome synthesis, *rpoBC* genes for that of RNAP, and *bulk* genes determine the production of all cellular proteins. Hence, with these three gene classes, we were able to model the entire transcription-translation apparatus in *E. coli*. Notably, using the formalism presented in [115], we modelled the transcription

regulation due to free RNAP and the translation regulation exerted by free ribosomes, but we also took into account the specific growth rate dependent regulation which affects the sRNA and mRNA synthesis. Notably, we extended the model framework of [115] by introducing two Hill functions, which account for the transcriptional growth rate regulation. More precisely, this growth regulation determines how resources are allocated during the transcription process. This mechanism, which is probably due to the ppGpp molecule, works in the following fashion: at higher growth rate values transcription of *rnn* genes is favoured with respect to that of *bulk* genes, leading to an overproduction of sRNA with respect to mRNA synthesis. This because, at higher growth rates, more ribosomes (sRNA are the limiting factor of ribosome assembling) are needed to speed up the translation of proteins.

Then, the GEM dynamical model, which consists of five variables and two algebraic equations, has been calibrated using some parameter values taken from literature, others calculated using genetic information about gene classes and biological formulas, and finally, the remaining ones, identified using experimental data of *E. coli* macromolecular composition at different steady state growth rate values from [31]. We showed that the calibrated GEM model (all parameter values are reported in Table 8.18) was able to consistently reproduce all the experimental data (see fitting results in Figure 8.3).

However, to give more consistency to the estimated parameters (reported in Table 8.17), an identifiability analysis should be conducted, as well as a validation of the model predictions on an alternative data set should be pursued. But these studies were beyond the scope of this work and they can be addressed in the future.

Moreover, the developed GEM model can also be used as a "tool" to calculate steady states of free RNAP and ribosomes, which are quantities difficult to measure in practice (Section 8.6). A reduced version of original model (8.43) has also been presented, in which mRNAs species have been assumed to be at their quasi steady state equilibrium (Section 8.7).

It is worth saying that, even though model (8.43) is a dynamical model, it is only valid when growth rate $\mu$ is constant, that is during the exponential phase of the bacterial growth. In other growth phases, a dynamical expression of $\mu$ has to be developed (see Chapters 4-6-7 for possible dynamical $\mu$ models), and probably also a recalibration of the model is required, possibly using dynamical experimental data.

To conclude, this model of the GEM of *E. coli* describes central regulatory mechanisms and allowed us to elucidate how the *E. coli* transcription-translation apparatus works on a global scale, pointing out that the transcriptional growth rate regulation plays a fundamental role in the overall regulation of the gene expression machinery. Moreover,

this GEM model could also be used as host-cell model where more complex/specific gene networks are embedded into it, or as a starting point for the design of synthetic genetic manipulations, which implement desired controls of the bacterial GEM.

# Chapter 9

# State estimation for gene networks with intrinsic and extrinsic noise: A case study on *E.coli* arabinose uptake dynamics

The work presented in this chapter is the result of a collaboration I had with Eugenio Cinquemani (Inria Ibis team, Grenoble). A reduced version of this chapter has been presented at the *European Control Conference* (ECC) in 2013 [39].

We address state estimation for gene regulatory networks at the level of single cells. We consider models that include both intrinsic noise, in terms of stochastic dynamics, and extrinsic noise, in terms of random parameter values. We take the Chemical Master Equation (CME) with random parameters as a reference modelling approach, and investigate the use of stochastic differential model approximations for the construction of practical real-time filters. To this aim we consider a Square-Root Unscented Kalman Filter (SRUKF) built on a Chemical Langevin Equation (CLE) approximation of the CME. Using arabinose uptake regulation in *Escherichia coli* bacteria as a case study, we show that performance is comparable to that of a (computationally heavier) particle filter built directly on the CME, and that the use of information about parameter uncertainty allows one to improve state estimation performance.

## 9.1 Introduction

A key player of single-cell gene network dynamics is noise [161]. A distinction is usually made between intrinsic noise, i.e. the uncertainty inherent in biochemical events (binding/unbinding of transcription factors, synthesis of mRNA or protein molecules, etc.), and extrinsic noise, such as the variability of individual features over an isogenic population (abundance of aspecific transcription/translation factors, local environmental conditions, etc.) [138].

Gene expression monitoring techniques with single-cell resolution opened the way to the identification of stochastic gene network models. The CME [147], describing the kinetics of the network species in terms of probabilistic reaction events, is a standard tool for the description of intrinsic noise. To account for extrinsic noise, similar to Mixed-Effects (ME) modelling [52], one approach is to describe the parameters of the network dynamics as random variables taking different values in different individuals (see e.g. [93]). Stochastic gene network models are nowadays fundamental tools for understanding the behaviour of cells in face of environmental and evolutionary challenges [137]. Most recently, they have also been considered for the real-time computer-based control of gene expression in single cells [128, 169].

This paper investigates state estimation from cell-level measurements for networks with intrinsic and extrinsic noise. State estimation is interesting *per se* for the reconstruction of network states that cannot be measured directly. In addition, it can be used as an intermediate step for identification, and plays a central role toward model-based control.

We start from the CME as the reference ("true") model of a cell network. Inspired by the ME approach [52, 93], we include extrinsic noise in terms of variability of the model parameters. Since CME appears to be impractical for real-time filtering, we propose to use an asymptotic approximation, the Chemical Langevin Equation (CLE) [79], to implement filtering. First, we compare simulations of the CME and CLE models. Then, we use the latter to construct a Square-Root Unscented Kalman Filter (SRUKF) [101, 150, 156]. Using data generated from the true (CME) system, we compare performance of SRUKF with that of a (computationally heavier) particle filter built directly on the CME [63].

We develop our work on the case study of the network regulating the uptake of sugar arabinose in bacteria *Escherichia coli*. While relatively simple, this well characterized system (see [125] and references therein) is representative of the nature and complexity of the genetic feedback mechanisms regulating bacterial response to environmental stress. Different from e.g. [93, 182], where the observations consist of time series of the empirical distribution of gene expression obtained via flow cytometry, we consider the case where

the expression in every cell of a small population is observed over time, as it can be obtained e.g. by fluorescence microscopy (see e.g. [169]).

Bayesian inference, such as parameter and state estimation, for biological networks has been considered before, see e.g. [78, 178]. Here, we focus on state estimation under parameter uncertainty. First, we show that, despite the known limitations for small molecule numbers, the CLE is a viable CME approximation for the construction of computationally affordable filters coping with intrinsic noise (stochastic dynamics) and extrinsic noise (random parameters). Second, we show that the use of ME-type models, accounting for parameter variability, may improve state estimation performance.

## 9.2 Stochastic modelling of genetic networks

Consider a biochemical reaction network involving $n$ species and $m$ possible reactions among them. For gene regulatory networks, the species are typically proteins, mRNAs, transcription factors, etc., while reactions are e.g. binding/unbinding events, formation of complexes, degradation, and, at a higher level of abstraction, gene expression.

Assume that the reaction volume is uniform. For cells or cell nuclei, this assumption is still accepted in many contexts, as long as spatial resolution is not central. Let $X = (X_1, \ldots, X_n) \in \mathbb{N}^n$, where $X_i$ denotes the number of elements of species $i$, with $i = 1, \ldots, n$. Let $\nu_j \in \mathbb{Z}^n$ be the stoichiometry of reaction $j$, with $j = 1, \ldots, m$. That is, element $i$ of $\nu_j$, denoted $\nu_{j,i}$ is the number of elements of species $i$ produced or consumed in reaction $j$. Assume that reactions occur stochastically with propensities $a_j \in \mathbb{R}_{\geq 0}$ generally depending on $X$.Then $X$ is the random state vector of a Markovian jump process. For times $t \geq 0$, say, the probability $p(Z,t) = \text{Prob}(X(t) = Z)$, $Z \in \mathbb{N}^n$, obeys the CME [147]

$$\frac{dp(Z,t)}{dt} = \sum_{j=1}^{m} a_j(Z - \nu_j)p(Z - \nu_j, t) - a_j(Z)p(Z,t). \tag{9.1}$$

whose solution is fully determined given the distribution of $X(0)$. The CME is a linear but infinite-dimensional differential equation. For all but the simplest systems, the exact solution cannot be computed in practice. Simulated sample trajectories can be obtained by the Gillespie and related algorithms [147]. Under appropriate conditions on the process $X$, typically satisfied for large numbers $X_i$, the jump process $X$ is well approximated by a continuous process with state $x \in \mathbb{R}^n_{\geq 0}$ that satisfies the so-called

Langevin equation [79], i.e. the system of stochastic differential equations

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{m} \nu_{j,i} a_j\big(x(t)\big) + \nu_{j,i}\sqrt{a_j\big(x(t)\big)}\Gamma_j(t), \tag{9.2}$$

with $i = 1, \ldots, n$, where, for $j = 1, \ldots, m$, the $\Gamma_j(t)$ are mutually uncorrelated white noise processes. Here $x$ plays the role of a continuous approximation of the molecule count $X$. Eq. (9.2) equally describes the evolution of molar concentrations $x = X/(\Omega N_A)$[M], where $\Omega$ is the reaction volume, $N_A$ is Avogadro's number and [M] denotes molar (moles/liter) units, provided appropriate rescaling of the reaction propensities and their parameters. From now on we assume $x = X/(\Omega N_A)$ and omit symbol [M] where no confusion may arise.

**Inter-individual variability** Similar to ME-modelling in pharmacokinetics [52], variability of reaction dynamics among different cells (extrinsic noise) can be described in terms of inter-individual variability of the parameters of a common kinetic model [93]. Using a Bayesian approach, one assumes that individual parameters are concentrated around a known population average (so-called fixed-effects) but deviate from it by a quantity modelled as a random variable with a given prior. This prior, characteristic of the population, is inferred from a set of representative individuals. Then, deviations of new individuals from the population average are treated as random outcomes from the same prior. In our context, let $a_j^\theta$ denote that reaction propensity $a_j$ depends on a parameter vector $\theta$. Let $X^\ell$ (resp. $x^\ell$) and $\theta^\ell$ be the state (resp. the state of the Langevin approximation) and the parameters of the $\ell$th cell in a population of $N$ cells. Then $X^\ell$ (resp. $x^\ell$) evolves according to the dynamics determined by $a_j^{\theta_\ell}$. To model individual variations from population average, we assume that $\theta^\ell, \ldots, \theta^N$ are independent identically distributed (i.i.d.) outcomes of the random variable $\theta$ defined by

$$\theta = \bar{\theta} + \delta, \tag{9.3}$$

where $\bar{\theta}$ is fixed and $\delta$ is a random variable with known distribution $F_\delta$. Provided suitable definition of $F_\delta$, this model includes the case where some entries of $\theta$ are fixed.

**Observation model** We consider the case where noisy measurements from individual cells are collected over time. Let $\mathscr{T} = t_k : k = 0, 1, 2, \ldots$, with $t_k < t_{k+1}$ for all $k$, be a set of measurement times. After standard preprocessing (such as e.g. background removal in fluorescent gene reporter systems) let $y^\ell(t)$ be the measurement at time $t \in \mathscr{T}$

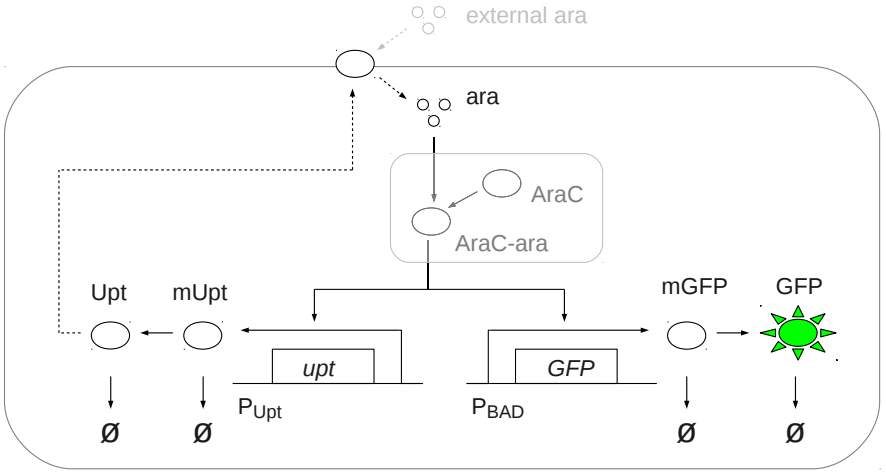



FIGURE 9.1: Arabinose uptake regulatory network in (modified) *E.coli* cells (inspired from [125]). In absence of glucose, global transcriptional regulators enable the expression of arabinose import enzymes, here lumped into a "virtual" enzyme Upt, and metabolization enzyme AraBAD, under the control of promoter $P_{BAD}$. In the modified strains we refer to, gene *araBAD* is deleted from the DNA and replaced by the $P_{BAD}$-controlled *GFP* (Green Fluorescent Protein) gene on plasmids. When arabinose is present, it forms complexes with AraC molecules, which promote transcription of the above genes, thus increasing the amount of arabinose transporters and fluorescent molecules. More transporters imply faster arabinose uptake and hence higher transporter synthesis rate, in a positive feedback loop. Parts in grey will not be modelled explicitly.

for cell $\ell$, with $\ell = 1, \ldots, N$. We assume that

$$y^\ell(t) = CX^\ell(t) + e^\ell(t), \qquad t \in \mathscr{T}, \tag{9.4}$$

where the output matrix $C$, typically selecting one of the system states, is known, and the $e^\ell(t)$ are i.i.d. noise samples of appropriate dimension from a Gaussian distribution $\mathscr{N}(0, R)$, with $R > 0$ known. More specific measurement models depend on the details of the experimental setup and are not pursued in this paper. Under the Langevin approximation, we will replace $X^\ell$ by the process $x^\ell$ from (9.2) and rescale quantities accordingly.

## 9.3 Case study: *E.coli* arabinose uptake dynamics

We are interested in the network that regulates the uptake of arabinose in *Escherichia coli*, a well characterized bacterium. Upon exhaustion of primary environmental carbon sources (glucose), *E.coli* activates adaptation mechanisms triggering the uptake of less favourable carbon sources such as arabinose (see [125], and references therein). A simplified representation of the system is in Figure9.1.

| Synthesis | Rate $a_j$ | Degradation | Rate $a_j$ |
|---|---|---|---|
| $\varnothing \xrightarrow{a_1}$ ara | $v_1 x_3$ | ara $\xrightarrow{a_2} \varnothing$ | $\gamma_1 x_1$ |
| $\varnothing \xrightarrow{a_3}$ mUpt (basal) | $v_2^0$ | mUpt $\xrightarrow{a_5} \varnothing$ | $\gamma_2 x_2$ |
| $\varnothing \xrightarrow{a_4}$ mUpt (regulated) | $v_2 \dfrac{x_1^3}{K_u^3 + x_1^3}$ | | |
| mUpt $\xrightarrow{a_6}$ mUpt+Upt | $v_3 x_2$ | Upt $\xrightarrow{a_7} \varnothing$ | $\gamma_3 x_3$ |
| $\varnothing \xrightarrow{a_8}$ mGFP (basal) | $v_4^0$ | mGFP $\xrightarrow{a_{10}} \varnothing$ | $\gamma_4 x_4$ |
| $\varnothing \xrightarrow{a_9}$ mGFP (regulated) | $v_4 \dfrac{x_1^3}{K_u^3 + x_1^3}$ | | |
| mGFP $\xrightarrow{a_{11}}$ mGFP+GFP | $v_5 x_4$ | GFP $\xrightarrow{a_{12}} \varnothing$ | $\gamma_5 x_5$ |

FIGURE 9.2: Reactions of the stochastic model of the system of Figure9.1 and corresponding propensities. An arrow from (to) symbol $\varnothing$ means synthesis (degradation, including dilution effects due to cell growth).

We consider a model inspired by [125]. The model consists of the $n = 5$ species ara (arabinose), mUpt (Upt messenger RNA), Upt (Upt protein), mGFP (GFP messenger RNA), GFP (GFP protein), interacting via the $m = 12$ "lumped" reactions reported in Figure9.2 with the corresponding propensities $a(x) = \big[a_1(x), \ldots, a_{12}(x)\big]$, where $x = \big[x_1, x_2, x_3, x_4, x_5\big]$ denotes amounts of ara, mUpt Upt, mGFP and GFP, in the same order. Nominal parameter values $\bar{\theta}$, mostly derived from [125], are listed in Figure9.3. In this model, the rate of transcription of the Upt and GFP genes (resp. $a_4$ and $a_9$) is described as a switch-like (Hill) function of the internal arabinose abundance, with a threshold parameter that depends on the concentration of unmodelled regulators (notably AraC, see Figure 9.1 and [125]). The observed variable is a fluorescence level proportional to the amount of GFP, i.e. $C = [0\ 0\ 0\ 0\ K]$, for simplicity we take $K = 1$. The CME and CLE models follow from replacing the stoichiometries $\nu_1, \ldots, \nu_{12}$ and the propensities of the model of Figure 9.2 into (9.1) and (9.2). In particular, the CLE can be written in the matrix form

$$\dot{x} = V\ a(x) + H\ V\ \mathrm{diag}\left(\sqrt{a(x)}\right)\ \Gamma \tag{9.5}$$

where $V = [\nu_1, ..., \nu_{12}]_{5\times 12}$, $\mathrm{diag}\left(\sqrt{a(x)}\right)$ is the diagonal matrix having the square root of the entries of vector $a(x)$ on the diagonal, $\Gamma = [\Gamma_1, ..., \Gamma_m]^T$ and $H = 1/\sqrt{N_A \Omega}$.

To get an insight into the accuracy of the CLE approximation of the CME, we analyse numerical simulations of the two processes. To simulate the CLE (9.5) we used a modified version of the Euler-Maruyama method [112] (with sampling time of 0.1[min]) which shuts down a reaction channel when the amount of any its reactants reaches zero, for preserving non-negativity of the system state [51]. To simulate the CME (9.1) we used a customized version of software *StochKit* [149]. Simulations are started from the state $X^- = (0, 0, 43, 0, 155)$ which is (up to integer round-off) the expected state of equilibrium before the arabinose uptake mechanism kicks in. In practice, this value is computed

| $\bar{\theta}$ | Values (CME) | Values (CLE) |
|---|---|---|
| $v_1$ | $120\, min^{-1}$ | $120\, min^{-1}$ |
| $v_2^0$ | $0.05\,\#\, min^{-1}$ | $3.9643 \cdot 10^{-11}\, M\, min^{-1}$ |
| $v_2$ | $4.95\,\#\, min^{-1}$ | $3.9643 \cdot 10^{-9}\, M\, min^{-1}$ |
| $K_u$ | $58541.79\,\#$ | $4.6416 \cdot 10^{-5}\, M$ |
| $v_3$ | $4.16\, min^{-1}$ | $4.16\, min^{-1}$ |
| $v_4^0$ | $0.05\,\#\, min^{-1}$ | $3.9643 \cdot 10^{-11}\, M\, min^{-1}$ |
| $v_4$ | $4.95\,\#\, min^{-1}$ | $3.9643 \cdot 10^{-9}\, M\, min^{-1}$ |
| $v_5$ | $5\, min^{-1}$ | $5\, min^{-1}$ |

FIGURE 9.3: Nominal parameter values for the CME (in molecule number units – symbol # denotes number of molecules) and the CLE (in concentration units – normalization factor $1/N_A\Omega \simeq 7.9 \cdot 10^{-10}\mathrm{M}/\#$). Degradation rates, equal for CME and CLE models, are $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (0.0139, 0.347, 0.0139, 0.116, 0.0139) min^{-1}$.

by approximately solving the equation $0 = Va(x^-)$ subject to $x_1^{-1} = 0$, with $x^- = X^-/(\Omega N_A)$, with respect to $X^- \in \mathbb{N}^n$. The simulated states $X(t)$ and $x(t)$ are recorded at times $\mathscr{T} = \{t_k = k \cdot T :\ k = 0, \ldots, 20\}$, with $T = 5$[min]. Results from 100 simulated trajectories are reported in Figure 9.4(a). Distributions look similar, but the Langevin approximation appears to be slightly biased. At every time $k$, the hypothesis of equal distributions was tested by applying a standard two-sample Kolmogorov-Smirnov test first on the simulated states, then on the same data but with their means equalized. Results in Figure 9.4(b) show that, while the hypothesis is often rejected before bias correction, this is no longer the case after mean equalization, except for small molecule numbers (see e.g. mUpt and mGFP around time 10) where the CLE is known to perform worse [79]. We will discuss the implications of this for filtering in the next section.

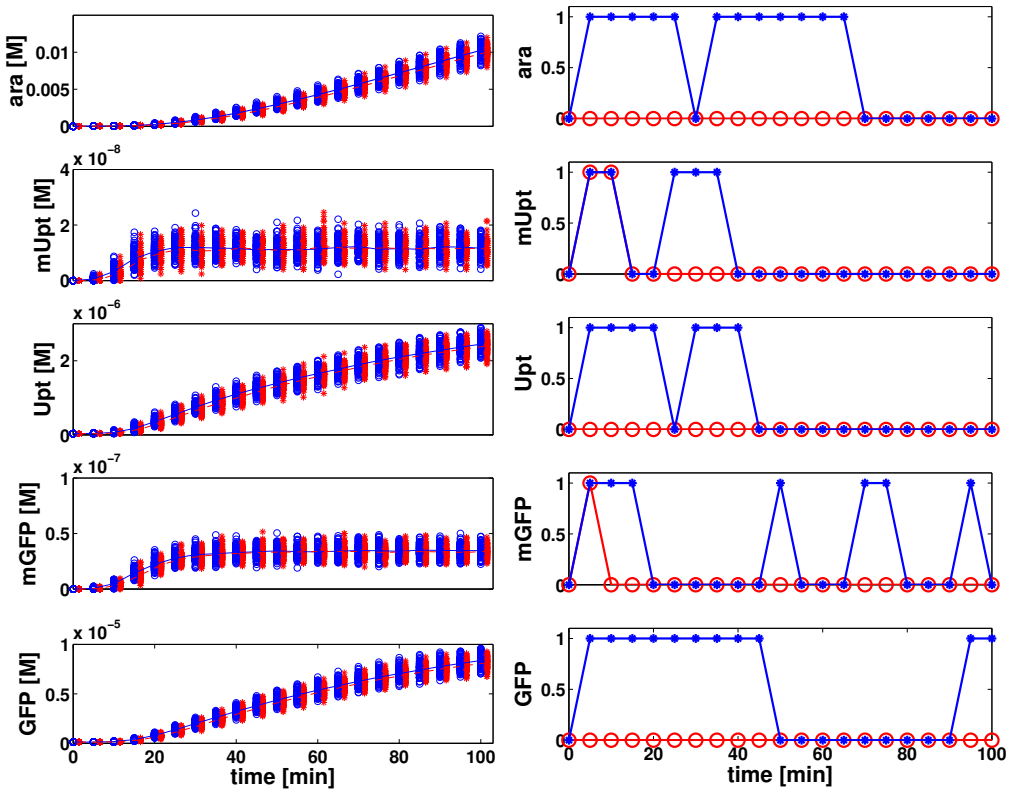
## 9.4 Gene network state estimation

Consider a cell population model specified as in Section 9.2. Given an initial distribution $p(\cdot, 0)$ at initial time $t_0 = 0$, the dynamics of $X^\ell$ in every individual $\ell = 1, \ldots, N$ are described by the propensities $a_j^{\theta_\ell}$ of the reactions $\nu_j$, with $j = 1, \ldots, m$, and the individual parameters $\theta_\ell$ follow (9.3) with assigned distribution $F_\delta$. Let measurements (9.4) be available for all individuals. We consider the following real-time state estimation problem in all individuals.

*Problem* 1. Let $\mathscr{Y}^\ell(t) = \{y^\ell(t_k) :\ t_k \leq t\}$. For $t \geq 0$, compute

$$\mathbb{E}[X^1(t), \ldots X^N(t) | \mathscr{Y}^1(t), \ldots, \mathscr{Y}^N(t)].$$

Expectation is taken with respect to the process laws and the random deviations of the individual parameters from population average. Since these quantities are assumed to be




(a) (b)

FIGURE 9.4: (a) Comparison of 100 simulations of the CME model (red, stars) and the CLE approximation (blue, circles) at times $\mathscr{T}$. At each time point, samples from the two simulations are plotted next to each other (slightly off the corresponding time point) for visual comparison. Lines indicate simulation mean for CME (red dashed) and CLE (blue solid). (b) Results of the Kolmogorov-Smirnov test for the CME and CLE simulations in Figure 9.4(a) (blue), and for the same data after equalization of the means (red). The test is applied, separately for every state entry, to the simulated states at times $\mathscr{T}$. A value of 0 (resp. 1) means that the hypothesis of equivalence of the distributions has been accepted (resp. rejected) with 95% confidence level.

statistically independent across individuals, the problem splits into $N$ estimation problems, one per individual. Focusing on the problem of filtering, the problem reformulates as follows.

*Problem* 2. For every $\ell = 1, \ldots, N$, and $k \in \mathbb{N}$, compute $\hat{X}^\ell[k|k] = \mathbb{E}[X^\ell(t_k)|\mathscr{Y}^\ell(t_k)]$.

Being the problem identical for every individual, from now on we drop superscript "$\ell$" from the notation and look at the generic individual (process) $X$ with parameters $\theta$ obeying (9.3).

Given the known difficulty of solving the CME, the problem of computing $\hat{X}^\ell[k|k]$ appears quite challenging and is further complicated by the randomness of parameters $\theta$. Possible but computationally demanding approaches include Markov Chain Monte Carlo (MCMC) sampling [78] and particle filtering [63, 178]. We consider that computationally more effective solutions may be obtained based on the CLE approximation (9.2).

We saw in Section 9.3 that, at least in our case study, the CLE provides a viable approximation of the CME, with the exception of a small bias. For state estimation purposes, the closed-loop correction of the estimates as new measurements become available is expected to compensate for this. Since, conditionally on $\theta$, process $x$ in (9.2) is an approximation of process $X$, it is natural to see the augmented process $\xi = (x, \theta)$ as an approximation of $(X, \theta)$, and to approximate the solution of Problem 2 by computing $\hat{x}[k|k] = \mathbb{E}[x(t_k)|Y(t_k)]$ in place of $\hat{X}[k|k] = \mathbb{E}[X(t_k)|Y(t_k)]$. Following a well-known approach (see e.g. [60]), the dynamics of $\xi$ can be written by treating $\theta$ as an invariant state $\theta(t)$ with random initial condition. Combining $\dot{\theta}(t) = 0$ with Eq. (9.5) leads to the augmented Langevin system

$$\dot{\xi}(t) = \begin{bmatrix} Va(\xi(t)) \\ 0 \end{bmatrix} + \begin{bmatrix} HV\,\mathrm{diag}\left(\sqrt{a(\xi(t))}\right)\Gamma \\ 0 \end{bmatrix} \tag{9.6}$$

with a priori distribution of $\theta(t)$ at time $t_0$ given by (9.3). Based on this, we address the following problem, slightly more general than the computation of $\mathbb{E}[x(t_k)|Y(t_k)]$.

*Problem* 3. Compute $\hat{\xi}[k|k] = \mathbb{E}[\xi(t_k)|\mathscr{Y}(t_k)]$, $\forall k \in \mathbb{N}$.

In the next section we will provide one solution based on a version of the so-called Unscented Kalman Filter (UKF) [101]. The goodness of the resulting estimates $\hat{x}[k|k]$ relative to a direct solution of Problem 2 will be assessed in simulation by comparison with a particle Filter (PF), the Bootstrap filter, built on model (9.1). This PF is implemented in accordance with [63, §1.3.3] , i.e. sampling the particle dynamics by Gillespie simulation of (9.1), via customized *StochKit* [149] software and Matlab.

### 9.4.1 The Square-Root Unscented Kalman Filter

To solve Problem 3, we present a continuous-discrete SRUKF. A square-root version of UKF has been chosen to improve numerical stability and also to speed up filtering via the direct use of the matrix roots from UKF. The filter is built for system (9.6) with measurements (9.4) rewritten as $y(t_k) = \bar{C}\xi(t_k) + e(t_k)$, with $\bar{C} = [C\ 0]$, where the $0s$ account for the extension of the state.

The SRUKF utilizes a deterministic sampling approach for the approximate computation of $\hat{\xi}[k|k]$. Let $L$ be the dimension of $\xi$. So-called sigma vectors $\mathcal{X}_i$, with $i = 0, \ldots, 2L$, are chosen after each measurement update based on a square-root decomposition of the estimation covariance $P$ and mean $\mu = \hat{\xi}[k|k]$, and used in a two steps (*prediction* and *correction*) algorithm to compute weighted mean and covariance approximating the true conditional distribution of $\xi$ at each $t_k$. Let $A(t) = \mathrm{chol}(P(t))$, that is $A$ is computed

as the lower triangular Cholesky factor of the covariance $P$. Following [150], one defines $\mathcal{X}_0 = \mu$, $\mathcal{X}_i = \mu + (\sqrt{c}A)_i$ for $i = 1, ..., L$ and $\mathcal{X}_i = \mu - (\sqrt{c}A)_i$ for $i = L+1, ..., 2L$, and corresponding weights $W_0^{(\mu)} = \lambda/c$, $W_i^{(\mu)} = W_i^{(c)} = \lambda/(2c)$, for $i = 1, ..., 2L$, and $W_0^{(c)} = W_0^{(\mu)} + (1 - \alpha^2 + \beta)$. The parameters $c = \alpha^2(L + \kappa)$ and $\lambda = c - L$ are scaling parameters with positive constants $\alpha$, $\beta$ and $\kappa$ used to tune the SRUKF (in our applications we set $\alpha = 0.17$, $\kappa = 200$, $\beta = 2$). Define $\mathcal{X} = [\mathcal{X}_0, \ldots, \mathcal{X}_{2L}]$, $w_m = \left[W_0^{(\mu)} \ ... \ W_{2L}^{(\mu)}\right]^T$, $W = H \operatorname{diag}\left(W_0^{(c)} \ ... \ W_{2L}^{(c)}\right) H^T$, with $H = (I - [w_m \ ... \ w_m])$,

$$\underbrace{\begin{bmatrix} Va(\mathcal{X}_1) & \cdots & Va(\mathcal{X}_{2L}) \\ 0 & \cdots & 0 \end{bmatrix}}_{\triangleq F(\mathcal{X})}, \quad \underbrace{\begin{bmatrix} HV \operatorname{diag}\left(\sqrt{a(\mathcal{X}_i)}\right) \\ 0 \end{bmatrix}}_{\triangleq G(\mathcal{X}_i)}.$$

**Prediction** From time $t_{k-1}$ to $t_k$, the SRUKF prediction equations [150, 156] can be written in terms of sigma vectors as $B(t) = \sqrt{c}\begin{bmatrix} 0 & A(t)\Phi(M(t)) & -A(t)\Phi(M(t)) \end{bmatrix}$, $M(t) = A^{-1}(t)[\mathcal{X}(t)WF^T(\mathcal{X}(t)) + F^T(\mathcal{X}(t))W\mathcal{X}(t) + \sum_{i=0}^{2L} W_i^{(\mu)} G(\mathcal{X}_i(t))G^T(\mathcal{X}_i(t))]A^{-T}(t)$,

$$d\mathcal{X}_i(t)/dt = F(\mathcal{X}(t))w_m + B_i(t), \quad i = 0, ..., 2L,$$

where $\Phi(\cdot)$ is a function defined as: $\Phi_{ij}(M) = M_{ij}$ $\forall i > j$, $\Phi_{ij}(M) = 1/2 M_{ij}$ $\forall i = j$, $\Phi_{ij}(M) = 0$ otherwise. In practice, the above equations are integrated numerically as follows. Choosing a discretization interval of $\delta t$ (we set $\delta t = 0.005$[min]) and dividing the interval between measurements into $J = (t_k - t_{k-1})/\delta t$ subintervals, one computes $\mathcal{X}_i(t + \delta t) = \mathcal{X}_i(t) + [F(\mathcal{X}(t))w_m + B_i(t)]\,\delta t, \quad i = 0, ..., 2L$ iteratively from $t_{k-1}$ to $t_k$. At each iterate, one extracts $A(t)$ from the current $\mathcal{X}(t)$ and updates $F(\mathcal{X}(t))$ and $B(t)$ accordingly. This eventually yields new $\mathcal{X}(t_k)$ and $A(t_k)$, from which the a priori moments are given by $\hat{\xi}[k|k-1] = \mathcal{X}(t_k)w_m$, $P[k|k-1] = A(t_k)A^T(t_k)$.

**Measurement update** For a new measurement $y(t_k)$, $\hat{\xi}[k|k]$ and $P[k|k]$ are computed from the above $\hat{\xi}[k|k-1]$ and $P[k|k-1]$ by a standard Kalman update step according to our linear measurement model.

**SRUKF Initialization** We set $\hat{\xi}[0|-1] = \mathbb{E}[\xi(t_0)]$ and $P[0|-1] = Var[\xi(t_0)]$, where the statistics of $\xi$ at $t_0$ are determined by the priors on $x(t_0)$ and $\theta$.

## 9.5 State estimation: Simulation results for the *E.coli* arabinose uptake system

### 9.5.1 Comparison of SRUKF and PF

To evaluate the CLE approximation of the CME for filtering performance, we compare the (CLE-based) SRUKF with a (CME-based) PF using $P = 1000$ particles. This comparison is carried out with model parameters fixed to the nominal values of Table 9.3. Data are generated by simulating the "true" CME model. We consider two scenarios.

**Scenario 1** We consider simulations always starting from state $X^-$ of Section 9.3), and initialize the filters at $X^-$ and $x^-$, assuming this state is known. This choice is relevant to experiments where the beginning of arabinose uptake occurs at a known time, e.g. at the delivery of arabinose in a glucose-poor medium. We simulated 100 trajectories $X(t)$ and produced corresponding fluorescence measurements $y(t)$ at times $\mathscr{T}$ by corrupting the simulated values of $X_5$ with zero-mean Gaussian noise with standard deviation $\sigma$ fixed to $4 \cdot 10^{-7}$[M], which is approximately 10% of the mean observed value of $X_5/(N_A\Omega)$. On each trajectory, we ran SRUKF and PF from the true initial state and null variance.

**Scenario 2** We assume that the system has started the arabinose uptake mechanism earlier than expected. This is relevant to experiments where the start of arabinose uptake is somewhat undetermined, e.g. for bacteria placed in an arabinose-rich medium where the switch to arabinose depends on depletion of environmental glucose. To simulate this we considered the same 100 simulations of the previous dataset, discarding the first two measurement times (thus taking $t = 10$[min] as the initial time) and extending the simulations to include two additional measurements. For each trajectory, we ran both SRUKF and PF initialized as in the first scenario, in the (wrong) belief that the system is observed starting from an equilibrium state.

Let $\hat{x}[k|k]$ and $\hat{X}[k|k]$ be the estimates of the SKRUF and of the PF, respectively. Denote with $\mu[k]$ the *a priori* mean of the process $X$ (computed from 10000 separate Gillespie simulation runs). Note that $\mu[k]$ can be thought of as the *a priori* state estimator. To compare performance, from the 100 simulations we compute empirically the estimation error time series (in the concentrations domain)

$$\begin{aligned}
\bar{e}[k|k] &= \mathbb{E}\big[\,|\,X[k]/(N_A\Omega) - \hat{x}[k|k]\,|\,\big] && \text{(SRUKF)},\\
\bar{e}_{PF}[k|k] &= \mathbb{E}\big[\,|\,X[k] - \hat{X}[k|k]\,|\,\big]/(N_A\Omega) && \text{(PF)},\\
\bar{e}_{\mu}[k] &= \mathbb{E}\big[\,|\,[X[k] - \mu[k]\,|\,\big]/(N_A\Omega) && (\textit{a priori}).
\end{aligned}$$

For reference, we also compute $\bar{x}[k] = \mathbb{E}[X(k)]/(N_A\Omega)$, the empirical mean of the 100 simulated trajectories on which filtering is performed.

In Scenario 1 filters performed nearly identically: For all $k$ and all state entries $i = 1, \ldots, 5$, $|\,\bar{e}_i[k|k] - \bar{e}_{PF,i}[k|k]\,|\,/\bar{x}_i[k] \leq 10^{-2}$. However, both filters improved upon $\mu[k]$ only in the 5th (observed) state component (on average over time $\bar{e}_{\mu,5}/\bar{x}_5 = 0.083$ whereas $\bar{e}_5/\bar{x}_5 \simeq \bar{e}_{PF,5}/\bar{x}_5 \simeq 0.054$). The SRUKF appears to compensate for the bias of the CLE approximation (Section 9.3), while neither SRUKF nor PF decreased the uncertainty on the unobserved states. In Scenario 2, Figure 9.5 shows plots of the relative estimation errors $\bar{e}[k|k]/\bar{x}[k]$, $\bar{e}_{PF}[k|k]/\bar{x}[k]$ and $\bar{e}_{\mu}[k]/\bar{x}[k]$ over time. Both filters

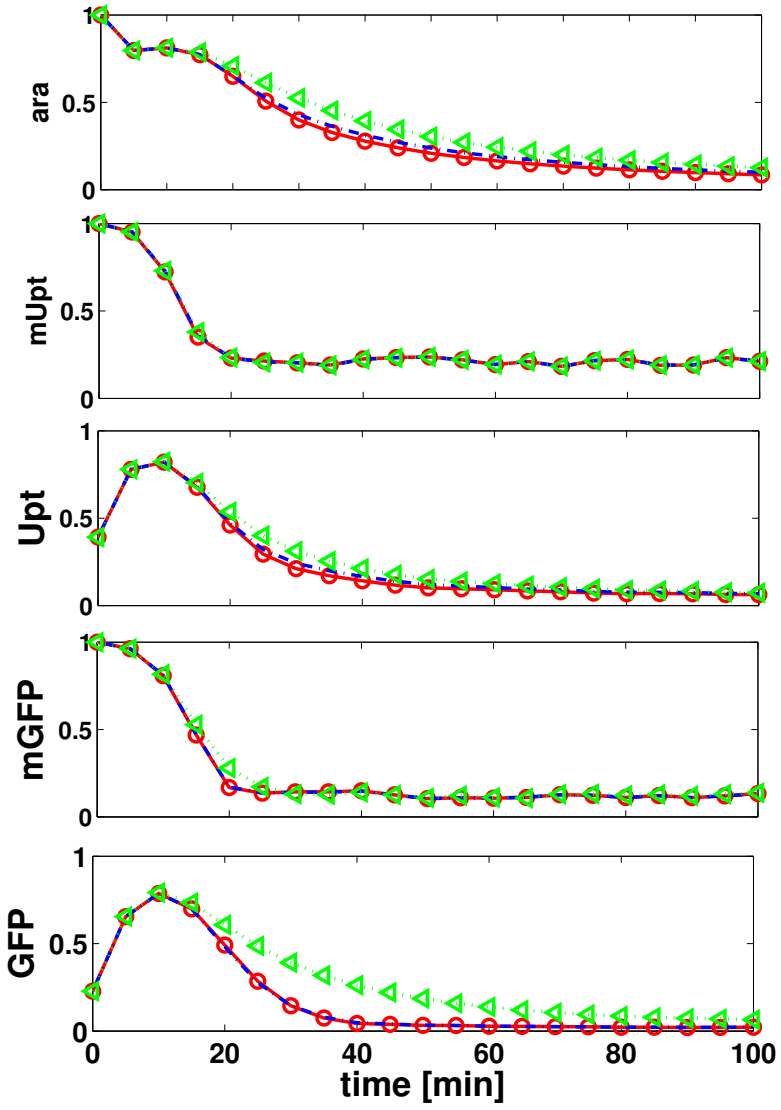



FIGURE 9.5: Comparison of estimation errors $\bar{e}[k|k]/\bar{x}[k]$ (red, solid), $\bar{e}_{PF}[k|k]/\bar{x}[k]$ (blue, dash-dotted) and $\bar{e}_{\mu}[k]/\bar{x}[k]$ (green, dotted).

improve significantly upon $\mu[k]$ in state entries 1, 3 and 5, and, to a lesser extent, in entries 2 and 4, showing that measurements are successfully exploited to compensate

for the wrong belief on the initial state. Again, SRUKF and PF behave very similarly. We interpret these results as follows. While a more advanced PF (e.g a larger number of particles $P$) could be considered, at least for our case study, the modelling error introduced by approximating the CME with the CLE has no practical effect on estimation performance. In addition, in our implementation, the SRUKF is much faster than PF (filtering one trajectory takes more than 15min for PF and less than 2min for the SRUKF on a 64bits 3.20GHz 6-core 6Gb-RAM Linux workstation). Although the PF computational burden may be reduced e.g. by the use of tau-leaping [147], these results make CLE-based SRUKF an appealing alternative for real-time applications.

### 9.5.2 Performance of the SRUKF in presence of extrinsic noise

We now evaluate performance of the SRUKF when some parameters are random. We again consider two scenarios.

**Scenario 1** All parameters are known and fixed to the nominal values except for $K_u$. This is the threshold of the sigmoidal function that determines the switch-like behaviour of the system via regulation of mUpt synthesis, and may represent e.g. individual-dependent concentrations of unmodelled regulators (such as AraC, see Figure 9.1 and [125]).

**Scenario 2** We fix all parameters to nominal values except $v_4$, the maximal regulated synthesis rate of mGFP. Variability of $v_4$ may represent e.g. different number of promoters for GFP (unequal number of plasmids carrying the reporter) in different cells.

In both cases, we consider 100 values of the variable parameter sampled from a Gaussian distribution with mean equal to the nominal value and standard deviation equal to 20% of the mean (for consistency, we ensure that all samples are positive). For each of the 100 parameter values, we simulate the system once starting from $X^-$ (which, by its definition, does not depend on $K_u$ and $v_4$) and record state and noisy measurements at times $\mathscr{T}$, with noise distributed as in the previous section. SRUKF is run on each simulated trajectory initialized with the true statistics of the parameters and null-variance initial state estimate $X^-/(N_A\Omega)$. To assess the performance gain in using the information on parameter variability, we also run a SRUKF using the wrong belief that parameters are all fixed to nominal values, and refer to it as nominal SRUKF (nSRUKF). Let $\Xi[k]$ denote either $(X[k], K_u)$ or $(X[k], v_4)$, depending on the scenario. Let $\hat{\xi}[k|k]$ and $\hat{x}[k|k]$ denote the estimates from SRUKF and nSRUKF, in the same order. Let $\mu[k]$ denote the *a priori* mean of process $\Xi[k]$ with nominal parameters (*a priori* estimator, with

last entry now fixed to the nominal parameter value). From the 100 filtering results we compute the (empirical) statistics

$$\begin{aligned}
\bar{e}[k|k] &= \mathbb{E}\big[\,|\,\Xi[k]/(N_A\Omega) - \hat{\xi}[k|k]\,|\,\big] && \text{(SRUKF)},\\
\bar{e}_n[k|k] &= \mathbb{E}\big[\,|\,X[k]/(N_A\Omega) - \hat{x}[k|k]\,|\,\big] && \text{(nSRUKF)},\\
\bar{e}_\mu[k] &= \mathbb{E}\big[\,|\,[\Xi[k] - \mu[k]\,|\,\big]/(N_A\Omega) && (\textit{a priori}).
\end{aligned}$$

Figure 9.6 reports plots of the estimation errors $\bar{e}[k|k]$ and $\bar{e}_n[k|k]$ relative to $\bar{e}_\mu[k]$. In both scenarios, both SRUKF and nSRUKF improve upon the prior knowledge on $X[k]$ in at least some components. In Scenario 1, uncertainty about $K_u$ is reduced around times $20 - 30$min, where the threshold is crossed by the increasing concentrations of intracellular arabinose, more markedly for SRUKF. This leads to a transient improvement of the estimation of arabinose concentration. For the remaining times, where the specific value of $K_u$ is inessential (saturation of the nonlinearity), the contribution of filtering is not apparent. In Scenario 2, SRUKF clearly outperforms nSRUKF in the estimation of mGFP and GFP concentrations, i.e. the states more directly related to $v_4$. Overall, results show that exploiting the prior on extrinsic noise (parameter variability) not only enables estimation of the individual parameter value, but also improves estimation of unobserved states. This supports the use of ME-type models for state estimation and control applications.

## 9.6 Conclusions

We investigated filtering of single-cell biochemical regulatory networks with intrinsic and extrinsic noise. Simulation results on a relevant case study show that approximation of the reference CME model via CLE is an appealing approach to construct practical real-time state estimators. Moreover, the use of prior information on parameter uncertainty led to improved estimation results, showing the potential of extrinsic noise modelling for state estimation and control applications. Directions of investigation include extensive performance comparisons and applications to real data.

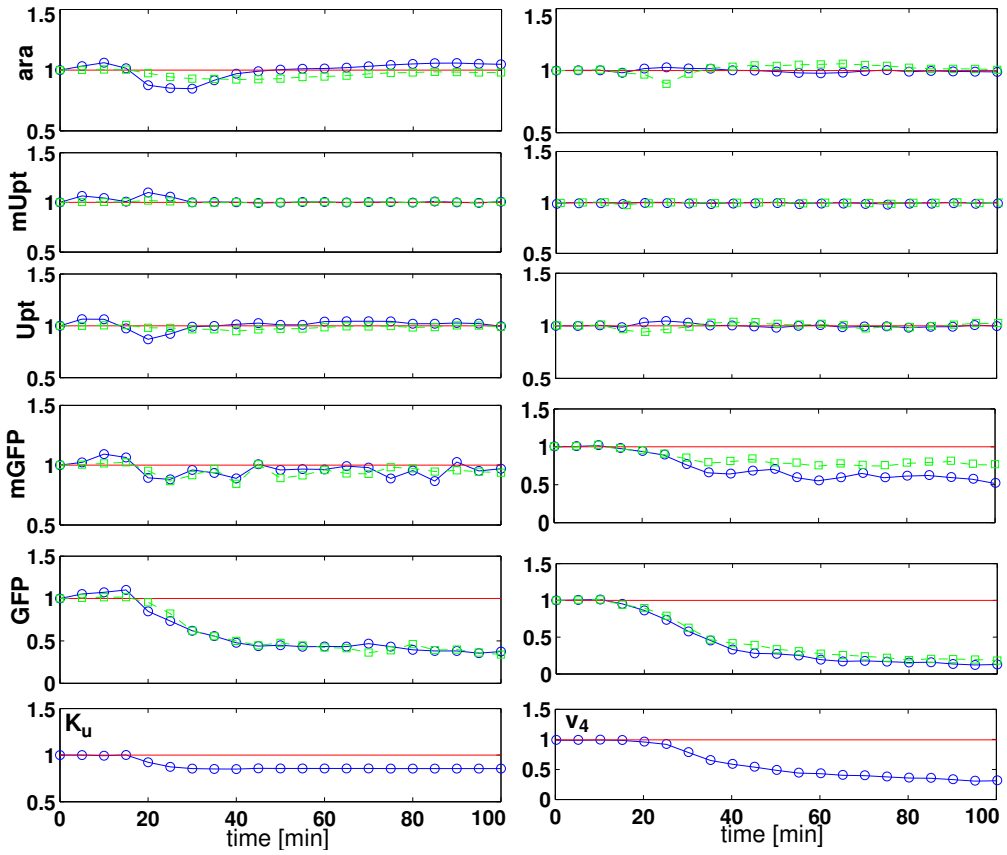



FIGURE 9.6: Estimation error ratios $\bar{e}_i[k|k]/\bar{e}_{\mu,i}[k]$ (blue, solid) and $\bar{e}_{n,i}[k|k]/\bar{e}_{\mu,i}[k]$ (green, dashed) for the state entries $i = 1, \ldots, 5$ (top to bottom, first five rows) and $\bar{e}_6[k|k]/\bar{e}_{\mu,6}[k]$ (unknown parameter, last row). The smaller the value, the more accurate the state estimate. Reference value 1 (no improvement relative to prior knowledge) is indicated by a horizontal line (red). Left: Variability on $K_u$; Right: Variability on $v_4$.

# Chapter 10

# Conclusions and Perspectives

The research work presented in this thesis aimed at applying both classical and less classical methods and tools from Systems and Control Theory to Systems and Synthetic Biology. In particular, we focused on different mathematical formalisms, i.e. qualitative/quantitative as well as deterministic/stochastic, for modelling gene regulatory networks in different experimental and/or biological contexts. Then, to gather predictive information and insights from these models, as well as to study their dynamical properties, we analysed them using bifurcation diagrams, phase planes and computational Boolean tools. Moreover, we addressed the problems of designing qualitative control strategies to tune bacterial growth rate, identifying model parameters and filtering of single-cell biochemical regulatory networks with intrinsic and extrinsic noise.

## 10.1 Qualitative models

In the first part of this thesis we mainly focused on qualitative models. In fact, qualitative formalisms turn out to be very useful when little is known about the underlying molecular processes, model parameter values or when the aim is that of speeding up the computation. Notably, we addressed the problem of developing new qualitative formalisms to model gene expression dependent on dilution due to growth rate.

The two novel modelling approaches presented in this thesis can be considered as an extension of piecewise linear (PL) systems. In fact, we kept the use of step functions to model the interactions between the elements from PL systems and we added a growth rate expression to model the dilution effect. Notably, in Chapters 4-5 we considered the growth rate to be proportional to the concentration of a component of the gene expression machinery (as RNAP or ribosome), and this choice leaded to a piecewise quadratic (PQ)

system. Whereas, in Chapter 6, we model the growth rate as the weighted minimum of two limiting gene products (RNAP and Ribosome) responsible for transcription and translation, and so for bacterial growth too. In this case, the modelling formalism turn out to be a switched piecewise quadratic (SPQ) system. The dynamics of such systems were studied using Filippov's theory and criteria assessing the equilibria stability have been formulated.

Moreover, always in the vein of qualitative models, in Chapter 7 we studied several configurations of a model made up of two interconnected Boolean modules. Notably, one module describes a Boolean version of the model for nutritional stress response in [144], and the other one a basic cellular growth module. The aim was that of testing several dynamic model variants for bacterial growth rate that consider limitation by availability of the proteins needed for cell division (RNA polymerase for transcription, ribosomes for translation, or other "bulk" proteins), and analysing (qualitative) feasibility of the wiring network, as well as the logical coherence of each model variant. The main advantage in using the Boolean framework, and in particular the use of interconnection of two Boolean modules, was that of computing the attractors of a large network at a much lower cost than with classical graph theoretical tools. From a biological point of view, our qualitative analysis showed that a limitation of growth rate by the ribosomes is needed in order to correctly reproduce the asymptotic modes, as well as transient dynamics, of the original model [144].

## 10.2 Qualitative control strategies

With the advent of synthetic biology a number of new problems, that can be studied with mathematical models, have arisen. Such problems are related to the coupling of individual components, their regulation and control to produce a desired dynamical outcome. In this thesis, we addressed the problem of mathematically controlling the bacterial growth rate, acting on the GEM by appropriately introducing synthetic inducible promoters and rewiring gene network interactions. To this aim, using the PQ and SPQ formalisms, we presented in Chapters 4-5-6 some qualitative control strategies implementing open and closed loop configurations. Notably, the open-loop configurations (Chapters 4-5-6) have been obtained by putting the expression of component of the GEM (e.g. RNAP, Ribosome), which determines the bacterial growth, under the external control of an inducible promoter. Whereas, in Chapter 5, starting from the analysis of the open-loop system in the cases that cells are grown in glucose and in maltose, we developed a closed-loop system, obtained by rewiring the open-loop's gene interactions, which is able to mathematically show the *inverse diauxie*, namely reaching

an higher growth rate on maltose then on glucose. We studied these models by means of bifurcation diagrams and phase planes analysis and we showed that a mathematical control of the bacterial growth rate can be achieved. These qualitative control strategies represent an attempt to help guide biologists in the construction of synthetic gene networks, which in turn should force bacteria to modify their growth rate in a desired fashion.

## 10.3 Quantitative models

This thesis also dealt with the construction of more quantitative (detailed) models of GRN. In Chapter 8, we used the mathematical formalism developed in [115] to model in more detail the transcription and translation machinery of the bacterium *E. coli*. Notably, the model consists of 5 non-linear differential equations coupled with 2 algebraic equations. This algebraic-differential equation system model the dynamics of the transcriptional and translational products of three gene classes, i.e. *rnn* genes, *rpoBC* genes and the proxy *bulk* genes, which represent the entire *E. coli* genome. Where *rnn* genes account for ribosome synthesis, *rpoBC* genes for that of RNAP, and *bulk* genes determine the production of all cellular proteins. In addition, by means of two Hill functions, we also took into account the growth rate dependent regulations which affect the sRNA and mRNA synthesis. More precisely, this growth regulation determines how resources are allocated during the transcription process. Quantitative models, once their parameter values have been identified, can be a useful tool to perform a large numbers of *in silico* experiments at little cost. In fact, *in silico* experiments can explore experimental conditions too costly or too long/complicated to carry out in the lab. Moreover, they can help scientists develop insights into the roles of different regulatory interactions, as well as discover inconsistencies in modelling assumptions when a GRN model is not able to reproduce certain experimental data.

## 10.4 Parameter estimation

In this thesis we also addressed the important problem of estimating the parameter values of a biological model. More precisely, in Chapter 4 we pursued a practical identifiability analysis, based on numerical simulations, of the derived expression for the growth rate as function of the input. In this study we showed that some issues may arise with noisy measurements. In fact, in this particular case, our analysis suggested that the original growth rates' measurements should be adimensionalized and unknown parameters grouped into a new set of "lumped" parameters in order to obtain local

identifiability. Moreover, we found that only the ratio between two estimated parameters can be recovered with sufficient precision in the case when only limited and noisy data are available. This identifiability analysis is a preliminary study which will be most useful to help dealing with and solving parameter estimation problems with real data sets.

Furthermore, in Chapter 8 we presented an algebraic-differential equation model of the gene expression machinery of *E. coli*, whose parameter values have been identified. Notably, this GEM dynamical model, which consists of five variables and two algebraic equations, has been calibrated using a sort of data integration. In fact, some parameter values were taken from literature, others specifically calculated using genetic information about gene classes and biological formulas, and finally, the remaining ones, identified using experimental data of *E. coli* macromolecular composition at different steady state growth rate values from [31]. We showed that the calibrated GEM model was able to consistently reproduce all the experimental data, since all model predictions were within the measurement confidence intervals (see fitting results in Figure 8.3).

## 10.5 Stochastic models and state estimation

The use of differential equations, or more generally deterministic models, for describing molecular processes makes certain assumptions that are not always valid. One of the assumption is that variables can assume continuous values, but this is only a simplification because molecules are discrete entities. Of course, while molecule numbers are sufficiently large their dynamics can be modeled by ODEs, but when they are few, the discreteness should be taken into account. Another important fact is that small molecule numbers give rise to random fluctuations that can not be captured by ODEs. Hence, when dealing with population based models (as for example the model in Chapter 8) differential equations can be successfully employed, but when one addresses the problem of modelling GRN at the single cell level, stochastic formalisms might be more appropriate to model random molecular reaction events, due to the small molecular counts present in the single cell volume. To this aim, the Chemical Master Equation (CME) [147], describing chemical kinetics in terms of probabilistic reaction events among the network species, turns out to be a standard tool for the accurate description of this type of randomness, also called intrinsic noise. However, the stochasticity in reaction events may not be the only source of noise, other noise may be due to the variability of individual features over an isogenic population (abundance of aspecific transcription/translation factors, local environmental conditions, etc.) [138], usually referred as extrinsic noise.

In Chapter 9, we used CME and its stochastic differential equation approximation, i.e. Chemical Langevin Equation (CLE), to model intrinsic noise in the GRN context. Whereas, to account for extrinsic noise, similar to Mixed-Effects (ME) modelling [52], we described the parameters of the network dynamics as random variables taking different values in different individuals (see e.g. [93]). Moreover, we investigated the problem of filtering biochemical regulatory network system with intrinsic and extrinsic noise, given single-cell fluorescent data. To this aim we consider two filters: a particle filter (PF) built on a CME and a Square-Root Unscented Kalman Filter (SRUKF) built on a CLE. Simulation results on a relevant case study of the arabinose uptake dynamics in *E.coli* showed that approximation of the reference CME model via CLE is an appealing approach to construct practical real-time state estimators. Moreover, the use of prior information on parameter uncertainty led to improved estimation results, showing the potential of extrinsic noise modelling for state estimation and control applications.

## 10.6 Perspectives

Some prospective applications as well as research directions are discussed below.

### 10.6.1 Qualitative control: application to real data

Our collaborators (Jérôme Izard and Hans Geiselmann [1]) are performing an ongoing experiment on a synthetic *E. coli*, implementing the open-loop model presented in Chapters 4-5. Notably, their experiments relate the level of growth rate during the exponential phase of the synthetic bacteria to the amount of the inducer. Their preliminary (unpublished) results are very promising and, in the future, these dose-response curves can be used to calibrate and validate the growth rate expression presented in Chapters 4. Moreover, a growth rate model—as a function of control input—could be also derived for the SPQ system (Chapter 6) where a weighted minimum model is used, and hence fitting this expression to experimental data too. Then, an evaluation of the predictive power of the two growth rate models could help in discriminating the best growth model.

Therefore, a real implementation of the gene network rewiring suggested by the closed-loop model in Chapter 5, and collection of experimental data would, of course, help in confirming the theoretical results on the *inverse diauxie* and/or suggesting new directions of investigations both at biological and mathematical levels.

[1] *Laboratoire Adaptation et Pathogénie des Microorganismes*, (CNRS UMR 5163), Université Joseph Fourier, La Tronche, France

### 10.6.2 Identifiability, sensitivity analysis and validation of GEM model

The quantitative gene expression machinery model of *E. coli* presented in Chapter 8 can be further investigated and analysed. For example, to give more consistency to the estimated parameters, an identifiability analysis should be conducted [174], as well as confidence intervals of the estimated parameters should be computed. Moreover, a sensitivity analysis of the model parameters could also indicate which regulatory mechanisms are fundamental or not to correctly reproduce experimental data, and hence help modellers in finding new directions to simplify/reduce the model. Finally, validation of the model predictions on an alternative data set would give more consistency to the model and to the estimated parameter values.

Another point that would be worth exploring is that of comparing/testing the qualitative behaviour of the reduced models of Chapters 4-5-6 with that of the more quantitative model in Chapter 8. This, would either confirm the main modelling assumptions used in both formalisms, or suggest possible improvements for the qualitative models (Chapters 4-5-6) as well as viable strategies to reduce/simplify the quantitative model (Chapter 8).

### 10.6.3 Combining qualitative and quantitative formalisms for control purposes

Another possible direction of investigation could be that of combining, in a unique model, the qualitative and quantitative formalisms. In fact, one could first develop control laws using qualitative models, which turn out to be easier to construct, and then adapt these control strategies to larger quantitative models. For example, a first attempt could be that of plugging the qualitative control laws developed for the open and closed-loop models in Chapters 4-5-6 into the quantitative GEM model of Chapter 8, and hence studying the effects of these interactions.

In a more general context, this GEM model could also be used as host-cell model where more complex/specific gene networks are embedded into it, or as a starting point for the design of synthetic genetic manipulations, which implement desired controls of the bacterial GEM.

### 10.6.4 Further investigation of dynamical growth rate models

In the first part of this thesis we have focused on modelling bacterial growth rate as a function of components of GEM. However, we only presented qualitative expressions

of such model. Hence, further investigations—starting from the insights given by the Boolean model of Chapter 7—on how to model the bacterial growth rate in more details, i.e. more quantitatively, could be conducted. To this aim, the quantitative GEM model of Chapter 8 will help for searching more precise and dynamical growth rate formulas. Eventually, once a satisfactory dynamical growth rate expression will be available, the quantitative model of Chapter 8 may evolve towards an even more refined *E. coli* GEM model, which will be valid in each phase of the bacterial growth, and not only in the exponential phase when the growth rate is simply constant.

### 10.6.5 Filtering applications of GRN models

A natural follow-up of the filtering method proposed in Chapter 9 would be that of application on real single-cell fluorescent data. Moreover, this filtering approach could also be employed as part of a tool for parametric identification of GRN with intrinsic and extrinsic noise or, more in general, for the identification of non-linear mixed-effects models with stochastic differential equations [62, 135]. Another possible application would be that of using this filtering technique for the control of single-cell biochemical network dynamics via model-based control strategies [169], which ultimately may also be useful for bacterial growth control.

# Appendix A

# List of Publications

1. M. Chaves and A. Carta. Attractor computation using interconnected Boolean networks: testing growth rate models in *E.coli*. *Theoretical Computer Science*, 2014. (To appear).

2. A. Carta, M. Chaves, and J.-L. Gouzé. Switched piecewise quadratic models of biological networks: application to control of bacterial growth. AUTOMATICA, 2014, Submitted.

3. A. Carta, M. Chaves, and J.-L. Gouzé. A class of switched piecewise quadratic systems for coupling gene expression with growth rate in bacteria. In *Proceedings of the 9th IFAC Symposium on Nonlinear Control Systems, NOLCOS 2013, Toulouse, France, September 2013*.

4. A. Carta and E. Cinquemani. State estimation for gene networks with intrinsic and extrinsic noise: A case study on *E.coli* arabinose uptake dynamics. In *Proceedings of the European Control Conference 2013 , Zurich, Switzerland, Jul 2013*.

5. A. Carta, M. Chaves, and J.-L. Gouzé. A simple model to control growth rate of synthetic *E.coli* during the exponential phase: Model analysis and parameter estimation. In D. Gilbert and M. Heiner, editors, *Computational Methods in Systems Biology*, Lecture Notes in Computer Science, pages 107–126. Springer Berlin Heidelberg, 2012, [BEST STUDENT PAPER AWARD].

6. A. Carta, M. Chaves, and J.-L. Gouzé. A coarse-grained dynamical model of *E. coli* gene expression machinery at varying growth rates. *Technical Report* - ANR GeMCo Project deliverable.

# Bibliography


[1] http://en.wikipedia.org/wiki/File:Average_prokaryote_cell-_en.svg.

[2] http://en.wikipedia.org/wiki/File:Animal_cell_structure_en.svg.

[3] http://en.wikipedia.org/wiki/File:EscherichiaColi_NIAID.jpg.

[4] http://en.wikipedia.org/wiki/File:Genetic_code.svg.

[5] http://en.wikipedia.org/wiki/File:DNA_chemical_structure.svg.

[6] http://en.wikipedia.org/wiki/File:Simple_transcription_elongation1.svg.

[7] http://en.wikipedia.org/wiki/File:Peptide_syn.png.

[8] http://en.wikipedia.org/wiki/File:Heatmap.png.

[9] http://en.wikipedia.org/wiki/File:Anti-lipoic_acid_immunoblot.png.

[10] http://en.wikipedia.org/wiki/File:Reporter_gene.png.

[11] R. Albert and H. G. Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *Journal of theoretical biology*, 223(1):1–18, 2003.

[12] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology, Third Edition*. Garland Science/Taylor & Francis Group, 2009.

[13] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 5 edition, 2007.

[14] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/CRC Mathematical & Computational Biology)*. Chapman and Hall/CRC, 1 edition, 2006.

[15] E. Andrianantoandro, S. Basu, D. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Molecular systems biology*, 2(1), 2006.

[16] J. Ang, S. Bagh, B. Ingalls, and D. McMillen. Considerations for using integral feedback control to construct a perfectly adapting synthetic gene network. *Journal of theoretical biology*, 266(4):723–738, 2010.

[17] A. P. Arkin and D. V. Schaffer. Network news: innovations in 21st century systems biology. *Cell*, 144(6):844–849, 2011.

[18] M. H. Asyali, D. Colak, O. Demirkaya, and M. S. Inan. Gene expression profile classification: a review. *Current Bioinformatics*, 1(1):55–73, 2006.

[19] I. Belgacem and J.-L. Gouzé. Analysis and reduction of transcription translation coupled models for gene expression. In *CAB (Computer applied to Biotechnology)*, Mumbai, India, 2013. IFAC.

[20] I. Belgacem and J.-L. Gouzé. Stability analysis and reduction of gene transcription models. In *52nd IEEE Conference on Decision and Control (CDC'13)*, Florence, Italy, 2013.

[21] S. A. Benner and A. M. Sismour. Synthetic biology. *Nature Reviews Genetics*, 6(7):533–543, 2005.

[22] G. Bernot, J.-P. Comet, A. Richard, M. Chaves, J.-L. Gouzé, and F. Dayan. Modeling and analysis of gene regulatory networks. In F. Cazals and P. Kornprobst, editors, *Modeling in Computational Biology and Biomedicine*, pages 47–80. Springer-Verlag Heidelberg, 2013.

[23] S. Berthoumieux, H. de Jong, G. Baptist, C. Pinel, C. Ranquet, D. Ropers, and J. Geiselmann. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular Systems Biology*, 9:634, 2013.

[24] K. Bettenbrock, T. Sauter, K. Jahreis, A. Kremling, J. W. Lengeler, and E.-D. Gilles. Correlation between growth rates, EIIACrr phosphorylation, and Intracellular Cyclic AMP levels in Escherichia coli K-12. *J. Bacteriol.*, 189(19):6891–6900, 2007.

[25] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, et al. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–1462, 1997.

[26] H. Bolouri. *Computational modeling of gene regulatory networks: a primer*. World Scientific, 2008.

[27] J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 6. Macmillan London, 1976.

[28] M. Branicky. Stability of switched and hybrid systems. In *Decision and Control, 1994., Proceedings of the 33rd IEEE Conference on*, volume 4, pages 3498 –3503 vol.4, dec 1994.

[29] H. Bremer and P. Dennis. Feedback control of ribosome function in *Escherichia coli*. *Biochimie*, 90(3):493–499, 2008.

[30] H. Bremer, P. Dennis, and M. Ehrenberg. Free RNA polymerase and modeling global transcription in Escherichia coli. *Biochimie*, 85(6):597–609, 2003.

[31] H. Bremer, P. Dennis, et al. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology*, 2:1553–1569, 1996.

[32] H. Bremer and M. Ehrenberg. Guanosine tetraphosphate as a global regulator of bacterial RNA synthesis: a model involving RNA polymerase pausing and queuing. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1262(1):15–36, 1995.

[33] F. Cacace, P. Paci, V. Cusimano, A. Germani, and L. Farina. Stochastic modeling of expression kinetics identifies messenger half-lives and reveals sequential waves of co-ordinated transcription and decay. *PLoS Comput Biol*, 8(11):e1002772, 11 2012.

[34] L. Calzone, L. Tournier, S. Fourquet, D. Thieffry, B. Zhivotovsky, E. Barillot, and A. Zinovyev. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS computational biology*, 6(3):e1000702, 2010.

[35] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, 121(9):4059–4067, 2004.

[36] S. Cardinale, M. P. Joachimiak, and A. P. Arkin. Effects of genetic variation on the *E. coli* host-circuit interface. *Cell reports*, 4(2):231–237, 2013.

[37] A. Carta, M. Chaves, and J.-L. Gouzé. A simple model to control growth rate of synthetic E. coli during the exponential phase: Model analysis and parameter estimation. In D. Gilbert and M. Heiner, editors, *Computational Methods in Systems Biology*, Lecture Notes in Computer Science, pages 107–126. Springer Berlin Heidelberg, 2012.

[38] A. Carta, M. Chaves, and J.-L. Gouzé. A class of switched piecewise quadratic systems for coupling gene expression with growth rate in bacteria. In *9th IFAC Symposium Nonlinear Control Systems (NOLCOS'13)*, 2013.

[39] A. Carta and E. Cinquemani. State estimation for gene networks with intrinsic and extrinsic noise: A case study on e. coli arabinose uptake dynamics. In *Control Conference (ECC), 2013 European*, pages 3658–3663. IEEE, 2013.

[40] R. Casey, H. De Jong, and J. Gouzé. Piecewise-linear models of genetic regulatory networks: Equilibria and their stability. *Journal of Mathematical Biology*, 52(1):27–56, 2006.

[41] M. Chaves. GeMCo: Model reduction, experimental validation, and control for the gene expression machinery in *E. coli*. http://www-sop.inria.fr/members/Madalena.Chaves/ANR-GeMCo/main.html.

[42] M. Chaves and A. Carta. Attractor computation using interconnected boolean networks: testing growth models in *E. Coli*. *Theoretical Computer Science*, 2014. To appear.

[43] M. Chaves and J.-L. Gouzé. Piecewise affine models of regulatory genetic networks: review and probabilistic interpretation. In J. Lévine and P. Mullhaupt, editors, *Advances in the Theory of Control, Signals and Systems, with Physical Modelling*, volume 407 of *Lecture Notes in Control and Information Sciences*, pages 241–253. Springer, 2010.

[44] M. Chaves and J.-L. Gouzé. Exact control of genetic networks in a qualitative framework: The bistable switch example. *Automatica*, 47(6):1105 – 1112, 2011. Special Issue on Systems Biology.

[45] M. Chaves and L. Tournier. Predicting the asymptotic dynamics of large biological networks by interconnections of boolean modules. In *Proc. $50^{th}$ Conf. Decision and Control and European Control Conf.*, Orlando, Florida, USA, December 2011.

[46] M. Chaves, L. Tournier, and J. L. Gouzé. Comparing Boolean and piecewise affine differential models for genetic networks. *Acta Biotheoretica*, 58(2):217–232, 2010.

[47] O. Chis, J. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PloS one*, 6(11):e27755, 2011.

[48] F. Corblin, S. Tripodi, E. Fanchon, D. Ropers, and L. Trilling. A declarative constraint-based method for analyzing discrete genetic regulatory networks. *BioSystems*, 98(2):91–104, 2009.

[49] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms*. MIT Press and McGraw-Hill, 2001.

[50] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[51] S. Dana and S. Raha. Physically consistent simulation of mesoscale chemical kinetics: The non-negative FIS-$\alpha$ method. *Journal of Computational Physics*, 230(24):8813–8834, 2011.

[52] M. Davidian and D. M. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8(4):387–419, 2003.

[53] M. I. Davidich and S. Bornholdt. Boolean network model predicts cell cycle sequence of fission yeast. *PloS one*, 3(2):e1672, 2008.

[54] H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[55] H. de Jong, J. Geiselmann, C. Hernandez, and M. Page. Genetic network analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics*, 19(3):336–344, 2003.

[56] H. De Jong, J. Gouzé, C. Hernandez, M. Page, T. Sari, and J. Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bulletin of Mathematical Biology*, 66(2):301–340, 2004.

[57] H. de Jong, C. Ranquet, D. Ropers, C. Pinel, and J. Geiselmann. Experimental and computational validation of models of fluorescent and luminescent reporter genes in bacteria. *BMC Systems Biology*, 4(1):55, 2010.

[58] D. A. Dean, J. Reizer, H. Nikaido, and M. H. Saier. Regulation of the maltose transport system of Escherichia coli by the glucose-specific enzyme III of the phosphoenolpyruvate-sugar phosphotransferase system. characterization of inducer exclusion-resistant mutants and reconstitution of inducer exclusion in proteoliposomes. *Journal of Biological Chemistry*, 265(34):21005–10, 1990.

[59] P. P. Dennis, M. Ehrenberg, and H. Bremer. Control of rRNA synthesis in *Escherichia coli*: a systems biology approach. *Microbiology and Molecular Biology Reviews*, 68(4):639–668, 2004.

[60] D. Dochain. State and parameter estimation in chemical and biochemical processes: a tutorial. *Journal of Process Control*, 13:801–818, 2003.

[61] D. Dochain and P. Vanrolleghem. *Dynamical Modelling and Estimation in Wastewater Treatment Processes*. IWA Publishing, 2001.

[62] S. Donnet and A. Samson. A review on estimation of stochastic differential equations for pharmacokinetic/pharmacodynamic models. *Advanced Drug Delivery Reviews*, 65(7):929 – 939, 2013. Mathematical modeling of systems pharmacogenomics towards personalized drug delivery.

[63] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, 2001.

[64] D. M. Dykxhoorn, R. St Pierre, and T. Linn. Synthesis of the $\beta$ and $\beta'$ subunits of *Escherichia coli* RNA polymerase is autogenously regulated in vivo by both transcriptional and translational mechanisms. *Molecular microbiology*, 19(3):483–493, 1996.

[65] L. Edelstein-Keshet. *Mathematical models in biology*, volume 46. Siam, 1988.

[66] E. Eden, N. Geva-Zatorsky, I. Issaeva, A. Cohen, E. Dekel, T. Danon, L. Cohen, A. Mayo, and U. Alon. Proteome half-life dynamics in living human cells. *Science*, 331(6018):764–768, 2011.

[67] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[68] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[69] B. M. Emerson. Specificity of gene regulation. *Cell*, 109(3):267–270, 2002.

[70] F. Fages, S. Soliman, and N. Chabrier-Rivier. Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. *J. Biological Physics and Chemistry*, 4(2):64–73, 2005.

[71] C. Fall, E. Marland, J. Wagner, and J. Tyson. *Computational Cell Biology*. Interdisciplinary applied mathematics. Springer, 2002.

[72] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

[73] A. Filippov. *Differential equations with discontinuous righthand sides*. Mathematics and its Applications Series. Kluwer Academic Publishers, 1988.

[74] N. Flores, S. Flores, A. Escalante, R. de Anda, L. Leal, R. Malpica, D. Georgellis, G. Gosset, and F. Bolívar. Adaptation for fast growth on glucose by differential expression of central carbon metabolism and *gal* regulon genes in an *Escherichia coli* strain lacking the phosphoenolpyruvate: carbohydrate phosphotransferase system. *Metabolic engineering*, 7(2):70–87, 2005.

[75] A. Gallant. Nonlinear regression. *The American Statistician*, 29(2):73–81, 1975.

[76] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403(6767):339–342, 2000.

[77] O. George and R. Danny. A unified theory of gene expression. *Cell*, 108(4):439 – 451, 2002.

[78] C. Gillespie and A. Golighty. Bayesian inference for the chemical master equation using approximate models. In *Proceedings of the WCSB*, ULM Germany, 2012.

[79] D. Gillespie. The chemical langevin equation. *Journal of Chemical Physics*, 2000.

[80] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403 – 434, 1976.

[81] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[82] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.

[83] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[84] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[85] L. Glass and S. A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103 – 129, 1973.

[86] A. Goelzer and V. Fromion. Bacterial growth rate reflects a bottleneck in resource allocation. *Biochim. Biophys. Acta*, 1810(10):978–988, 2011.

[87] D. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-wesley, 1989.

[88] A. Gonzalez, A. Naldi, L. Sànchez, D.Thieffry, and C. Chaouiya. GINsim: a software suite for the qualitative modelling, simulation and analysis of regulatory networks. *BioSystems*, 84(2):91–100, 2006.

[89] R. L. Gourse, T. Gaal, M. S. Bartlett, J. A. Appleman, and W. Ross. rRNA transcription and growth rate-dependent regulation of ribosome sysnthesis in Escherichia coli. *Annual Review of Microbiology*, 50(1):645–677, 1996. PMID: 8905094.

[90] J. Gouzé and T. Sari. A class of piecewise linear differential equations arising in biological models. *Dynamical systems*, 17(4):299–316, 2002.

[91] F. Grognard, J.-L. Gouzé, and H. de Jong. Piecewise-linear models of genetic regulatory networks: theory and example. pages 137–159, 2007.

[92] T. Hardiman, K. Lemuth, M. Keller, M. Reuss, and M. Siemann-Herzberg. Topology of the global regulatory network of carbon limitation in *Escherichia coli*. *J. Biotechnology*, 132:359–374, 2007.

[93] J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics*, 2011.

[94] J. Hasty, D. McMillen, F. Isaacs, and J. J. Collins. Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genetics*, 2(4):268–279, 2001.

[95] R. Heinrich and S. Schuster. *The regulation of cellular systems*. Springer, 1996.

[96] W. S. Hlavacek and M. A. Savageau. Rules for coupled expression of regulator and effector genes in inducible circuits. *Journal of molecular biology*, 255(1):121–139, 1996.

[97] F. Horn and R. Jackson. General mass action kinetics. *Archive for Rational Mechanics and Analysis*, 47(2):81–116, 1972.

[98] J. Huisman and F. J. Weissing. Biodiversity of plankton by species oscillations and chaos. *Nature*, 402(6760):407–410, 1999.

[99] S. Jamshidi, H. Siebert, and A. Bockmayr. Comparing discrete and piecewise affine differential equation models of gene regulatory networks. In M. Lones, S. Smith, S. Teichmann, F. Naef, J. Walker, and M. Trefzer, editors, *Information Processing in Cells and Tissues*, volume 7223 of *LNCS*, pages 17–24. Springer, 2012.

[100] D. J. Jin, C. Cagliero, and Y. N. Zhou. Growth rate regulation in *Escherichia coli*. *FEMS microbiology reviews*, 36(2):269–287, 2012.

[101] S. Julier and J. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[102] M. Kaern, W. Blake, and J. Collins. The engineering of gene regulatory networks. *Annual Review of Biomedical Engineering*, 5(1):179–206, 2003.

[103] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein. Random boolean network models and the yeast transcriptional network. *Proceedings of the National Academy of Sciences*, 100(25):14796–14799, 2003.

[104] S. A. Kauffman. Emergent properties in random complex automata. *Physica D: Nonlinear Phenomena*, 10(1):145–156, 1984.

[105] J. Keener and J. Sneyd. *Mathematical Physiology: I: Cellular Physiology*, volume 1. Springer, 2010.

[106] I. M. Keseler, A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen, and P. D. Karp. Ecocyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1):D605–D612, 2013.

[107] A. Khalil and J. Collins. Synthetic biology: applications come of age. *Nature Reviews Genetics*, 11(5):367–379, 2010.

[108] H. K. Khalil and J. Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, 2002.

[109] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

[110] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[111] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2008.

[112] P. Kloeden, E. Platen, and H. Schurz. *Numerical solution of SDE through computer experiments*, volume 1. Springer, 1994.

[113] S. Klumpp and T. Hwa. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51):20245–20250, 2008.

[114] S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–1375, 2010.

[115] A. Kremling. Comment on mathematical models which describe transcription and calculate the relationship between mrna and protein expression ratio. *Biotechnology and Bioengineering*, 96(4):815–819, 2007.

[116] E. Krin, O. Sismeiro, A. Danchin, and P. N. Bertin. The regulation of Enzyme IIAGlc expression controls adenylate cyclase activity in Escherichia coli. *Microbiology*, 148(5):1553–1559, 2002.

[117] P. LeDuc, W. Messner, and J. Wikswo. How do control-based approaches enter into biology? *Annual Review of Biomedical Engineering*, 13:369–396, 2011.

[118] S. Y. Lee. *Systems biology and biotechnology of Escherichia coli*. Springer, 2009.

[119] F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781–4786, 2004.

[120] S. Li, S. M. Assmann, and R. Albert. Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS biology*, 4(10):e312, 2006.

[121] D. Liberzon and A. Morse. Basic problems in stability and design of switched systems. *Control Systems, IEEE*, 19(5):59 –70, oct 1999.

[122] D. Longo and J. Hasty. Dynamics of single-cell gene expression. *Molecular systems biology*, 2(1), 2006.

[123] T. Lorenz, H. Siebert, and A. Bockmayr. Analysis and characterization of asynchronous state transition graphs using extremal states. *Bull. Mathematical Biology*, 75(6):920–938, 2013.

[124] A. G. Marr. Growth rate of *Escherichia coli*. *Microbiological Reviews*, 55(2):316–333, 1991.

[125] J. Megerle, G. Fritz, U. Gerland, K. Jung, and J. Rädler. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophysical journal*, 95(4):2103–2115, 2008.

[126] F. Menolascina, M. di Bernardo, and D. di Bernardo. Analysis, design and implementation of a novel scheme for in-vivo control of synthetic gene regulatory networks. *Automatica*, 47(6):1265 – 1270, 2011. Special Issue on Systems Biology.

[127] T. Mestl, E. Plahte, and S. Omholt. A mathematical framework for describing and analysing gene regulatory networks. *Journal of Theoretical Biology*, 176(2):291–300, 1995.

[128] A. Milias-Argeitis, S. Summers, J. Stewart-Ornstein, I. Zuleta, D. Pincus, H. El-Samad, M. Khammash, and J. Lygeros. In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology*, (29):1114–1116, 2011.

[129] J. Monod. The growth of bacterial cultures. *Annual Review of Microbiology*, 3(1):371–394, 1949.

[130] B. A. Morgan, E. Kellett, and R. S. Hayward. The wild-type nucleotide sequence of the rpoBC-attenuator region of *Escherichia coli* DNA, and its implications for the nature of the rifd18 mutation. *Nucleic acids research*, 12(13):5465–5470, 1984.

[131] S. Mukherji and A. Van Oudenaarden. Synthetic biology: understanding biological design from synthetic circuits. *Nature Reviews Genetics*, 10(12):859–871, 2009.

[132] A. Naldi, E. Rémy, D. Thieffry, and C. Chaouiya. Dynamically consistent reduction of logical regulatory graphs. *Theor. Comput. Sci.*, 412(21):2207–18, 2011.

[133] A. Narang and S. S. Pilyugin. Bacterial gene regulation in diauxic and non-diauxic growth. *Journal of Theoretical Biology*, 244(2):326 – 348, 2007.

[134] L. Notley-McRobb, A. Death, and T. Ferenci. The relationship between external glucose concentration and cAMP levels inside *Escherichia coli*: implications for models of phosphotransferase-mediated regulation of adenylate cyclase. *Microbiology*, 143(6):1909–1918, 1997.

[135] R. V. Overgaard, N. Jonsson, C. W. Tornøe, and H. Madsen. Non-linear mixed-effects models with stochastic differential equations: implementation of an estimation algorithm. *Journal of pharmacokinetics and pharmacodynamics*, 32(1):85–107, 2005.

[136] O. Paliy and T. S. Gunasekera. Growth of *E. coli* BL21 in minimal media with different gluconeogenic carbon sources and salt contents. *Applied microbiology and biotechnology*, 73(5):1169–1172, 2007.

[137] J. Paulsson. Control, exploitation and tolerance of intracellular noise. *Nature*, 420:231–237, 2002.

[138] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005.

[139] A. Polynikis, S. Hogan, and M. di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

[140] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[141] É. Remy, P. Ruet, and D. Thieffry. Graphic requirements for multistability and attractive cycles in a boolean dynamical framework. *Advances in Applied Mathematics*, 41(3):335–350, 2008.

[142] A. Richard. Positive circuits and maximal number of fixed points in discrete dynamical systems. *Discrete Applied Mathematics*, 157(15):3281–3288, 2009.

[143] A. Richard. Negative circuits and sustained oscillations in asynchronous automata networks. *Advances in Applied Mathematics*, 44(4):378–392, 2010.

[144] D. Ropers, H. de Jong, M. Page, D. Schneider, and J. Geiselmann. Qualitative simulation of the carbon starvation response in *Escherichia coli*. *Biosystems*, 84(2):124 – 152, 2006.

[145] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*, volume 707. John Wiley & Sons, 2011.

[146] J. Saez-Rodriguez, L. Simeoni, J. A. Lindquist, R. Hemenway, U. Bommhardt, B. Arndt, U.-U. Haus, R. Weismantel, E. D. Gilles, S. Klamt, and B. Schraven. A logical model provides insights into T cell receptor signaling. *PLoS Comput. Biol.*, 3(8):e163, Aug 2007.

[147] H. E. Samad, M. Khammash, L. Petzold, and D. Gillespie. Stochastic modelling of gene regulatory networks. *Intl J Rob Nonl Control*, pages 691–711, 2005.

[148] L. Sánchez and D. Thieffry. A logical analysis of the *drosophila* gap-gene system. *J. Theor. Biol.*, 211:115–141, 2001.

[149] K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold. Stochkit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, 2011.

[150] S. Sarkka. On unscented kalman filtering for state estimation of continuous-time nonlinear systems. *Automatic Control, IEEE Transactions on*, 52(9):1631–1641, 2007.

[151] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[152] G. Seber and C. Wild. *Nonlinear regression*, volume 503. LibreDigital, 2003.

[153] I. H. Segel. *Enzyme kinetics*. Wiley New York, 1993.

[154] V. Sevim, X. Gong, and J. Socolar. Reliability of transcriptional cycles and the yeast cell-cycle oscillator. *PLoS Comput. Biol.*, 6:e1000842, 2010.

[155] I. Shachrai, A. Zaslaver, U. Alon, and E. Dekel. Cost of unneeded proteins in *E. coli* is reduced after several generations in exponential growth. *Molecular Cell*, 38(5):758 – 767, 2010.

[156] H. Singer. Continuous-discrete unscented kalman filtering. Technical Report 384, FernUniversität Hagen, Germany, 2006.

[157] P. Smolen, D. A. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks—methods, recent results, and future directions. *Bulletin of mathematical biology*, 62(2):247–292, 2000.

[158] E. Sontag. Some new directions in control theory inspired by systems biology. *Systems biology*, 1(1):9–18, 2004.

[159] A. D. Tadmor and T. Tlusty. A coarse-grained biophysical model of *E. coli* and its application to perturbation of the rRNA operon copy number. *PLoS Comput Biol*, 4(5):e1000038, 05 2008.

[160] C. Tan, P. Marguet, and L. You. Emergent bistability by a growth-modulating positive feedback circuit. *Nature chemical biology*, 5(11):842–848, 2009.

[161] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *PNAS*, 98(15):8614–8619, 2001.

[162] I. Thiele, N. Jamshidi, R. M. Fleming, and B. Ø. Palsson. Genome-scale reconstruction of *Escherichia coli*'s transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization. *PLoS computational biology*, 5(3):e1000312, 2009.

[163] R. Thomas. Boolean formalization of genetic control circuits. *Journal of Theoretical Biology*, 42(3):563–585, 1973.

[164] R. Thomas and R. d'Ari. *Biological feedback*. CRC Press, 1990.

[165] M. Tigges, T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457(7227):309–312, 2009.

[166] L. Tournier and M. Chaves. Interconnection of asynchronous boolean networks, asymptotic and transient dynamics. *Automatica*, 49(4):884–893, 2013.

[167] L. Tournier and J.-L. Gouzé. Hierarchical analysis of piecewise affine models of gene regulatory networks. *Theory Biosci.*, 127:125–134, 2008.

[168] F. Trueba and C. Woldringh. Changes in cell diameter during the division cycle of *Escherichia coli*. *Journal of bacteriology*, 142(3):869–878, 1980.

[169] J. Uhlendorf, A. Miermont, T. Delaveau, G. Charvin, F. Fages, S. Bottani, G. Batt, and P. Hersen. Long-term model predictive control of gene expression at the population and single-cell levels. *Proceedings of the National Academy of Sciences*, 109(35):14271–14276, 2012.

[170] S. Vajda, H. Rabitz, E. Walter, and Y. Lecourtier. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, 83(1):191–219, 1989.

[171] P. van Ham. How to deal with more than two levels. In R. Thomas, editor, *Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems*, volume 29 of *Lecture Notes in Biomathematics*, pages 326–343. Springer, 1979.

[172] F. Viel, E. Busvelle, and J. Gauthier. Stability of polymerization reactors using i/o linearization and a high-gain observer. *Automatica*, 31(7):971 – 984, 1995.

[173] B. Volkmer and M. Heinemann. Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. *PLoS ONE*, 6(7):e23126, 07 2011.

[174] É. Walter and L. Pronzato. *Identification of parametric models from experimental data*. Communications and control engineering. Springer, 1997.

[175] R.-S. Wang, A. Saadatpour, and R. Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical biology*, 9(5):055001, 2012.

[176] D. B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[177] S. Widder, J. Schicho, and P. Schuster. Dynamic patterns of gene regulation i: simple two-gene systems. *Journal of theoretical biology*, 246(3):395–419, 2007.

[178] D. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall/CRC, 2006.

[179] G. Yagil. Quantitative aspects of protein induction. *Current topics in cellular regulation*, 9:183–236, 1974.

[180] G. Yagil and E. Yagil. On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophysical Journal*, 11(1):11–27, 1971.

[181] R. Yang, S. Lenaghan, J. Wikswo, and M. Zhang. External control of the GAL network in S. cerevisiae: A view from control theory. *PloS one*, 6(4):e19353, 2011.

[182] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *PNAS*, 2012.