Modeling, analysis and reduction of biological systems

Stefano Casagranda

PhD THESIS

prepared at
Inria Sophia Antipolis - Méditerranée

and presented at the
Université Côte d'Azur
Graduate School of Information and Communication Sciences
École Doctorale STIC

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

DOCTOR OF SCIENCE
Specialized in Control, Signal and Image Processing

Modeling, analysis and reduction of biological systems

Stefano Casagranda

Directed by Jean-Luc Gouzé and Delphine Ropers

Defended on June 30th 2017 in front of the jury composed by:

Supervisor	Jean-Luc Gouzé	DR Inria Sophia Antipolis, France
Co-Supervisor	Delphine Ropers	CR Inria Grenoble Rhône Alpes, France
Reviewers	Riccardo Bellazzi	Prof. Università degli Studi di Pavia, Italy
	Béatrice Laroche	DR INRA Unité MaIAGE Jouy, France
Examiners	Gregory Batt	DR Inria Saclay Ile-de-France, France
	Gilles Bernot	DR Laboratoire I3S, France
	Frédéric Dayan	Founder ExactCure, France
	David Rouquié	DR Bayer CropScience, France
Invited	Eugenio Cinquemani	CR Inria Grenoble Rhône Alpes, France
	Suzanne Touzeau	CR INRA ISA Sophia Antipolis, France

UNIVERSITÉ CÔTE D'AZUR

Inria informatiques mathématiques

École Doctorale STIC
(Sciences et Technologies de l'Information et de la Communication)
Unité de recherche: Inria
Équipe: Biocore

Thèse de doctorat

Présentée en vue de l'obtention du grade de docteur en
Automatique, Traitement du Signal et des Images
de l'Université Côte d'Azur

par

Stefano Casagranda

Modélisation, analyse et réduction des systèmes biologiques

Dirigée par Jean-Luc Gouzé *et co-encadrée par* Delphine Ropers

Soutenue le 30 Juin 2017

Devant le jury composé de:

Directeur de thèse	Jean-Luc Gouzé	DR Inria Sophia Antipolis, France
Co-Encadrante de thèse	Delphine Ropers	CR Inria Grenoble Rhône Alpes, France
Rapporteurs	Riccardo Bellazzi	Prof. Università degli Studi di Pavia, Italie
	Béatrice Laroche	DR INRA Unité MaIAGE Jouy, France
Examinateurs	Gregory Batt	DR Inria Saclay Ile-de-France, France
	Gilles Bernot	DR Laboratoire I3S, France
	Frédéric Dayan	Fondateur ExactCure, France
	David Rouquié	DR Bayer CropScience, France
Invité	Eugenio Cinquemani	CR Inria Grenoble Rhône Alpes, France
	Suzanne Touzeau	CR INRA ISA Sophia Antipolis, France

Abstract

This thesis deals with modeling, analysis and reduction of various biological models, with a focus on gene regulatory networks in the bacterium *E. coli*. Different mathematical approaches are used. In the first part of the thesis, we model, analyze and reduce, using classical tools, a high-dimensional transcription-translation model of RNA polymerase in *E. coli*. In the second part, we introduce a novel method called *Principal Process Analysis* (PPA) that allows the analysis of high-dimensional models, by decomposing them into biologically meaningful processes, whose *activity* or *inactivity* is evaluated during the time evolution of the system. Exclusion of processes that are *always inactive*, and *inactive* in one or several time windows, allows to reduce the complex dynamics of the model to its core mechanisms. The method is applied to models of circadian clock, endocrine toxicology and signaling pathway; its robustness with respect to variations of the initial conditions and parameter values is also tested. In the third part, we present an ODE model of the gene expression machinery of *E. coli* cells, whose growth is controlled by an external inducer acting on the synthesis of RNA polymerase. We describe our contribution to the design of the model and analyze with PPA the core mechanisms of the regulatory network. In the last part, we specifically model the response of RNA polymerase to the addition of external inducer and estimate model parameters from single-cell data. We discuss the importance of considering cell-to-cell variability for modeling this process: we show that the mean of single-cell fits represents the observed average data better than an average-cell fit.

Résumé (en français)

Cette thèse porte sur la modélisation, l'analyse et la réduction de modèles biologiques, notamment de réseaux de régulation génique chez la bactérie *E. coli*. Différentes approches mathématiques sont utilisées. Dans la 1ère partie de la thèse, on modélise, analyse et réduit avec des outils classiques un modèle de transcription-traduction de grande dimension de l'ARN polymérase (RNAP) chez *E. coli*. Dans la 2de partie, l'introduction d'une nouvelle méthode appelée *Analyse de Processus Principaux* (PPA) nous permet d'analyser des modèles de haute dimension, en les décomposant en processus biologiques dont l'activité est évaluée pendant l'évolution du système. L'exclusion des processus inactifs réduit la dynamique du modèle à ses principaux mécanismes. La méthode est appliquée à des modèles d'horloge circadienne, de toxicologie endocrine et de voie de signalisation; on teste également sa robustesse aux variations des conditions initiales et des paramètres. Dans la 3ème partie, on présente un modèle ODE de la machinerie d'expression génique de cellules d'*E. coli* dont la croissance est contrôlée par un inducteur de la synthèse de RNAP. On décrit notre contribution au développement du modèle et analyse par PPA les mécanismes essentiels du réseau de régulation. Dans une dernière partie, on modélise spécifiquement la réponse de RNAP à l'ajout d'inducteur et estime les paramètres du modèle à partir de données de cellules individuelles. On discute l'importance de considérer la variabilité entre cellules pour modéliser ce processus: ainsi, la moyenne des calibrations sur chaque cellule apparaît mieux représenter les données moyennes observées que la calibration de la cellule moyenne.

Résumé étendu (en français)

La vie est l'un des phénomènes les plus complexes dans l'univers [60]. En ce qui concerne la biologie, l'étude d'une seule unité de vie, la cellule, est une tâche infiniment compliquée.

Au cours du siècle dernier, les mécanismes cellulaires ont été étudiés dans différentes perspectives par des biologistes, des mathématiciens, des ingénieurs: par des expérimentations dans différentes conditions, par la modélisation mathématique du comportement cellulaire, par la calibration de ces modèles en utilisant des données expérimentales, par l'analyse et la réduction des structures de ces modèles à des fins différentes.

Toutes ces différentes études ont créé un domaine, un grand ensemble de connaissances, appelé biologie des systèmes [58], où différents auteurs ont contribué dans des directions diverses. L'objectif de cette thèse est d'y ajouter une brique.

Motivations

Un sujet majeur de la biologie des systèmes est la modélisation et l'analyse des réseaux cellulaires.

La création de modèles biologiques et leurs simulations dans différentes conditions sont déterminantes pour comprendre comment l'adaptation des organismes vivants aux signaux environnementaux résulte de grands réseaux de métabolites, d'ARN, de protéines et de leurs interactions mutuelles.

De plus en plus grands modèles cinétiques de réseaux cellulaires sont aujourd'hui publiés, comme résultat de décennies de travail en biologie, de progrès récents dans les biotechnologies [26, 60] et des progrès dans la modélisation et les approches d'estimation de paramètres (par exemple, voir [23] et [64]). La grande taille de ces modèles et leur non linéarité (en raison de boucles de rétroaction complexes) rendent leur calibration et leur analyse dynamique plutôt difficiles. Plus précisément, il est extrêmement difficile de relier le comportement global du système au fonctionnement de processus cellulaires spécifiques (par example La transcription de l'ARN, la phosphorylation des protéines

ou la formation de complexes), alors que ce gain de connaissances est essentiel pour identifier quels sont les processus cellulaires clés pour l'adaptation environnementale et quand ils sont en jeu.

Dans cet esprit, notre travail aborde différentes façons d'obtenir des informations sur le fonctionnement des cellules, en particulier sur la bactérie *Escherichia coli* [6]. Le réseau de régulation des gènes et la croissance de cet organisme modèle sont une grande source d'intérêts pour la communauté scientifique et pour l'industrie. En outre, l'expérimentation et la modélisation de *E. coli* sont l'un des principaux intérêts de l'équipe Inria Ibis et du groupe de Hans Geiselmann à l'Univ. Grenoble-Alpes avec lequel j'ai collaboré.

Approche

Les méthodes de réduction jouent un rôle central dans la conception des modèles. La description de la synthèse et de la consommation des composants biologiques d'un réseau peut conduire à un gros ensemble d'équations différentielles ordinaires (ODE): les approches de modèles classiques comme des *quasi-equilibrium approximation* ou des *quasi-steady-state approximation* (QSSA) [103] aident à réduire la dimension du modèle, à travers la séparation des échelles de temps. Cependant, la réduction du modèle avec ces approches n'est pas une tâche facile, en particulier pour les systèmes avec des boucles de rétroaction, que l'on trouve souvent dans les systèmes biologiques. Pour cette raison, dans la première partie de la thèse, nous montrons la réduction d'un modèle ODE de grande dimension, décrivant l'activité de l'ARN polymérase dans *E. coli*, qui favorise sa propre transcription. En utilisant la théorie des systèmes monotones et les arguments d'échelles de temps, nous pouvons le réduire à un modèle avec deux variables (ARN polymérase et son ARNm). Nous analysons le modèle réduit, en particulier la relation entre le taux de production RNAP, la quantité de ribosome et le taux de croissance cellulaire.

Ces outils classiques ont permis le développement de modèles plus grands, dont les formes réduites conservent encore de nombreuses équations et boucles de rétroaction. Si l'on considère le mécanisme complet d'expression de gènes de *E. coli*, par exemple, il ne

comprend pas seulement l'ARN polymérase, mais aussi les ribosomes, les protéines cellulaires et les métabolites, ainsi que leurs interactions régulatrices mutuelles. Pour analyser ces modèles cellulaires complexes, dans la deuxième partie de la thèse, nous présentons une nouvelle approche numérique appelée *Analyse de Processus Principaux* (PPA) qui permet à la fois l'analyse et la réduction des systèmes biologiques sans modifier leur structure principale. Basé sur la décomposition de la dynamique du système en processus biologiques *actifs* ou *inactifs* par rapport à une certaine valeur seuil, PPA apporte la connaissance des processus clés impliqués lors de l'évolution du système dans différentes fenêtres temporelles. Dans chaque fenêtre temporelle, le système est réduit à ses principaux mécanismes, négligeant les processus biologiques considérés comme *étant inactifs*. Cette approche est une méthode simple à utiliser, qui constitue un outil supplémentaire et utile pour analyser le comportement dynamique complexe des systèmes biologiques. La réduction de modèle qui en résulte n'entraîne pas une perte d'information ou de changements significatifs de la structure du modèle, comme cela se produit avec d'autres techniques de réduction. Pour tester la qualité de notre approche, nous appliquons la PPA sur différents systèmes biologiques à grande dimension: modèles d'horloges circadiennes, toxicologiques, et de voies de signalisation. Chaque analyse donne des informations biologiques importantes et, pour la plupart, nous obtenons un sous-modèle pour chaque fenêtre de temps proposée. En fait, la PPA peut être appliqué à n'importe quel modèle biologique exprimé par ODEs et il a été récemment utilisé par d'autres équipes de recherche [88, 95] à des fins d'analyse et de réduction, obtenant des résultats intéressants. Parce que notre approche est basée sur la connaissance a priori des trajectoires du système, elle dépend des paramètres et des valeurs de condition initiale: nous avons également testé la robustesse de la PPA aux valeurs des paramètres en utilisant l'analyse de sensibilité globale et les valeurs de condition initiale à l'aide d'une méthode ayant des similitudes avec un formalisme *piece-wise linear*.

Après avoir testé notre technique sur différents modèles biologiques, dans la troisième partie de la thèse, nous l'appliquons pour l'analyse d'un modèle, conçu par Delphine Ropers de l'équipe Inria Ibis, qui décrit le fonctionnement du mécanisme d'expression des gènes. En tant que tel, le modèle étend avec d'autres modules le modèle de transcription-traduction de l'ARN polymérase décrite dans la première partie. Il est également capable

de décrire le contrôle externe de la croissance de *E. coli* par un inducteur externe (IPTG) agissant sur la transcription des ARNm de la sous-unité de l'ARN polymérase.

La dernière partie de la thèse étudie en outre le contrôle externe de *E. coli* par IPTG, au moyen d'un modèle beaucoup plus simple de ce système, axé sur les processus clés nécessaires pour reproduire des observations biologiques sur l'expression des gènes dans des cellules individuelles, avec ou sans IPTG. Ce système est l'occasion d'aborder le problème de l'estimation des paramètres, qui suit immédiatement celui de la réduction du modèle. Dans le cas présent, nous calibrons le modèle simple en utilisant des données de gènes rapporteurs et des données de croissance obtenues dans des cellules individuelles traitées ou non avec IPTG [51]. Nous montrons que la calibration du modèle sur chaque cellule est préférable à la calibration d'un modèle moyen aux données moyennes, en raison de la grande variabilité entre les cellules et bien que cette variabilité soit incluse dans la procédure de calibration comme une erreur de mesure.

Organisation du manuscrit et contributions

Le manuscrit est organisé comme suit. Dans le premier chapitre introductif, nous décrivons brièvement la biologie cellulaire de *E. coli* et des méthodes pour contrôler sa croissance (Chapitre 3). Dans le deuxième chapitre introductif, nous présentons différents formalismes classiques pour concevoir, analyser et réduire les systèmes de réseau de régulation des gènes (GNR) (Chapitre 4).

Dans le Chapitre 5, nous nous concentrons sur le modèle de transcription-traduction de l'ARN polymérase dans *E. coli*. Mes contributions sont: effectuer des simulations de modèles complets et réduits avec un nouvel ensemble de paramètres pour obtenir des résultats plus réalistes d'un point de vue biologique; comparer un modèle classique de polymérase RNAP au modèle réduit obtenu, y compris une étude de sensibilité par rapport au nombre de ribosomes; concevoir et étudier un système réduit incluant un taux de croissance variable. Une version de ce chapitre a été soumise au journal *Bulletin of Mathematical Biology*, dans lequel je suis le deuxième auteur.

Dans le Chapitre 6, nous présentons l'*analyse de processus principaux* et la notion de poids relatif associés aux processus afin de les comparer. Nous appliquons la PPA à un modèle de rythmes circadiens chez les mammifères [73]: les erreurs relatives globales sont utilisées pour tester la qualité de la réduction, tandis que l'application de l'analyse de sensibilité globale nous permet de tester l'influence des paramètres du modèle sur ces erreurs. Les résultats obtenus prouvent la robustesse de notre méthode. J'ai développé en détail cette approche numérique à partir d'une version préliminaire développée par Jean-Luc Gouzé, avec la collaboration de ma co-encadrante Delphine Ropers. J'ai effectué l'analyse de sensibilité globale avec l'aide de Suzanne Touzeau (Biocore et INRA). Une version en papier de journal de ce chapitre a été soumise à *Journal of Theoretical Biology* dans laquelle je suis le premier auteur.

Les premières applications de PPA sur un modèle circadien de *Drosophila* [72] et un modèle de voie de signalisation [68] sont présentés en Annexe B: nous ne les insérons pas dans un chapitre ordinaire pour éviter les redondances avec le Chapitre 6. Ce travail a été présenté au *23ème Méditerranée Conférence sur le contrôle et l'automatisation* MED, tenue à Torremolinos, en Espagne, du 16 au 19 juin 2015 (avec des relectures par des pairs) et a été accepté comme un papier de conférence dans lequel je suis le premier auteur.

Le Chapitre 7 traite de la robustesse du modèle aux conditions initiales: nous évaluons la qualité de la PPA sur un ensemble de valeurs initiales possibles. Par souci de simplicité, et parce que les ordres de grandeur peuvent être importants dans les modèles biologiques, nous considérons les conditions initiales dans les rectangles représentant un ordre de grandeur et nous limitons cette approche à la dimension deux. Le plan est divisé en une grille logarithmique et nous appliquons (sous certaines hypothèses concernant la monotonie des processus) la PPA en calculant une limite maximale pour les poids de chaque processus. Nous conservons les processus *actifs* qui ont un poids dynamique plus élevé qu'un seuil fixe. Avec ce travail, nous démontrons la robustesse de notre méthode aux variations des conditions initiales. J'ai effectué ce travail en collaboration avec mon directeur de thèse Jean-Luc Gouzé et il sera présenté au Congrès mondial IFAC 2017

(avec relectures par les pairs) et a été accepté comme un papier de conférence dans lequel je suis le premier auteur.

Dans le Chapitre 8, nous appliquons la PPA sur un modèle déterministe conçu par Bayer CropScience [79], qui décrit les effets toxicologiques d'un fongicide sur les souris mâles. Nous voulons vérifier si les processus du modèle calibré sont *actifs* dans l'ordre attendu, connaissant la série d'événements clés proposés pour cette substance [96]. Pour cela, nous utilisons, en tant que critère de comparaison, les valeurs absolues des modèles biologiques et un seuil variable qui dépend des valeurs maximales et minimales des processus dans chaque variable. Nous appelons cette approche l'*Analyse Absolue des Processus Principaux* (APPA). Le travail a été réalisé en collaboration avec David Rouquié, *senior researcher* au centre de recherche en toxicologie de Bayer CropScience et avec Frédéric Dayan, fondateur d'ExactCure et ancien chef d'équipe de R&D chez Dassault Systèmes. Le système a été modélisé en 2014 par un stagiaire Bayer CropScience, Benjamin Miraglio, sous la supervision de David Rouquié et Frédéric Dayan. Ce travail fera partie d'un article de journal futur.

Après avoir établi PPA et testé la méthode sur différents modèles, nous l'utilisons maintenant pour étudier un nouveau modèle mathématique dans *E. coli*. Le modèle a été conçu par ma co-encadrante Delphine Ropers et décrit les mécanismes d'expression des gènes de la bactérie dans les détails ainsi que l'effet de l'inducteur IPTG sur celui-ci. Dans le Chapitre 9 nous présentons notre contribution au développement du modèle pour la description du taux de croissance cellulaire. Nous appliquons ensuite une PPA sur le modèle GEM pour analyser ses mécanismes de base et nous étudions l'effet de l'addition d'IPTG au milieu de culture sur la croissance de la bactérie. Ce travail fera partie d'un article de journal futur.

Dans le Chapitre 10 nous poursuivons l'étude du contrôle externe de *E. coli* par IPTG avec un modèle plus simple calibré avec des données expérimentales de surface cellulaire et de fluorescence, obtenu par Jérôme Izard lors de sa thèse de doctorat dans le laboratoire Adaptation et Pathogénie des Micro-organismes (Univ. Grenoble-Alpes). Nous comparons la calibration sur chaque cellule individuelle et la calibration de la cellule moyenne et montrons comment adapter le modèle à chaque cellule individuelle au lieu

d'une cellule moyenne, permettant d'obtenir plus d'informations sur la variabilité entre les cellules et donnant une meilleure qualité de calibration. J'ai effectué cette analyse en collaboration avec Eugenio Cinquemani de l'équipe Ibis et Delphine Ropers. Ce travail fera partie d'un article de journal futur.

Les conclusions de ces travaux de recherche ainsi que les perspectives sont données au Chapitre 11.

Cette thèse a été dirigée par Jean-Luc Gouzé (Inria Biocore) et co-encadrée par Delphine Ropers (Inria Ibis). Elle a bénéficié du support financier du Conseil Régional PACA et du projet RESET (ANR-11-BINF-0005) du programme Investissements d'Avenir Bioinformatique.

Acknowledgements

First of all I want to thank my supervisor Jean-Luc Gouzé to have picked me to be part of his team three and a half years ago: I have a lot of great memories about all the discussions we have done together about science in your office and I will always thank you for all the opportunities you gave me to learn more and more at Inria and around the world. It was a big pleasure to work with you. A big thanks goes to my co-supervisor Delphine Ropers: you have dedicated so much time to my scientific and professional growth. I will always remember the days in the laboratory in Grenoble when you "transformed" me into a biologist, performing real experiments on *E. coli*: since that week we had a really strong collaboration together and I will always carry these memories with me. A very special gratitude goes to Eugenio Cinquemani and Suzanne Touzeau: it was a pleasure to have worked with you and to have shared this experience with people so strong in science as you. You have taught me a lot. I am grateful to Hidde de Jong who introduced me in the RESET project. It was an honor to have been a part of your meetings in Grenoble: I have always admired your humility and kindness in listening to others, despite your large scientific knowledge. Special thanks go to Riccardo Bellazzi, Béatrice Laroche, Frédéric Dayan, David Rouquié, Gregory Batt and Gilles Bernot to have taken the time to read my work and to be here today. Thanks to my friends (and colleagues) Nicolas, Francesco, Pierre-Olivier, Natacha, David, Alfonso, Elsa, Camille, Riccardo, Lucie, Ivan, Carlos, Melaine, Ignacio, Ismail, Diego, Marjorie, Eleni, Claudia, Quentin, Sofia, Bapan and the babyfoot for the wonderful time spent together inside and outside Inria! Merci beaucoup to Stephanie to have helped me all these years with my missions and my continuous requests! Thanks to Olivier, Frédéric, Madalena, Francis, Marie-Line for having always made me feel very comfortable in the Biocore team. Un grazie gigante a Marco: I remember the day you came in my office looking for Italians! Since that day a wonderful friendship started: thank you for everything, I won't never forget it. A big efharistò to my best porcellini Christos and Konstantinos: you are great and kind people and you make me really understand that "una faccia una razza" and that team Porco Rosso is just the best. Grazie Lamberto for all the good talks at Inria: it was a pleasure meeting you, even for a short time! Thanks also to other great people I

met at Inria, namely Dmitry, Matthias, Adam, Dora, Nathalie, Valeria, David, Laurent and Juliette. Il ringraziamento più grande di tutti va a te, Susi, per tutto il sostegno che mi hai dato in questi mesi e per tutte le bellissime emozioni che mi fai provare: sei una persona meravigliosa, piena di passione per la vita e di energie inesauribili. Sei una fiamma vivacissima dalle mille sfumature e sono onorato di poter condivere tanto assieme a te! Grazie a voi mamma e papà per tutti gli aiuti che mi avete dato in questi anni e per avere sempre creduto in me e in quello che faccio! Un grazie anche a Franca per essere stata sempre gentile e ospitale con me. Grazie agli amici che ci sono e che c'erano nella nostra Nizza: a Matteo, Pietro, Massimo, Marta, Greta, Andrea, Massimo, Andrea, Enrico, Laura e Giulia. Due coincidenze fortunate al karaoke dell'Akathor e davanti alla porta sul retro del Tapaloca hanno portato ad una vita di bellissimi momenti che non dimenticherò mai. Grazie agli amici di Pavia: al mio dude Luca, a Marco, Giò, Sandro, Amer, Izio, Rassa, Ruffo, Cisco, Rich, Paul, Antonio, Federico, Ferro, Duca e Pelle. Ogni volta che ritorno mi sento sempre uno di voi e sembra che il tempo non stia volando affatto.

Contents

Chapter 1

Introduction

Life is one of the most complex phenomena in the universe [60]. When it comes to biology, the study of even a single unit of life, the cell, is not at all an easy task.

In the last century, cellular mechanisms were studied from different perspectives by biologists, mathematicians, engineers: through experimentations in different conditions, mathematical modeling of cell behavior, calibration of these models using experimental data, analysis and reduction of model structures for different purposes.

All these different studies created a field, a big wall of knowledge, called systems biology [58], where different minds contributed in their own way.

The aim of this thesis is to add a brick to it.

1.1 Motivations

A major topic of systems biology is in fact the modeling and analysis of cellular networks.

The creation of biological models and their simulations in different conditions are determinant in understanding how adaptation of living organisms to environmental cues results from large networks of metabolites, RNAs, proteins, and their mutual interactions.

Larger and larger kinetic models of cellular networks are nowadays published, as a results of decades of work in biology, recent advances in high throughput technologies [26, 60] and progress in modeling and parameter estimation approaches (for example, see [23] and [64]). The large size of these models and their non linearity due to complex feedback loops make their calibration and dynamical analysis rather difficult. More specifically,

it is extremely difficult to relate the global behavior of the system to the functioning of specific cellular processes (e.g. RNA transcription, protein phosphorylation, or complex formation), while this gain of knowledge is crucial to identify what are the key cellular processes for the environmental adaptation and when they are at play.

In this spirit, our work addresses different ways to gain information on cell functioning, especially on the bacterium *Escherichia coli* [6]. The gene regulatory network and the growth of this model organism is a source of interest in the scientific community and industry. Furthermore experimental and modeling of *E. coli* are one of the main interest of the Inria Ibis team and of the group of Hans Geiselmann at the Univ. Grenoble-Alpes with which I collaborated.

1.2 Approach

Reduction methods play a pivotal role in model designing. Describing the synthesis and the consumption of the biological components of a network can lead to a large set of ordinary differential equations (ODEs): classical model approaches as *quasi-equilibrium* approximations or *quasi-steady-state* approximations (QSSA) [103] help to reduce consistently the model dimension, through time scale separation. However model reduction with these approaches is not an easy task, in particular for systems with feedback loops, as often found in biological systems. For this reason, in the first part of the thesis, we show the reduction of a high dimensional ODE model, describing the activity of RNA Polymerase in *E. coli*, which promotes its own transcription. Using monotone system theory and time-scale arguments we are able to reduce it to a model with two variables (RNA polymerase and its mRNA). We analyze the reduced model with a specific focus on the relation between the RNAP production rate, ribosome quantity and cellular growth rate.

These classical tools have allowed the development of larger models, whose reduced forms still retain many equations and feedback loops. If we consider the full gene expression machinery of *E.coli*, for instance, it does not only include the RNA polymerase, but also the ribosomes, cell proteins and metabolites, as well as their mutual regulatory interactions. For analyzing such complex cellular models, in the second part of the thesis, we present a new numerical approach called *Principal Process Analysis* (PPA) that allows both the analysis and the reduction of biological systems without changing their main structure. Based on the decomposition of the system dynamics into biological processes that are *active* or *inactive* with respect to a certain threshold value, PPA brings the knowledge of which are the key processes involved during the system evolution in different time windows. In each time window the system is reduced at its core mechanisms,

neglecting the biological processes that are considered to be *inactive*. This approach is a simple-to-use method, which constitutes an additional and useful tool for analyzing the complex dynamical behavior of biological systems. The resulting model reduction does not lead to a loss of information or significantly changes of the model structure as can happen with other reduction techniques. To test the quality of our approach, we apply PPA on different high dimensional biological systems: circadian clocks, toxicological and signaling pathway models. Each analysis gives important biological information and for most of them we obtain a sub-model for each proposed time window. In fact PPA can be applied to any biological model expressed by ODEs and it has been recently used by other research teams [88, 95] for analysis and reduction purposes, obtaining interesting results. Because our approach is based on the *a priori* knowledge of system trajectories, it depends on parameter and initial condition values: we have also tested the robustness of PPA to parameter values using global sensitivity analysis and to initial condition values using a method, which shares similarity with *piece-wise linear* formalism.

After having tested our technique on different biological models, in the third part of the thesis, we apply it to analyze a model, designed by Delphine Ropers from the Inria Ibis team, that describes the functioning of the gene expression machinery. As such, the model extends with other modules the transcription-translation model of RNA polymerase described in the first part. It is also able to describe the external control of the growth of *E. coli* through an external inducer (IPTG) acting on the transcription of RNA polymerase subunit mRNAs.

The last part of the thesis further studies the external control of *E. coli* growth by IPTG, by means of a much simpler model of this system, centered around the key processes needed to reproduce biological observations on gene expression in single cells, with or without IPTG. This system is an occasion to tackle the problem of parameter estimation, which immediately follows that of model reduction. In the present case, we calibrate the simple model using reporter gene data and growth data obtained in single cells treated or not with IPTG [51]. We show that single-cell calibration of the model is preferable over fitting a mean model to the average data, due to the large cell-to-cell variability and despite its inclusion into the calibration procedure as a measurement error.

1.3 Organization of the manuscript and contributions

The manuscript is organized as follows. In the first introductory chapter we describe in a nutshell the cell biology of *E. coli* and methods to control its growth (Chapter 3). In the second introductory chapter we present different classical formalisms to design, analyze and reduce gene regulatory network (GNR) systems (Chapter 4).

In Chapter 5 we focus on the transcription-translation model of RNA polymerase in *E. coli*. My contributions are: performing simulations of the full and reduced models with a new set of parameters to have more realistic results from a biological point of view; comparing a classical model of RNA polymerase to the reduced model obtained, including a sensitivity study with respect to the number of ribosomes; designing and studying a reduced system including a variable growth rate. A version of this chapter has been submitted to the journal *Bulletin of Mathematical Biology*, in which I am second author.

In Chapter 6 we introduce *principal process analysis* and the notion of relative weights associated with processes in order to compare them. We apply PPA to a model of circadian rhythms in mammals [73]: global relative errors are used to test the quality of the reduction, while applying global sensitivity analysis allows us to test the influence of the model parameters on these errors. The results obtained prove the robustness of our method. I developed in detail this numerical approach from a preliminary version developed by Jean-Luc Gouzé, with the collaboration of my co-supervisor Delphine Ropers. I performed the global sensitivity analysis with the help of Suzanne Touzeau. A journal paper version of this chapter has been submitted to *Journal of Theoretical Biology* in which I am first author.

The first applications of PPA on a *Drosophila* Circadian model [72] and a signaling pathway model [68] are presented in Appendix B: we do not insert them in a regular chapter to avoid redundancy with Chapter 6. This work has been presented at the *23rd Mediterranean Conference on Control and Automation* MED, held in Torremolinos, Spain, on June 16th-19th, 2015 (with peer reviewed proceedings) and has been accepted as a conference paper in which I am first author.

Chapter 7 deals with the model robustness to the initial conditions: we assess the quality of PPA on an entire set of possible initial values. For the sake of simplicity, and because the orders of magnitude can be large in biological models, we consider initial conditions in rectangles representing one order of magnitude and we limit this approach to dimension two. The plane is divided in a logarithmic grid and we apply (under some assumptions concerning the monotonicity of the processes) PPA by computing a maximal bound for the weights of of each process. We retain the *active* processes that have a dynamical weight higher that a fixed threshold. With this work we prove the robustness of our method to variations of initial conditions. I performed this work in collaboration with my supervisor Jean-Luc Gouzé and it will be presented at the IFAC 2017 World Congress (with peer reviewed proceedings) and has been accepted as a conference paper in which I am first author.

In Chapter 8 we apply PPA on a deterministic model designed by Bayer CropScience [79], that mimics the toxicological effects of a fungicide on male mice. We want to verify if the processes of the calibrated model get *active* in the order expected, knowing the series of key events that have been proposed for this substance [96]. For this purpose we use, as a comparison criteria, the absolute values of the biological models and a varying threshold that depends on the maximum and minimum values of the processes in each variable. We call this approach *Absolute Principal Process Analysis* (APPA). The work has been done in collaboration with David Rouquié, senior researcher at the toxicology research center of Bayer CropScience, and with Frédéric Dayan, ExactCure founder and former R&D team leader at Dassault Systèmes. The system was modeled in 2014 by a Bayer CropScience intern, Benjamin Miraglio, under the supervision of David Rouquié and Frédéric Dayan. This work will be a part of a future journal paper.

Having established PPA and tested the method on various models, we now use it to study a new mathematical model in *E. coli*. The model has designed by my co-supervisor Delphine Ropers and describes the gene expression machinery of the bacterium in details as well has the effect of the inducer IPTG on it. In Chapter 9 we present our contribution to the model development for the description of cell growth rate. We then apply PPA on the GEM model to analyze its core mechanisms and we study the effect of IPTG addition to the culture medium on the bacterium growth. This work will be a part of a future journal paper.

In Chapter 10 we continue the study of the external control of *E. coli* by IPTG with a simpler model calibrated with experimental data of cellular area and fluorescence, obtained by Jérôme Izard during his PhD thesis in the laboratoire Adaptation et Pathogénie des Micro-organismes (Univ. Grenoble-Alpes). We compare single-cell calibration and average-cell calibrations, and show how fitting the model to each individual cell instead of an average cell, leads to more information about cell-to-cell variability and results in a better calibration quality. I performed this analysis in collaboration with Eugenio Cinquemani from the Ibis team and Delphine Ropers. This work will be a part of a future journal paper.

Conclusions for these research works together with perspectives are given in Chapter 11.

Chapter 2

Introduction (en français)

Cette thèse porte sur la modélisation, l'analyse et la réduction de modèles biologiques, notamment de réseaux de régulation génique chez la bactérie *E. coli*. Différentes approches mathématiques sont utilisées. Dans la 1ère partie de la thèse, on modélise, analyse et réduit avec des outils classiques un modèle de transcription-traduction de grande dimension de l'ARN polymérase (RNAP) chez *E. coli*. Dans la 2de partie, l'introduction d'une nouvelle méthode appelée *Analyse de Processus Principaux* (PPA) nous permet d'analyser des modèles de haute dimension, en les décomposant en processus biologiques dont l'activité est évaluée pendant l'évolution du système. L'exclusion des processus inactifs réduit la dynamique du modèle à ses principaux mécanismes. La méthode est appliquée à des modèles d'horloge circadienne, de toxicologie endocrine et de voie de signalisation; on teste également sa robustesse aux variations des conditions initiales et des paramètres. Dans la 3ème partie, on présente un modèle ODE de la machinerie d'expression génique de cellules d'*E. coli* dont la croissance est contrôlée par un inducteur de la synthèse de RNAP. On décrit notre contribution au développement du modèle et analyse par PPA les mécanismes essentiels du réseau de régulation. Dans une dernière partie, on modélise spécifiquement la réponse de RNAP à l'ajout d'inducteur et estime les paramètres du modèle à partir de données de cellules individuelles. On discute l'importance de considérer la variabilité entre cellules pour modéliser ce processus: ainsi, la moyenne des calibrations sur chaque cellule apparaît mieux représenter les données moyennes observées que la calibration de la cellule moyenne.

Chapter 3

Notes on molecular cell biology

Various biological systems have been studied during this PhD thesis, from the bacterium *Escherichia coli* to the fly *Drosophila*. Rather than describing these systems in detail, we will introduce in this chapter important concepts of cell biology in the case of the bacterium *E. coli*. For more details, see [6] and [7].

3.1 *Escherichia coli*

Because cells descend from a common ancestor, studying properties of one organism can help to understand the properties of others [6]: usually these model organisms are chosen for their easy genetic manipulation and cultivation in the laboratory or because they can survive under certain conditions of stress.

The bacterium *Escherichia coli* is a model organism for **prokaryotic cells**. It is commonly found in the lower intestine of warm-blooded organisms and was one of the first organisms to have its complete genome sequenced [25]. The bacterium has a rod-shaped form and is typically 2 μM long. Its cell wall consists of an outer membrane and an inner membrane containing only one compartment with cytoplasm and generally, no organelles. The cytoplasm contains most of the cell components: DNA, RNAs, proteins, metabolites... It is the place where most cellular processes take place: the metabolism, DNA replication, gene expression processes for instance (see Figure 3.1). Complex molecular machineries also present in the cytoplasm catalyze these processes: for instance, the ribosomes, responsible for the production of proteins, and the RNA polymerase, involved in RNA synthesis.

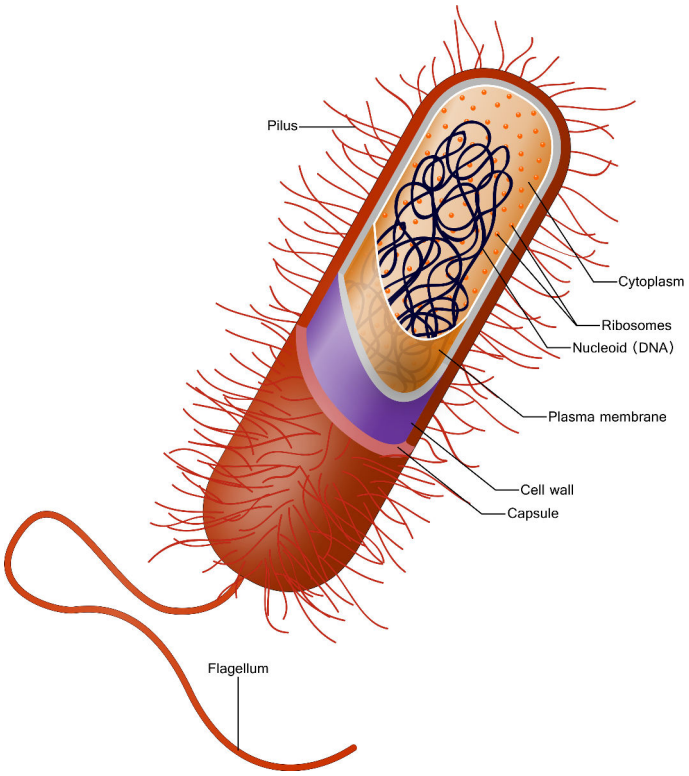


FIGURE 3.1: **Schematic representation of a prokaryotic cell**. Intracellular components (proteins, DNA and metabolites) are located within the cytoplasm, protected by the cell wall composed of two membranes. On their surface, *E. coli* cells carry a lash-like appendage called flagellum, useful to move in a fluid-like environment and to detect concentration gradients and other signals (picture taken from [1]).

3.2 Growth of *E. coli*

E. coli reproduce asexually by a process called binary fission [84], involving an orderly increase in the quantity of cellular constituents: in terms of cell mass and number of ribosomes, followed by a duplication of the bacterial chromosome, the synthesis of new cell walls, the partitioning of the two chromosomes, the septum formation, and the cell division.

In the laboratory, bacterial growth can be studied from two different perspectives [122]:

- At the level of the single cell, where the increase in cell length or cell volume is monitored;
- At the population level, with the monitoring of the population size (expressed in number of cells or total biovolume).

If N is the population size or the cell volume, we can define the bacterial growth rate as:

$$\frac{dN(t)}{dt} = \mu(t) \cdot N(t) \tag{3.1}$$

where μ is the specific growth rate.

When cells are grown in population in a batch culture, that is, when the environment changes over time, the growth can be decomposed in four different phases (see also Figure 3.2):

- The Lag Phase: bacteria adjust to the new environmental conditions by adapting gene expression in order to resume growth;
- The Log phase or exponential phase: cell grow and divide at a constant rate such that the number of cells doubles with each consecutive time period. In this phase the growth rate expressed by Equation (3.1) is maximal and constant (cells are in a quasi-steady-state growth);
- The Stationary phase: the depletion of a growth-limiting factor such as a nutrient arrests growth. In this phase growth rate and death rate are equal. *E. coli* and other bacteria produce secondary metabolites, such as **antibiotics**, during this phase [84]. The growth rate expressed by Equation (3.1) is null;
- The Death phase: bacteria die.

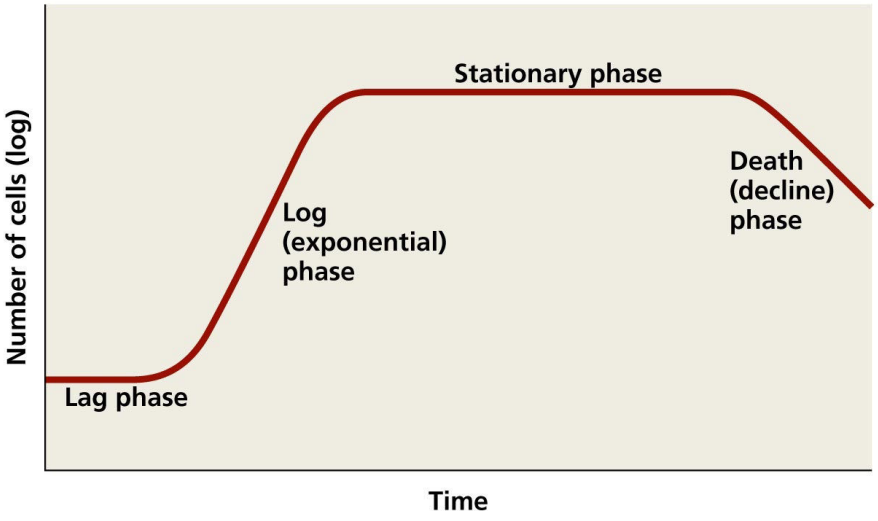


FIGURE 3.2: **Bacterial growth curve**. The evolution of the size of the bacterial population is represented along time on a logarithmic scale (picture taken from [2]).

The generation time of *E. coli* bacteria depends widely on the environmental conditions, from 20 minutes to several hours. An example is shown in Figure 3.3, where *E. coli* bacteria were grown in two different growth media containing either glucose as a carbon source, or a mixture of glucose and amino acids. I did myself the experiment in the group of Hans Geiselmann at the Univ. Grenoble-Alpes, associated with the Ibis project-team. Amino acids in the second growth medium can be used directly as building blocks for

the synthesis of proteins, as a result of which bacteria in the second medium grow faster than in the presence of glucose only. The size of their population increases to reaches a plateau when glucose is depleted. Cells subsequently start a new phase of growth where they use the amino acids as a source of carbon.

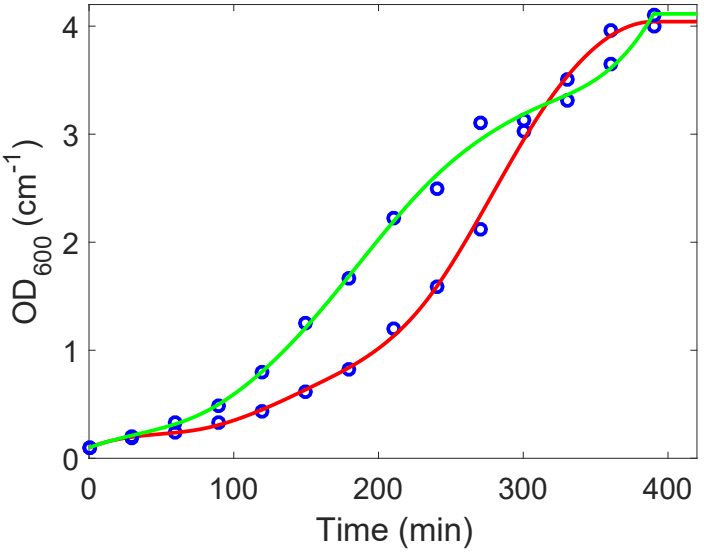


FIGURE 3.3: **Growth kinetics of *E. coli*.** The *E.coli* strain K12 BW25113 was inoculated in two minimal media M9 supplemented with 0.3% glucose, in the absence or presence of 0.1% casamino acids (CAA). Samples were taken every 30 minutes during 420 minutes, and their optical density at 600 nm (OD_{600}) was measured. Optical density is generally proportional to the number of bacteria in the sample. The continuous lines are spline fits of the data: red, resp. green, line in absence, resp. presence, of CAA. I performed this experiment within the group of Hans Geiselmann (Univ. Grenoble-Alpes), associated with the Ibis project-team.

3.3 Gene expression

The adaptation of *E. coli* to different environmental conditions is done through the reprogramming of **gene expression**. This process leads to the synthesis of proteins, starting from the information stored inside genes. Proteins have regulatory and structural functions needed to form new bacteria cells.

The genome of *E. coli* is a circular double-stranded DNA of approximately 4.6 million nucleotide pairs, coding for 4288 different proteins (see Figure 3.4).

In this section, we describe the main phases of gene expression: the transcription of genes into **RNAs** and the translation of mRNAs (messenger ribonucleic acids) into proteins.

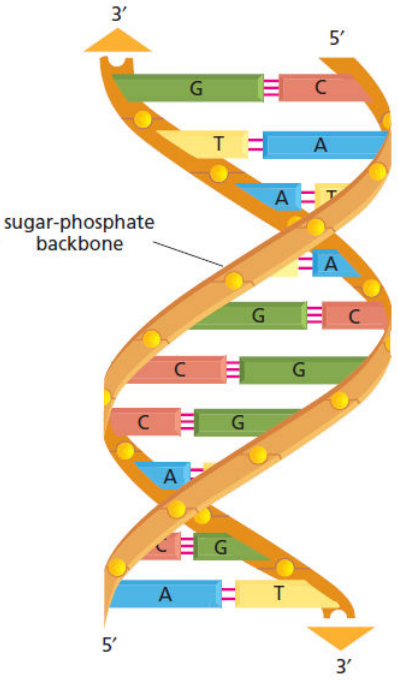


FIGURE 3.4: **DNA molecule**. DNA is built with four types of nucleotides, each of them is composed of a sugar-phosphate covalently linked to a base (adenine (A), cytosine (C), guanine (G) or thymine (T)). They are also linked together through a sugar-phosphate backbone forming a polynucleotide chain. Two chains, held together by hydrogen bonds between the paired bases, form a DNA helix (picture taken from [6]).

3.3.1 Transcription

During transcription, the information is copied in another chemical form, but still in the language of nucleotides: RNAs are linear polymers made of a single-stranded helix containing ribonucleotides.

The enzyme that catalyzes transcription is called **RNA polymerase** (RNAP): one of its sub-unit, called σ factor, recognizes and binds to the promoter region of a gene. Once bound, RNAP moves stepwise along the DNA, using energy to open the double helix of DNA and adding ribonucleotides one by one to the growing RNA, which are complementary to one of the two DNA strands. Transcription stops when RNAP meets a termination site. There are two possible mechanisms, ρ dependent or ρ independent, depending on whether the protein ρ binds to the transcription terminator pause site or not. The transcription process is over and RNAP halts, releasing the RNA molecule (see Figure 3.5).

Different types of RNAs can be produced in *E. coli*:

- RNA molecules that are transcribed from genes coding for proteins are called **messenger RNAs** (mRNAs);

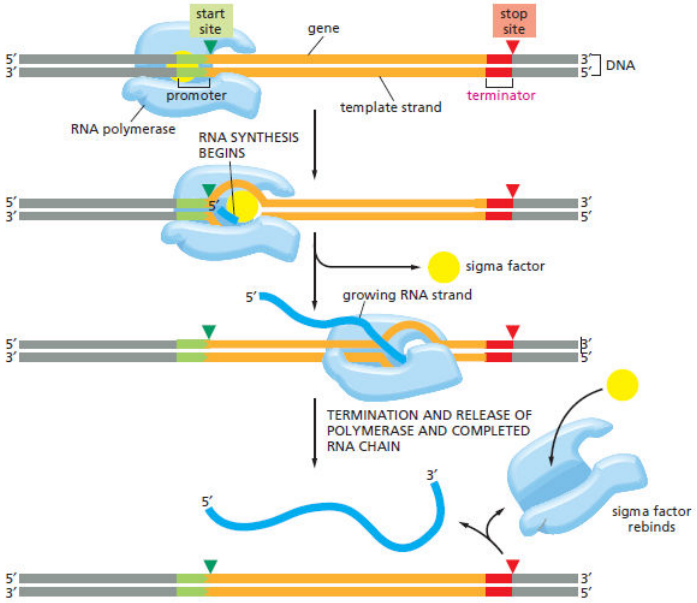


FIGURE 3.5: **Transcription**. The σ subunit of RNAP (RNA Polymerase) identifies the promoter (green in the figure) in the DNA and allows the binding of the enzyme. RNAP opens the double helix of DNA and transcription starts: the σ factor is released and RNAP synthesizes the RNA, by adding each time a ribonucleotide to the chain (the bases of ribonucleotides are called adenine (A), guanine (G), cytosine (C) and uracil (U)). Transcription stops when RNAP meets the terminator signal of DNA (red in the figure). At this point, RNAP halts and releases both the DNA template and the newly-made RNA. The enzyme then binds again to the σ factor, searching for a new DNA promoter to bind (picture taken from [6]).

- RNA molecules that form the ribosome, essential for translation process, are called **ribosomal RNAs** (rRNAs);
- RNA molecules that carry an amino acid to the ribosome for protein synthesis are called **transfer RNAs** (tRNAs);
- **Small RNAs** that are non-coding RNA sequences with regulatory functions within cells.

The total amount of rRNAs and tRNAs is called **stable RNAs** (sRNAs).

3.3.2 Translation

While DNA and RNA are chemically and structurally similar, RNA and proteins differ in composition: proteins are made of amino acids covalently linked during translation. The mRNA sequence is decoded in sets of three nucleotides, called codons, each coding for an amino acid. Due to the redundancy of the genetic code, there are $4^3 = 64$ possible

combinations of three nucleotides, even though only 20 amino acids are commonly found in proteins.

The process of translation is described in Figure 3.6. It is catalyzed by ribosomes, large macromolecular complexes made of three ribosomal RNAs and more than fifty ribosomal proteins. The ribosomes assemble on the ribosome binding site of mRNAs, from which they move three nucleotides by three nucleotides to allow an accurate and rapid translation of the genetic code. Specific incorporation of amino acids in nascent proteins is ensured by tRNAs: they carry an amino acid and possess a sequence called anti-codon, complementary to a mRNA codon.

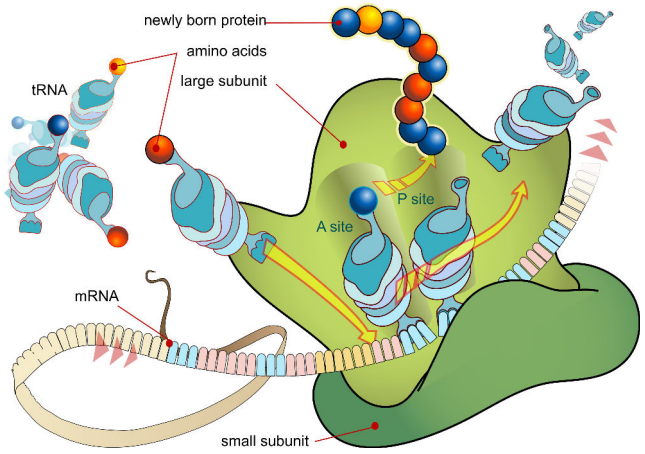


FIGURE 3.6: **Translation process**. In the initialization phase, the ribosome assembles onto the mRNA and the first tRNA binds to the start codon. In the elongation phase, the tRNA transfers an amino acid to the tRNA corresponding to the next codon. The ribosome then moves to the next mRNA codon to continue the process, creating an amino-acid chain. In the termination phase, when a stop codon is reached, the ribosome releases the polypeptide (picture taken from [3]).

3.3.3 mRNA degradation

While proteins are usually stable and essentially consumed through growth dilution, mRNA are labile and can degrade: in *E. coli* they are actively degraded by enzymes, like the **Ribonuclease E (RNAse E)** [76]. When mRNAs are not used in translation (and thus not protected by translating ribosomes), they have a much higher probability to be degraded. Contrary to mRNAs, stable RNAs like tRNAs and rRNAs are much more stable due to their three dimensional structure.

3.4 Regulation of gene expression in *E. coli*

Every type of cell, including *E. coli*, is able, through a wide range of mechanisms to increase or decrease the production of a specific gene product. These regulations allow cells to adjust gene expression levels to external signals, for example, according to the food sources that are available in the environment [33, 85, 87].

Cells control gene regulation at different levels, by [6]:

- Controlling when and how frequently a given gene is transcribed;
- Selectively degrading certain mRNA molecules;
- Selecting which mRNA are translated by ribosomes;
- Selectively activating or inactivating proteins following their synthesis.

Here we will focus on the regulation of transcription. The promoter of a gene contains a binding site for the RNA polymerase, as well as binding sites for transcription factor(s) if its expression is regulated. These factors can be:

- a **repressor** protein if, in its active form, it blocks the binding of RNAP to the promoter, thus switching genes off;
- an **activator** protein if, in its active form, it switches some genes on by binding nearby the promoter and recruiting RNAP to the promoter to initiate transcription.

In addition to these specific regulations, **global effects** such as the abundance of ribosomes and RNA polymerase contribute to adjust gene expression to the environmental conditions [61]. These effects are growth-rate dependent: for instance the number of ribosomes and RNAP vary with the growth rate, which directly affects the rates of transcription and translation.

A vivid example of regulation of gene expression is the glucose-lactose diauxie. If a culture medium contains both glucose and lactose, *E. coli* cells will preferentially use glucose by blocking the transport and metabolism of lactose through the transcriptional inhibition of the *lac* operon. This first phase of growth on glucose stops with the depletion of the carbon source. After some time during which bacteria express the enzymes needed for growth on lactose, they resume growth on this nutrient.
The choice of using glucose or lactose is regulated by two mechanisms controlling the expression of the *lac* operon. One mechanism is the **carbon catabolite repression**

[114]: depletion of glucose is accompanied by the production of high levels of a small molecule, cAMP (cyclic adenosine monophosphate), which binds to the catabolite repressor protein (CRP). The complex CRP-cAMP is active in transcription: it stimulates transcription of the *lac* operon by binding near the *lac* promoter, which helps to recruit RNAP onto the promoter region. The second regulatory mechanism informs bacteria about the presence of lactose in the growth medium through the accumulation of allolactose, a product of lactose metabolism within cells (Figure 3.7). Allolactose binds to the lactose repressor, LacI, which makes the protein unable to bind to the operator sequence next to the *lac* promoter and relieves the transcriptional inhibition of the operon. Thus, when glucose is absent and lactose is present, the two regulatory mechanisms ensure that RNAP binds to the promoter region and maximally transcribe the *lac* genes whose products are needed for cells to start growing on lactose.

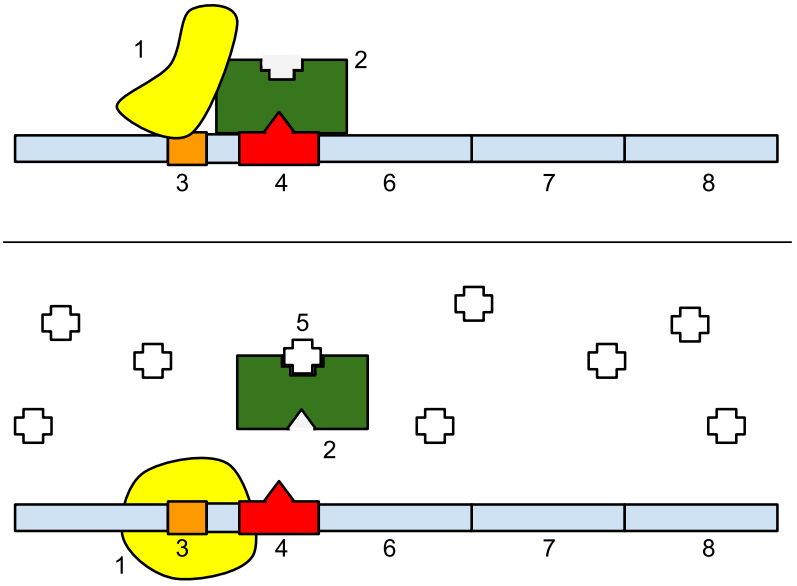


FIGURE 3.7: **Lac operon**. The *lac* operon includes three genes: *lacZ* (6) coding for the β-galactosidase cleaving lactose into glucose and galactose; *lacY* (7) whose product is a permease involved in the transport of lactose, and *lacA* (8) coding for a β-galactoside transacetylase which is not directly involved in lactose metabolism. In the top panel, RNAP (1) cannot bind to the promoter (3) due to the repressor (2) binding to the operator (4). In the bottom panel, the allolactose (5) binds to the repressor, so that RNAP can bind to the promoter region of the *lac* operon (picture taken from [4]).

The regulation of the *lac* operon has inspired various applications, in which a synthetic *lac* promoter is used to control the transcription of a gene of interest. One example concerns the modification of *E. coli* to create a strain whose growth rate can be controlled: this topic is of interest in bio-technologies, where the arrest and re-start of bacterial growth allows to maximize the production of products of interest. For example, in [51], the growth rate of *E. coli* has been artificially modulated, by controlling the transcription of the *rpoBC* genes coding for the $\beta\beta'$ sub-units of RNAP, through the replacement of their natural gene promoter by a synthetic *lac* promoter. In this case, **isopropyl**

***β*-D-1-thiogalactopyranoside (IPTG)** is used as a synthetic inducer that mimics allolactose, without being metabolized by the cell:

- When IPTG is added to the culture medium, it enters the cells where it binds to the repressor LacI and allows transcription of *rpoBC* genes. These conditions allow the synthesis of new RNAP, that induce the expression of proteins needed by cells to grow and divide;
- When IPTG is absent, LacI binds to its operator in the *rpoBC* promoter region. RNAP is no longer expressed, other cell proteins are no longer synthesized and cells stop growing.

We will come back to this application in Chapters 9 and 10.

Chapter 4

Modeling genetic regulatory network systems

As explained in Chapter 3, the growth and adaptation to the environmental conditions in *E.coli* are due to gene expression and its control. To study and understand the connections through positive and negative loops between genes, mRNA, proteins and other cell elements, within **gene regulatory networks** (GRNs), mathematical and computer tools are necessary [31, 43, 108].

In this chapter, we present the most well known formalism to model a GRN (ordinary differential equations), classical tools to analyze graphically and mathematically the stability of the system (phase plane analysis and Jacobian matrix), reduction methods to simplify the structure of the model (quasi-steady-state assumptions and piece-wise formalism), methods to study the uncertainty of the system (parameter sensitivity analysis) and to calibrate the model (least-square fitting).

We illustrate these methods with two small regulatory circuits: one including the interaction between the mRNA and the protein of a generic gene and one with two proteins mutually inhibiting the expression of their gene.

We describe these techniques because they are applied in the following chapters. Quasi-steady-state approximation is used to reduce a high dimension model of the transcription-translation of RNA Polymerase. Jacobian matrix and phase plane analysis are used to calculate the steady states of the reduced system and to visualize them graphically (Chapter 5). These methods are then briefly compared to our technique called **principal process analysis (PPA)** that allows both the analysis and the reduction of biological systems (Chapter 6 and Appendix B). Global parameter sensitivity analysis

is applied in Chapter 6 to verify the robustness of PPA. Some ideas from the piecewise linear formalism, like regular domains, switching domains and transition graphs, are combined to PPA to extend our reduction methodology to biological models with initial conditions spanning several orders of magnitude (Chapter 7). Then in Chapter 10, parameter fitting is applied to calibrate single cell models and average cell models.

4.1 Ordinary differential equation models

Ordinary differential equation (ODE) systems are the mostly used formalism to model gene regulatory networks. Example of biological models involving the ODE formalism can be found in [36, 44, 56].

The ODE formalism models the concentration of mRNAs, proteins and other cell elements which are represented by non-negative continous time variables. Regulatory interactions take the form of functional and differential relations between the concentration variables. More specifically, gene regulation is modeled by **reaction-rate equations** expressing the rate of production of a gene product - a protein or mRNA - as a function of the concentrations of other elements of the system [31].

Reaction-rate equations have the mathematical form:

$$\frac{dx_i}{dt} = f_i(x),\ x_i \geq 0,\ 1 \leq i \leq n \tag{4.1}$$

where $x = (x_1, \ldots, x_n)^t \in \mathbb{R}^n_+$ is the concentration of n molecular species in the system (mRNAs, proteins, metabolites). If we distinguish the positive contribution to the molecular species x_i as the production or the synthesis process ($g_i(x) \geq 0$) and the negative contribution as the dilution, degradation or transformation in other species ($d_i(x) \geq 0$) Equation (4.1) becomes [20]:

$$\frac{dx_i}{dt} = g_i(x) - d_i(x). \tag{4.2}$$

4.1.1 Modeling transcription-translation

Transcription and translation can be modeled with the formalism of Equation 4.2, by taking into account the activator and repressor proteins that enhance/reduce transcription and translation rates.

Let us call A the activator protein, m the number of proteins A and D the promoter site of a gene. When A binds to D we have the complex C through the reaction:

$$D + m\,A \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C \tag{4.3}$$

where k_1 and k_2 are the **reaction rates**, which indicate how quickly or slowly a reaction takes place. We can model this reaction in a set of ODEs using the **law of mass-action**, where the rate of a chemical reaction is directly proportional to the product of the activities or concentrations of the reactants [38, p.3]:

$$\begin{aligned} \dot{C} &= k_1\,D\,A^m - k_2\,C, \\ \dot{D} &= -\dot{C}. \end{aligned} \tag{4.4}$$

Figure 4.1 shows the formation of the complex C.

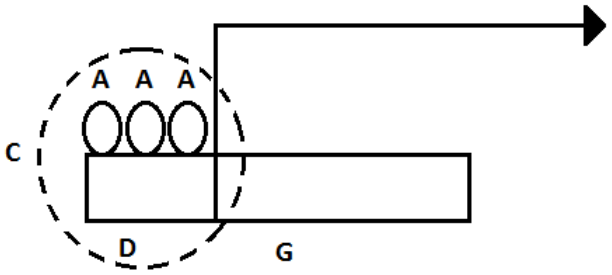


FIGURE 4.1: **Formation of complex *C*.** $m = 3$ activator proteins bind to the promoter D of the gene G to form the complex C.

Applying the **law of conservation of mass** - that states that for any system closed to all transfers of matter and energy, the mass of the system must remain constant over time - the equation $D+C = D_T$ is set: the total amount of promoter sites, free or bound, remains constant. Knowing that the binding processes are faster than transcription we suppose that $\dot{C} \approx 0$ (see the quasi-steady-state assumption in Section 4.1.2). We then obtain the equations:

$$\begin{aligned} C &= D_T \frac{A^m}{\theta_A^m + A^m}, \\ D = D_T - C &= D_T \frac{\theta_A^m}{\theta_A^m + A^m}, \end{aligned} \tag{4.5}$$

with $\theta_A = (\frac{k_2}{k_1})^{\frac{1}{m}}$. The amount of mRNA produced depends both on the concentration of free DNA sites and the concentration of DNA sites bound to an activator or repressor [20]: supposing that the effect of repressors and activators can be modeled independently, that the production of mRNA is linearly dependent on D and C, and that the mRNA

degrades at constant rate, the equation for transcription is:

$$\dot{M} = \alpha_0\, D + \alpha_1\, C - \gamma_M\, M. \tag{4.6}$$

We can have two distinct cases. In the case of an activator, the contribution of C on mRNA is much larger than that of D (then $\alpha_1 \gg \alpha_0$). Setting the basal activity $\kappa_0 = \alpha_0\, D_T$ and the parameter $\kappa_1 = (\alpha_1 - \alpha_0)\, D_T$, we obtain for the activator case:

$$\dot{M} = \kappa_0 + \kappa_1 \frac{A^m}{\theta_A^m + A^m} - \gamma_M\, M. \tag{4.7}$$

In the case of a repressor, the contribution of C to mRNA production is much smaller than that of D ($\alpha_1 \ll \alpha_0$). Setting the basal activity $\kappa_0 = \alpha_1\, D_T$ and $\kappa_1 = (\alpha_0 - \alpha_1)\, D_T$:

$$\dot{M} = \kappa_0 + \kappa_1 \frac{\theta_A^m}{\theta_A^m + A^m} - \gamma_M\, M. \tag{4.8}$$

The function $h^+(A,\theta,m) = \frac{A^m}{\theta_A^m + A^m}$ in its positive form and in its negative form $h^-(A,\theta,m) = \frac{\theta_A^m}{\theta_A^m + A^m}$ is called the **Hill function**: $h^+(A,\theta,m)$ describes a curve that starts from zero and approaches unity [92] and $h^-(A,\theta,m)$ describes the opposite case. The parameter θ is the *expression threshold* of the protein A necessary to produce a significant increase of mRNA and the parameter m is called Hill coefficient. It controls the steepness of the Hill functions (the higher is m, the more step-like is the Hill function): if $m = 1$ the function is then called the **Michaelis-Menten equation**. Figure 4.2 shows the steepness of Hill functions at different values of m.

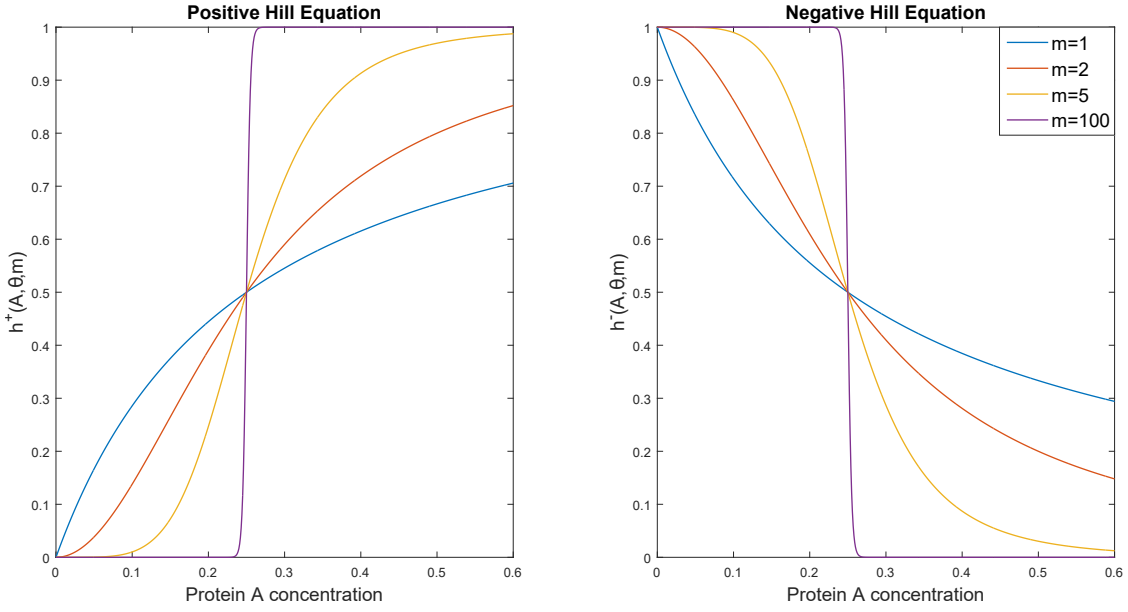


FIGURE 4.2: **Hill function**. Positive (resp. negative) Hill function on the left (resp. right) for different Hill coefficients: m=1 (Michaelis-Menten case), 2, 5, 100. The expression threshold of protein A is set at $\theta = 0.25$.

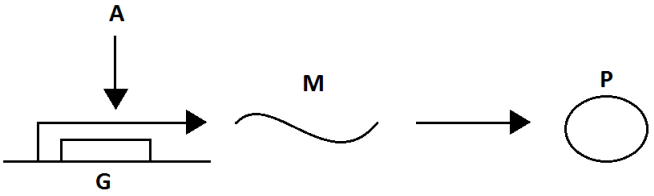


FIGURE 4.3: **Classical model of gene regulation**. The regulation of the gene G by the protein A.

Translation can be modeled as a linear function of mRNA concentration with a degradation term as well [20]:

$$\dot{P} = \kappa_2 M - \gamma_P P. \tag{4.9}$$

Therefore the classical model of gene regulation is:

$$\begin{cases} \dot{M} = \kappa_0 + \kappa_1 h^+(A, \theta, m) - \gamma_M M, \\ \dot{P} = \kappa_2 M - \gamma_P P. \end{cases} \tag{4.10}$$

The gene G is transcribed in the M mRNA and the latter is translated in the protein P. The transcription of the gene is regulated by the protein A, see Figure 4.3.

4.1.2 Quasi-steady-state assumption of mRNA concentration

It is possible to further simplify the system using the **quasi-steady-state-assumption** (QSSA) [103]. QSSA is an well-known approximation method in biochemical kinetics an and other fields, simplifying the ODE systems with two relevant time scales (fast and slow scale).

Most of the time mRNA dynamics in GRNs is much faster than protein dynamics, i.e. the mRNA concentration reaches its equilibrium faster than that of the protein (typical mRNA half-lives are 2-6 minutes, while those of proteins are on the order of hours [8]).

So, in System (4.10), the mRNA M is degrading faster than the protein P ($\gamma_M \gg \gamma_P$): because mRNA concentrations reaches its **equilibrium point** - the point where $\dot{M} = 0$ - on a time scale much quicker than the concentration of the protein, we can apply the QSSA in System (4.10).

We now consider the case with an *activator* protein [20]. We do a time variable change ($\tau = \gamma_P t$) and we obtain the scaled system:

$$\frac{dM}{d\tau} = \frac{\kappa_0}{\gamma_P} + \frac{\kappa_1}{\gamma_P} \frac{A^m}{\theta_A^m + A^m} - \frac{\gamma_M}{\gamma_P} M, \tag{4.11}$$

$$\frac{dP}{d\tau} = \frac{\kappa_2}{\gamma_P} M - P. \tag{4.12}$$

The Tikhonov's theorem (see Appendix C.2), for a fixed value of the concentration of the activator protein, can be applied, setting $y = M$, $x = P$, $\epsilon = \frac{\gamma_P}{\gamma_M}$ and with $f(x, y, \epsilon) = \frac{\kappa_2}{\gamma_P} y - x, g(x, y, \epsilon) = \frac{\kappa_0}{\gamma_M} + \frac{\kappa_0}{\gamma_M} + \frac{\kappa_1}{\gamma_m} \frac{A^n}{\theta_A^n + A^n} - y$.

If we substitute the quasi-steady-state equation for M ((4.11), where $\frac{dM}{d\tau} = 0$) in the Equation (4.12) for P rewritten in the original time variable, we obtain:

$$\dot{P} = \tilde{\kappa}_0 + \tilde{\kappa}_1 \frac{A^m}{\theta_A^m + A^m} - \gamma_P P, \tag{4.13}$$

where $\tilde{\kappa}_0 = \frac{\kappa_2 \kappa_0}{\gamma M}$ and $\tilde{\kappa}_1 = \frac{\kappa_2 \kappa_1}{\gamma M}$.

4.2 Analysis of a genetic bistable switch

Once we obtain one single ODE Model (4.13), it is possible to study the case where two proteins are inhibiting mutually their expression. The protein P_1 is the repressor of the protein P_2 and vice-versa. We model the System (4.14):

$$\begin{aligned} \dot{P}_1 &= \kappa_{10} + \kappa_{11} \frac{\theta_2^{m_2}}{\theta_2^{m_2} + P_2^{m_2}} - \gamma_1 P_1, \\ \dot{P}_2 &= \kappa_{20} + \kappa_{21} \frac{\theta_1^{m_1}}{\theta_1^{m_1} + P_1^{m_1}} - \gamma_2 P_2. \end{aligned} \tag{4.14}$$

Figure 4.4 shows this mutual inhibition.

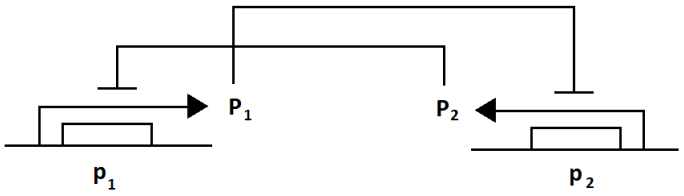


FIGURE 4.4: **Mutual inhibition between gene p_1 and p_2.** The protein P_1 inhibits the transcription of the gene p_2 and the protein P_2 inhibits the transcription of gene p_1.

4.2.1 Phase plane analysis

It is possible to represent the interaction between P_1 and P_2 by plotting one concentration against one another.

This graphical representation is called system's **phase plane** or **phase portrait**: it shows the trajectory that starts at the initial starting point or **initial values** of the system ($P_1(0)$,$P_2(0)$) and, because the concentration of molecular species are not negative, the solutions of the system $P_1(t)$ and $P_2(t)$ evolves in the region of space called **positive orthant** where all coordinates are zero or positive. In the case of System (4.14) the positive orthant has 2 dimensions ($\mathbb{R}^2_+ = \mathbb{R}_+ \times \mathbb{R}_+$).

It is possible to use vectors to indicate the direction and speed at each point of phase plane: the overall plot is called **direction field** that can be determined directly from the differential equation model. For System (4.14), written in its general form,

$$\begin{aligned} \frac{d}{dt}P_1(t) &= f_1(P_1(t), P_2(t)), \\ \frac{d}{dt}P_2(t) &= f_2(P_1(t), P_2(t)), \end{aligned} \tag{4.15}$$

the motion in the phase plane at any given point (P_1,P_2) is given by the vector $(f_1(P_1, p_2), f_2(P_1, P_2))$ [47].

An important feature of phase portrait are the points where the trajectories change direction with respect to one axis, more precisely where one of the two variables reaches a local minimun or maximum with respect to time. These points constitute the system **nullclines**. The set of points (P_1, P_2), where $\dot{P}_1 = f_1(P_1, P_2) = 0$ is called P_1-nullcline and where $\dot{P}_2 = f_2(P_1, P_2) = 0$ is called P_2-nullcline. In general the nullcline for the i coordinate is:

$$\Gamma_i = x \in \mathbb{R}^2_+ : f_i(x) = 0. \tag{4.16}$$

The points of intersection of nullclines are called the equilibria or **steady-state** of the system, where:

$$x^* = (x_1^*, x_2^*) \in \mathbb{R}^2_+ : f_1(x_1^*, x_2^*) = 0 \text{ and } f_2(x_1^*, x_2^*) = 0. \tag{4.17}$$

The steady state is a configuration of the system where both variables in the system remain constant. The region of the phase plane from which trajectories converge to each steady state is called the **basin of attraction** (the set of points $x_0 \in \mathbb{R}^2_+$ such as the solution $x(t, x0)$ converges to x^* as time approaches infinity).

It is possible to verify the *stability* of the equilibrium points by applying a small perturbation to the initial condition $x(0) = x^*$. From an intuitive point of view, if the solution returns at the value x^*, it means that this point is **stable** and if the solution does not

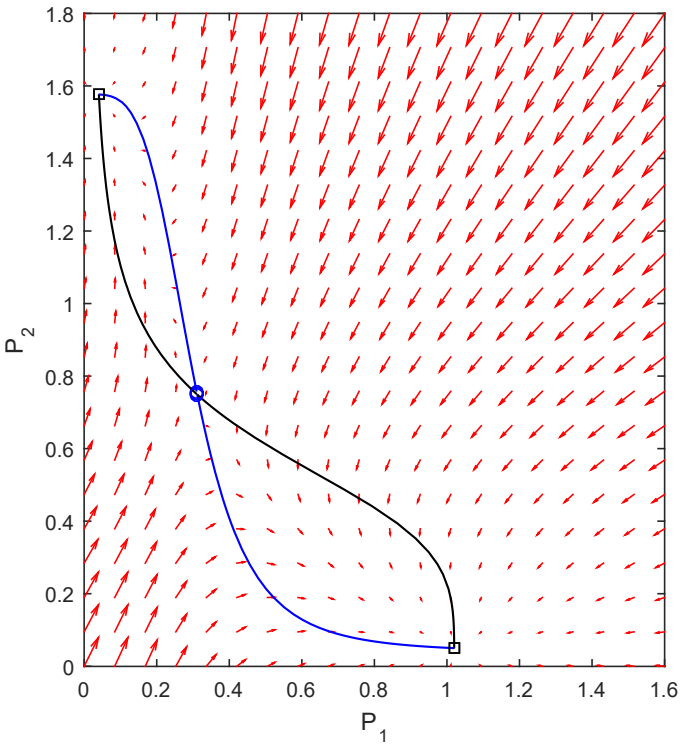


FIGURE 4.5: **Phase portrait**. Phase plane for System (4.14), in the case of bistability. The red arrows compose the direction field of the system. Each region is delimited by the nullclines ($f(P_1, P_2) = 0$ in black and $g(P_1, P_2) = 0$ in blue). The stable steady states are marked by a open rectangle and the unstable steady state is marked by a black circle. The parameter of the systems are [20]: $\kappa_{10} = 0.02$, $\kappa_{11} = 1$, $\theta_1 = 0.3$, $\gamma_1 = 1$, $\kappa_{20} = 0.05$, $\kappa_{21} = 2$, $\theta_2 = 0.6$, $\gamma_1 = 1.3$, $m_1 = m_2 = 4$.

return to the original point it means that this point is **unstable**. This notion is formalized by Lyapunov stability in [57, Ch.4]. There could be more than one equilibrium point in one system: for example, System (4.14) has two distinct stable steady-state points (**bistable system**) and one unstable point: the biological reason is that if one of the proteins is present at high concentration, it inhibits the transcription of the other gene and this implies that the second protein is present at very low concentration (and vice versa).

This example is called **bistable switch** because only external stimulus can force the system to switch from one steady state to the other.

4.2.2 Jacobian matrix

Phase portrait is an important tool to determine the stability of a system, but there is a technique for stability analysis, called **Jacobian matrix**, that does not rely on graphical representation and is not restricted to two species networks. It is the matrix of all first order partial derivatives of a vector-valued function:

$$J(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{pmatrix}.$$

The steady state of the system x^* is locally stable if all the eigenvalues of $J(x^+)$ have a strictly negative part [20]: for the two dimensional case, if the trace of the matrix $J(^*)$ is strictly negative and the determinant strictly positive, the steady state is locally stable. In Example (4.14), the equation of each nullcline $f_i(P_1, P_2)$, implicity defines a function $P_2 = \tilde{f}_i(P_1)$. The Jacobian matrix for this system is:

$$J(x) = \begin{pmatrix} -\frac{\partial f_1}{\partial P_2}\frac{d\tilde{f}_1}{dP_1} & \frac{\partial f_2}{\partial P_2} \\ -\frac{\partial f_2}{\partial P_2}\frac{d\tilde{f}_2}{dP_1} & \frac{\partial f_2}{\partial P_2} \end{pmatrix}.$$

and $det(J) = \frac{\partial f_1}{\partial P_2}\frac{\partial f_2}{\partial P_2}(\frac{d\tilde{f}_2}{dP_1} - \frac{d\tilde{f}_1}{dP_1})$. So the trace and the determinant for J^* are: $tr(J^*) = -(\gamma_1 + \gamma_2)$ and $det(J^*) = \gamma_2\kappa_{11}\frac{m_2\theta_2^{m_2}(P_2^*)^{m_2-1}}{(\theta_2^{m_2}+(x_2^*)^{m_2})^2}\big(\frac{d\tilde{f}_2}{dx_1}(P^*) - \frac{d\tilde{f}_1}{dx_1}(P^*)\big)$.

The trace is always negative and in the steady states near to P_1 and P_2 axis the determinant is positive and the stable states are positive. The middle steady state is unstable and the determinant negative (see Figure 4.5).

4.2.3 Piece-wise affine linear system

If more than two proteins are present in a gene network model, the application of phase plane analysis could be difficult. The **piece-wise linear (PL)** formalism helps to overcome this problem [41]: for example, it leads to a qualitative description of the bistable switch.

Let us take the PL general form:

$$\dot{x}_i = f_i(x) - \gamma_i\, x_i, \quad 1 \le i \le n, \tag{4.18}$$

where the functions $f_i(x)$ describe the activation/inhibition of the expression of gene i by the activation/repression of the protein x_j and the term $\gamma_i x_i$ describes the degradation of the product of gene i. The function f_i: $\mathbb{R}_+^n \to \mathbb{R}_+$ can be detailed as:

$$f_i(x) = \sum_{l \in I} \kappa_{il}\, b_{il}(x), \tag{4.19}$$

where κ_{il} is a rate parameter and b_{il} is a boolean-valued regulation function and I is a index set. The conditions for the synthesis of the protein of gene i, captured by b_{ij}, are

modeled as the sum of product of step functions s^+, s^-. The latter are derived from the Hill functions of ODE models by letting the Hill coefficients tend to infinity ($m \to \infty$). So:

$$\begin{aligned} \lim_{m\to\infty} h^+(x;\theta,m) = s^+(x;\theta) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ if } x < \theta \end{cases} \\ \lim_{n\to\infty} h^-(x;\theta,m) = s^-(x;\theta) = \begin{cases} 1 \text{ if } x < \theta \\ 0 \text{ if } x > \theta \end{cases} \end{aligned} \tag{4.20}$$

Assuming that $\kappa_{10} = \kappa_{20} = 0$ and $\lim_{n\to\infty}$, System (4.14) becomes:

$$\begin{aligned} \dot{P}_1 &= \kappa_{11}\, s^-(P_2,\theta_2) - \gamma_1\, P_1, \\ \dot{P}_2 &= \kappa_{21}\, s^-(P_1,\theta_1) - \gamma_2\, P_2. \end{aligned} \tag{4.21}$$

The system is defined inside the set $\Omega = [0, \frac{\kappa_{11}}{\gamma_1}] \times [0, \frac{\kappa_{21}}{\gamma_2}]$ and is now divided in four boxes or **regular domains** [20]:

$$\begin{aligned} B_{00} &= \left\{P \in \mathbb{R}^2_+ : 0 < P_1 < \theta_1, 0 < P_2 < \theta_2\right\}, \\ B_{01} &= \left\{P \in \mathbb{R}^2_+ : 0 < P_1 < \theta_1, \theta_2 < P_2 < \frac{\kappa_2}{\gamma_2}\right\}, \\ B_{10} &= \left\{P \in \mathbb{R}^2_+ : \theta_1 < P_1 < \frac{\kappa_1}{\gamma_1}, 0 < P_2 < \theta_2\right\}, \\ B_{11} &= \left\{P \in \mathbb{R}^2_+ : \theta_1 < P_1 < \frac{\kappa_1}{\gamma_1}, \theta_2 < P_2 < \frac{\kappa_2}{\gamma_2}\right\}. \end{aligned} \tag{4.22}$$

In each box the system is simple to study (linear decoupled form) and its solution can move in one of the adjacent boxes, crossing the box threshold (or switching domains): the succession of the possible transitions is called **transition graph**.

For example in $B_{0,0}$ the system becomes $\dot{P}_1 = \kappa_1 - \gamma_1 P_1$, $\dot{P}_2 = \kappa_2 - \gamma_2 P_2$: the steady state of the solution is $(P_1^*, P_2^*) = (\kappa_1/\gamma_1, \kappa_2/\gamma_2)$. If $\theta_i < \kappa_1/\gamma_i$, the system crosses the domain and the solution switches to another system.

In Figure 4.6 is shown System (4.21) with its equilibrium points $\phi_1 = (\kappa_1/\gamma_1, 0)$, $\phi_2 = (0, \kappa_2/\gamma_2)$, $\phi_3 = (\theta_1, \theta_2)$. ϕ_1 and ϕ_2 belong to the boundary of their respective domains (B_{10} and B_{01}), so any trajectory entering in one of these boxes remains there. Instead the ones starting in B_{00} and B_{11} will switch to another domain. This behavior is summarized in Figure 4.7.

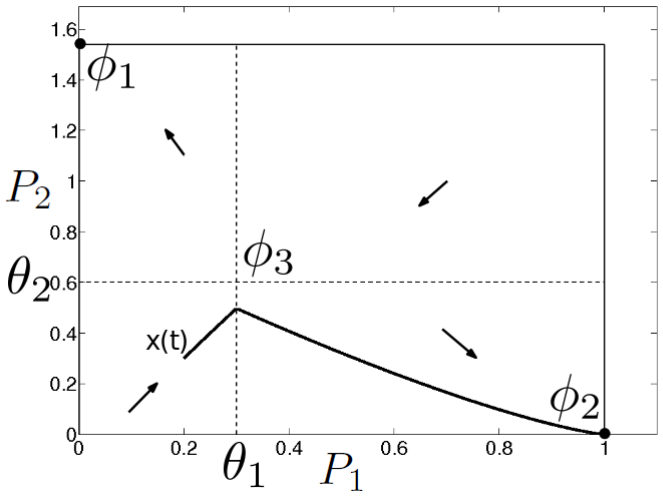


FIGURE 4.6: **Phase plane for piecewise linear System (4.21).** The threshold values θ_1, θ_2 divide the plane into four rectangular regions, where the vector field is constant. There are two stable steady states (ϕ_1, ϕ_2) and an unstable equilibrium point (ϕ_3). One solution $x(t) = [P_1(t), P_2(t)]$ is shown in black. Picture taken from [20] and label modified to match our notations.

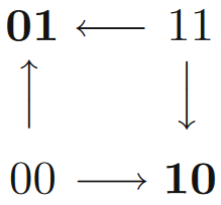


FIGURE 4.7: **Graph of transition of System (4.21).** It is a qualitative description of the behavior of system 4.21. ϕ_1 is represented by 10, ϕ_2 by 01, ϕ_3 is located in the middle, at the boundary of the four regular domains. Picture taken from [20].

4.3 Parameter sensitivity analysis

Another important type of model analysis focuses on how the model behavior depends on parameter values: the study of this dependence is called **parameter sensitivity analysis**.

In general, a sensitivity analysis is the study of how the uncertainty variability in the output of a mathematical model can be apportioned to different sources of uncertainty variability in its input [97, 99]: for our applications the inputs of the system, i.e. the model components whose influence on the output is to be investigated, are the parameter values. This analysis is useful for different purposes [89], for example: to check that the model output behaves as expected when parameters vary, to identify which parameters have a small or a large influence on the output, to simplify the model finding and

removing the parameters that have no effect on the outputs, to detect and quantify parameter interactions.

There are different ways to define sensitivity of a model with respect to its inputs. We can distinguish two types of analysis: local and global sensitivity analyses.

4.3.1 Local sensitivity analysis

A **local sensitivity analysis** is based on the local derivatives of the output $\hat{Y}$ of the model $\hat{Y} = f(Z)$ with respect to the input factors, for instance the parameters Z. It indicates how fast the output increases or decreases locally around a given value of Z. The absolute local sensitivity coefficient $S_i(z_k)$ is defined as the partial derivative of the output variable $\hat{Y}$ with respect to factor Z_i, calculated for $Z = z_k$, where z_k is a set of parameter values:

$$S_i(z_k) = \frac{\partial f(Z)}{\partial Z_i}\Big|_{z_k}. \tag{4.23}$$

As an example, let us consider Equation (4.14) and, for the sake of simplicity, let us assume that P_2 has a fixed concentration $\bar{P}_2$:

$$\dot{P}_1 = \kappa_{10} + \kappa_{11}\frac{\theta_2^{m_2}}{\theta_2^{m_2} + \bar{P}_2^{m_2}} - \gamma_1 P_1. \tag{4.24}$$

The system output of interest ($\hat{Y}$) is the steady state of P_1, P_1^*:

$$P_1^* = \frac{\kappa_{10}}{\gamma_1} + \frac{\kappa_{11}}{\gamma_1}\frac{\theta_2^{m_2}}{\theta_2^{m_2} + \bar{P}_2^{m_2}}. \tag{4.25}$$

The absolute local sensitivity coefficient for the steady state P_1^* in Equation (4.25) with respect to κ_{11}, is:

$$S_{\kappa_{11}} = \frac{\partial P_1^*}{\partial \kappa_{11}} = \frac{\partial}{\partial \kappa_{11}}\left(\frac{\kappa_{10}}{\gamma_1} + \frac{\kappa_{11}}{\gamma_1}\frac{\theta_2^{m_2}}{\theta_2^{m_2} + \bar{P}_2^{m_2}}\right) = \frac{\theta_2^{m_2}}{\gamma_1(\theta_2^{m_2} + \bar{P}_2^{m_2})}. \tag{4.26}$$

Setting the parameter values as follows: $\kappa_{10} = 0.02$, $\kappa_{11} = 1$, $\theta_2 = 0.6$, $m_2 = 4$, $\bar{P}_2 = 0.6$, $\gamma_1 = 1$, we obtain the following sensitivity coefficient: $S_{\kappa_{11}} = 0.5$.

The local sensitivity analysis can also be applied to the system dynamics. Let us consider the state-space equation [121, p.154]:

$$\frac{d}{dt}x = f(x, Z),\ x(0) = x_0(Z). \tag{4.27}$$

It is possible to calculate the time derivative of the dynamical local sensitivity index, S_i:

$$\frac{d}{dt}S_i(t,z_k) = \frac{\partial f(x,Z)}{\partial x}\,S_i + \frac{\partial f(x,Z)}{\partial Z_i},\ S_i(0,z_k) = \frac{\partial x_0(Z)}{\partial Z_i}\Big|_{z_k},\ i=1,\ldots,s \tag{4.28}$$

where $S_i(t,z_k) = \frac{\partial x}{\partial Z_i}$.
Taking as an example Equation (4.24), the time derivative for the dynamical sensitivity index of κ_{11} is:

$$\frac{d}{dt}S_{\kappa_{11}}(t,z_k) = -\gamma_1\,S_{\kappa_{11}} + \frac{\theta_2^{m_2}}{\theta_2^{m_2}+\bar{P}_2^{m_2}}. \tag{4.29}$$

As $S_{\kappa_{11}}(0,z_k)=0 \Longrightarrow S_{\kappa_{11}}(t,z_k) = \frac{\theta_2^{m_2}}{(\theta_2^{m_2}+\bar{P}_2^{m_2})\gamma_1}\,(1-e^{-\gamma_1 t})$.
The advantage of these local sensitivity analyses is that one can obtain analytical expressions of the sensitivity coefficients. However, they are local methods, that focus on one parameter at a time (for more details, see [116]). To study the effect of several parameters on model outputs on a larger parameter space, global sensitivity analysis methods are more relevant.

4.3.2 Global sensitivity analysis

In a **global sensitivity analysis** the output variability is evaluated when the inputs factors vary in a given range. Inputs can vary one at a time (for instance in the *Morris* method [82]), as in local analyses, but most global methods have inputs varying simultaneously to account for and estimate parameter interactions. Although global sensitivity analyses are numerical analyses, these two advantages (global and input interactions) make them particularly relevant to study the model behavior to input uncertainty or variability. There are different methods to perform a global sensitivity analysis (for details of all methods see [98]). The core methodology of parameter sensitivity analysis is the same for most methods [80]. The first step is to identify the inputs of the analysis, i.e. s parameters, and for each parameter Z_i to quantify its uncertainty, an interval value $[Z_{min(i)}, Z_{max(i)}]$ around its nominal value $z_{0,i}$. This defines the parameter space. The second step is to generate N scenarios to explore the parameter space. An input scenario z_k is a combination of input factor levels: $z_k = (z_{k,1},\ldots,z_{k,s})$, $k=1,\ldots,N$. The N scenarios depend on the sensitivity analysis method chosen. The third step is to compute the model output for each scenario $f(z_k)$, $k=1,\ldots,N$. The fourth step is the analysis of the output distribution, which also depends on the method chosen.
In this section we present major variance-based methods.

Factorial design based method

This method evaluates simultaneously the influence of a large number of parameters, that can be quantitative or qualitative. The input space is discretized: each input factor

i can be set to l values or levels (which can depend on i). In a *full factorial design*, having s input factors and l levels per factor, there are l^s distinct input scenarios: the full l^s factorial design consists in running simulations for all these scenarios exactly once. To reduce the number of simulations required, a *fractional factorial design* can be implemented. It is a subset of the full design, chosen according to the model terms one wishes to estimate (main effects, two-way interactions,...). For factorial design based methods, the analysis of variance (ANOVA) is used to estimate the contribution of each factor and each interaction between factors to the output variability, assuming that there is a linear statistical model linking factors and output.

Supposing we have s parameters Z_i and one output Y, we can implement a fractional factorial design in order to estimate all main effects α_i and two-way interactions β_{ij} of the following linear statistical model linking the output Y to the parameters Z_i :

$$Y = \mu + \sum_i \alpha_i + \sum_i \sum_{i \neq j} \beta_{ij} + \epsilon \tag{4.30}$$

where μ is the grand mean and ϵ the residual.

After performing l^{s-f} simulations, where f describes the size of the fraction of the full factorial used, one can estimate thanks to an ANOVA the sum of squares associated with each factorial term for the output Y: the main effect SS_i (α_i), or the two-way interaction $SS_{i,j}$ ($\beta_{i,j}$). According to the sparsity-of-effects principle, a system is usually dominated by main effects and low order interactions, so neglecting third order and higher interactions can still provide good estimates. Denoting by SS_T the total sum of squares, the total sensitivity index of parameter Z_i is defined as follows:

$$tSI_i = \frac{SS_i + \sum_{i \neq j} SS_{i,j}}{SS_T}. \tag{4.31}$$

In Chapter 6 we will use fractional factorial design as a method to evaluate the influence of every parameter on the global relative errors. These errors assess the quality of the model reduction by PPA applied to a well-known model of circadian ryhthms in mammals [73]. Through global sensitivity analysis we will prove the robustness of our method.

Other variance-based methods

The most common variance-based methods are the *Sobol* and *FAST* methods. Compared to the method based on factorial design, they only consider quantitative inputs and they generally require a fairly larger number of simulations.

The *Sobol* method [109] considers input that vary continuously within the uncertainty interval, which can be scaled to [0,1]. The parameter space is explored by *Monte Carlo*

sampling or *latin hypercube sampling* (LHS) [99], with samples that only vary for a limited number of parameters. The *Sobol* method hence requires a large number of simulations, but being a stochastic method, it provides confidence intervals for the sensitivity index estimations.
The *Fourier amplitude sensitivity test* (FAST) method also considers quantitative input parameters that vary within given intervals. Simulation scenarios are selected regularly (periodic sampling) along one or several search trajectories that are designed to explore the input space. Each trajectory scans each parameter range, with frequency that vary among parameters. Sensitivity indices are estimated based on a Fourier decomposition of the variance. The *original FAST* method [30] is a method for estimating essentially the first-order sensitivity index for every parameter, while the *extend FAST* (eFAST) method [100] allows the estimation of the first-order and the total sensitivity indices. The *FAST* method hence requires less simulations than the *Sobol* method, but the index approximations may be biased.

4.4 Parameter fitting

Another important topic in modeling is to find the appropriate parameter values for a biological dynamical model: this task is called **model calibration** or **parameter fitting**. There are parameters that can be measured directly like the degradation rate of a mRNA from observations of its half-lives but others are not: their values are assigned by fitting the model behavior to corresponding **observations** of the system behavior [47].

The goal of parameter fitting is to determine the parameter values for which model simulation best matches the data. The accuracy of the model can be assessed by comparing model predictions to each of the experimental observations, by focusing on **residuals** that are the difference between an observed value and the fitted value provided by the model. One measure of how well the model fits the data is the sum of squared errors, which sums up the squares of the residuals for each data point, thus giving a single measure of the quality of the fit.

The measures can be compared during the whole dynamic of the system: in Figure 4.8, for example, the least square method is calculated through the sum of the square difference between the dynamics of the variable $P_1(t)$ and the experimental observations p_j $(j = 1 \dots 5)$ of the concentration of P_1 at times t_j. The sum of square is:

$$SSE = \sum_{j=1}^{5} (P(t_j) - p_j)^2. \tag{4.32}$$

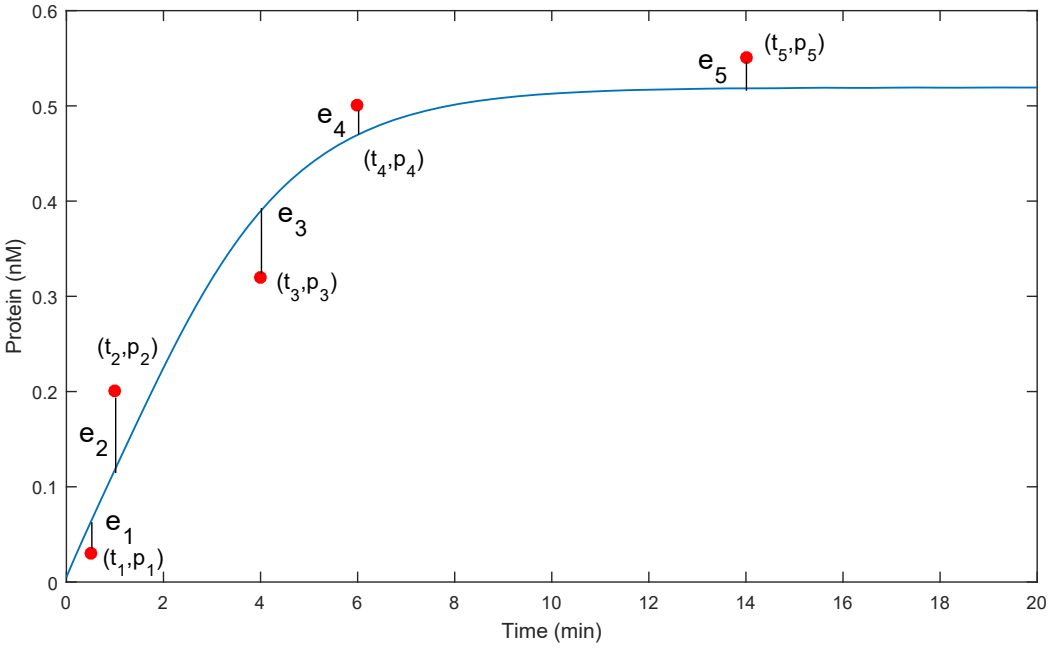


FIGURE 4.8: **Fitting of the curve P_1.** A fitting of the curve based of experimental data of the curve P_1 and the *residuals* e_1, e_2, e_3, e_4, e_5: the sum of square is $SSE = \sum_{j=1}^{5}(P(t_j) - p_j)^2 = (0.06082 - 0.03)^2 + (0.1397 - 0.2)^2 + (0.3877 - 0.32)^2 + (0.4737 - 0.5)^2 + (0.52 - 0.55)^2 = 0.0108$.

The method called **least-square fitting** calculates the fit that corresponds to the parameter values minimizing the sum of squared errors: this parameter set can be found by numerical function-minimization techniques. For example, MATLAB software has a nonlinear programming solver, called *fmincon*: knowing the objective function (for example the SSE), the parameter initial guess and parameter boundaries, it finds the parameter values that minimize the objective function [5].

In Chapter 10 we will use parameter fitting to calibrate a deterministic model that describes the growth arrest of *E. coli* through the transcriptional control of RNAP genes.

Chapter 5

Reduction and stability analysis of a transcription-translation model of RNA polymerase

The aim of the work presented in this chapter is to analyze the dynamical behavior of models of gene transcription and translation, in the case of RNA polymerase synthesis. This is an example of positive feedback loop, where RNA polymerase is needed to transcribe its own gene. We write a full model of high dimension based on mass-action laws. Using monotone system theory and time-scale arguments, we reduce it to a model with two variables (RNA polymerase and its mRNA). We show that it has either a single globally stable trivial equilibrium in (0, 0), or it has an unstable zero equilibrium and a globally stable positive one. We give generalizations of this model, in particular with a variable growth rate. The dynamical behavior can be related to biological observations on the bacterium *Escherichia coli*.

A first draft of this work was written by Ismail Belgacem, former PhD student of the BIOCORE team. My contributions were: performing simulations of the full and reduced models with a new set of parameters to have more realistic results from a biological point of view; comparing a classical model of RNA polymerase to the reduced model that we obtained, including a sensitivity study with respect to the number of ribosomes; designing and studying a reduced system including a variable growth rate.

A journal paper version of this chapter was submitted to *Bulletin of Mathematical Biology*, in which I am second author (see Appendix A).

5.1 Introduction

The central dogma of molecular biology argues that "DNA makes RNA and RNA makes proteins", which are the primary components of cells, see [8]. As we have seen in Chapter 3 gene expression starts with transcription, where the gene is copied into a messenger RNA (mRNA) by the RNA polymerase. The mRNA is then translated into proteins by ribosomes. In prokaryotic cells like bacteria, transcription and translation take place in the same compartment. As a consequence, ribosomes can translate nascent mRNAs being elongated by the RNA polymerase.

Classical models of gene expression often disregard the effect of RNA polymerase and ribosome concentration on the accumulation of RNAs and proteins, which is assumed non limiting. Yet, some works emphasize the important role of the global machinery for gene expression (see [22] for an example). It is therefore interesting to build detailed models involving the main actors of the transcription-translation processes, such as RNA polymerase and ribosomes: some partial detailed models have been developed, see [63] for an example. We develop in this chapter a complete and detailed model of RNA polymerase. From the point of view of Control Theory, it is also a nice example of a positive feedback loop, where RNA polymerase is needed to transcribe its own gene.

Based on mass-actions laws, we first write a detailed mechanistic model of transcription and translation, where every event (binding, release,...) is accounted for. The high dimension of the resulting model makes it too difficult to handle: we reduce it into a much simpler system by time-scale arguments and we study the mathematical properties of the reduced model. To investigate the stability of the fast subsystem and of the reduced system, we use monotone system theory and concavity properties.

Monotone systems form a class of dynamical systems such that the partial order in dimension n between two solutions is conserved (see [107]); see Appendix C.1 for more details about monotone systems. These tools are well adapted to analyze the stability of biological models [112]. They have strong properties of convergence towards equilibria, and cannot (for example) exhibit stable periodic oscillations. The second tool is related to the concavity of functions used in differential equations [105]. In our opinion, these tools are particularly appealing because they are qualitative (they do not depend too much on the values of parameters), and they give very strong results about the global dynamical behavior of the system [46]. These tools have been applied to biological systems already: population dynamics [106], chemical networks [71]... J.-L. Gouzé and I. Belgacem have worked with monotone systems theory on metabolic-genetic networks [15] and on detailed models for gene expression, without any loop [14, 16]. Yet, to our knowledge, the theorem on concave and monotone systems has not been used in the context of detailed gene expression models, where functions can be given by rather complex algebraic expressions resulting from mass balance.

Using monotone system theory, we are able to prove the global stability of the fast subsystem with respect to the equilibrium; the global result is difficult to prove by other techniques, and not proven or proven only locally in most similar works [44, Chapter 4]. We present the original model and its reduction in the next section. Using parameters built from the literature, we analyze the different time scales in which evolve the variables of the full system. We decompose the full system into fast and slow subsystems and verify in Section 5.4 that the fast subsystem satisfies the conditions for applicability of the Tikhonov's theorem. This allows us to put the fast subsystem at its quasi-steady state and obtain a reduced model with a similar dynamical behavior. In Section 5.6, we verify that the concavity and monotonicity assumptions hold for the reduced model. We show that the trivial equilibrium is either globally stable (in that case no other equilibrium exists) or locally unstable, and that it implies the existence and uniqueness of a positive equilibrium, which is globally stable with respect to the positive orthant. We provide the biological condition for this alternative. We then investigate a generalization of the model with a variable growth rate, compare it with simpler models, and finally give conclusions from a modeling and a biological point of view.

5.2 The coupled transcription-translation model of RNA polymerase

5.2.1 Description of the model

Figure 5.1 shows the transcription-translation model for the synthesis of RNA polymerase in a single cell; for simplification, we consider it to be encoded from a single gene. This model is inspired from those given in [63]. Transcription is initiated by the specific binding of RNA polymerase to the promoter region D onto the DNA, a process promoted by an initiating factor called σ. The RNA polymerase clears the promoter (with a constant rate k_c) and moves along the DNA (with a constant rate k_t). Complexes Y and Y^i describe the elongating RNA polymerase, which adds nucleotides one by one. Addition of the last nucleotide completes the full length mRNA, which is released from the RNA polymerase. The completed RNA molecule is either subject to degradation (with a constant rate k_m) or it is used by ribosomes as a template for the synthesis of a new RNA polymerase[1]. Translation starts with the ribosome R forming a complex $RRNA'$ with the free ribosome binding site RNA' on the newly synthesized mRNA. After clearance of the ribosome binding site (with a constant rate k_w), the elongating

[1]The process of translation can be initiated from every nascent mRNA as shown in [63]. For simplicity, we suppose that proteins are synthesized from completed mRNAs only. This is consistent with recent observations on the lack of coupling between transcription and translation in *E. coli* cells [11].

$$
\begin{aligned}
\sigma + P + D &\underset{k_+}{\overset{k_-}{\leftrightarrows}} PD \\
PD &\xrightarrow{k_c} Y + D + \sigma \\
Y + Nu &\xrightarrow{k_t} Y^1 \\
Y^1 + Nu &\xrightarrow{k_t} Y^2 \\
&\vdots \\
Y^{L-1} + Nu &\xrightarrow{k_t} P + RNA \\
RNA &\xrightarrow{k_m} \phi \\
R + RNA' &\underset{k'_+}{\overset{k'_-}{\leftrightarrows}} RRNA' \\
RRNA' &\xrightarrow{k_u} X + RNA' \\
X + tRNA^* &\xrightarrow{k'_t} tRNA + X^1 \\
X^1 + tRNA^* &\xrightarrow{k'_t} tRNA + X^2 \\
&\vdots \\
X^{H-1} + tRNA^* &\xrightarrow{k'_t} tRNA + R + P \\
P &\xrightarrow{k_p} \phi
\end{aligned}
$$

FIGURE 5.1: Reaction scheme of the transcription-translation model.

form of the ribosome X starts synthesizing the protein: amino acids carried by tRNAs ($tRNA^*$) are transferred one by one to the nascent protein (with a constant rate k'_t), giving the complexes $X^1, X^2, ..., X^i$. After addition of the last amino acid, the protein is completed and released by the ribosome. The newly synthesized RNA polymerase is able to start transcribing its own gene and other cellular genes.

The protein and its mRNA are also subject to degradation (with a constant rate k_p and k_m, respectively), and dilution by growth due to the augmentation of cell volume (at a rate μ). In a first step, we consider a constant growth rate μ and, for simplicity, the sum of degradation and dilution will be expressed by only one parameter: k'_p for the protein and k'_m for the mRNA. In a second step (Section 5.9) we will consider the degradation rate and the dilution rate separately, where the latter varies in function of the RNA polymerase concentration.

5.2.2 Full equation

To write the full system, some assumptions have to be done:

- The nucleotides, the amino acids and the sigma factor are non limiting and their concentrations are included in the parameters;
- The free and bound forms of RNA polymerase and mRNA are considered to be degraded at the same rate;
- The degradation of the bound forms (PD and $RRNA'$) releases the promoter and the ribosome;
- The free mRNA corresponds to the mRNA with a free ribosome binding site RNA'.

Using classical mass action kinetics laws we obtain the system:

$$\begin{aligned}
\dot{c} &= k_+ p\, d - k_- c - k_c c - k_p^{'} c \\
\dot{d} &= -k_+ p\, d + k_- c + k_c c + k_p^{'} c \\
\dot{p} &= -k_+ p\, d + k_t\, y^{L-1} + k_- c + k_t^{'} x^{H-1} - k_p^{'} p \\
\dot{y} &= k_c c - k_t\, y - k_p^{'} y \\
\dot{y}^1 &= k_t\, y - k_t\, y^1 - k_p^{'} y^1 \\
\dot{y}^2 &= k_t\, y^1 - k_t\, y^2 - k_p^{'} y^2 \\
&\vdots \\
\dot{y}^{L-1} &= k_t\, y^{L-2} - k_t\, y^{L-1} - k_p^{'} y^{L-1} \\
\dot{w} &= k_+^{'} r\, m - k_-^{'} w - k_w w - k_m^{'} w \\
\dot{m} &= -k_+^{'} r\, m + k_-^{'} w + k_w w + k_t\, y^{L-1} - k_m^{'} m \\
\dot{r} &= -k_+^{'} r\, m + k_-^{'} w + k_t^{'} x^{H-1} + k_m^{'} w \\
\dot{x} &= k_w w - k_t^{'} x \\
\dot{x}^1 &= k_t^{'} x - k_t^{'} x^1 \\
&\vdots \\
\dot{x}^{H-1} &= k_t^{'} x^{H-2} - k_t^{'} x^{H-1}
\end{aligned} \tag{5.1}$$

where p, d, c, y, y^i and m are the concentrations of P, D, PD, Y, Y^i and mRNA respectively, and where w, r, x and x^i are the concentrations of $RRNA'$, R, X and X^i respectively. L and H are the lengths of the mRNA and the protein, respectively[2]. The promoter D remains intact during the degradation of the complex between the RNA polymerase and the promoter, which means that the following mass conservation relation holds for the total concentration of promoter:

$$d = d_0 - c$$

[2] H is equal to $L/3$, because the translation of three combined nucleotides in the mRNA gives one amino acid.

where d_0 is the initial concentration of promoter. We can reduce System (5.1) by replacing the three first equations with the following ones:

$$\dot{c} = k_+ p\,(d_0 - c) - k_-\, c - k_c\, c - k_p^{'}\, c$$
$$\dot{p} = -k_+ p\,(d_0 - c) + k_t\, y^{L-1} + k_-\, c + k_t^{'}\, x^{H-1} - k_p^{'}\, p$$

The total number of ribosomes (R_0) is also conserved:

$$\frac{d}{dt}(r + w + x + x^1 + \cdots + x^{H-1}) = 0$$
$$r + w + x + x^1 + \cdots + x^{H-1} = R_0$$

System (5.1) has a high dimension due to the large size (L) of the mRNA (up to several thousands nucleotides) and of the protein (H), see Table 5.1 for their values. The system dimension is $L + H + 5$[3]. In addition, the system is non linear and non monotone (for example $\frac{\partial \dot{m}}{\partial r} = -k_+^{'}\, m$ is negative), which makes it difficult to study its properties: in this case we need a simplification of the full system.

5.3 Time-scale reduction (fast-slow behavior)

5.3.1 Parameter values for the coupled transcription-translation models of RNA polymerase

The values of parameters in Tables 5.1[4], 5.2 and 5.3 have been carefully built from the literature based on classical papers such as [29]. In the next section, we will show that System (5.1) has two different time scales - fast and slow - and therefore it can be approximated by a reduced system by applying Tikhonov's theorem (see Appendix C.2 and [57] for the statement of the theorem).

[3] In the ODE System (5.1) there are L differential equations for $\dot{y}, \dot{y}^1 ... \dot{y}^{L-1}$ (L=8253 nucleotides), H for $\dot{x}, \dot{x}^1 ... \dot{x}^{H-1}$ (L=2751 amino acids) and 5 for $\dot{c}$, $\dot{p}$, $\dot{w}$, $\dot{m}$, $\dot{r}$, which gives a total of 11009 differential equations.

[4] In Table 5.1, the ratios $\frac{L}{k_t}$, and $\frac{H}{k_t^{'}}$ are constant (see [63]), so if we take another gene length L and H, the new values of k_t and, $k_t^{'}$ with respect to L and H respectively will be : $k_t = (2340/8253)L$, and $k_t^{'} = (1258/2751)H$. This rescaling is useful to reduce the number of equations under consideration.

TABLE 5.1: Values of the set of parameters (fast scale)

Parameter	Value	Unit
k_+	80	$\mu M^{-1} \cdot min^{-1}$
k_-	40	min^{-1}
k_c	1.5	min^{-1}
k_t	2340	$nucleotide \cdot min^{-1}$
L	8253	$nucleotide$
k'_+	11	$\mu M^{-1} \cdot min^{-1}$
k'_-	100	min^{-1}
k_w	55	min^{-1}
k'_t	1258	$aminoacid \cdot min^{-1}$
H	2751	$aminoacid$

TABLE 5.2: Values of the set of parameters (slow scale)

Parameter	Value	Unit
k_p	0.00048	min^{-1}
k_m	0.17	min^{-1}
μ	0.012	$\mu M \cdot min^{-1}$
k'_p	0.01248	min^{-1}
k'_m	0.182	min^{-1}

TABLE 5.3: Initial conditions

Parameter	Value	Unit
d_0	0.000347	μM
R_0	35	μM
z_0	0.5	μM

5.3.2 Separation of the full system into "fast" and "slow" variables

We choose the following slow variables:

$$z = c + p + y + y^1 + \ldots + y^{L-1} \tag{5.2}$$

$$q = m + w \tag{5.3}$$

They represent the total concentration (free and bound forms) of the RNA polymerase z and the total concentration of its mRNA q. The fast variables appear more easily if some variables are rescaled. So, to make the time scales more obvious and verify that the evolution of z and q is slow, we scale the variables $y, y^1, ..., y^{L-1}$ with respect to a scaling factor α, and the variables $x, x^1, ..., x^{H-1}$ with respect to a scaling factor β.

Consider $\overline{y} = \alpha y$, $\overline{y}^i = \alpha y^i$ and $\overline{x} = \beta x, \overline{x^i} = \beta x^i$, this gives:

$$
\begin{aligned}
\dot{c} &= k_+\, p\,(d_0 - c) - k_-\, c - k_c\, c - k_p^{'}\, c \\
\dot{p} &= -k_+\, p\,(d_0 - c) + \frac{k_t}{\alpha}\,\overline{y}^{L-1} + k_-\, c + \frac{k_t^{'}}{\beta}\,\overline{x}^{H-1} - k_p^{'}\, p \\
\dot{\overline{y}} &= \alpha k_c\, c - k_t\,\overline{y} - k_p^{'}\,\overline{y} \\
\dot{\overline{y}}^1 &= k_t\,\overline{y} - k_t\,\overline{y}^1 - k_p^{'}\,\overline{y}^1 \\
\dot{\overline{y}}^2 &= k_t\,\overline{y}^1 - k_t\,\overline{y}^2 - k_p^{'}\,\overline{y}^2 \\
&\vdots \\
\dot{\overline{y}}^{L-1} &= k_t\,\overline{y}^{L-2} - k_t\,\overline{y}^{L-1} - k_p^{'}\,\overline{y}^{L-1} \\
\dot{w} &= k_+^{'}\, r\, m - k_-^{'}\, w - k_w\, w - k_m^{'}\, w \\
\dot{m} &= -k_+^{'}\, r\, m + k_-^{'}\, w + k_w\, w + \frac{k_t}{\alpha}\,\overline{y}^{L-1} - k_m^{'}\, m \\
\dot{r} &= -k_+^{'}\, r\, m + k_-^{'}\, w + \frac{k_t^{'}}{\beta}\,\overline{x}^{H-1} + k_m^{'}\, w \\
\dot{\overline{x}} &= \beta k_w\, w - k_t^{'}\,\overline{x} \\
\dot{\overline{x}}^1 &= k_t^{'}\,\overline{x} - k_t^{'}\,\overline{x}^1 \\
&\vdots \\
\dot{\overline{x}}^{H-1} &= k_t^{'}\,\overline{x}^{H-2} - k_t^{'}\,\overline{x}^{H-1}
\end{aligned}
\tag{5.4}
$$

where $\frac{k_t}{\alpha}$, $\frac{k_t^{'}}{\beta}$ are small compared to k_t and $k_t^{'}$ but where the first one is bigger than the second one ($\frac{k_t}{\alpha} = 200\text{min}^{-1}$ and $\frac{k_t^{'}}{\beta} = 10\text{min}^{-1}$)[5].

Finally the slow evolution part is given by the equation: $z = c + p + \frac{1}{\alpha}(\overline{y} + \overline{y}^1 + \ldots + \overline{y}^{L-1})$, then $\dot{z} = \frac{k_t^{'}}{\beta}\,\overline{x}^{H-1} - k_p^{'}\, z$. Similarly, $q = m + w$ therefore $\dot{q} = \frac{k_t}{\alpha}\,\overline{y}^{L-1} - k_m^{\prime}\, q$. Having introduced the two new variables z and q we return, for simplicity, to the original system scale (see System (5.5)).

[5] To obtain these values we choose $\alpha = 11.7$ and $\beta = 125.8$.

Finally, the full system writes (the slow terms are marked in bold):

$$
\begin{aligned}
\dot{\mathbf{z}} &= \mathbf{k_t^{'}}\,\mathbf{x^{H-1}} - \mathbf{k_p^{'}}\,\mathbf{z} \\
\dot{\mathbf{q}} &= \mathbf{k_t}\,\mathbf{y^{L-1}} - \mathbf{k_m^{'}}\,\mathbf{q} \\
\dot{c} &= k_+\,p\,(d_0 - c) - k_-\,c - k_c\,c - \mathbf{k_p^{'}}\,\mathbf{c} \\
\dot{p} &= -k_+\,p\,(d_0 - c) + k_t\,y^{L-1} + k_-\,c + \mathbf{k_t^{'}}\,\mathbf{x^{H-1}} - \mathbf{k_p^{'}}\,\mathbf{p} \\
\dot{y} &= k_c\,c - k_t\,y - \mathbf{k_p^{'}}\,\mathbf{y} \\
\dot{y}^1 &= k_t\,y - k_t\,y^1 - \mathbf{k_p^{'}}\,\mathbf{y^1} \\
\dot{y}^2 &= k_t\,y^1 - k_t\,y^2 - \mathbf{k_p^{'}}\,\mathbf{y^2} \\
&\vdots \\
\dot{y}^{L-1} &= k_t\,y^{L-2} - k_t\,y^{L-1} - \mathbf{k_p^{'}}\,\mathbf{y^{L-1}} \\
\dot{w} &= k_+^{'}\,r\,(q - w) - k_-^{'}\,w - k_w\,w - \mathbf{k_m^{'}}\,\mathbf{w} \\
\dot{r} &= -k_+^{'}\,r\,(q - w) + k_-^{'}\,w + k_t^{'}\,x^{H-1} + \mathbf{k_m^{'}}\,\mathbf{w} \\
\dot{x} &= k_w\,w - k_t^{'}\,x \\
\dot{x}^1 &= k_t^{'}\,x - k_t^{'}\,x^1 \\
&\vdots \\
\dot{x}^{H-1} &= k_t^{'}\,x^{H-2} - k_t^{'}\,x^{H-1}
\end{aligned}
\tag{5.5}
$$

We obtain the fast subsystem by neglecting the bold terms. We group the system in 2 sub-systems. The first one describes the dynamics of the RNA polymerase z: $\dot{z} = k_t^{'}\,x^{H-1} - k_p^{'}\,z$. The other fast variables $c, p, y, y^1, \ldots, y^{L-1}$ have the following dynamics:

$$
\begin{aligned}
\dot{c} &= k_+\,p\,(d_0 - c) - k_-\,c - k_c\,c \\
\dot{p} &= -k_+\,p\,(d_0 - c) + k_t\,y^{L-1} + k_-\,c \\
\dot{y} &= k_c\,c - k_t\,y \\
\dot{y}^1 &= k_t\,y - k_t\,y^1 \\
\dot{y}^2 &= k_t\,y^1 - k_t\,y^2 \\
&\vdots \\
\dot{y}^{L-1} &= k_t\,y^{L-2} - k_t\,y^{L-1}
\end{aligned}
\tag{5.6}
$$

and should leave the hyperplane $z = c + p + y + y^1 + \ldots + y^{L-1}$ invariant.

The second subsystem describes the dynamics of the mRNA q. With $\dot{q} = k_t\, y^{L-1} - k_m^{'}\, q$, the other fast variables $w, r, x, x^1, \ldots, x^{H-1}$ have the following dynamics:

$$\begin{aligned}
\dot{w} &= k_+^{'}\, r\, (q - w) - k_-^{'}\, w - k_w\, w \\
\dot{r} &= -k_+^{'}\, r\, (q - w) + k_-^{'}\, w + k_t^{'}\, x^{H-1} \\
\dot{x} &= k_w\, w - k_t^{'}\, x \\
\dot{x}^1 &= k_t^{'}\, x - k_t^{'}\, x^1 \\
&\vdots \\
\dot{x}^{H-1} &= k_t^{'}\, x^{H-2} - k_t^{'}\, x^{H-1}
\end{aligned} \tag{5.7}$$

and should leave the hyperplane $R_0 = w + r + x + x^1 + \ldots + x^{H-1}$ invariant.

In Section 5.4, we will show that Subsystem (5.6) - with the variables c, p, y, y^1,..., y^{L-1} - and Subsystem (5.7) - with variables w, r, x, x^1, x^{L-1} - converge to a quasi steady state and satisfy the conditions to apply the Tikhonov's theorem and reduce the full system. In particular we will establish that each of the two subsystems has a unique, globally stable equilibrium.

5.4 Verification of the applicability of the Tikhonov's theorem for the fast subsystems

To check the assumptions of Tikhonov's theorem, we need first to study the existence and uniqueness of the steady state and the global stability of Subsystems (5.6) and (5.7), which represent the fast part of the full system.
Considering the fast Subsystem (5.6), we see that $\dot{c} + \dot{p} + \dot{y} + \dot{y}^1 + ... + \dot{y}^{L-1} = 0$: the system is closed. Powerful theorems apply to this type of monotone system, as can be easily checked: the Jacobian matrix $J(c, p, y, y^1, \ldots, y^{L-1})$

$$= \begin{pmatrix}
-(k_- + k_c - k_+\, p) & k_+\, (d_0 - c) & 0 & \ldots & 0 \\
k_- + k_+\, p & -k_+\, (d_0 - c) & 0 & \ldots & k_t \\
k_c & 0 & -k_t & \ldots & 0 \\
0 & 0 & k_t & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & -k_t
\end{pmatrix} \tag{5.8}$$

is a compartmental matrix which means that

$$\begin{cases} J_{ii}(c,p,y,y^1,\ldots,y^{L-1}) \leq 0 & \text{for all } i, \\ J_{ij}(c,p,y,y^1,\ldots,y^{L-1}) \geq 0 & \text{for all } i \neq j, \\ -J_{jj}(c,p,y,y^1,\ldots,y^{L-1}) \geq \sum_{i\neq j} J_{ij}(c,p,y,y^1,\ldots,y^{L-1}) & \text{for all } j \end{cases}$$

We can also easily check that the graph of the Jacobian matrix is strongly connected. The interaction graph associated with the Jacobian matrix (5.8) is shown in Figure 5.2.

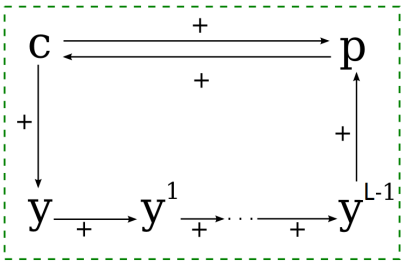


FIGURE 5.2: Interaction graph of System (5.6).

Therefore, we can apply Property 5 [13] to obtain this theorem.

Theorem 5.1. *Let* $z(c,p,y,y^1,\ldots,y^{L-1}) = c+p+y+y^1+\ldots+y^{L-1}$ *be the (fixed) total concentration of the closed system.*
Then for any $z > 0$, *hyperplane* $H_z = \{(c,p,y,y^1,\ldots,y^{L-1}) \in \mathbb{R}_+^{L+2} : z(c,p,y,y^1,\ldots, y^{L-1}) = z > 0\}$ *is forward invariant and contains a unique equilibrium, globally stable in* H_z.

The proof of this theorem can be obtained using a Lyapunov function (for more details of this proof see [52]).
The second Subsystem (5.7) is quite similar to Subsystem (5.6). The study of the equilibrium and the stability is exactly the same as before and will not be given here for the sake of brevity. We define the hyperplane
$G = \{(w,r,x,x^1,\ldots x^{H-1}) \in \mathbb{R}^{H+2} : w+r+x+x^1+\ldots+x^{H-1} = R_0\}$. The final result is that System (5.7) has a unique equilibrium on this invariant hyperplane, which is globally asymptotically stable.

5.5 Application of the Tikhonov's theorem

Since the assumptions of the Tikhonov's theorem are verified, we apply it and set the differential equations of the fast part to zero, which gives the following algebraic equations:

$$\begin{aligned} y^{L-1} = \ldots = y^1 = y = \frac{k_c}{k_t}\, c = \frac{k_c d_0}{k_t}\, \frac{p}{p+k_1}, \\ x^{H-1} = \ldots = x^1 = x = \frac{k_w}{k_t^{'}}\, w = \frac{k_w}{k_t^{'}}\, \frac{q\, r}{r+k_2}, \\ k_1 = \frac{k_- + k_c}{k_+}, k_2 = \frac{k_-^{'} + k_w}{k_+^{'}}. \end{aligned} \tag{5.9}$$

Moreover, using the two conservation relations on the hyperplanes, and the above algebraic equations, we obtain:

$$\begin{aligned} p + \frac{d_0\, p}{p+k_1} + l\, \frac{k_c}{k_t}\, \frac{d_0\, p}{p+k_1} = z, \\ r + \frac{q\, r}{r+k_2} + h \frac{k_w}{k_t^{'}} \frac{q\, r}{r+k_2} = R_0. \end{aligned} \tag{5.10}$$

The slow subsystem is always given by:

$$\begin{aligned} \dot{z} = k_t^{'}\, x^{H-1} - k_p^{'}\, z, \\ \dot{q} = k_t\, y^{L-1} - k_m^{'}\, q, \end{aligned} \tag{5.11}$$

which results in the reduced system:

$$\begin{aligned} \dot{z} = k_w \frac{q\, r(q)}{r(q)+k_2} - k_p^{'}\, z, \\ \dot{q} = k_c d_0\, \frac{p(z)}{p(z)+k_1} - k_m^{'}\, q. \end{aligned} \tag{5.12}$$

where $p(z)$ and $r(q)$ are calculated from the following algebraic equations:

$$\begin{aligned} z = \gamma \frac{p(z)}{p(z)+k_1} + p(z), \\ R_0 = \lambda \frac{q\, r(q)}{r(q)+k_2} + r(q). \end{aligned} \tag{5.13}$$

with $\gamma = (L\, \frac{k_c}{k_t} + 1)\, d_0$ and $\lambda = H \frac{k_w}{k_t^{'}} + 1$. These algebraic equations have only one positive solution $p(z)$ and $r(q)$, because for fixed z and q, the right members of System (5.13) are increasing functions of p and r.

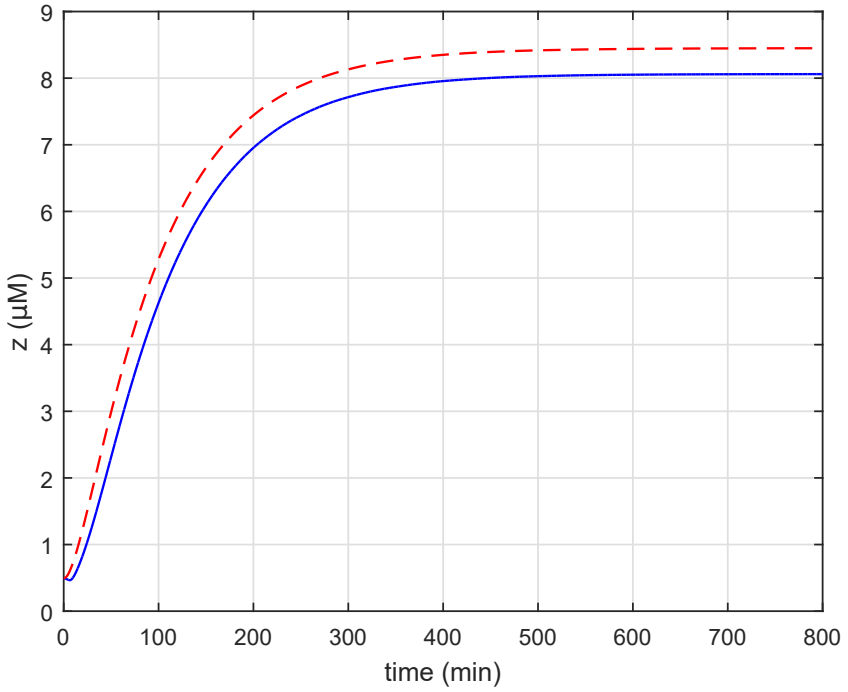


FIGURE 5.3: Time behavior of the total concentration of RNA polymerase z in the reduced system (dashed red line) and in the complete system (solid blue line), with initial conditions: $p_0 = z_0 = 0.5\ \mu M$, taking $L = l = 100$ nucleotides, $H = h = 33$ amino acids. At the equilibrium: $z^*_{reduced} = 8.449\ \mu M$, $z^*_{full} = 8.061\ \mu M$.

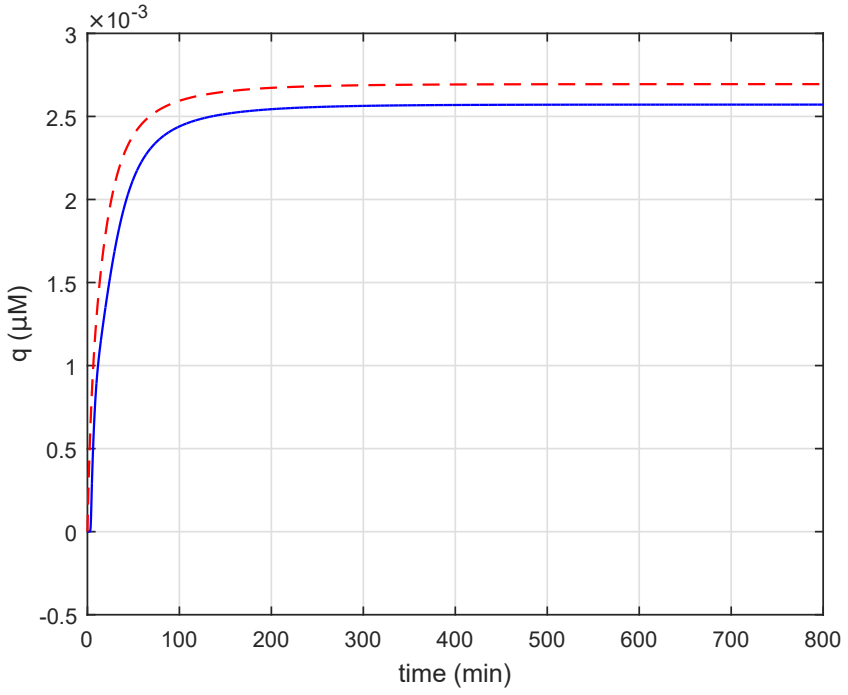


FIGURE 5.4: Time behavior of the concentration of mRNA q in the reduced system (dashed red line) and in the complete system (solid blue line), with $q_0 = 0\ \mu M$, $l = 100$ nucleotides, $h = 33$ amino acids. At the equilibrium: $q^*_{reduced} = 0.002694\ \mu M$, $q^*_{full} = 0.002571\ \mu M$.

5.6 Dynamical study of the reduced system

5.6.1 Simulations of the full and the reduced system

We performed simulations to study the similarity of the reduced system and the full system, as shown in Figure 5.3 and Figure 5.4 for the concentration of RNA polymerase and its mRNA, respectively. The dashed line represents the evolution of the reduced system, and the full line shows the evolution of the complete system. The behavior of the two systems are rather similar, the full system slightly oscillates a at the beginning, while the reduced system does not.

5.6.2 Equilibria of the reduced system

We get the following equations for the equilibrium:

$$\begin{aligned} z &= \frac{k_w}{k_p}\frac{q\, r(q)}{r(q)+k_2} = \frac{k_w}{k_p'\lambda}(R_0 - r(q)), \\ q &= \frac{k_c d_0}{k_m}\frac{p(z)}{p(z)+k_1} = \frac{k_c d_0}{k_m'\gamma}\,(z - p(z)). \end{aligned} \tag{5.14}$$

We therefore obtain z as a function of q, and q as a function of z. Consider the functions:

$$\begin{aligned} \xi(q) &= \frac{k_w}{k_p'}\frac{q\, r(q)}{r(q)+k_2}, \\ \varphi(z) &= \frac{k_c d_0}{k_m'}\frac{p(z)}{p(z)+k_1}. \end{aligned} \tag{5.15}$$

First we can notice that $\varphi(z)$ is bounded ($\varphi(z) < \frac{k_c d_0}{k_m'}$), because $\frac{p(z)}{p(z)+k_1} < 1$. Similarly, $r(q) = R_0 - \frac{k_p'\lambda}{k_w}\xi(q)$, should remain positive, which leads to $\xi(q) < \frac{k_w R_0}{k_p'\lambda}$. To determine the equilibria we have to study the intersections of the two above functions $\xi(q)$ and $\varphi(z)$. From the algebraic Equations (5.13), if we differentiate the first equation with respect to z, and the second with respect to q, we obtain:

$$\begin{aligned} \frac{\partial p(z)}{\partial z} &= \frac{(p(z)+k_1)^2}{\gamma k_1 + (p(z)+k_1)^2}, \\ \frac{\partial r(q)}{\partial q} &= -\frac{\lambda \frac{r(q)}{r(q)+k_2}}{\lambda q \frac{k_2}{(r(q)+k_2)^2} + 1}. \end{aligned} \tag{5.16}$$

$p(z)$ is positive and increasing, with $p(0) = 0$ and $p(z) \approx z$ when z is large. Similarly, $r(q)$ is positive and decreasing, with $r(0) = R_0$ and $r(q)$ tends toward 0 for large q.

From the second derivative of the algebraic Equations (5.13), the first equation with respect to z, and the second with respect to q, we obtain:

$$\begin{aligned}\frac{\partial^2 p(z)}{\partial^2 z} &= \gamma \frac{2k_1}{(p(z)+k_1)^3}\frac{\partial p(z)}{\partial z}^2, \\ \frac{\partial^2 r(q)}{\partial^2 q} &= \frac{-2\lambda \frac{k_2}{(r(q)+k_2)^2}\frac{\partial r(q)}{\partial q} + \lambda q \frac{2k_2}{(r(q)+k_2)^3}\frac{\partial r(q)}{\partial q}^2}{1+\lambda q \frac{k_2}{(r(q)+k_2)^2}}.\end{aligned} \tag{5.17}$$

Thus, $\frac{\partial^2 p(z)}{\partial^2 z}$ and $\frac{\partial^2 r(q)}{\partial^2 q}$ are positive, and we conclude that $p(z)$ and $r(q)$ are also convex. The derivative of $\xi(q)$ is $\xi'(q) = -\frac{k_w}{k'_p\lambda}\frac{\partial r(q)}{\partial q}$, which is positive because $\frac{\partial r(q)}{\partial q}$ is negative. We have also $\xi''(q) = -\frac{k_w}{k'_p\lambda}\frac{\partial^2 r(q)}{\partial q^2}$, which is negative, because $\frac{\partial^2 r(q)}{\partial q^2}$ is positive. Similarly, for $\varphi(z)$, we have $\varphi'(z) = \frac{k_c d_0}{k'_m}\frac{k_1}{(p(z)+k_1)^2}\frac{\partial p(z)}{\partial z} = \frac{k_c d_0}{k_m \gamma}(1 - \frac{\partial p(z)}{\partial z})$, which is positive, because $\frac{\partial p(z)}{\partial z}$ is positive. We also have $\varphi''(z) = -\frac{k_c d_0}{k'_m \gamma}\frac{\partial^2 p(z)}{\partial z^2}$, which is negative, because $\frac{\partial p(z)}{\partial^2 z}$ is positive.

The functions $\xi(q)$ and $\varphi(z)$ are therefore increasing, positive and concave, and are bounded ($\xi(q) < \frac{k_w R_0}{k_p \lambda}, \varphi(z) < \frac{k_c d_0}{k'_m}$).

In the phase space (q, z), two cases are possible, see Figure 5.5 and Figure 5.6: either (0,0) is the unique equilibrium, or there exists another unique, positive equilibrium (the point (0,0) is always an equilibrium for this system). The alternative between these two cases depends on the slope at the origin (0,0).

- If : $\xi'(q)|_{q=0} > \frac{1}{\varphi'(z)|_{z=0}} \Rightarrow k'_m k'_p < k_w \frac{k_c d_0}{\gamma + k_1}\frac{R_0}{R_0 + k_2}$, then there exist two equilibria which are $(0,0)$, and a unique, positive (z^*, q^*).
- If: $\xi'(q)|_{q=0} < \frac{1}{\varphi'(z)|_{z=0}} \Rightarrow k'_m k'_p > k_w \frac{k_c d_0}{\gamma + k_1}\frac{R_0}{R_0 + k_2}$, then there exists only one equilibrium for the system which is $(0,0)$.

5.6.3 Stability of equilibria

The stability study of the reduced System (5.12) gives the following results: it has either a single stable equilibrium in $(0,0)$ or two equilibria, one in zero (unstable) and another stable one (z^*, q^*).

Proposition 5.2. *If* $k'_m k'_p - k_w \frac{k_c d_0}{\gamma + k_1}\frac{R_0}{R_0 + k_2} > 0$ *then*

- $(0,0)$ *is the unique equilibrium of the system and it is globally stable in the non-negative orthant.*

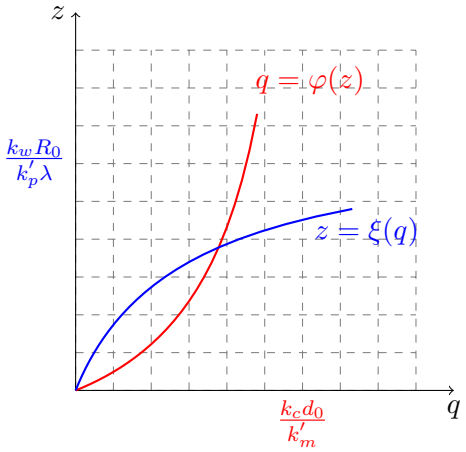


FIGURE 5.5: The phase space (q, z). Two equilibria exist: $(0, 0)$ and another one which is strictly positive (z^*, q^*) under the condition $\xi'(q)|_{q=0} > \frac{1}{\varphi'(z)|_{z=0}}$.

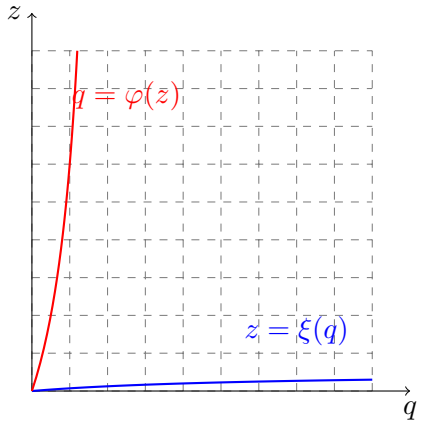


FIGURE 5.6: The phase space (q, z). One equilibrium exists, which is $(0, 0)$ if $\xi'(q)|_{q=0} < \frac{1}{\varphi'(z)|_{z=0}}$.

Proposition 5.3. *If* $k'_m k'_p - k_w \frac{k_c d_0}{\gamma + k_1} \frac{R_0}{R_0 + k_2} < 0$ *then*

- $(0, 0)$ *is unstable.*
- *The positive equilibrium* (z^*, q^*) *is globally stable in the positive orthant.*

For the proof of these propositions, we use the fact that System (5.12) is still monotone [107]. See [18] for more details. The biological interpretations of Proposition 1 and Proposition 2 are presented in the conclusion.

Figure 5.7 shows the simulation of the reduced system from two different initial conditions, $(q_{01}, z_{01}) = (0, 0.5)$ and $(q_{02}, z_{02}) = (0.1, 0.8)$. We observe that the two trajectories converge to the positive equilibrium.

The power of this approach is that it allows results from qualitative hypotheses, even

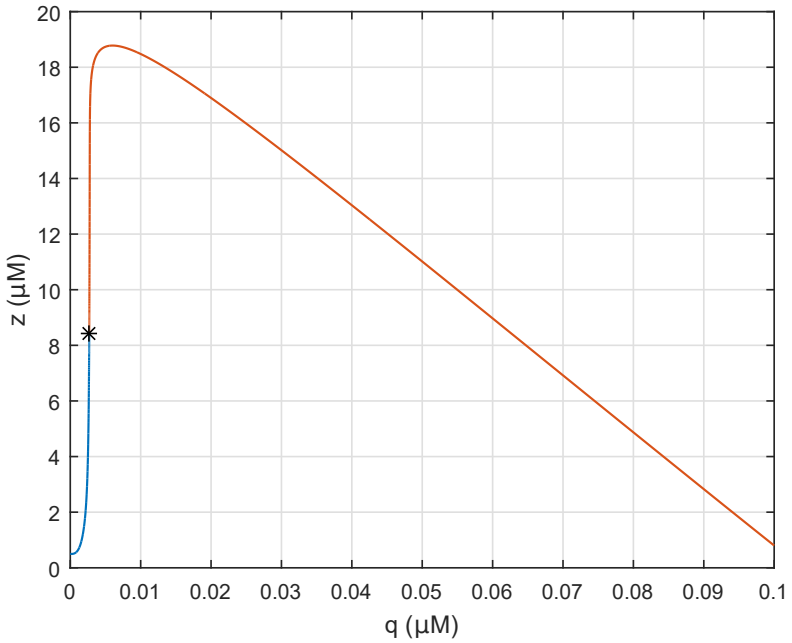


FIGURE 5.7: Simulation with the parameter values in Tables 5.1 and 5.2 from two different initial conditions: $q_{01}, z_{01} = (0, 0.5)$ (blue line) and $q_{02}, z_{02} = (0.1, 0.8)$ (red line). This corresponds to the case that two equilibria exist, $(0, 0)$ and $(0.002694, 8.449)$ (black star). The latter is globally stable in the positive orthant.

in high dimensions. Being qualitative, it can be easily applied to similar models with other assumptions. We consider below the case of different degradation rates.

5.7 Applications to other models

In the above model, we supposed that all forms of RNA polymerase are degraded at the same rate. This assumption could be changed, for example by supposing a degradation of the free form of the polymerase only, but no degradation for the bound forms. The new equations are:

$$\begin{aligned} z &= \gamma \frac{p}{p + k_1} + p, \\ R_0 &= \lambda \frac{q\, r}{r + k_2} + r, \\ \dot{z} &= k_w \frac{q\, r(q)}{r(q) + k_2} - k_p' \, p(z), \\ \dot{q} &= k_c d_0 \frac{p(z)}{p(z) + k_1} - k_m' \left(q - \frac{q\, r(q)}{r(q) + k_2} \right). \end{aligned} \tag{5.18}$$

The conservation equations are the same, but the diagonal elements of the Jacobian matrix change: yet, it is easy to verify that our approach is still applicable, and, similarly,

TABLE 5.4: Parameters of $\chi(t)$

Parameter	Value	Unit
υ_c	0.1251	$\mu M \cdot min^{-1}$
k_c	0.6452	μM
k_{pc}	0.00087264	min^{-1}

the global stability now depends only on the local stability at (0, 0) with the new inequality (for the trivial equilibrium) $k'_p\, k'_m > k_w \frac{k_c\, d_0\, R_0}{k_1\, k_2}$. Other generalizations with qualitative functions of Michaelis-Menten type ($M(p)$ is strictly increasing, concave, bounded, and such that $M(0) = 0$) are possible (see [17, 18]).

5.8 Comparison with a classical model of RNA polymerase

It is interesting to compare the new Model (5.12) with the more "classical" model used to describe gene expression in [8, chapter 2, p. 13], in the case of a protein activating its own transcription:

$$\dot{\chi} = \upsilon_c \frac{\chi}{\chi + k_c} - (k_{pc} + \mu)\, \chi, \tag{5.19}$$

where χ is the protein concentration and μ is the same growth rate as in Table 5.2. We first substitute in the model of Equations (5.18) k'_p with $k_p + \mu$ and k'_m with $k_m + \mu$. The dilution term (term with μ) of the equation $\dot{q}$ can be neglected because biological evidences demonstrate that $\mu \ll k_m$. To simplify our study, we assume that the dynamics of $\dot{z}$ is much slower than the one of $\dot{q}$, so that $\dot{q} \approx 0$ (by a new quasi-steady state approximation). We obtain System (5.20).

$$\begin{aligned} \dot{z} &= k_w \frac{q\, r(q)}{r(q) + k_2} - (k_p + \mu)\, z, \\ q &= \frac{k_c\, d_0}{k_m} \frac{p(z)}{p(z) + k_1}. \end{aligned} \tag{5.20}$$

We use then the points of the solution $z(t)$ to fit the model of Equation (5.19)[6] and find the values of the parameters υ_c, k_c, k_{pc}, shown in Table 5.4, so that the dynamical behavior of the solution $\chi(t)$ is the closest to the one of $z(t)$, where $\chi(0) = z(0)$. With these particular values of parameters, the two solutions are very close: the model of Equations (5.20) can fit classical simpler models.

Yet, from their comparison, we can observe phenomena of biological interest with our model that are not captured in the "classical" model, especially in the translation term.

[6] The fitting was done using the *fmincon* function of MATLAB. We thank Eugenio Cinquemani (Inria, IBIS) for his help with the optimization procedure.

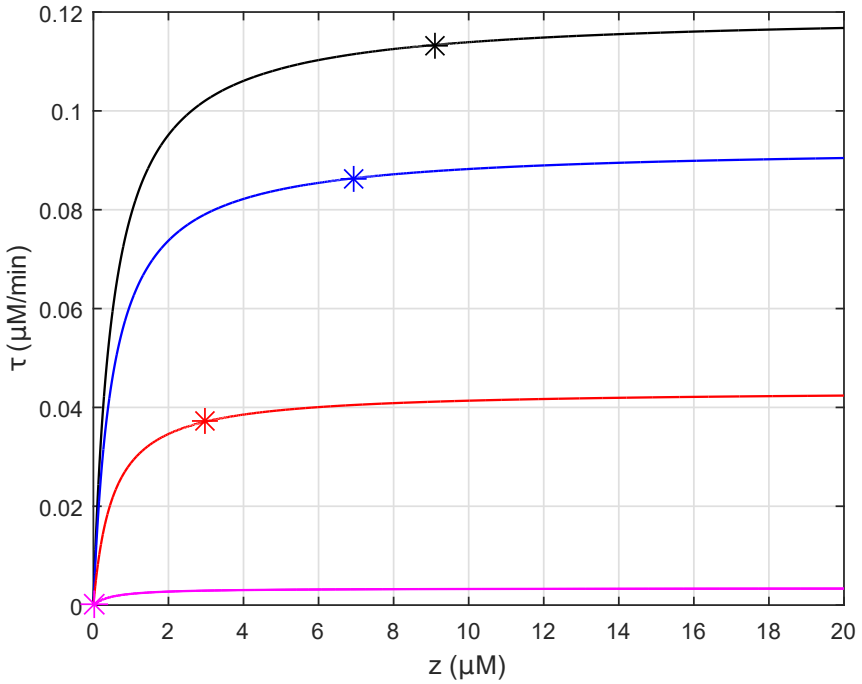


FIGURE 5.8: Variation of the translation process τ with the RNA polymerase concentration z, for different ribosome concentrations: R_0=35 μM (black line; in this case the dynamics of z and χ overlap), $R_0 = 12.5$ μM (blue line), $R_0 = 5$ μM (red line), and $R_0 = 0.3$ μM (magenta line). The star in each curve denotes the equilibrium points, respectively $z = 9.082$ μM, $z = 6.919$ μM, $z = 2.972$ μM, $z = 0$ μM.

.

If we substitute in System (5.20) the algebraic expression for q in the differential expression of z and isolate the translation term, we have:

$$\tau = k_w \frac{q\, r(q)}{r(q) + k_2}. \tag{5.21}$$

Contrary to the "classical" model, this model shows that the translation process τ is sensitive to R_0 (the number of ribosomes): the algebraic part of System (5.18) highlights the dependence between $r(q)$ and R_0 (for $q = 0$, $r(0) = R_0$). In Figure 5.8 are shown different curves of the translation process as a function of RNA polymerase concentration, due to different concentrations of R_0: the more ribosomes are available in the system, the more z can be translated. The first three curves from the top represent three systems that follow the case of Proposition 1 (one positive stable equilibrium) and the last curve, the case of Proposition 2 (0 is the only stable equilibrium). Obviously, these different curves cannot be generated by the simpler model of Equation (5.19).

TABLE 5.5: Growth Rate Parameters

Parameter	Value	Unit
K	8.85	μM
ν	0.0237	$\mu\text{M} \cdot \text{min}^{-1}$

5.9 System with a variable growth rate

In the above sections, we considered a fixed growth rate for the model of Equations (5.12). In the reality, bacteria experience a variable growth rate most of the time. What is the consequence of this variation on the dynamics of the model? To simplify, we study only a one-dimensional model, as in Section 5.8. We study the effect of introducing a variable growth rate, which depends on the concentration of the cell components, here the concentration of RNA polymerase z. The growth rate can be expressed by means of a function analogous to the widely used Monod equation in growth kinetic studies [81, p. 211]:

$$\mu = \nu \frac{z}{K + z}. \tag{5.22}$$

The previous System (5.20) becomes:

$$\begin{aligned} \dot{z} &= k_w \frac{q\, r(q)}{r(q) + k_2} - \left(k_p + \nu \frac{z}{K + z} \right) z, \\ q &= \frac{k_c\, d_0}{k_m} \frac{p(z)}{p(z) + k_1}. \end{aligned} \tag{5.23}$$

A way to find reasonable values for parameters ν and K is to substitute in the Equation (5.22) the equilibrium point that the variable z reaches in System (5.20), z^*, to obtain:

$$\mu(z^*) = \nu \frac{z^*}{K + z^*}. \tag{5.24}$$

We manually set parameter values in Table 5.5 so that $\mu(z^*)$ reaches the value of the constant μ in Table 5.2 and that System (5.22) reaches the same equilibrium point as System (5.20). The values in Table 5.5 allows System (5.24) to have the same equilibrium point of System (5.20). The simulation for z is shown in Figure 5.9 and the growth rate dynamics in Figure 5.10.

To analyze the new system, we notice that the first term of the equation $\dot{z}$, $G(z) = H(q(z)) = k_w \frac{q(z)\, r(q(z))}{r(q(z)) + k_2}$ is a function of q, which in turn is a function of z. Using the same notation as Equations (5.15): $G(z) = \xi(\varphi(z))$, the derivative of this function is $G'(z) = \xi'(\varphi(z))\, \varphi'(z)$. Both term are positive (as discussed in Section 5.6), so the derivative is positive. The second derivative is $G''(z) = \xi''(\varphi(z))\, \varphi'(z) \varphi'(z) + \xi'(\varphi(z))\, \varphi''(z)$. The

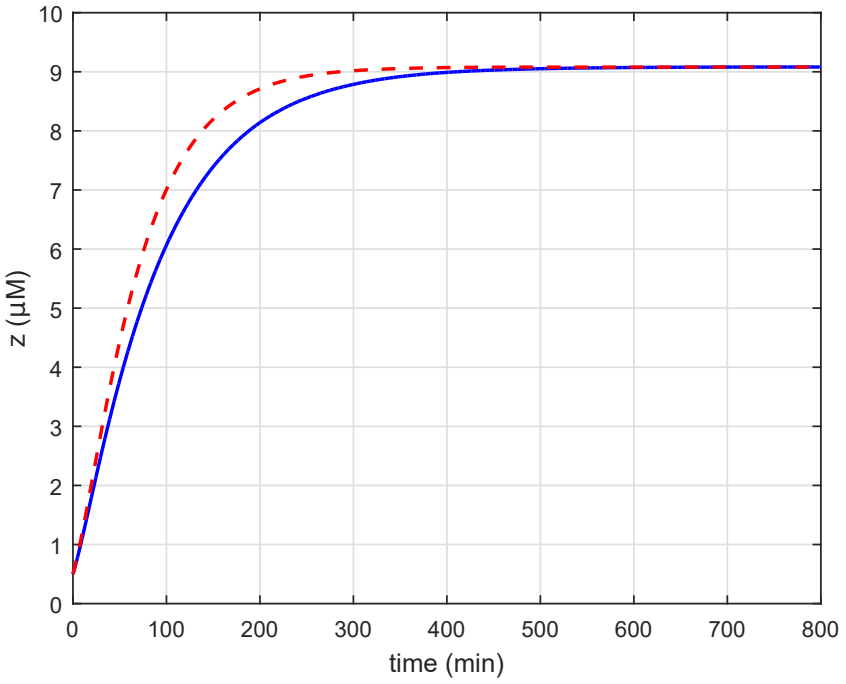


FIGURE 5.9: Evolution of the total concentration of RNA polymerase z in the case of a dependent growth rate (dashed red line), compared with system (5.20) (solid blue line). Initial conditions are: $p_0 = z_0 = 0.5\ \mu M$, $L = l = 100$ nucleotides and $H = h = 33$ amino acids. The final equilibrium is the same at 9.084 μM

.

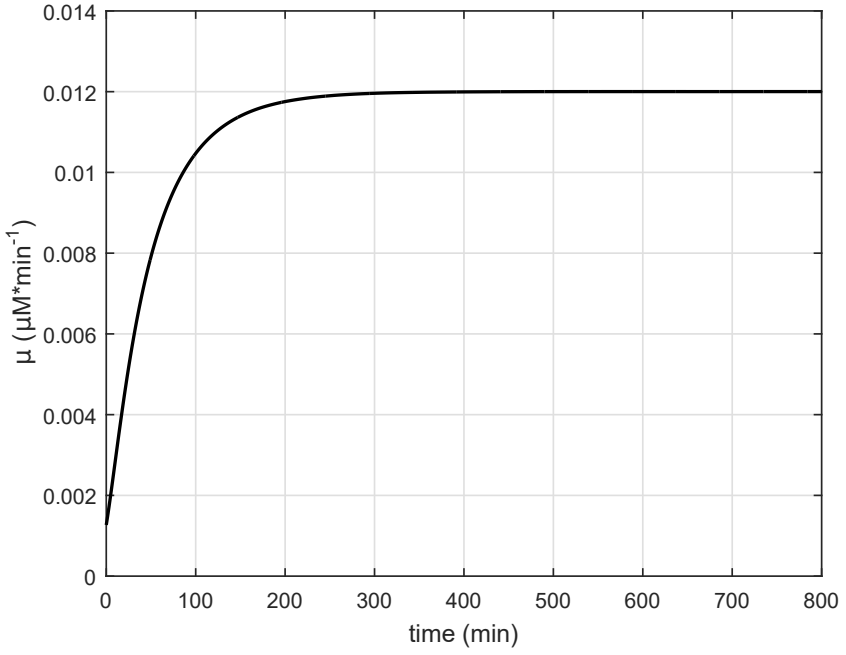


FIGURE 5.10: Growth rate dynamics with the parameters of Table 5.5.

first and last terms are negative and the others positive (as we discussed in Section 5.8), so the second derivative of the function is positive. $G(z)$ is therefore increasing, positive, concave and it is bounded ($\frac{k_c\, d_0\, k_w}{k_m}$). The second term of the equation $D(z) = k_p\, z + \nu \frac{z^2}{K+z}$ is positive, increasing and convex: depending on the values of the parameters, the functions $G(z)$ and $D(z)$ can have one or two equilibrium points.

- If : $D'(z)|_{z=0} < G'(z)|_{z=0} \Rightarrow k_m\, k_p < k_w \frac{k_c d_0}{\gamma + k_1} \frac{R_0}{R_0 + k_2}$, there exist two equilibria which is 0, and a unique, positive z^*.
- If: $D'(z)|_{z=0} > G'(z)|_{z=0} \Rightarrow k_m\, k_p > k_w \frac{k_c d_0}{\gamma + k_1} \frac{R_0}{R_0 + k_2}$, there exists only one equilibrium for the system which is 0.

Following the same rules as in Section 5.6.3, we see that in the first case the equilibrium point z^* is globally stable and 0 is globally unstable, and in the second case 0 is globally stable.

We notice that the growth rate, which depends only on the concentration of RNA polymerase z, does not play any role in the qualitative stability of the system. Of course, it changes the trajectories, as shown in Figure 5.9.

5.10 Conclusion

Several interesting conclusions can be made from this study. First, we demonstrate that tools from monotone theory are useful for proving stability for, e.g. the fast part of the system. Moreover, biological conclusions can be drawn. For example, computations lead to the fact that the system is stable depending on the sign of (see Proposition 1 and Proposition 2):

$k_m\, k_p - k_w \frac{k_c\, d_0}{\gamma + k_1} \frac{R_0}{R_0 + k_2}$.

If R_0 is large (many ribosomes), the zero equilibrium is unstable; if R_0 is small, the zero equilibrium is globally stable, and every variable tends to zero. These results are in agreement with several biological observations on the adaptation of living organisms to their environment. For instance, in the case of bacteria, the zero equilibrium corresponds to the situation of cells whose growth is arrested by harmful environmental conditions. Translation is halted in these cells, through an arrest of ribosome synthesis and the inactivation of the remaining ribosomes [104, 119]. As a consequence, the intracellular concentration of active ribosomes decreases, which lowers the concentration of RNA polymerase. The essential cell components can no longer be synthesized; cells eventually

die if the ribosomes and the RNA polymerase remains at so low concentrations. By contrast, when environmental conditions become favorable again, ribosome synthesis starts immediately and inactivated ribosomes become functional again [104, 119]. The concentration of ribosomes rises in the cell. According to the model, the zero equilibrium becomes unstable in these conditions. The consequence is a rapid accumulation of new pools of RNA polymerase and ribosomes, that are necessary for the cell to synthesize all the precursors needed to grow and divide again. Note that this very simple loop is not isolated from the rest of the cell. The simple model could be easily extended so as to include these regulatory mechanisms. As well, the reduced system could be included into more general models of the gene expression machinery.

Adding to the system a variable growth rate (depending only on the RNA polymerase concentration) does not change the stability of the curve, because for $z = 0 \Rightarrow \mu'(0) = 0$. We can reach the same equilibrium point of z with a constant or a variable growth rate as shown in Figure 5.9, if $\mu(z^*) = \mu$ of Table 5.2.

Moreover, it is important to notice that, in restricted cases, the dynamics of our model overlaps with a simpler dynamical model (see Section 5.8): this confirms the fact that we can explain the dynamics of RNA polymerase in a more precise way, detailing also the relationship with mRNA, without changing its performance with respect to more simpler and classical models.
Finally, an exciting perspective on which we are working is to add a control on this system, for example via the action of an inducer which activates or inhibits the transcription step.

Chapter 6

Principal process analysis and its robustness to parameter changes

In this chapter we discuss a work that has been written as a journal paper and submitted to *Journal of Theoretical Biology* in which I am first author (see Appendix A).

We design a method called *principal process analysis* (PPA) that aims at analyzing the key processes for the system behavior of dynamical networks of high dimension. The knowledge of the system trajectories allows us to decompose the system dynamics into processes that are *active* or *inactive* with respect to a certain threshold value. Process *activities* are graphically represented by boolean and dynamical process maps. We eliminate from the model processes that are always *inactive*, and *inactive* in one or several time windows. This reduces the complex dynamics of the original model to the much simpler dynamics of the core processes, in a succession of sub-models that are easier to analyze. In this chapter we apply the method to a well-known model of circadian rhythms in mammals [73] and we use global relative errors to assess the quality of the model reduction and apply global sensitivity analysis to test the influence of model parameters on the errors. The results obtained prove the robustness of our method. Analysis of the sub-model dynamics allows us to analyze the source of circadian oscillations. We find that the negative feedback loop involving proteins PER, CRY, CLOCK-BMAL1 is the main oscillator, in agreement with previous modeling and experimental studies. Hence, PPA is a simple-to-use method, which constitutes an additional and useful tool for analyzing the complex dynamical behavior of biological systems.

The parameter sensitivity analyses were performed with the supervision of Suzanne Touzeau, researcher at INRA and Inria of Sophia Antipolis.

In Appendix B we present our first use of PPA. We apply our method to a model that describes the circadian rhythm in *Drosophila* [72] and another one that describes the influence of RKIP on the ERK signaling pathway [68]. The work in Appendix B has been presented at the 23rd *Mediterranean Conference on Control and Automation* MED, held in Torremolinos, Spain, on June 16th-19th, 2015 (with peer reviewed proceedings) and has been accepted as a conference paper in which I am first author (see Appendix A).

6.1 Introduction

Mathematical modeling has been used for decades as an approach to address complex problems in several domains of biology. For example, it helps studying the large networks of metabolites, RNAs and proteins that allow cells to live and grow. Numerous models of these networks have been developed in computational biology, of increasing complexity due to advances in modeling and parameter estimation approaches (see [23] and [64] for an example). Complexity arises from the high dimension of these networks, the large number of biological processes involved and their non linearity due to complex feedback loops. This makes the analysis of network functioning rather difficult, notably in terms of key processes and regulatory mechanisms for the system dynamics.

Different techniques are classically used to reduce model complexity. The simplified models are easier to analyze, while retaining the main characteristics of the original models and their biological significance. Quasi-steady-state approximations are mostly used to reduce system dimension when different time scales are present. Replacement of some ordinary differential equations (ODEs) by algebraic expressions results in a differential-algebraic system of smaller dimension. However the reduced models may remain difficult to analyze [103]. Other approaches simplify the mathematical functions describing the molecular processes. For instance, piece-wise affine differential equations approximate by step functions the sigmoidal functions used to describe the regulation of gene expression. The dynamics of the simplified system can be easily analyzed by means of state transition graphs [32]. However, these simplifications are generally restricted to models of gene expression and are more difficult to apply to other types of networks [12].

Here we address the problem of high dimensional model analysis and reduction by developing a mathematical and numerical approach, based on the boolean concept of *activity/inactivity*. The method, called *principal process analysis* (PPA), determines the contribution of each cellular process to the output of the dynamical system, without

changing its main structure. We first identify processes that are not contributing significantly to the system dynamics and neglect the *inactive* ones. Some processes can switch from *active*, when they contribute most to the system dynamics, to *inactive*. In a second step, we thus define time windows in which processes are either always *active* or always *inactive*. We eventually create sub-models for each time window that only contain the *active* processes. This reduces the system to its core mechanisms. PPA is a general approach that can be easily applied to any biological system described by ODEs. For example, it has been recently used to reveal the correlation between C cycling and pesticide degradation in the detritusphere and fungal dynamics [88], or to reduce a dynamic metabolic model of lipid accumulation [95].

PPA shares some common features with a method focusing on the major model parameters rather than processes [9]. The exploration of parameter space leads to admissible system outputs. Parameters contributing most to the system dynamics are identified by cross sections of the admissible regions, whereas PPA uses dynamical weights and fixed thresholds to determine major processes. Parameters contributing less to the system dynamics are eventually removed. Another approach dedicated to chemical kinetics makes use of stoichiometric coefficients and chemical reactions to identify and remove chemical species that contribute less to the model output [91]. In this case, the problem is solved using optimization approaches (see also [24]).

The main objective in this work is that our method should neither change the structure of the model nor require additional and complicated computations as QSSA can do. The original and simplified models should remain close and the interpretation of results should be easy for the biologist.

In Appendix B, we started to develop PPA and applied it to two ODE models whose reduction preserved their dynamical behavior. Hence, the model of circadian rhythms in *Drosophila* [72] was reduced into two models, each describing the system dynamics during day light or darkness. The simpler models maintain a functional negative feedback loop responsible for the oscillatory behavior of clock proteins. In the second model describing the regulation of the ERK signaling pathway, the process found to be the most *active* influences the variable determinant for the system dynamics, as shown in [90]. Questions remained open though, concerning the scalability of the approach and its robustness: to which extent does PPA preserve model dynamics in systems of higher dimension, with many more biological processes involved and interlocked feedback loops? And since the approach requires an *a priori* knowledge of parameter values, how sensitive are process *activities* or *inactivities* to parameter values? In this study, we address these questions by studying a much more complex model of circadian rhythms in mammals, including 16 variables, 76 processes, and intertwined positive and negative

feedback loops [74]. Parameter sensitivity analysis of the global relative error between the original and reduced systems allowed us to assess the quality and robustness of our approach. To that aim, fractional factorial design was used to explore a large parameter space in a limited number of simulations.

Section 6.2 describes the principle of PPA and global sensitivity analysis. Section 6.3 introduces the model of mammalian circadian clock. We apply our approach to this complex model in Sections 6.4 to 6.6, and draw conclusions in Section 6.7.

6.2 Methodology

We describe below the basics of the method. We will use as a running example the 14^{th} variable of the mammalian circadian clock model (see Section 6.3 and Appendix D.1), which describes concentration changes of the nuclear form of protein BMAL1 ($B_N = x_{14}$):

$$\frac{dB_N}{dt} = -V_{3B}\frac{B_N}{K_p + B_N} + V_{4B}\frac{B_{NP}}{K_{dp} + B_{NP}} + k_5 B_C - k_6 B_N - k_7 B_N P C_N + k_8 I_N - k_{dn} B_N \tag{6.1}$$

6.2.1 Principal process analysis (PPA)

Consider the following ODE model of biological network:

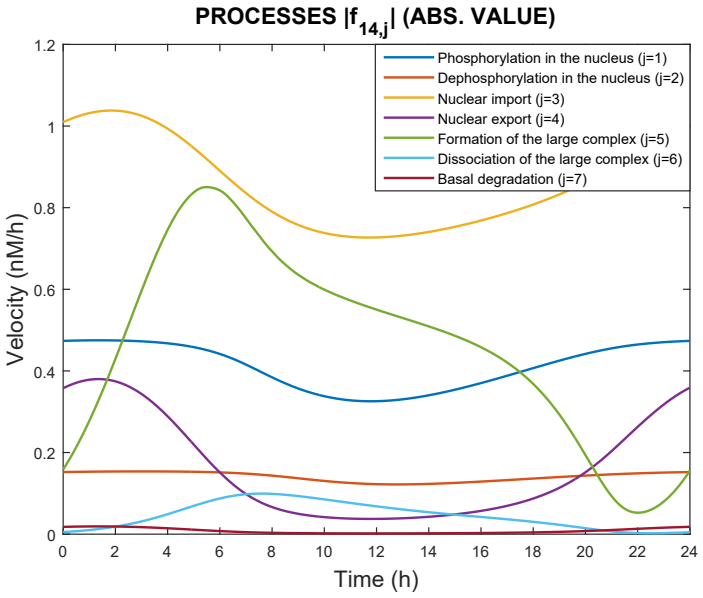
$$\dot{x} = f(x, p) \tag{6.2}$$

where $x = (x_1, x_2, ..., x_n) \ \epsilon \ \mathbb{R}^n$ is the vector of component concentrations, $x0 = (x0_1, x0_2, \ldots, x0_n) \ \epsilon \ \mathbb{R}^n$ the vector of their initial values and $p \ \epsilon \ \mathbb{R}^b$ the vector of parameters. Each equation is decomposed into a sum of biological processes:

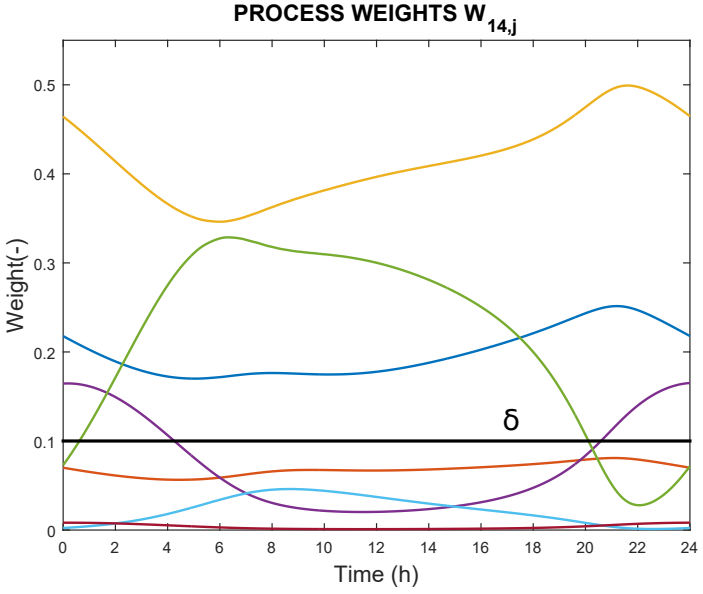
$$\dot{x}_i = \sum_j f_{ij}(x, p) \tag{6.3}$$

where f_{ij} represents the j^{th} process involved in the dynamical evolution of the i^{th} variable of the system over a period of time [0,T].

Example: *Equation (6.1) includes seven processes, each associated with a specific biological function. They take a positive or negative value, depending on whether they affect positively or negatively the variation of BMAL1 concentration. The equation of*



(A)



(B)

FIGURE 6.1: Dynamics of processes that change the nucleic concentration of protein BMAL1 (B_N, see Equations (6.1) and (6.4)) over a 24-hour time window. A: Absolute value of the processes along time (one color per process). B: Weights associated with the processes along time. The threshold δ is set at 0.1.

the protein is rewritten as:

$$\dot{x}_{14} = f_{14,1} + f_{14,2} + f_{14,3} + f_{14,4} + f_{14,5} + f_{14,6} + f_{14,7} \tag{6.4}$$

where $f_{14,1} = -V_{3B}\frac{B_N}{K_p+B_N}, ..., f_{14,7} = -k_{dn}B_N$.

Comparison criteria are needed to weigh the influence of the different processes f_{ij} on

the time evolution of each variable x_i. There are several alternatives. For instance, we can compare their absolute value $\left(|f_{ij}(x,p)|\right)$, scale it by the i^{th} initial condition $\left(\frac{|f_{ij}(x(t),p)|}{x_{0i}}\right)$, or scale it by the solution of the i^{th} ODE $\left(\frac{|f_{ij}(x(t),p)|}{x(t)_i}\right)$. In this work we associate a relative weight to each process to make it dimensionless:

$$W_{ij}(t,p) = \frac{|f_{ij}(x(t),p)|}{\sum_j |f_{ij}(x(t),p)|} \tag{6.5}$$

where $0 \leq W_{ij}(t,p) \leq 1$ and $\sum_j W_{ij}(t,p) = 1$.

Definition: *Let the continuous function $f_{ij}(x(t),p)$ be the j^{th} process of $\dot{x}_i(t)$ in $t \in [0,T]$ and let the threshold $\delta \in [0,1]$.*
We call a process $f_{ij}(x(t),p)$ always inactive when $W_{ij}(t,p) < \delta \ \forall \ t \in [0,T]$.
We call a process $f_{ij}(x(t),p)$ inactive at time t when $W_{ij}(t,p) < \delta$.
We call a process $f_{ij}(x(t),p)$ active at time t when $W_{ij}(t,p) \geq \delta$.
Switching time for a process $f_{ij}(x(t),p)$ is the time t^s_{ij} when $W_{ij}(t,p) = \delta$. A process can have $0,1,...,z$ switching times.
The switching time set S_i for the i^{th} variable contains all the switching times t^s_{ij} where $j = 1,..,k$ and $s = 1,...,z$.
The global switching time set S is the union of all S_i.

Example: *We set $\delta = 0.1$ (see Section 6.2.3 for the choice of this value) and apply Equation (6.5) to the seven processes of Equation (6.1). We obtain their dynamical weight:*

- *the weight of processes $W_{14,2}$, $W_{14,6}$, $W_{14,7}$ is always below δ and their related processes $f_{14,2}$, $f_{14,6}$, $f_{14,7}$ are always inactive;*
- *the processes $W_{14,1}$ and $W_{14,3}$ are always above δ and $f_{14,1}$ and $f_{14,3}$ are active during the whole dynamics;*
- *the weight of the processes $W_{14,4}$ and $W_{14,5}$ crosses the threshold twice and the switching times $t^1_{14,4} = 4.4h$, $t^2_{14,4} = 20.7h$, $t^1_{14,5} = 0.8h$ and $t^2_{14,5} = 20.3h$ are collected in the set S_{14}. Figure 6.1a shows the dynamics of the absolute values of processes involved in Equation (6.1) during a day, while relative weights defined in (6.5) are shown in Figure 6.1b.*

6.2.2 Visualization of process activities

For models as complex as the mammalian circadian clock model, it is advantageous to qualitatively visualize process *activities* or *inactivities* during the system dynamics.

PPA allows to visually summarize this information in one figure with the help of graphical tools. They are described below.

Boolean Process Map: shows the time-dependent *activity* of processes, ordered by variables, during the whole system dynamics $[t_0, T]$. *Active*, resp. *inactive*, processes are depicted by a white, resp. black, bar.

Dynamical Process Map: displays the *activity* of processes and their interactions with variables. In this map, we distinguish three types of process *activity* to take account of system equations sharing common processes. Variables (represented by boxes) are connected by processes (arrows), which can be either *inactive* (shown in black), *active* for all the variables involved (red) or, *active* for some variables involved and *inactive* for the others (yellow).

3-D Process Map: depicts qualitatively for each process, the time-dependent evolution of its intensity. Process *activities* are averaged per hour, leading to the discretization of time. Vertical bars represent process weights for each hour. Their color code represents the intensity of process weights relatively to the other weights.

Example: *Figure 6.2 shows the boolean process map for variable* x_{14} *with its specific switching times (Panel A), the dynamical process map for the time intervals between* $t^1_{14,4}$ *and* $t^1_{14,5}$ *(Panel B), and the 3-D process map with the evolution of seven processes during time, discretized for each hour (Panel C). The nuclear import of protein BMAL1 is the strongest process.*

6.2.3 First model reduction

The first step of PPA identifies *always inactive* processes and remove them from the original system.

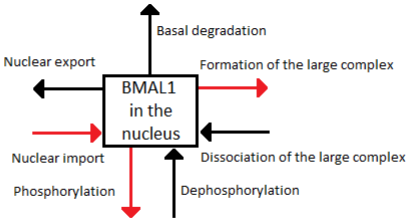
The threshold value δ must be chosen in the range [0,1], preferentially at a low value to avoid neglecting important processes. Otherwise the dynamics of the new system would change significantly. The objective is to obtain $g(x^r)$, the function approximating $f(x)$ and including less processes.

We introduce the ODE system (6.6), which approximates system (6.2):

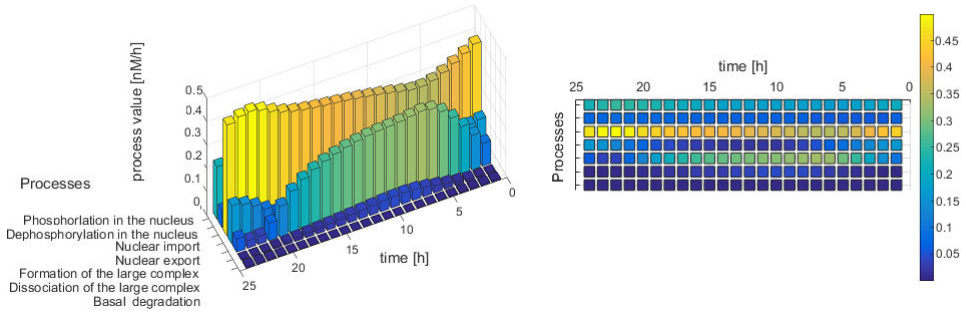
$$\dot{x}^r = g\left(x^r, p^r\right) \tag{6.6}$$

where $x^r = (x^r_1, x^r_2, ..., x^r_n) \; \epsilon \; \mathbb{R}^n$ is the vector of component concentrations, $x0 = (x0_1, x0_2, ..., x0_n) \; \epsilon \; \mathbb{R}^n$ the vector of their initial values, and $p^r \; \epsilon \; \mathbb{R}^c$, where $c \leq b$ is the vector of parameters. The model reduction approach relies basically on the following theorem: if the vector fields of two systems are close ($f(x) \approx g(x)$), then the

(A)



(B)



(C)

FIGURE 6.2: A. *Boolean process map*, B. *Dynamical process map* between times $t^1_{14,4}$ and $t^1_{14,5}$, C. *3-D process map* for the variable x_{14} and its corresponding 2-D version.

solutions of the original and approximated systems are close during some time interval under the assumptions on the Lipschitz conditions listed in [57, p. 96, Th. 3.4].

At this stage, dynamical weights have been assigned to every process and a value has been set for the threshold δ. We can now apply the following rule to define $g(x^r, p^r)$:
if $W_{ij}(x(t), p) < \delta \ \forall \ t \in [0,T]$ *then* $g_{ij} = 0$;
if not, $g_{ij} \equiv f_{ij}$.
We thus define x^r as an approximation of x and p^r as a subset of p.
Example: *Because* $f_{14,2}$, $f_{14,6}$, $f_{14,7}$ *are always inactive,* $g_{14,2} = 0$, $g_{14,6} = 0$, $g_{14,7} = 0$ *and* $g_{14,1} \equiv f_{14,1}$, $g_{14,3} \equiv f_{14,3}$, $g_{14,4} \equiv f_{14,4}$, $g_{14,5} \equiv f_{14,5}$. *The resulting ODE for* x^r_{14} *is:*

$$\frac{dB^r_N}{dt} = -V_{3B}\frac{B^r_N}{K_p + B^r_N} + V_{4B}\frac{B^r_{NP}}{K_{dp} + B^r_{NP}} - k_6 B^r_N - k_7 B^r_N PC^r_N - k_{dn} B^r_N. \tag{6.7}$$

To assess the quality of the reduced model $g(x^r)$, we numerically compute the global

relative error between the original and the reduced models on the six outputs of the system: the concentrations of *Per* mRNA (M_P), *Cry* mRNA (M_C), *Bmal1* mRNA (M_B), total PER protein (P_{Tot}), total CRY protein (C_{Tot}), total BMAL1 protein (B_{Tot})[1]. If y_h and y_h^r are the h^{th} output of the original and the reduced systems respectively, one possible form of global relative error is:

$$e_h = \frac{\int |y_h(t) - y_h^r(t)|dt}{\int |y_h(t)|dt} \tag{6.8}$$

6.2.4 Creation of chains of sub-models

The second step of PPA consists in defining sub-models. The time period during which the system evolves can be split into time intervals using the switching times t_b (with $b = 1, \ldots, d$) previously grouped in set S and sorted in ascending order: this allows creating a succession of sub-models for each time window, which contain the core mechanisms in that period of time. To avoid large chains of sub-models, we reduce the number of time windows by grouping closer switching times with the easy-to-compute k-means clustering [54]. Given our global switching time set $S = [t_1, t_2, ..., t_d]$, this leads us to group the d switching times into z ($\leq d$) clusters C=$\{C_1, C_2, ..., C_z\}$, so as to minimize the within-cluster sum of square (or within-cluster inertia):

$$\text{argmin}_C \sum_{v=1}^{z} \sum_{t \epsilon C_v} ||t - \mu_v||^2 \tag{6.9}$$

where μ_v is the mean of the switching times in C_v. We assume that processes with a switching time in cluster C_v switch together at time $t_v^r = \mu_v$, the mean switching time in cluster C_v. There is no precise rule to choose the number of clusters z, but it can be related to the difference between the maximum and minimum number of *active* processes during the time evolution of the system: if the difference is low, z should be chosen low as well. Such an approach could be:

$$z \approx \frac{\max_v(n_{act}^v) - \min_v(n_{act}^v)}{2}, \tag{6.10}$$

where n_{act}^v denotes the number of *active* processes in the v^{th} time window.

We eventually end up with a chain of $z + 1$ sub-models in the time interval $[0, T]$, the first one being valid in $[0, t_1^r]$, while the last is valid in $[t_z^r, T]$. To test the quality of this second model reduction in each time window, we compute the error (6.8) between the original model and each sub-model. The global error can be calculated with or without

[1]The outputs P_{Tot}, C_{Tot} and B_{Tot} are: $P_{Tot} = P_C + P_{CP} + PC_C + PC_N + PC_{CP} + PC_{NP} + I_N$, $C_{Tot} = C_C + C_{CP} + PC_C + PC_N + PC_{CP} + PC_{NP} + I_N$, $B_{Tot} = B_C + B_{CP} + B_N + B_{NP} + I_N$.

the propagation error: in the first case, for each time window $[t^r_{v-1}, t^r_v]$ ($v = 1, \ldots, z+1$ with $t^r_0 = 0$ and $t^r_{z+1} = T$), the initial values of the h outputs of the sub-model SM_v are equal to the final values at t^r_{v-1} of the sub-model SM_{v-1}; in the second case, they are equal to the values of the original model at t^r_{v-1}.

6.2.5 Global sensitivity analysis

The simplification method described above is performed for a fixed set of parameters and initial conditions. In Chapter 7 we will study the robustness of PPA with respect to variations of the initial conditions of the system. Here we focus instead on the effect of varying parameter values on the quality of the method.

To that aim, we performed a global sensitivity analysis on the global relative errors (6.8) between the original and reduced models. In a first analysis, we considered the errors defined as previously, for the six model outputs (e_{M_P},e_{M_C},e_{M_B},$e_{P_{Tot}}$,$e_{C_{Tot}}$,$e_{B_{Tot}}$); then, in a more detailed analysis, we computed the global relative error for each variable ($e_i, i = 1, ..., 16$). The method we used is based on a factorial design on the uncertain parameters [35, Ch.3, pp. 69-209], analysis of variance (ANOVA) and principal component analysis (PCA) [69].

First, we explored the parameter space using factorial design. We varied $N_f = 51$ parameters of the model [74] (see Section 6.3). We chose $N_l = 2$ levels for each parameter p_f (or factor): $p_f^- = 0.8\, p_f$ and $p_f^+ = 1.2\, p_f$. A full factorial design, defined as all possible combinations of the parameter levels, would be necessary to estimate the main effects and interactions of all parameters. Such a full design corresponds to $N_l^{N_f} = 2^{51}$ parameter combinations and would necessitate the same number of model simulations to compute the corresponding outputs, which are far too many. Thus we implemented a fractional factorial design [27], which is a subset (fraction) of the full design, chosen in order to estimate all main effects α_f and two-way interactions β_{fk} of the following linear statistical model linking the error e_h to the parameters p_f :

$$e_h = \mu + \sum_f \alpha_f + \sum_f \sum_{k \neq f} \beta_{fk} + \epsilon_h \tag{6.11}$$

where μ is the grand mean and ϵ_h the residual. The fractional factorial design is obtained using the R package planor[2] and consists of 2^{12} parameter combinations, yielding as many simulations.

By means of an ANOVA based on the linear model and the simulations described above, one can estimate the sum of squares associated with each factorial term for each error e_h:

[2]Generation of Regular Factorial Designs https://cran.r-project.org/web/packages/planor/

the main effect SS_f^h or the two-way interaction $SS_{f,k}^h$. According to the sparsity-of-effects principle, a system is usually dominated by main effects and low order interactions, so neglecting third order and higher interactions can still provide good estimates. Denoting by SS_T^h the total sum of squares, the total sensitivity index of parameter p_f is defined as follows:

$$tSI_f^h = \frac{SS_f^h + \sum_{k \neq f} SS_{f,k}^h}{SS_T^h}. \tag{6.12}$$

It represents the fraction of the variance of the error ($\sigma_{e_h}^2$) explained by parameter p_f. However, as an ANOVA requires a scalar output, separate sensitivity indices were hence computed for each error e_h. To compare the parameter influence on the different errors e_h, we used non-normalized indices, obtained by multiplying each tSI_f^h by the variance of the error:

$$tSI_f^{h'} = \sigma_{e_h}^2 tSI_f^h. \tag{6.13}$$

To obtain sensitivity indices that represent the global output variance for all 16 variables, a decomposition of the multivariate output (e_h) using PCA was performed (without normalizing e_h). As a result, an inertia proportion ω_l was attributed to each component l. It represents the variability among simulations carried by the component. Only the N_c first components whose cumulated inertia added up to 95% or more were retained. Moreover, each simulation was given a *score* on each component, a scalar representing the projection of the simulation on the component. Then, for each component retained, an ANOVA was performed on the *scores* and total sensitivity indices tSI_f^l were computed, as described in Equation (6.12). Finally, a total generalized sensitivity index was calculated for each parameter p_f as the sum of the total sensitivity indices on each PCA component, weighted by the inertia of the component:

$$tGSI_f = \sum_{l=1}^{N_c} w_l \, tSI_f^l. \tag{6.14}$$

We used the Multisensi R package[3] for this analysis.

6.3 Model description

Periodic fluctuations of the environment subject living organisms to biological rhythms. The latter are endogenous by nature, but entrained by environmental variations. For instance, circadian rhythms are generated by a molecular clock within cells, which synchronizes daily physiological variations to the day-night alternance. The model we study

[3] http://cran.r-project.org/web/packages/multisensi/index.html

here describes the circadian clock in mammals [73, 74]. We summarize it below. In the model, the clock forms a complicated network of intertwined positive and negative feedback loops involving four clock genes: *Per*, *Cry*, *Bmal1*, and *Clock*. Their mRNA and protein produce sustained oscillations with a period of 24 hours. Light affects expression of gene *Per* at the transcriptional level: the first twelve hours of day light increases its transcription rate (up to 1.8 [μM/h]), while it is lowered in the next twelve hours of darkness (down to 1.5 [μM/h]). The system functions as follows (for the complete schema, see Figure 6.3):

- Transcription of genes *Per*; *Cry* and *Bmal1* occurs in the nucleus. The newly synthesized mRNAs are exported into the cytosol.
- In the cytosol, the mRNAs can be either degraded or translated into proteins, which ones are subsequently phosphorylated (the process is reversible). Unphosphorylated proteins PER and CRY form the complex PER-CRY, which reversibly enters the nucleus. The nuclear and cytosolic forms of the complex can be phosphorylated. Likewise, protein BMAL1 is reversibly phosphorylated and reversibly enters the nucleus, but sole its unphosphorylated form makes a complex with protein CLOCK. Phosphorylated proteins and complexes in the nucleus or the cytosol are subject to degradation;
- In the nucleus, the complex CLOCK-BMAL1 activates the transcription of *Per* and *Cry* genes. Activation is stopped by binding of the PER-CRY complex to CLOCK-BMAL1, which indirectly inhibits *Per* and *Cry* transcription;
- The concentration of CLOCK protein is not a variable in the model because it is constitutively expressed at high levels and considered to be not limiting [74].

The 16 model equations, 56 parameter and 16 initial condition values are shown in Appendix D.1. The model dynamics is difficult to analyze though, as the circadian clock involves numerous processes, including interlocked positive and negative feedback loops responsible for the oscillatory behavior of the clock proteins. Reducing the original model around its core *active* processes can facilitate the model analysis, without changing significantly the original dynamics, in particular the sustained oscillations of the solutions.

6.4 Principal process analysis and first reduction

We applied PPA to identify major processes along the system dynamics. We decomposed each ordinary differential equation in processes, as shown in Equation (6.4) for BMAL1.

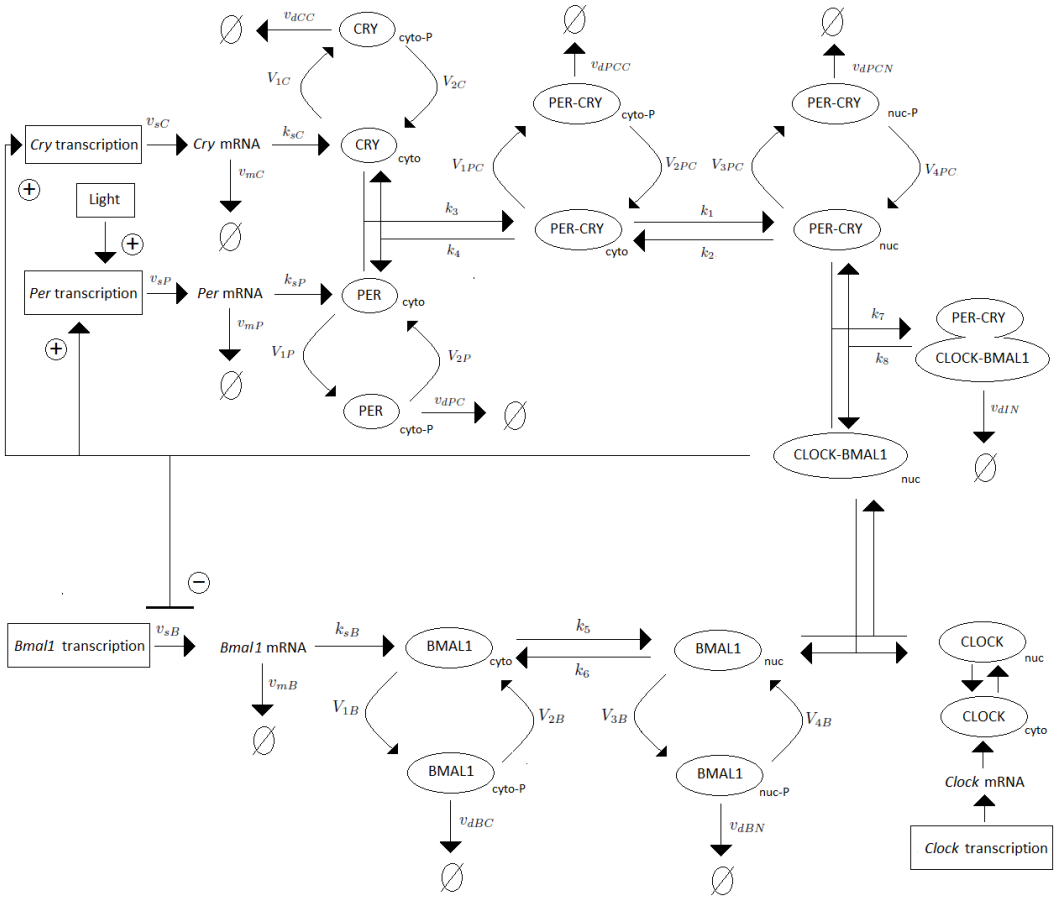


FIGURE 6.3: Schematic representation of the mammalian circadian clock. Light stimulates the transcription of gene *Per*. The complex CLOCK-BMAL1 inhibits the transcription of gene *Bmal1* and activates the transcription of genes *Cry* and *Per*. Notations: $\emptyset$: degradation product; the different forms of a given protein are noted *cyto*: cytosolic form, *nuc*: nuclear form; *P*: phosphorylated form.

Each process has a biological interpretation and corresponds to a regulatory process or a biochemical reaction.

We then calculated the relative weight of each process using Equation (6.5) and set a low threshold $\delta = 0.1$ (see Section 6.2.3 for the choice of this value).

We collected the *switching times* (values given in Appendix D.2) and then built a *boolean process map* to visualize the *activity/inactivity* of each process (see Figure 6.4). We obtained a first reduction of the model by neglecting 24 out of 76 processes, which were *always inactive* (32% of all processes). They correspond to mRNA and protein basal degradations; cytosolic dephosphorylations of CRY, BMAL1, and PER-CRY; PER-CRY-CLOCK-BMAL1 dissociation in the nucleus; and BMAL1 dephosphorylation in the nucleus. The list of neglected processes is shown in Appendix D.3.

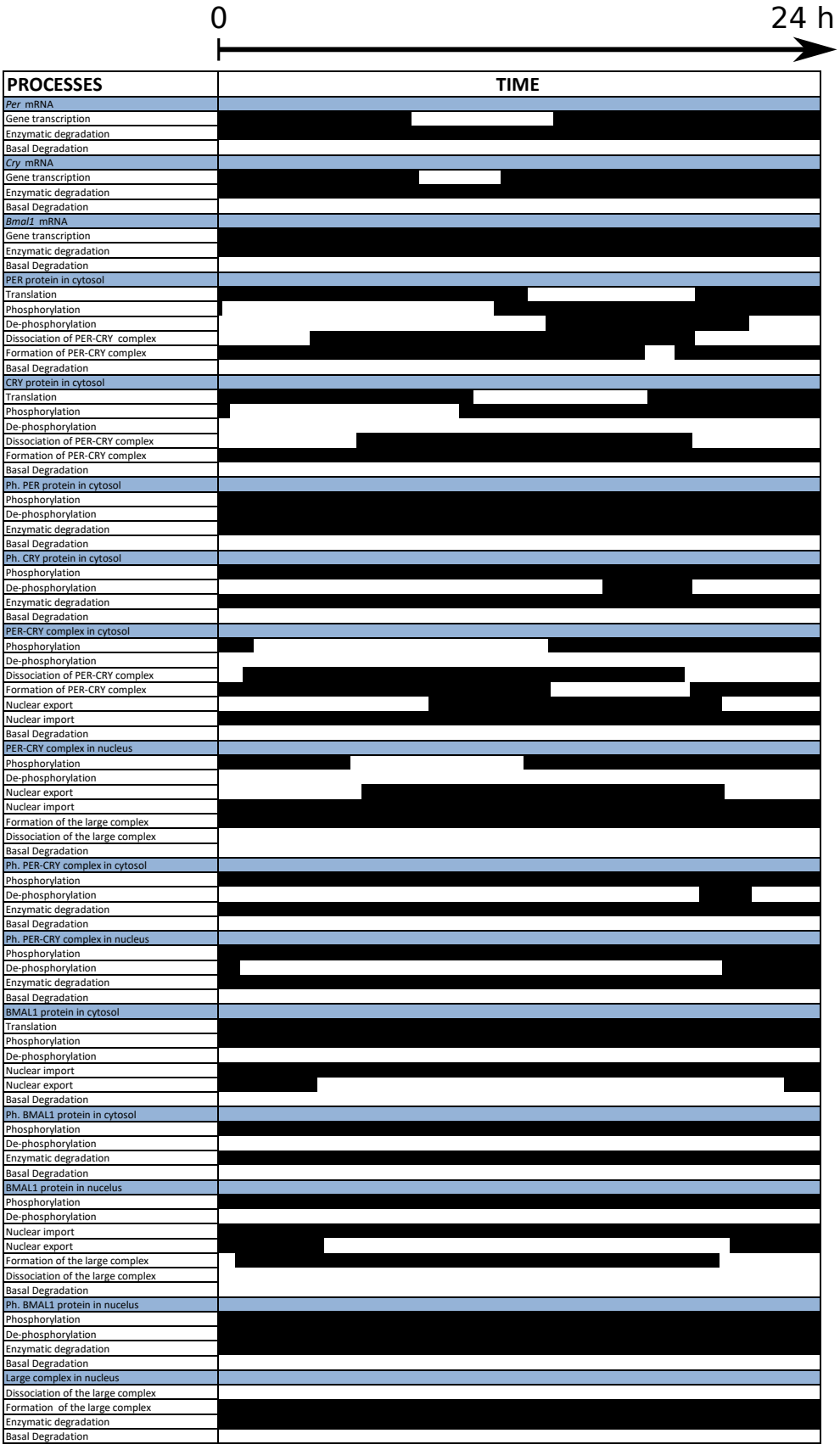


FIGURE 6.4: *Activity* of the 76 model processes during a 24-hour period. Processes are listed in the first column (white background), ordered by variable (blue background). Their *activity* is depicted in the second column between 0 and 24 h: a horizontal black, resp. white, bar when the process is *active*, resp. *inactive*. Values for the *switching times* are given in Appendix D.2.

TABLE 6.1: Global relative errors between the original and reduced models for the six outputs.

Global Relative Error						
Output	M_P	M_C	M_B	P_{Tot}	C_{Tot}	B_{Tot}
Error	0.2499	0.2148	0.1535	0.2648	0.1326	0.2053

We then determined the global relative errors between the original and reduced models using Equation (6.8) for all six outputs (see Table 6.1). The dynamics of the two models are compared in Figure 6.7a. The reduced model preserves qualitatively the trend of the original solutions, as well as their sustained oscillations. The most noticeable difference concerns the peak of the total concentration of protein PER (P_{Tot}), which corresponds also to the highest error in Table 6.1 (26.48 %): the peak is lower with the reduced model, which also explains the delay between the original and reduced solutions.

6.5 Creation of sub-models

The simplified model obtained above can be further reduced if we also neglect processes that are sometimes *inactive* during the system dynamics. Based on the *boolean process map* and the collected *switching times*, we identified between 38 and 45 *active* processes along time (Figure 6.5) and a total of 46 *switching times* (see Figure 6.6a). Clustering the *switching times* into 4 clusters (Figure 6.6) allowed us to generate the five sub-models described below. The number of clusters was chosen according to Equation (6.10).

- $SM1$, valid from $t_0^r = 0$ to $t_1^r = 0.9$ h: neglected processes for this model are *always inactive* (32% of the total). It therefore corresponds to the reduced model obtained in Section 6.4.

- $SM2$, from $t_1^r = 0.9$ h to $t_2^r = 6$ h: 46% of the processes are neglected. In addition to the *always inactive* processes listed in Section 6.4, we have the following *inactive* processes in this model: cytosolic dephosphorylation of PER, CRY, and PER-CRY; cytosolic dissociation of PER-CRY; nuclear dephosphorylation of PER-CRY; PER-CRY export from the nucleus; and formation of the large complex PER-CRY-CLOCK-BMAL1.

- $SM3$, from $t_2^r = 6$ h to $t_3^r = 12.5$ h, in which 50% of processes are neglected. In addition to the *always inactive* processes listed in Section 6.4, *inactive* processes are in this case: transcription of *Per* and *Cry* mRNAs; cytosolic phosphorylations

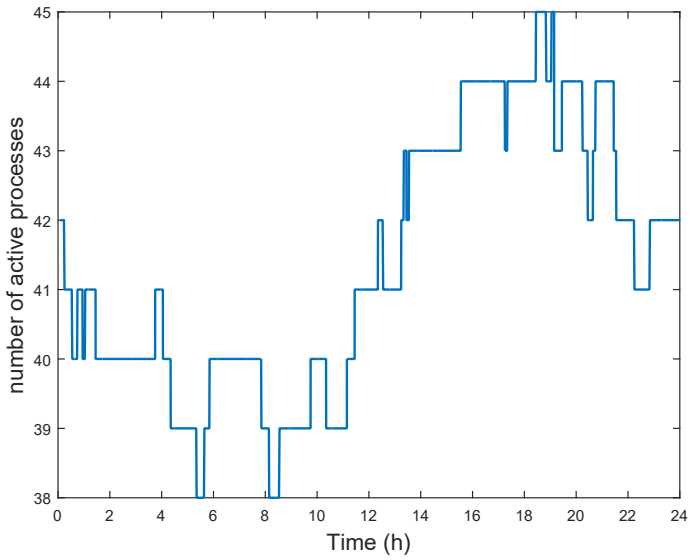


FIGURE 6.5: Evolution of the number of *active* processes as a function of time. The function increases or decreases at *switching times*, listed in Appendix D.2.

and dephosphorylations of PER and CRY; cytosolic dephosphorylation of PER-CRY; nuclear phosphorylation and dephosphorylation of PER-CRY; and nuclear export of BMAL1.

- $SM4$, from $t_3^r = 12.5$ h to $t_4^r = 20$ h, which neglects 42% of processes. The processes include the processes *always inactive* listed in Section 6.4, as well as: PER and CRY translation; formation of the PER-CRY complex in the cytosol; PER-CRY dephosphorylation in the cytosol and the nucleus; and export of BMAL1 from the nucleus.

- $SM5$, from $t_4^r = 20$ h to $t_5^r = 24$ h, in which 46% of the processes are neglected. With the *always inactive* processes listed in Section 6.4, other neglected processes are: cytosolic dephosphorylation of PER and CRY; PER-CRY dissociation in the cytosol; export of PER-CRY; PER-CRY dephosphorylation both in the cytosol and the nucleus; and PER-CRY-CLOCK-BMAL1 formation.

See also Appendix D.3 for the list of neglected processes in each sub-model.

Table 6.2 gives the global relative error without propagation error, between the original model and the sub-models for the six outputs. Figure 6.7b illustrates the six models outputs for the original model and the sub-models without propagation errors, while Figure 6.7c compares the coupled sub-models with and without propagation error. The simplified models preserve the oscillatory behavior of the total concentrations of PER, CRY, and BMAL1, albeit with some discrepancies in the amplitude of the oscillations. It is in the third time window that the approximated solution differs the most from the

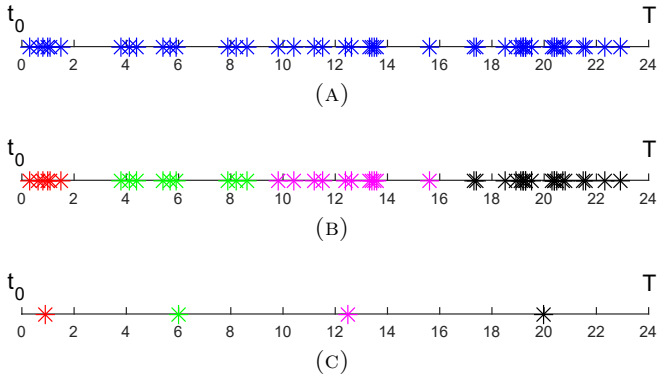


FIGURE 6.6: *Switching time* clustering. A: *switching times* t_b, $b = 1, ..., 46$ (also listed in Appendix D.2). B: the four *switching time* clusters (red, green, pink, black) obtained by the k-means method. C: the four reduced *switching times* $(t_v^r, v = 1, ..., 4)$, corresponding to the mean *switching time* within each cluster.

original one (Table 6.2). This is visible in Figure 6.7b where the total concentrations of PER and CRY form a much higher peak in the reduced solution. Note that this error is not an issue, since our objective is primarily the qualitative analysis of the model. It is sufficient that the remaining processes in the reduced model produce a dynamical behavior qualitatively similar and close to the original model. This shows their important contribution to the system dynamics.

Applying a *dynamical process map* to the third sub-model (Figure 6.8; see also Section 6.2.2) shows that the transcription of *Per* and *Cry* genes is *inactive* (black arrow) and that both PER and CRY phosphorylations in the cytosol and in the nucleus are not fully *active* (they are not *active* for all the variables in which they are involved, yellow arrow). In the other time windows these processes are always fully *active* (red arrows). This probably explains why we had an higher error in Table 6.2 for the variable M_P, M_C, P_{Tot} and C_{Tot} in SM3. The global sensitivity analysis, presented in the next section, confirmed the validity of this assumption.

Since the dynamics of the coupled sub-models remain close to the original one, we can further analyze the behavior of the network reduced to its core processes. We use the *dynamical process maps* for the different sub-models (Appendix D.4), together with the process *activities* in Figure 6.4 and the model outputs in Figure 6.7. The simplified models preserve the three main interlocked feedback loops described in the original model, one positive and two negative loops. The functioning of these loops is directly affected by changes of process *activities*. Among the two negative feedback loops, which one is the main oscillator? One negative feedback loop involves the inhibition of *Bmal1* transcription by the nuclear form of BMAL1 associated to the protein CLOCK. If this mechanism is the main source of oscillations, we should observe wide changes in process *activities* controlling BMAL1 levels. The total concentration of the protein does not

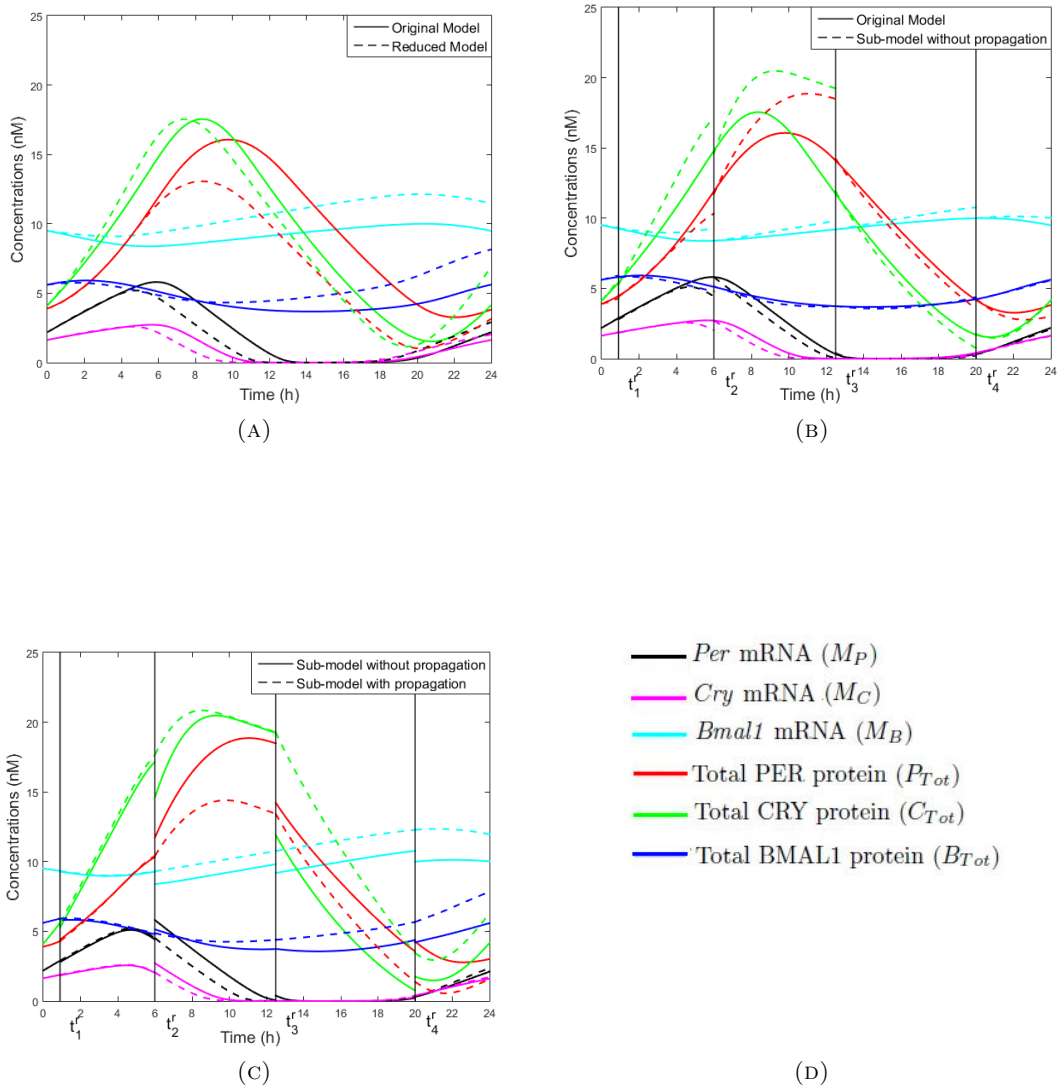


FIGURE 6.7: Model outputs along time for: (A) the original model (solid lines) and the reduced model (dashed lines); (B) the original model (solid lines) and the coupled sub-models without propagation errors (dashed lines); (C) the coupled sub-models, with (dashed lines) and without (solid lines) propagation errors. The equations for the total concentration of protein PER (P_{Tot}), CRY (C_{Tot}) and BMAL1 (B_{Tot}) are: $P_{Tot} = P_C + P_{CP} + PC_C + PC_N + PC_{CP} + PC_{NP} + I_N$, $C_{Tot} = C_C + C_{CP} + PC_C + PC_N + PC_{CP} + PC_{NP} + I_N$, $B_{Tot} = B_C + B_{CP} + B_N + B_{NP} + I_N$.

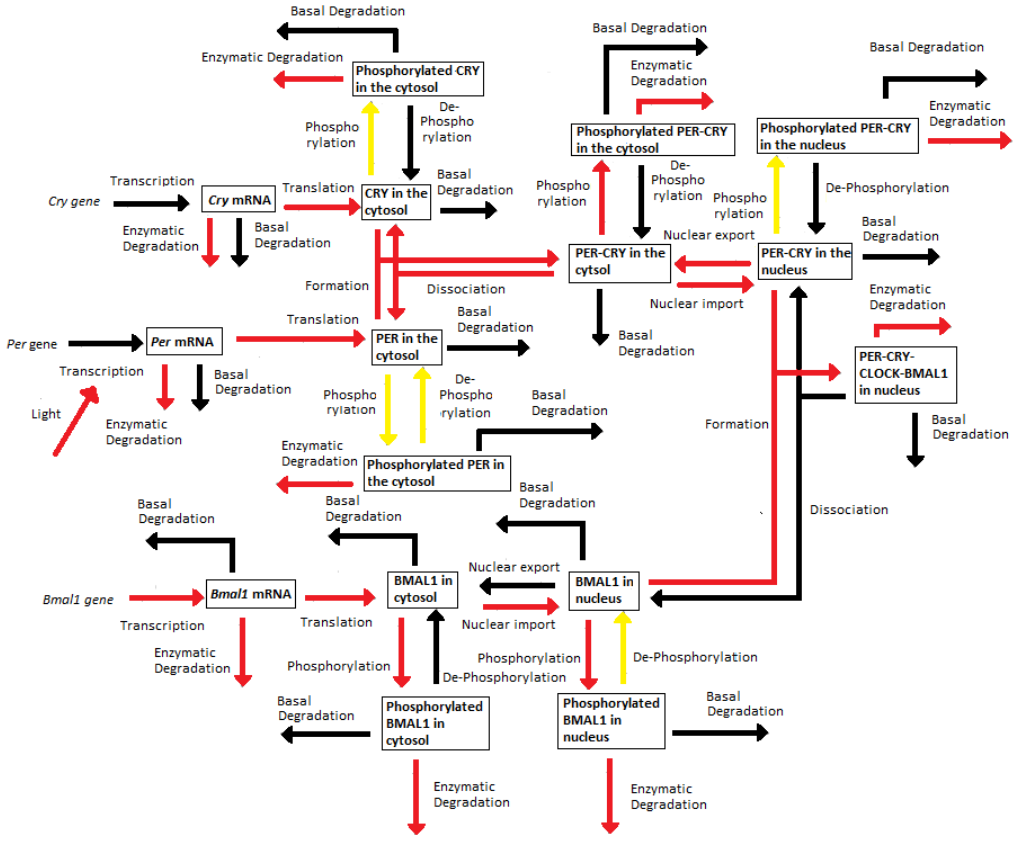


FIGURE 6.8: *Dynamical process map* for the third time window. Variables (boxes) and processes (arrows) are represented, as well as the process *activities*: *inactive* (black); *active* for all variables involved (red); *active* for some variables involved (yellow).

vary much in amplitude (Figure 6.7). It mainly decreases in SM2 and SM3, when the concentration of PER-CRY is also high and forms a complex with CLOCK-BMAL1, which is subsequently degraded. This degradation process is *active* most of the time (Figure 6.4 and Appendix D.4), but variations of the total BMAL1 concentration do not modify strongly the transcription of *Bmal1* mRNA, which remains always *active*. As well, the other processes of translation, phosphorylation and degradation for this variable almost never switch between *inactive* and *active* states over time (Figure 6.4 and Appendix D.4). Overall, this suggests that the negative feedback loop involving CLOCK-BMAL1 is not the main oscillator. This is consistent with analysis results of the original model in [74].

The other negative feedback loop inhibits *Per* and *Cry* transcription through the titration of CLOCK-BMAL1 by PER-CRY to form the inhibitory complex PER-CRY-CLOCK-BMAL1. The total concentration of BMAL1 peaks before that of PER and CRY, as can be seen in Figure 6.7 for SM2 and SM3. When its concentration is maximal in SM1 and SM2, the nuclear form of the protein associated to the protein CLOCK, stimulates

the transcription of *Per* and *Cry* genes, in conditions where light has also a stimulatory effect on the transcription of these two genes. The processes of transcription and translation of *Per* and *Cry* are *active* in both models, as a result of which levels of PER and CRY raise to reach their maximal concentration in SM3. As can be seen from the process *activities* in Figure 6.4 and the *dynamical process maps* in Appendix D.4 conditions are favorable for the accumulation of high levels of complexes PER-CRY and CLOCK-BMAL1-PER-CRY in the nucleus. For instance, numerous processes decreasing PER, CRY and PER-CRY concentrations in the cytosol and the nucleus are *inactive*: their phosphorylation is reduced (the process is *inactive* for the dephosphorylated forms but still *active* for the phosphorylated ones), which limits their degradation, and the nuclear import of PER-CRY is always *active*. During the same period of time, the formation of the large complex CLOCK-BMAL1-PER-CRY, which is *active* for both CLOCK-BMAL1 and PER-CRY (Figure 6.4 and Appendix D.4), suggests that the nuclear forms of PER-CRY and CLOCK-BMAL1 bind as soon as they accumulate in the nucleus. The large complex is immediately degraded since its degradation process is always *active* and its dissociation, *always inactive*.

In SM2 and SM3, the degradation of the large complex is not compensated for by other mechanisms allowing BMAL1 accumulation in the nucleus: the cytosolic form of the protein is *actively* phosphorylated and then degraded, while its dephosphorylation is *inactive*, which reduces the quantity of protein to be imported in the nucleus (see Figure 6.4 and the *dynamical process maps* in Appendix D.4). In this compartment, the absence of *active* dephosphorylation, together with the *active* protein phosphorylation, also contribute to decrease pools of CLOCK-BMAL1 complexes (Figure 6.4, Section D.4). This halts transcription of *Per* and *Cry* mRNAs in SM3 (the processes are *inactive* and light is also switched off towards the end of SM3). This also affects the translation of PER and CRY, which becomes *inactive* in SM4. Altogether these observations suggest that the negative feedback loop inhibiting *Per* and *Cry* transcription via the complex CLOCK-BMAL1-PER-CRY is the main source of circadian oscillations. This is consistent with conclusions in [74], which could obtain a second oscillator based on the auto-inhibition of BMAL1 for specific parameter values only. These results are also consistent with the observation of arrhythmic behaviors in mutant mice with double knock-out of the *Per* and *Cry* genes [118, 123].

The positive feedback loop activates *Per* and *Cry* transcription through a control of protein stability mediated by the phosphorylation processes. In the model, sole the phosphorylated forms of the proteins are degraded. We observed that the reversible phosphorylation reactions are often displaced in the forward sense, as dephosphorylation processes are often found *inactive*. In particular, they contribute to decrease the concentration of PER, CRY and PER-CRY, which also diminishes the concentration of

TABLE 6.2: Global relative error between the original model and each sub-model without propagation error for the six outputs.

Global Relative Error						
Output	M_P	M_C	M_B	P_{Tot}	C_{Tot}	B_{Tot}
Error SM1	0.0044	0.0044	0.0044	0.0208	0.0195	0.0073
Error SM2	0.0519	0.0434	0.0453	0.0397	0.1832	0.0402
Error SM3	0.2059	0.2951	0.0360	0.1427	0.2233	0.0356
Error SM4	0.0143	0.0377	0.0389	0.0678	0.1164	0.0210
Error SM5	0.0146	0.0032	0.0230	0.1150	0.0237	0.0053

the large complex CLOCK-BMAL1-PER-CRY and thus relieves the inhibition exerted by the complex on transcription of *Per* and *Cry* genes. Kinetic modeling of the circadian clock in *Drosophila* has shown the importance of this positive feedback loop for circadian rhythms [117].

6.6 Parameter influence

To check the robustness of the five sub-models, we performed a global sensitivity analysis on the output errors (e_h) without propagation error for each time window. We varied 51 among the 56 parameters of the model: the Hill coefficients m and n are kept fixed because they represent the degree of cooperativity in gene repression/activation, while k_{stot}, v_{stot}, V_{phos} are function of other parameters (see Appendix D.1). We hence computed the non normalized total sensitivity indices for all parameters according to Equation (6.13) (see Figure 6.9, first column). Because the last three outputs ($P_{Tot}, C_{Tot}, B_{Tot}$) are the sum of model variables that interact, some processes have no impact on these outputs and the information on the parameter influence is lost. Thus we also performed the global sensitivity analysis on the 16 global relative errors between the original model and the sub-model variables without propagation (see Figure 6.9, second column). The complex PER-CRY plays an important role in every time window: its variability is due mostly to its maximal phosphorylation velocity (V_{1PC}) and its degradation parameter (v_{dPCC}). In the third and fourth time window the other important variation is due to the CRY protein: in SM3 the variation is mostly due to the binding constants in the transcription of *Per* and *Cry* mRNAs (K_{AP} and K_{AC}) and in SM4, to the maximal translation rate of BMAL1 (k_{sB}) that stimulates *Per* and *Cry* mRNA transcription. In the last time window, lots of variables contribute to the system variation: the most important parameter for the variability of the outputs is the maximal velocity of BMAL1 phosphorylation in the nucleus (V_{3B}).

To get a more global view, we calculated for each parameter set and for each time window, the average error (averaged over the 16 variables) between the original model and the sub-model variables as follows:

$$\bar{e} = \frac{1}{16}\sum_{i=1}^{16} e_i \tag{6.15}$$

Results are shown in Figure 6.10: the variability is higher in the third and four sub-model: anyway, the difference between the lower and upper quartiles is low in all the sub-models.

Then, for each time-window, we computed the total generalized sensitivity indices according to Equation (6.14), which represents the fraction of error variability explained by each parameter when parameter values vary. The results are shown in Figure 6.11: we obtain similar results to the ones in Figure 6.9 (column 2), where in SM1 and SM2 the maximal phosphorylation velocity (V_{1PC}) and degradation (v_{dPCC}) of PER-CRY complex play the main role, in SM3 the binding constants of *Per* and *Cry* protein (K_{AP} and K_{AC}), in SM4 the translation of BMAL1 protein (k_{sB}) and in SM5 the maximal phosphorylation velocity of BMAL1 protein in the nucleus (V_{3B}). To check whether the error variations between the original model and the sub-models are due to parameters appearing in neglected processes, we calculated the following ratio: $R_h = \frac{\sum_{f\in\{inactive\ processes\}} tGSI_f^h}{\sum_f tGSI_f^h}$. We only used the 10 most informative parameters, with higher $tGSI$, as they explained most variability. We chose a conservative option: if a parameter is neglected in an *inactive* process but still appeared in other *active* processes, we still considered that it belongs to the neglected process parameters (worst case). Results are shown in Table 6.3. In most time windows, the variability is mainly due to parameters still contained in the reduced sub-models, i.e. the parameters of the *active* processes. In the third time-window, however, parameters appearing in neglected processes generate more than 50 % of the variability. It is consistent with Figure 6.7b: the peaks of the total concentration of PER and CRY are overestimated by the sub-model and some of the most important parameters that lead to the output variability for this time window are the translation rate of PER and CRY proteins, the maximal phosphorylation velocity of PER-CRY complex in the cytosol and nucleus (as it has been shown in Figure 6.11). This confirms what we have supposed when applying the *dynamical process map* to SM3 (see discussion about Figure 6.8 at the end of Section 6.5).

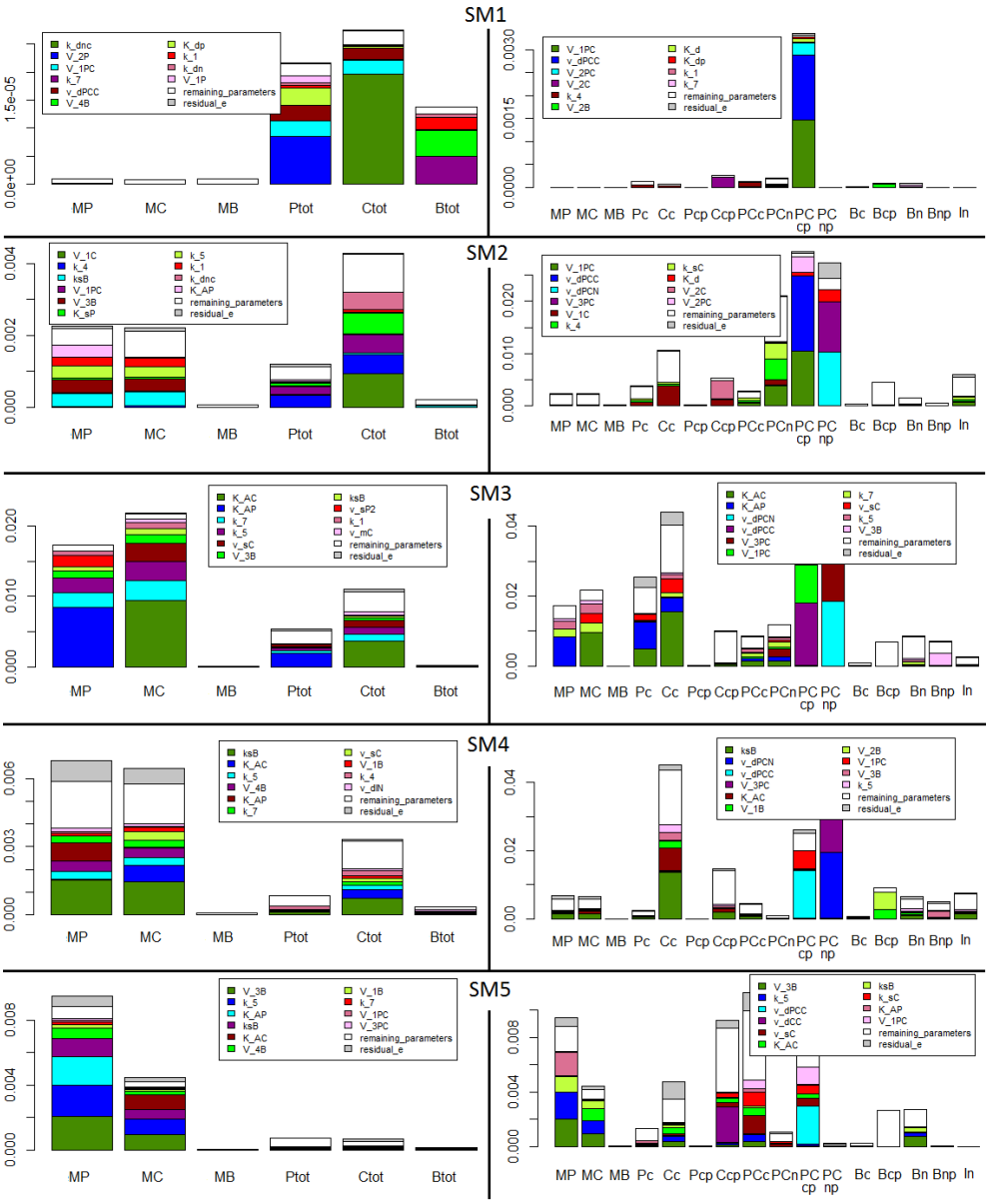


FIGURE 6.9: Global sensitivity analysis on the output (left column) or variable (right column) errors between the original model and the sub-models without propagation error for each time window (lines). Non-normalized total sensitivity indices are represented for each error (one bar per error) and for: (i) the 10 most influential parameters (color-coded); (ii) the remaining parameters (white). The residual is also represented (grey). For the biological meaning of the variables in the second column, see the equations in Appendix D.1.

TABLE 6.3: Percentage of *tGSI* for parameters contained in *inactive* processes .

% *tGSI* ***inactive***					
SM	$SM1$	$SM2$	$SM3$	$SM4$	$SM5$
$R_h(\%)$	19.11	15.55	59.54	0	0

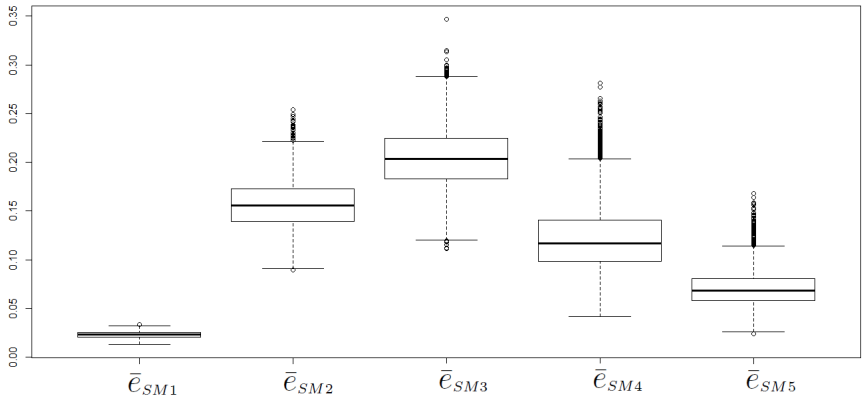


FIGURE 6.10: Average error between the original model and the sub-model variables calculated in each time window according to Equation (6.15). Variability (box-plots) within each sub-model (or time window) is due to the various parameter sets designed for the sensitivity analysis.

6.7 Conclusion

Model reduction approaches have been used to analyze biochemical network models since long, but there is no ideal method for models as complex as the mammalian circadian clock model. The main challenge when analyzing this type of network is to gain knowledge on key processes: ideally, one would like to identify the major processes, quantify and then understand their contribution to the system dynamics. PPA has been developed with this objective in mind, and with the final goal of reducing the original model in one or several sub-models around core *active* processes that are easier to analyze. Questions remained open though concerning the scalability and robustness of this approach.

In this chapter we applied PPA on a model of high dimension, which incorporates numerous processes and complex interlocked feedback loops responsible for oscillatory behaviors. Reduction of the original system dynamics to as much as 50% of its processes in five coupled sub-models helped us relate the dynamics of the simplified models to the system components and their *active* interactions. We hence observed that the negative feedback loop controlling *Per* and *Cry* transcription through the formation of the large complex PER-CRY-CLOCK-BMAL1 is the main oscillator, in agreement with previous experimental and modeling studies [74, 118, 123].

The quantification of the global errors allowed us conclude that the simplified models are good approximations of the original ones. Even in the case of the largest errors observed on the model output, did the simplified models preserve the oscillations of the clock proteins. Since PPA is based on the a priori knowledge of the model parameters, it was

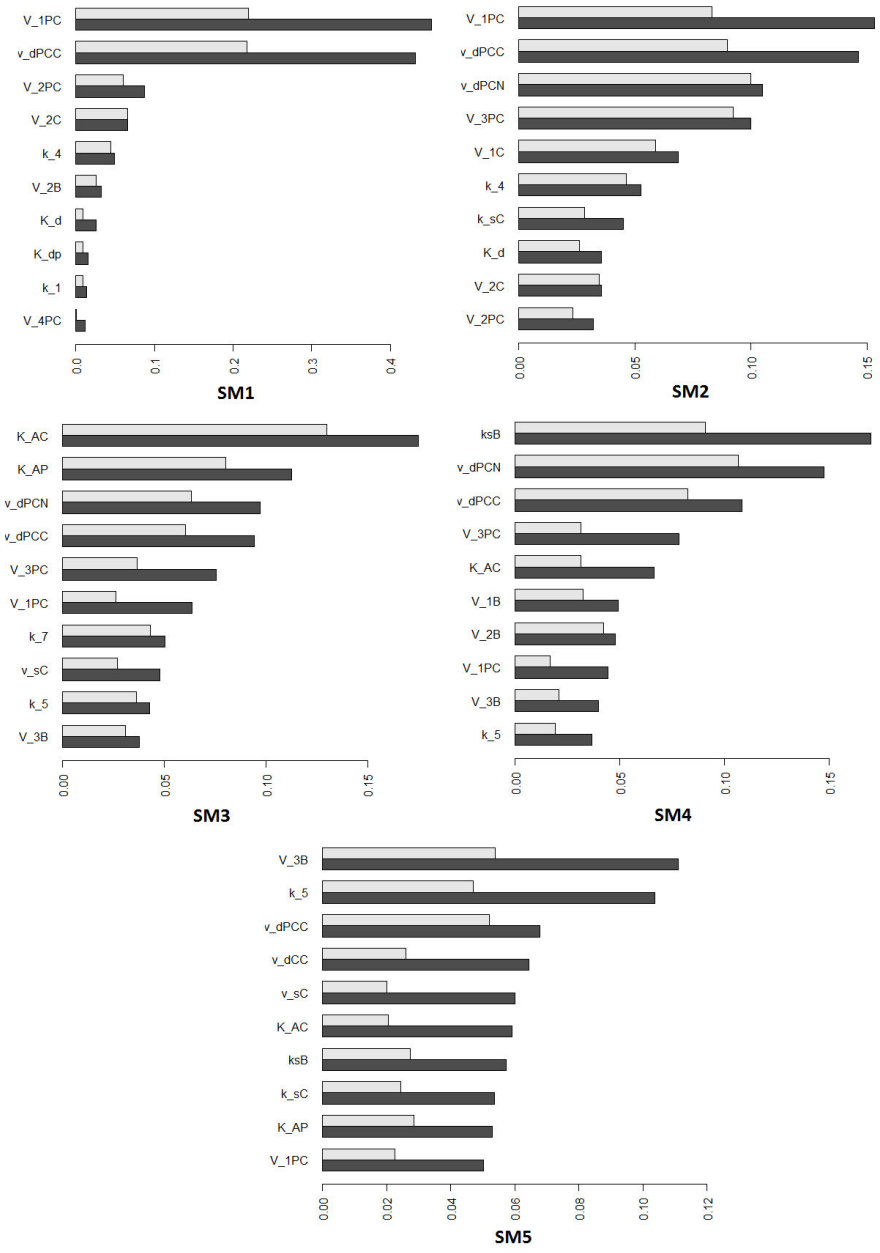


FIGURE 6.11: Generalized sensitivity indices (GSI) computed for each sub-model on the errors between the original model and the sub-model variables. The 10 most influential parameters on the errors are retained: main effect (grey bar) and total *GSI* (black bar).

important to assess the robustness of the approach to uncertainties on these parameter values. Through a global sensitivity analysis, we studied the impact of variations of parameter values on the error between the original model and the reduced sub-models. Not only was the variation of the error small, but it was mostly due to parameters of the neglected processes. With this analysis, we proved the robustness of PPA to parameter uncertainty. In a Chapter 7, we will show the robustness of PPA to initial conditions.

In Chapter 9 we will present a refinement of PPA by considering three different levels of *activities* (*inactive*, *moderately active*, *fully active*), defined by two different thresholds in order to improve the quality of model analysis.

Model analysis is the primary goal of PPA, but the method could be used as well for model reduction purposes: this requires to obtain better reduced models. Another possible extension is to apply PPA on the full coupled system of equations instead of working on each equation separately: this would help to analyze *activities* or *inactivities* of processes shared by several equations.

Chapter 7

Principal process analysis and reduction of biological models with different orders of magnitude

In this chapter we discuss a work that will be presented at IFAC 2017 World Congress (with peer reviewed proceedings) and has been accepted as a conference paper in which I am first author (see Appendix A).

This work is an extension of what we have presented in Chapter 6, testing the robustness of *principal process analysis* based on a change in initial values. First, we decompose the model into biological meaningful processes and then study their *activity* or *inactivity* during the time evolution of the system. Then the structure of the model is reduced to the core mechanisms involving only the *active* processes. The initial conditions are supposed to lie in some rectangle, that could represent one order of magnitude for the variables. Keeping only the *active* processes, we obtain the principal processes in the rectangle and then in the adjacent rectangles where the trajectories may have a transition. Finally we obtain a partition of the space with a reduced model within each rectangle. We apply these techniques to a classical model of gene expression with a protein and a messenger RNA.

7.1 Introduction

In the previous Chapter 6 and in Appendix B we applied *principal process analysis* (PPA) on different models: the results were valid for a fixed set of initial values and parameter values, and we tested the robustness of this approach with respect to a change

of parameter values in the Mammalian Clock Model of Chapter 6. In this chapter, we rather focus on how the choice of the initial conditions affects the *activity* of the processes, resulting into a possible change of the reduced model: this work has similarities with the work in [19] and other qualitative approaches based on phase-space partition (e.g. see [48–50, 66]), but our approach is not mainly oriented toward reduction and applies to general systems.

Instead of a single initial point we consider the PPA on an entire set of possible initial values. For the sake of simplicity and brevity, and because the orders of magnitude of the variables are very important in biological models, we consider initial conditions in rectangles representing one order of magnitude (e.g., the variables are between 1 and 10, or 10 and 100...) and we limit this first approach to the dimension two. It is however clear that it could be applied to any rectangular grid, and to any dimensions (but the notations would be more cumbersome). The plane (x_1,x_2) is therefore divided into a logarithmic grid, and we apply (under some assumptions concerning the monotonicity of the processes) our method by computing a maximal bound for the weight of each process within the rectangle. We only retain the *active* processes, having a dynamical weight higher than a fixed threshold δ.

The chapter is organized as follows: in Section 7.2, we present the technique of reduction for a fixed initial condition, then compute the weights in the rectangle and finally within every rectangle of the space that can be reached from the initial rectangle. In Section 7.3 we present the gene model and in Sections 7.4 and 7.5 we apply the technique presented in the previous sections. The conclusions are presented in Section 7.6.

7.2 Methodology

7.2.1 Principal process analysis and model reduction

We briefly remind PPA (see Chapter 6) for fixed initial conditions and parameter value. Consider the following ODE system that models a biological network (for example an intracellular network):

$$\dot{x} = f(x, p) \tag{7.1}$$

where $x = (x_1, x_2, .., x_n) \in \mathbb{R}^n$ is the vector of concentrations of the components, $x_0 = (x_{01}, x_{02}, ..., x_{0n}) \in \mathbb{R}^n$ is the vector of initial conditions and $p \in \mathbb{R}^b$ is the vector of

parameters. It is possible to decompose each equation into a sum of biological processes:

$$\dot{x}_i = \sum_j f_{i,j}(x,p) \tag{7.2}$$

where $f_{i,j}(x,p)$ represents the j^{th} process involved in the dynamical evolution of the i^{th} variable of the system over a period of time [0,T].
The first equation in the gene expression model in Section 7.3, Equation (7.9), is taken as an example: it represents the variation of the concentration of mRNA. It contains four different processes, each of which with a specific biological meaning. They can be positive or negative:

$$\dot{x}_1 = f_{1,1} + f_{1,2} + f_{1,3} + f_{1,4} \tag{7.3}$$

where $f_{1,1} = \kappa_1$, $f_{1,2} = \kappa_2 \frac{\alpha_P^m}{\alpha_P^m + P^m}$, $f_{1,3} = -\gamma_M$, $f_{1,4} = -\gamma_P$.
In order to compare the influence of the different processes $f_{i,j}(x,p)$ in the evolution of each variable x_i, we associate to them a dimensionless relative weight:

$$W_{i,j}(t,p) = \frac{|f_{i,j}(x(t),p)|}{\sum_j |f_{i,j}\,(x(t),p)\,|} \tag{7.4}$$

where $0 \leq W_{i,j}(t,p) \leq 1$ and $\sum_j W_{i,j}(t,p) = 1$.

Definition: *Let the continuous function $f_{i,j}(x(t),p)$ be the j^{th} process of $\dot{x}(t)_i$ for $t \in [0,T]$ and let the threshold $\delta \in [0,1]$.*
We call a process $f_{i,j}(x(t),p)$ always inactive when $W_{i,j}(t,p) < \delta \;\forall\; t \in [0,T]$.
We call a process $f_{i,j}(x(t),p)$ inactive at time t when $W_{i,j}(t,p) < \delta$.
We call a process $f_{ij}(x(t),p)$ active at time t when $W_{i,j}(t,p) \geq \delta$.

The first step of the PPA is to identify the *always inactive* processes and delete them from the original System (7.1). The threshold value δ must be chosen between the range [0,1]: a low threshold avoids neglecting important processes.
The goal is to obtain a function $g(x^r)$ which approximates the function $f(x)$, that contains a minor number of processes. Let consider the ODE system $g(x^r)$ which approximates the System (7.1):

$$\dot{x}^r = g\,(x^r, p^r) \tag{7.5}$$

where $x^r = (x_1^r, x_2^r, .., x_n^r) \epsilon \mathbb{R}^n$ is the vector of concentration of the components, x_0 is the vector of their initial values and $p \epsilon \mathbb{R}^c$, where $c \leq b$ is the vector of the parameters. The basic idea of the proposed model reduction method is based on the following classical theorem: if the vector fields of two systems are close ($f(x) \approx g(x)$), then the solutions

of the original and approximated systems are close during some time interval under the assumptions on the Lispchitz conditions listed in [57, p. 96, Th. 3.4].
After having assigned dynamical weights to every process and a value to the threshold δ , we follow this rule to obtain $g(x^r)$:
if $W_{i,j}(x(t),p) < \delta \ \forall \ t \ \epsilon \ [0,T]$ *then* $g_{i,j}(x(t),p) = 0$*;*
if not, $g_{i,j}(x(t),p) = f_{i,j}(x(t),p)$.
For example, applying this rule to Equation (7.3) we find that the processes $f_{1,1}$ and $f_{1,3}$ are *always inactive* (see Section 7.3). To test the quality of the reduced model $g(x^r)$, we numerically compute the global relative error between the original and reduced model for each variable. It is defined for the i^{th} variable as:

$$e_i = \frac{\int |x_i(t) - x_i^r(t)|dt}{\int |x_i(t)|dt} \tag{7.6}$$

where $x_i(t)$ and $x_i^r(t)$ are respectively the solutions of the original and reduced systems. This method strongly depends on the initial condition.

7.2.2 Principal process analysis and model reduction based on initial conditions in a rectangle

To increase the robustness of the method, the initial condition is chosen in some region, then we compute if the *activity/inactivity* of the process $f_{i,j}$ - and consequently the reduced system $g(x^r)$ - changes.
We divide the variable space into rectangles, and then apply the technique in each domain: for simplicity, we consider in this paper a system with two variables (x_1, x_2) and a logarithmic subdivision, corresponding to order of magnitude from the modeling point of view. The grid is shown in Figure 7.1.
Every point $\theta^{m,n} = (\theta_1^m, \theta_2^n)$ corresponds the value $(10^m, 10^n)$: for example the point $\theta^{2,0} = (10^2, 10^0) = (100, 1)$. We call $B_{m,n}$ the rectangle delimited by the four vertices $\theta^{m,n}$, $\theta^{m+1,n}$, $\theta^{m+1,n+1}$ and $\theta^{m,n+1}$ (shown in Figure 7.2): inside of it, every process $f_{i,j}(x,p)$ is limited horizontally $f_{i,j}(\theta_1^{m,n},p) < f_{i,j}(x,p) < f_{i,j}(\theta_1^{m+1,n},p)$ and vertically $f_{i,j}(\theta_2^{m,n},p) < f_{i,j}(x,p) < f_{ij}(\theta_2^{m,n+1},p)$.

To compute a global bound for the weights in the rectangle, we need the following assumption for the processes. Below, all the functions are supposed to be locally Lipschitz; by "fixed sign", we mean that the functions are either non-negative, or non-positive, or zero.

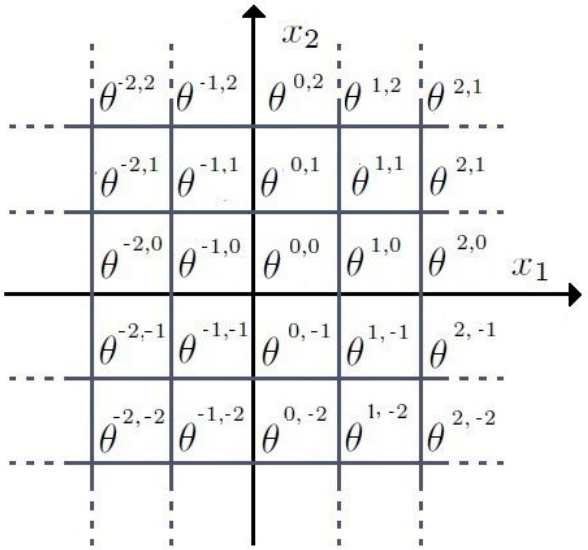


FIGURE 7.1: The variable space (coordinates x_1, x_2), divided in domains. Each vertex $\theta^{m,n}$ corresponds to the value $(10^m, 10^n)$.

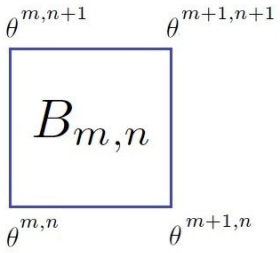


FIGURE 7.2: A generic rectangle $B_{m,n}$ delimited by the vertices $\theta^{m,n}$, $\theta^{m+1,n}$, $\theta^{m+1,n}$, $\theta^{m+1,n+1}$.

Assumption: *$\partial f_{ij}/\partial x_k$ has a fixed sign in B_{mn}, $\forall i, k \in \{1, \ldots, n\}, \forall j$; moreover for a given i and k, all the $\partial f_{ij}/\partial x_k$ have the same sign.*

In words, it means that all the processes for the velocity $\dot{x}_i$ have a derivative of a fixed sign with respect to any variable. This assumption is verified for many models. For example in Equation (7.3) $f_{1,1}$ is a constant and $f_{1,2}$, $f_{1,3}$, $f_{1,4}$ are decreasing processes. Because of Equation (7.2), it easily implies the following corollary.

Corollary: *The Jacobian matrix $J = Df(x,p)$ of the System (7.1) has a fixed sign inside the rectangle $B_{m,n}$.*

Remark that the Jacobian matrix is signed, but all the signs may be different (therefore the system is not monotone in the sense of conservation of partial order between trajectories). This assumption allows to study the behavior of the process $f_{i,j}(x,p)$ inside the full rectangle $B_{m,n}$, knowing only the behavior of the process at the vertices $\theta^{m,n}$,

$\theta^{m+1,n}$, $\theta^{m+1,n+1}$ and $\theta^{m,n+1}$. Indeed, the monotonicity of each process with respect to any variable implies that:

Corollary: *In the rectangle $B_{m,n}$, each process $f_{i,j}$ takes its maximum and minimum on the vertices of the rectangle.*

We note $S_{i,j}^{m,n}$ the vertex of $B_{m,n}$ where the process f_{ij} is maximum, and $s_{i,j}^{m,n}$ the vertex of $B_{m,n}$ where the process f_{ij} is minimum in $B_{m,n}$.

Inside $B_{m,n}$, a worst-case version of the general weight in Equation (7.4) is:

$$WW_{i,j}^{B_{m,n}}(p) = \frac{|f_{i,j}(S_{i,j}^{m,n}, p)|}{\sum_j |f_{i,j}(s_{i,j}^{m,n}, p)|} \tag{7.7}$$

and normalizing these weights to proportions summing to one we obtain:

$$W_{i,j}^{B_{m,n}}(p) = \frac{|f_{i,j}(S_{i,j}^{m,n}, p)|}{\sum_j |f_{i,j}(S_{i,j}^{m,n}, p)|}. \tag{7.8}$$

The reduction method in $B_{m,n}$ is now similar to the previous one: if this weight is smaller than some threshold δ, then the process is considered as *inactive* in $B_{m,n}$, and discarded. A reduced model is obtained within each domain $B_{m,n}$ by keeping the principal processes. For example, we find that the processes $f_{1,1}$, $f_{1,2}$, $f_{1,4}$ of the Equation (7.9), of the model described in Section 7.3, are *always inactive* in the rectangle $B_{0,0}$ because their weight, represented by Formula (7.8), are below the threshold δ.

7.2.3 Possible transitions between domains

Furthermore, our assumption has strong consequences concerning the possible transitions between rectangles. These transitions are conditioned by the vector field on the boundary (the edges) of each rectangle $B_{m,n}$. This analysis is valid for the full or the reduced model.

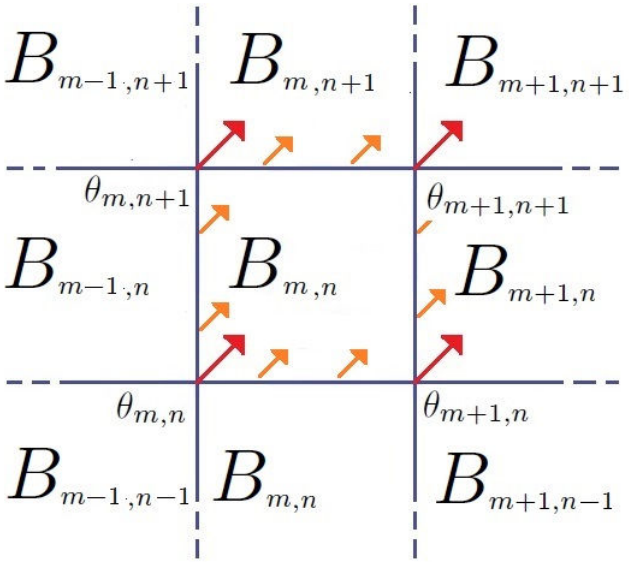
Proposition: *For $i \in \{1, 2, ..., n\}$, if $\dot{x}_i$ is positive (resp. negative) on two adjacent vertices, then it is positive (resp. negative) on the edge between these two vertices. If the signs are opposite, then the vector field will cancel somewhere on the edge.*

Proof: By Corollary (2), the Jacobian matrix is signed, and therefore each component of the vector field is increasing or decreasing along an edge.

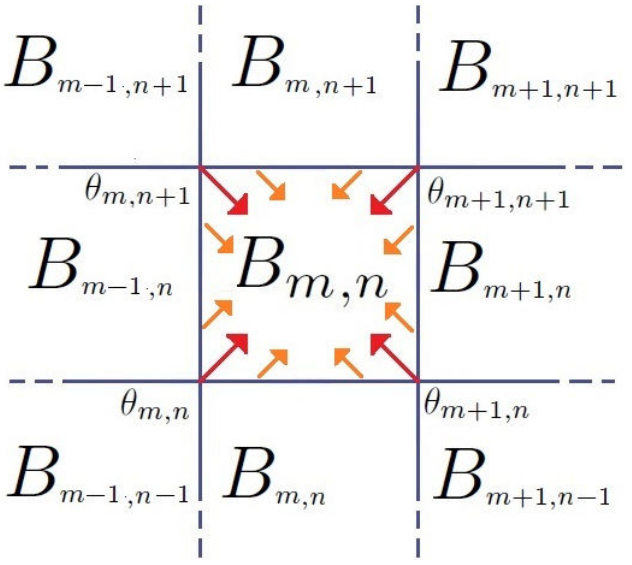
Therefore, knowing the values of the processes at the vertices of the rectangle, allows to study the possible transitions of solutions $x(t)$ from a rectangle $B_{m,n}$ into another adjacent rectangle or into the same rectangle.

As an example, in Figure 7.3 two different situations are illustrated. In Figure 7.3a, the solutions of the system x moves in the rectangles $B_{m+1,n}$ and $B_{m,n+1}$. In Figure 7.3b,

the solutions stay in the rectangle $B_{m,n}$, which is invariant.



(A)



(B)

FIGURE 7.3: In Figure 7.3a (top) the vector field leads the solutions of the system to move from $B_{m,n}$ to the adjacent rectangles. In Figure 7.3b (bottom) the vector field leads the solutions to stay inside the rectangle $B_{m,n}$.

Therefore, having the reduced model in some rectangle $B_{m,n}$, the vector field on the vertices gives the possible transitions toward adjacent rectangles. Then the model is reduced in these rectangles. Finally a graph of transition is obtained between rectangles, each rectangle having a reduced model. The biologist may follow the possible sequence of reduced models with respect to the different orders of magnitude. This sequence is not deterministic because most of the time a rectangle has transitions in several rectangles. Moreover, from the sequence of reduced models, if some process is *always inactive* for

every rectangle, it is possible to obtain a global reduced model valid on the whole pathway of rectangles.

7.3 The gene expression model

We apply this technique on a classical deterministic model in which the protein P inhibits its own mRNA [20, p.57].
The amount of mRNA produced (variable M) depends on its basal activity κ_1 and on the transcription activity, based in turn on the concentration of its DNA sites bound to the repressor P and on the amount of the free DNA sites, $\kappa_2 \frac{\alpha_P^m}{\alpha_P^m + P^m}$. The mRNA can degrade with a degradation term γ_M and be diluted due to growth rate (μ). The translation process leads the production of the protein P and it is designed as a linear function of the mRNA ($\kappa_3\, M$) . The protein can also degrade with a degradation term γ_P and be diluted due to growth rate (μ).

$$\frac{d}{dt}M = \kappa_1 + \kappa_2 \frac{\alpha_P^m}{\alpha_P^m + P^m} - (\gamma_M + \mu)\, M \tag{7.9}$$

$$\frac{d}{dt}P = \kappa_3\, M - (\gamma_P + \mu)\, P\,. \tag{7.10}$$

In Table 7.1 are presented the model parameters with their units. We used the information contained in database for biological numbers, BioNumbers, to give to model parameters reasonable values [78]. The criteria we used to choose the parameter values are:

- An average protein concentration at steady state would be in the order of μM. In our model we choose P^*=1 μM and $M^* = 0.015$ μM;
- We consider a cell doubling time of 1 hour, which gives the growth rate $\mu = 0.0116$ min^{-1};
- Proteins are usually stable. We consider an half life of 5 hours for protein P, which gives $\gamma_P = 0.0023$ min^{-1}. On the contrary, mRNAs are not stable, having half-lives of a few minutes. We consider 4 minutes here, which gives: $\gamma_M = 0.1733$ min^{-1};
- At steady state, $\kappa_3\, M^* = (\gamma_P + \mu)\, P^*$. We replace known parameters and variables by their value to estimate κ_3, which gives: $\kappa_3 = 1.39$ min^{-1};
- We choose α_P of the same order of magnitude as P such that the Hill function term plays a role in the system dynamics (we choose a reasonable valure for m=2).

From the Equation (7.9) at its steady-state and from the reasonable assumption that $\kappa_2 = 2\kappa_1$, we obtain: $\kappa_1 = 9.25 \times 10^{-4}\,\mu\text{M min}^{-1}$ and $\kappa_2 = 0.00185\,\mu\text{M min}^{-1}$.

TABLE 7.1: Parameters

Parameter	Value	Unit
κ_1	0.000925	$\mu M min^{-1}$
κ_2	0.00185	$\mu M min^{-1}$
κ_3	1.39	–
γ_M	0.1733	min^{-1}
γ_P	0.0023	min^{-1}
m	2	–
α_P	20	μM
μ	0.0166	$\mu M min^{-1}$

7.4 Model reduction from an initial condition

As a first step, we decompose the ODE System (7.9)-(7.10) in the following processes:

- the mRNA derivative ($\frac{d}{dt}M$) can be divided into its basal activity process $f_{1,1} = \kappa_1$, into transcription process $f_{1,2} = \kappa_2 \frac{\alpha_P^m}{\alpha_P^m + P^m}$, into degradation process $f_{1,3} = \gamma_M M$ and into dilution process $f_{1,4} = \mu M$;
- the protein derivative ($\frac{d}{dt}P$) can be divided into translation process $f_{2,1} = \kappa_3 M$, into degradation process $f_{2,2} = \gamma_P P$ and into dilution process $f_{2,3} = \mu P$.

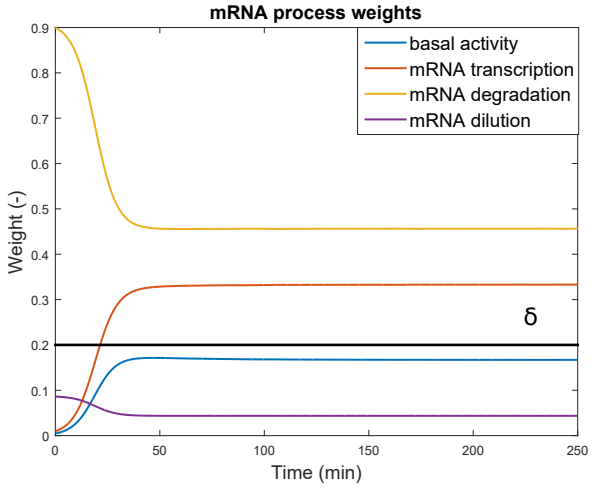
We calculate the process weights of the system using Formula (7.4), having a initial conditions vector: $x_0 = [\theta_1^0, \theta_2^0]$. In Figure 7.4 are shown the plots of the weights of the processes for a fix threshold of $\delta = 0.2$.

From this analysis, the processes resulting *always inactive* are $f_{1,1}$, $f_{1,3}$, $f_{2,2}$. The new system $g(x^r)$ is:

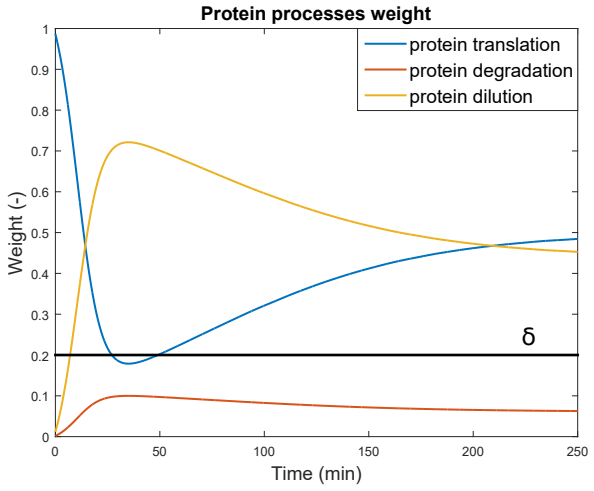
$$\frac{d}{dt}M^r = \kappa_2 \frac{\alpha_P^m}{\alpha_P^m + (P^r)^m} - \gamma_M\, M^r \tag{7.11}$$

$$\frac{d}{dt}P^r = \kappa_3\, M^r - \mu\, P^r\,. \tag{7.12}$$

In Figure 7.5 is shown the solution of the variable P in the original system $f(x)$ and the reduced one $g(x^r)$.



(A)



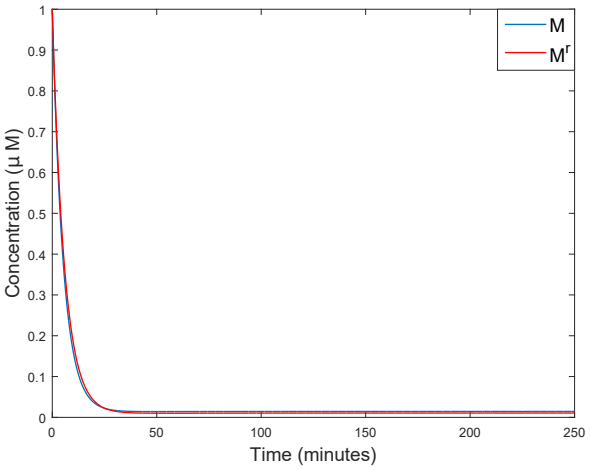
(B)

FIGURE 7.4: 7.4a. The evolution in time of mRNA process weights: the basal activity process and the degradation process are *always inactive* because the dynamics are always under the threshold δ. 7.4b. The evolution in time of protein process weights: the dilution process is *always inactive* because the dynamic is always under the threshold δ.

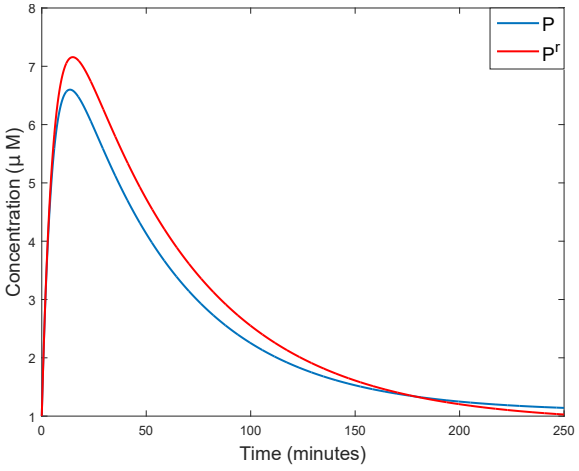
The global relative errors are shown in Table 7.2.

TABLE 7.2: Global Relative Errors

Variable	values
e_M	0.14
e_P	0.10



(A)



(B)

FIGURE 7.5: A. The solution of the variable M of the original model in blue and the solution the variable M^r of the reduced model in red. B. The solution of the variable P of the original model in blue and the solution the variable P^r of the reduced model in red.

7.5 Model reduction in a rectangle

We extend our method to the entire rectangle $B_{0,0}$ that has the vertices ($\theta^{0,0}$, $\theta^{0,1}$, $\theta^{1,1}$, $\theta^{1,0}$). We first verify the assumption on the monotonicity of the processes and the Jacobian matrix, written as:

$$J = \begin{bmatrix} \frac{df_1}{dM} & \frac{df_1}{dP} \\ \frac{df_2}{dM} & \frac{df_2}{dP} \end{bmatrix} = \begin{bmatrix} -(\gamma_M + \mu) & -\kappa_2 \frac{\alpha_P^m P^m m}{(\alpha_P^m + P^m)^2 P} \\ \kappa_3 & -(\gamma_P + \mu) \end{bmatrix}$$

We compute the vector field for the rectangle $B_{0,0}$, $\theta_1^0 < x_1 < \theta_1^1$ and $\theta_2^0 < x_2 < \theta_2^1$ at the 4 vertices:

- At $\theta^{0,0}$: $\sum_j f_{1,j}(\theta_1^0, \theta_2^0, p) < 0$, $\sum_j f_{2,j}(\theta_1^0, \theta_2^0, p) > 0$
- At $\theta^{0,1}$: $\sum_j f_{1,j}(\theta_1^0, \theta_2^1, p) < 0$, $\sum_j f_{2,j}(\theta_1^0, \theta_2^1, p) > 0$
- At $\theta^{1,0}$: $\sum_j f_{1,j}(\theta_1^1, \theta_2^0, p) < 0$, $\sum_j f_{2,j}(\theta_1^1, \theta_2^0, p) > 0$
- At $\theta^{1,1}$: $\sum_j f_{1,j}(\theta_1^1, \theta_2^1, p) < 0$, $\sum_j f_{2,j}(\theta_1^1, \theta_2^1, p) > 0$

Because of the monotonicity of the Jacobian matrix, we can deduce the behavior of the processes on the edges of the rectangle. The result is presented in Figure 7.6. Based on the direction of the arrows, the solutions move to the rectangles $B_{-1,0}$ and $B_{0,1}$.
Using the Formula (7.8), it is possible to compute the weight for the entire rectangle $B_{0,0}$, based on the worst case. In Table 7.3, for the rectangle $B_{0,0}$ we show the maximum value that the process $f_{i,j}(x,p)$ can reach and its weights: setting the value of the threshold δ at 0.2, we can neglect the processes $f_{1,1}$, $f_{1,2}$, $f_{1,4}$, $f_{2,2}$, $f_{2,3}$.

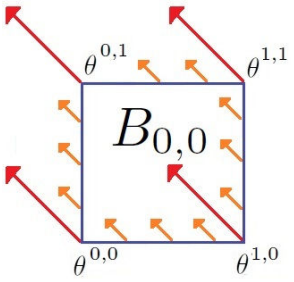


FIGURE 7.6: Vector field on the edges. The solutions are moving into the adjacent rectangles $B_{-1,0}$ and $B_{0,1}$.

TABLE 7.3: Processes in $B_{0,0}$

Process	Max. Value (μM)	Weight
κ_1	0.000925	0.00048638
$\kappa_2 \frac{\alpha_P^m}{\alpha_P^m + P^m}$	0.0018455	0.00097
$\gamma_M M$	1.733	0.91125
μM	0.166	0.087287
$\kappa_3 M$	13.9	0.9866
$\gamma_P P$	0.023	0.0016
μP	0.166	0.0118

The valid sub-model $g(x^r)$ for $B_{0,0}$ is:

$$\frac{d}{dt} M^r = -\gamma_M M^r \tag{7.13}$$

$$\frac{d}{dt} P^r = \kappa_3 M^r \tag{7.14}$$

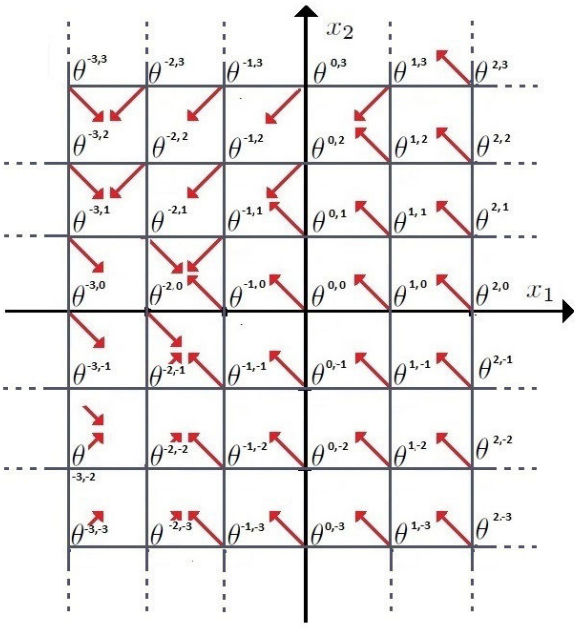


FIGURE 7.7: Vector field on the edges in the plane.

The sub-Model (7.13)-(7.14) is only valid in the rectangle $B_{0,0}$. To study the dynamics of the process weights over the whole time $[0, T]$ as we did for $\theta^{0,0}$ in Figure 7.4, we need to know the pathway of the solutions x in the different rectangles.

Extending the PPA as we did in Figure 7.6 to the full domain, we obtain the result shown in Figure 7.7: from any initial value x_0 the solutions are moving into the final rectangles $B_{-2,0}$ $B_{-2,-1}$. Starting from an initial value inside the rectangle $B_{0,0}$ we can have different solution pathways: the solutions can move into rectangle $B_{-1,0}$ or $B_{0,1}$, then from $B_{-1,0}$ they can move into $B_{-2,0}$ or $B_{-1,1}$ and from $B_{0,1}$ to $B_{0,2}$ or $B_{-1,1}$. Every pathway eventually ends in the space occupied by the rectangles $B_{-2,0}$ and $B_{-2,-1}$.

To explain the application of our technique, we perform it on one of the possible pathways that starts from the rectangle $B_{0,0}$: $B_{0,0} \Longrightarrow B_{0,1} \Longrightarrow B_{-1,1} \Longrightarrow B_{-2,1} \Longrightarrow B_{-3,1} \Longrightarrow B_{-3,0} \Longrightarrow B_{-2,0}$. The process weights of mRNA and protein, using Formula (7.7) in each rectangle (or region) are plotted in Figure 7.8. Neglecting the *always inactive processes* we obtain the global reduced model described by Equations (7.15)-(7.16).

$$\frac{d}{dt}M^r = \kappa_2 \frac{\alpha_P^m}{\alpha_P^m + (P^r)^m} - \gamma_M M^r \tag{7.15}$$

$$\frac{d}{dt}P^r = \kappa_3 M^r - \mu P^r . \tag{7.16}$$

The reduced model has the same structure of the one describes by Equations (7.11)-(7.12) with the difference that the first reduced model describes the dynamics of the system starting from a single initial value $\theta^{0,0}$ while the second one describes the dynamics of the system starting from any point of an entire region of initial values $B_{0,0}$ -

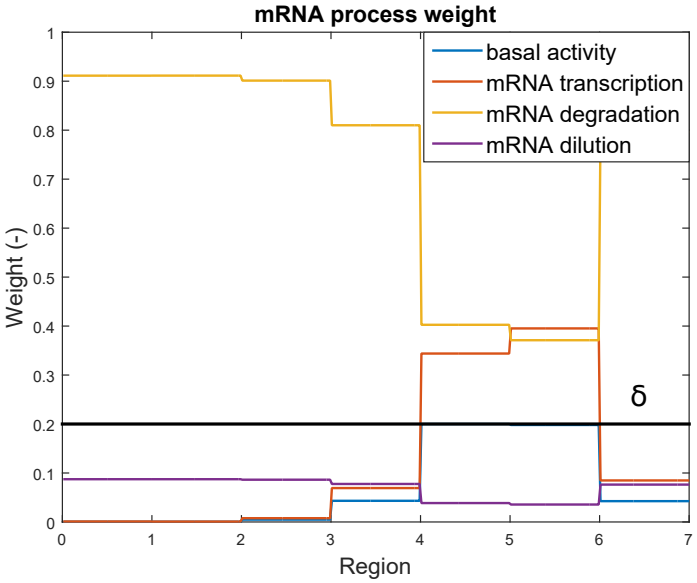
in which the point $\theta^{0,0}$ is included - and follows a pathway till it arrives in the rectangle $B_{-2,0}$ which contains the steady-state of the solutions x of the original System (7.9)-(7.10).
Figure 9 represents a graphic way to obtain quickly the knowledge of the *activity/inactivity* of each process (black means *active* and white *inactive*) in each rectangle:

- regarding mRNA processes it is possible to see that the basal activity is *active* when the mRNA has very low values and the protein has low values; the transcription is *active* only for small concentration of M and high concentration of P while the degradation is always an *active* process, in every rectangle;
- regarding the protein processes it is possible to see that while the degradation process is *always inactive* in every rectangle, the translation process is *active* in the rectangle where the protein has a small concentration and the dilution is *active* when the protein has an high concentration.

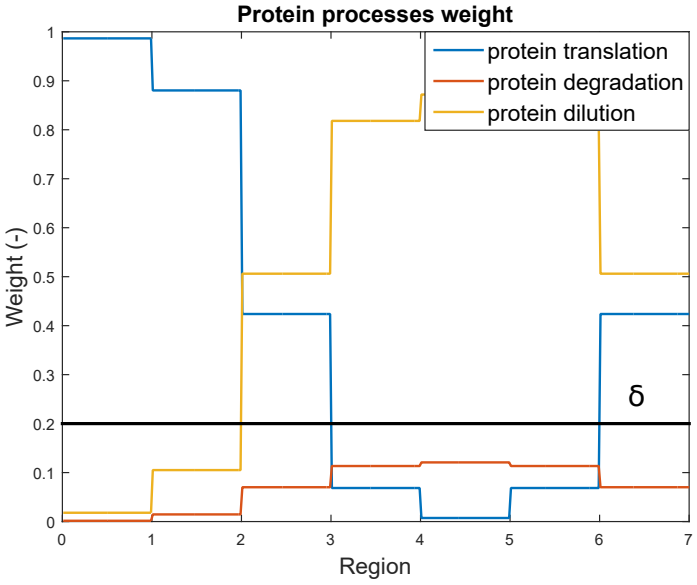
This information is very useful for the biological analysis of the system.

7.6 Conclusion

In this chapter we proved the robustness of our technique in relation with a variation of the initial conditions of the System (7.9)-(7.10). In fact we have obtained a reduced model described by Equations (7.11)-(7.12) applying our method to the original model that had an initial value $x_0 = [1, 1]$ and then we have obtained the same reduced model of Equations (7.15)-(7.16) choosing a space $B_{0,0}$ of initial values that contains $x_0 = [1, 1]$, a range of one order of magnitude in each coordinate and that follows a pathway close to the evolution of System (7.9)-(7.10), starting from x_0 and ending in the steady state point x^*.
Furthermore, in every rectangle $B_{m,n}$ - that represents a different order of magnitude of the system - we obtain a meaningful reduced model in which we can obtain the knowledge of the *activity/inactivity* of each process as we presented in Figure 7.9. The biological interpretation of this table can be very fruitful. We used a grid, in which every boundary differs of one order of magnitude in relation to the previous one: a different grid can be chosen.
At first we have tested the robustness of our method on a model of two dimensions for simplicity reasons and to describe easily the applications: a future work will verify the robustness of our method on models of higher dimensions. Furthermore we can extend our analysis applying the same method as in Section 7.5 to the model parameters. Finally

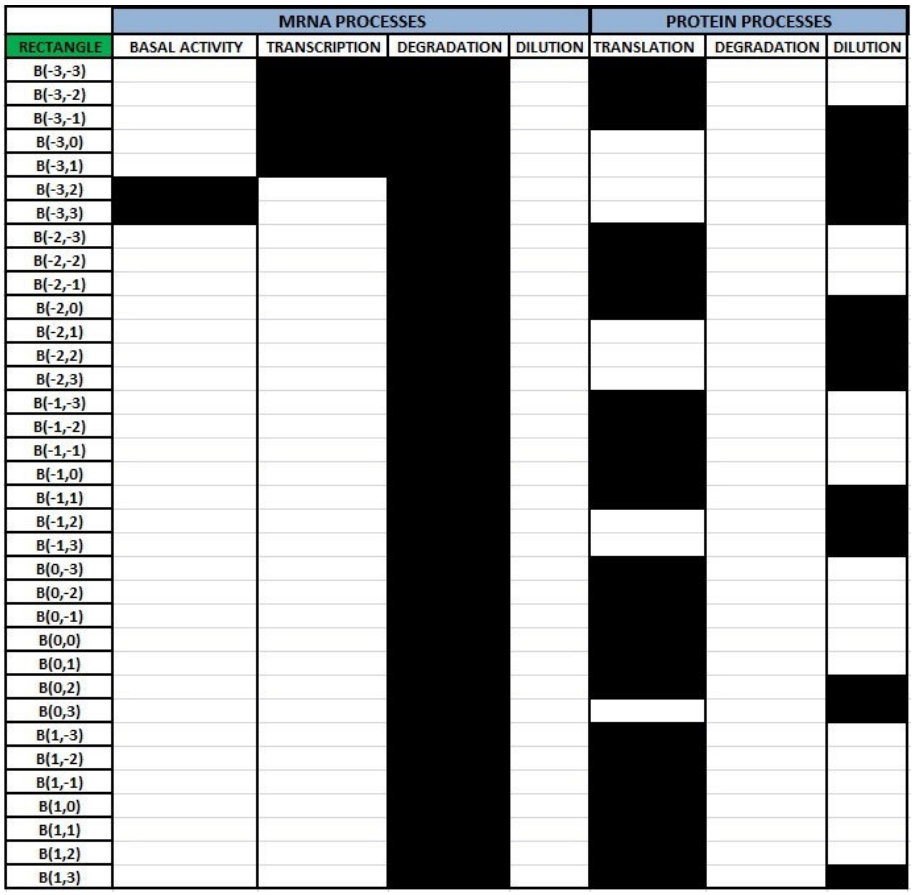


(A)



(B)

FIGURE 7.8: A. The evolution in each rectangle of the mRNA process weights: the basal activity and mRNA dilution are *always inactive* because the dynamic is always under the threshold δ. The regions correspond respectively to the rectangles $B_{0,0}, B_{0,1}, B_{-1,1}, B_{-2,1}, B_{-3,1}, B_{-3,0}, B_{-2,0}$. B. The evolution in each rectangle of the protein process weights: the degradation is *always inactive* because the dynamic is always under the threshold δ. The regions correspond respectively to the rectangles $B_{0,0}, B_{0,1}, B_{-1,1}, B_{-2,1}, B_{-3,1}, B_{-3,0}, B_{-2,0}$.



	MRNA PROCESSES				PROTEIN PROCESSES		
RECTANGLE	BASAL ACTIVITY	TRANSCRIPTION	DEGRADATION	DILUTION	TRANSLATION	DEGRADATION	DILUTION
B(-3,-3)		■	■		■		
B(-3,-2)		■	■		■		
B(-3,-1)		■	■				■
B(-3,0)		■	■				■
B(-3,1)			■				■
B(-3,2)	■		■				■
B(-3,3)	■		■				■
B(-2,-3)			■		■		
B(-2,-2)			■		■		
B(-2,-1)			■		■		
B(-2,0)			■				■
B(-2,1)			■				■
B(-2,2)			■				■
B(-2,3)			■				■
B(-1,-3)			■		■		
B(-1,-2)			■		■		
B(-1,-1)			■		■		
B(-1,0)			■		■		
B(-1,1)			■				■
B(-1,2)			■				■
B(-1,3)			■				■
B(0,-3)			■		■		
B(0,-2)			■		■		
B(0,-1)			■		■		
B(0,0)			■		■		
B(0,1)			■		■		
B(0,2)			■		■		■
B(0,3)			■				■
B(1,-3)			■		■		
B(1,-2)			■		■		
B(1,-1)			■		■		
B(1,0)			■		■		
B(1,1)			■		■		
B(1,2)			■		■		
B(1,3)			■		■		■

FIGURE 7.9: The *activity/inactivity* of every process in each rectangle: black means that the process is *active* in that rectangle, white means that it is *inactive*.

a further method of reduction could be applied: in Section 7.2.3 we consider that the possible reduction is done independently for each component of the vector field. We could also consider a more global reduction on the sum of the components.

Chapter 8

Principal process analysis applied to a model of endocrine toxicity induced by Fluopyram

This chapter is confidential. The results can not be used without the express written consent by the authors.

In this chapter we discuss an ongoing work about the application of *principal process analysis* to a deterministic model that describes the toxicological effect of a fungicide, called Fluopyram, in rodents (produced by Bayer). The work is in collaboration with David Rouquié, senior researcher at the toxicology research center of Bayer CropScience, and with Frédéric Dayan, ExactCure founder and former R&D team leader at Dassault Systèmes. The system was modeled in 2014 by a Bayer CropScience intern, Benjamin Miraglio, under the supervision of David Rouquié and Frédéric Dayan. In this chapter we present the results we have so far and the future applications are detailed in Section 8.7. This work will be a part of a future journal paper.

8.1 Introduction

Fluopyram is a broad spectrum fungicide developed by Bayer CropScience for the control of fungi such as white mold, black dot and botrytis. It inhibits the succinate dehydrogenase (complex II) within the fungal mithocondrial respiratory chain. This compound was shown to be a weak inducer of thyroid follicular cell tumors in male mice following life-time exposure [96].

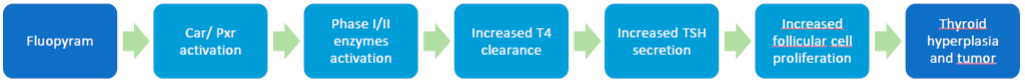


FIGURE 8.1: **Proposed mode of action (MoA) for Fluopyram**. The liver-mediated thyroid toxicity mode of action of Fluopyram. Picture taken from [79]

It was important to investigate the characteristics of this toxicological effect in order to market the new molecule: being able to predict this adverse effect earlier was really important for the cost of development process of the product.

The investigation was oriented to identify the mode of action (MoA) of thyroid toxicity: a MoA is constituted of a series of key events inducing cancer or other adverse effects (the conceptual framework for evaluating an animal mode of action for chemical carcinogenesis can be found in [111]).

The proposed MoA for Fluopyram underlines that the fungicide does not cause directly the thyroid cell proliferation: its MoA consists of an initial effect on the liver by activating the constitutive androstane (Car) and pregnane (Pxr) nuclear receptors causing increased elimination of thyroid hormones followed by an increased secretion of thyroid stimulating hormones (TSH). This change in TSH secretion results in an increase of thyroid follicular cell (TFC) proliferation which leads to hyperplasia and eventually adenomas after chronic exposure [79] (see Figure 8.1).

Furthermore, to obtain the full use of the fungicide in the North American market, it was necessary to demonstrate that the carcinogenic effect was induced by a threshold mechanism: that means that a no-observed adverse effect level (NOAEL), the highest experimental point that is without adverse effect, can be identified and then an acceptable daily intake (ADI) can be inferred. In general, $ADI = \frac{NOAEL}{100}$ to account for the differences between test animals and humans (factor of 10) and possible differences in sensitivity between humans (another factor of 10) [59, pp.92-94].

Once every threshold is determined for every key event of MoA, temporal and dose-response must be assessed: temporal concordance means that every event X must happen before event Y if the event X influences the event Y in the MoA of a given toxicity; the dose-reponse concordance is the proportionality that must be observed between the dose of the studied compound and the key events [79].

In order to prove the defined MoA and the threshold-dependent toxicity of Fluopyram, in [96] were run a set of mechanistic studies on male mice that showed the validity of the dose and the temporal concordance of the specific key events and, using Car/Pxr *knock out* mice it was confirmed that the activation of Car and Pxr is the initial molecular event and the thyroid effects were secondary to increased metabolism and elimination of

thyroid hormones: that means that Fluopyram thyroid toxicity is mediated by activation of hepatic Car/Pxr receptors. Furthermore, NOELs could be identified for each of key events, which provided evidence that Fluopyram acts through a threshold-dependent MoA.

Lately an important amount of data generated in mechanistic toxicity studies were used to design and calibrate a deterministic model able to quantitatively predict the evolution of the thyroid follicular proliferation given by the toxicological effect of Fluopyram. Because *knock out* experiments are very expensive and time-consuming, *in silico* modeling could bring an extra-argument for regulatory authorities. In this context, *principal process analysis* (PPA) could be an additional proof of the temporal ranking/hierarchy of key events: in fact through PPA we can verify if the temporal order of the *activation* of processes maintains the same sequence of the temporal concordance of the MoA of Fluopyram. If *in silico* results are in agreement with MoA hypothesis and if those results are robust, then this analysis can save a lot of economic resources dedicated to product marketing.

8.2 Methodology

In the previous applications with PPA, we associated a dynamical relative weight to each process in order to compare them: if they were over a fixed threshold they were declared *active* over a fixed threshold and *inactive* otherwise. In this chapter, application of PPA is oriented towards system analysis and parameter setting. We thus modify our usual approach and use abolute values as criteria to compare the processes during the system dynamics. A threshold depending of the values of the absolute values of the processes is set for every variable. We call this new methodology *Absolute Principal Process Analysis* (APPA).

8.2.1 Absolute principal process analysis

Considering the ODE System (8.1) that models a biological network:

$$\dot{x} = f(x, p) \tag{8.1}$$

where $x = (x_1, x_2, \ldots, x_n) \; \epsilon \; R^n$ is the vector of concentration of components, $x0 = (x0_1, x0_2, ..., x0_n) \; \epsilon \; R^n$ is the vector of their initial values and $p \; \epsilon \; R^b$ is the vectors of parameters. It is possible to decompose each equation into a sum of biological processes:

$$\dot{x}_i = \sum_j f_{ij}\,(x, p) \tag{8.2}$$

where f_{ij} represents the j^{th} process involved in the dynamical evolution of the i^{th} variable of the system over a period of time [0,T]. In order to weigh the influence of the different processes f_{ij} in the time evolution of each variable x_i, we use the dynamic of their absolute values as a criterion to compare them:

$$A(t,p)_{ij} = |f(t)_{ij}(x,p)|. \tag{8.3}$$

Then for each variable we choose a dynamical threshold:

$$\delta_i = \min_j \min_{t\epsilon[0,T]} |f_{ij}| + \frac{\max_j \max_{t\epsilon[0,T]} |f_{ij}| - \min_j \min_{t\epsilon[0,T]} |f_{ij}|}{c} \tag{8.4}$$

where $\min_j \min_{t\epsilon[0,T]} |f_{ij}|$ is the lowest of the local minima of the j absolute values of the processes in $[0,T]$ for the variable i, $\max_j \max_{t\epsilon[0,T]} |f_{ij}|$ is the highest of the local maxima of the j absolute values of the processes in $[0,T]$ for the variable i and c is a parameter of our choice ($c > 1$ and the higher is c the lower is the threshold). For this work we set $c = 5$. Therefore δ_i is depending on the values of the processes, on the time of the simulation and on an arbitrary parameter.

Definition: *Let the continuous function $f_{ij}(x(t),p)$ be the j^{th} process of $\dot{x}_i(t)$ in $t \epsilon [0,T]$ and let the threshold δ_i be the threshold associated to the variable x_i*
We call a process $f_{ij}(x(t),p)$ always inactive when $A_{ij}(t,p) < \delta_i \ \forall \ t \epsilon$ [0,T].
We call a process $f_{ij}(x(t),p)$ inactive at time t when $A_{ij}(t,p) < \delta_i$.
We call a process $f_{ij}(x(t),p)$ active at time t when $A_{ij}(t,p) \geq \delta_i$.
The switching time for a process $f_{ij}(x(t),p)$ is the time t_{ij}^s when $A_{ij}(t,p) = \delta$. A process can have $s = 0,1,...,z$ switching times.
The switching time set S_i for the i^{th} variable contains all the switching times t_{ij}^s where $j = 1,..,k$ and $s = 1,...,z$.
The global switching time set S is the union of all S_i.

8.2.2 Visualization of the process activity

An important aspect of APPA is to qualitatively visualize, with the help of graphical tools, the *activity/inactivity* of the processes during the system dynamics, with the advantage to summarize this information in one picture. In this chapter we apply two visualization tools.
Temporal Process Map: it allows to visualize the temporal *activity* of the absolute values of the processes (including their switching times), ordered by variable during the whole system dynamics $[0,T]$. Every process bar is in black, resp. white, color when the

respective process is *active*, resp. *inactive*.
Heat Process Map: it allows to study qualitatively the evolution of the intensity of the *active* process *activity* using colors. Values along the rows (the absolute process values $A_{i,j}(t,p)$) are standardized

$$SA_{i,j}(t,p) = \frac{A_{i,j}(t,p) - \bar{A}_{i,j}}{\sigma_{Ai,j}}, \tag{8.5}$$

where $\bar{A}_{i,j}$ is the mean of all the values of the processes $SA_{i,j}(t,p)$ in the time window $[0,T]$ and $\sigma_{Ai,j}$ its standard deviation. The standardized absolute processes assume a red color (*active*) if their value at a generic instant of time $t = \tau\epsilon[0,T]$ is above the mean, black if their value is equal to the mean and green (*inactive*) if their value is below the mean of a column across all rows (the mean of all the standardized processes $SA_{i,j}(t,p)$ at time τ). If the red (or green) color is lighter it means that the process is more *active* (*inactive*).

8.3 Hierarchical graph

This tool is not a part of PPA but it helps understanding the relationship between variables and, in our case, which variables can be affected by a parameter change (for details, see [86, Chapter 3]). This method consists in a decomposition of an interaction graph (graph that represents the structure of the Jacobian matrix) of the model considering the strongly connected components.
If we consider an oriented graph G, a strongly connected component C of a graph G is a maximal sub-set of vertices such that any two of them are connected by a path:

- if $a \epsilon C$, so $\forall\ y \epsilon C$, it exists a circuit containing a and y,
- if $a \epsilon C$, so $\forall\ z \epsilon G \backslash C$, it doesn't exist a circuit containing a and z,

A path is a sequence $(a_0, a_1, \ldots, a_{n-1}, a_n)$ of vertices of G such that any two consecutive vertices are connected by an arc G. a_0 and a_n are respectively the origin and the end of the path. A circuit is a path where the origin and the end are identical. It is possible to create a hierarchy in a graph, re-organizing the graph by level. The higher level includes the components that are not influenced by the others. The lower level are only influenced by the higher levels. The lowest level includes all the components that do not have any influence on the other components.

8.4 Model

For more information about the model see [79].

The pharmacokinetics of the fungicide and its effect in the mouse bodies are described through a multicompartimental model. This model integrates the different markers measured by Bayer CropScience. The model units are nanomoles, liters and hours. Every variable has a different notation depending on the compartment in which it is contained: 'b' means that the variable is in the blood compartment, 'l' in the liver, 'h' in the brain and 't' in the thyroid.

The full model is shown in Figure 8.2.

8.4.1 Blood compartment

Because the mice were orally exposed to Fluopyram, the daily intake is considered as linear. The degradation of the fungicide in the blood is a linear function. Fx denotes the concentration of Fluopyram.

$$\frac{dFx_b}{dt} = k_{1r} - k_{1f}\,Fx_b + k_{2r}\,Fx_l - k_{2f}\,Fx_b \tag{8.6}$$

In the blood compartment are also present the thyroid hormones, triiodothyronine (T3) and its prohormone, thyroxine (T4): they are tyrosine-based hormones produced by the thyroid gland that are primarily responsible for regulation of metabolism. In this model the effects of T3 and T4 are considered as one effect by only considering T4. In this model, the variable $T4$ is the concentration of triiodothyronine and thyroxine.

$$\frac{dT4_b}{dt} = k_{12r}\,T4_l - k_{12f}\,T4_b + k_{13f}\,T4_h - k_{13f}\,T4_b + k_{21r}\,T4_t - k_{21f}\,T4_b \tag{8.7}$$

Finally in the blood there is the thyroid-stimulating hormone (TSH) that is a pituitary hormone that stimulates the thyroid gland to produce thyroxine (T4), and then triiodothyronine (T3). The variable TSH denotes the concentration of TSH.

$$\frac{dTSH_b}{dt} = k_{19r}\,TSH_h - k_{19f}\,TSH_b + k_{20r}\,TSH_t - k_{20r}\,TSH_b - k_{23}\,TSH_b \tag{8.8}$$

8.4.2 Liver compartment

Once in the liver, Fluopyram binds reversibly to hepatic CAR/PXR receptors. The constitutive androstane receptor (CAR) and the pregnane X receptor (PXR) function

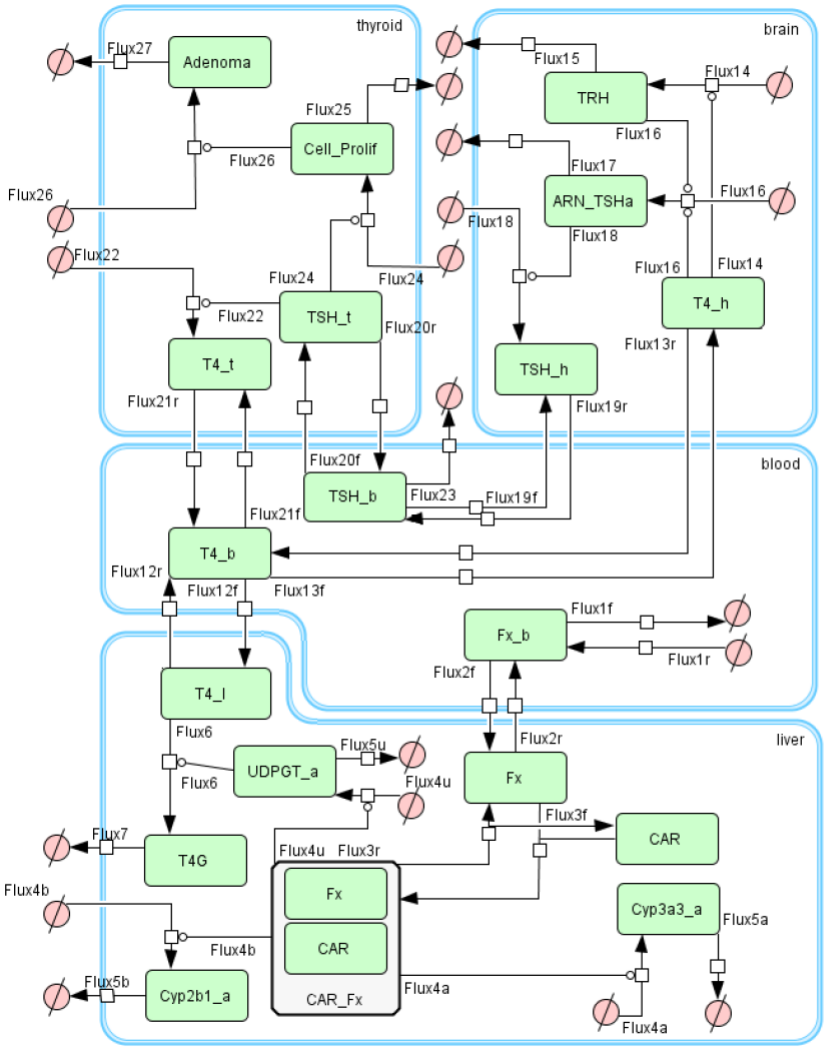


FIGURE 8.2: **Fluopyram MoA Model built by Benjamin Miraglio**. The picture was made by using *CellDesigner*. The different compartments are represented by the blue boxes; the different species are in green with the reactions represented by arrows. The arrows pointing to or from an "empty space" symbol indicate respectively the degradation or the creation of species. The name of the fluxes corresponds to the subscript of coefficients located in the different equations. Figure taken from [79].

as a sensor of endobiotic and xenobiotic substances. In response, expression of proteins responsible for the metabolism and excretion of these substances is upregulated. Hence, CAR and PXR play a major role in the detoxification of foreign substances such as Fluopyram. In the model these two receptors are simplified in one entity concentration CAR and the concentration of the complex between CAR and Fx_l in the liver is label as CAR_Fx_l.

$$\frac{dFx_l}{dt} = k_{2f}\, Fx_b - k_{2r}\, Fx_l - k_{3f}\, Fx_l\, CAR + k_{3r}\, CAR\_Fx_l \tag{8.9}$$

$$\frac{dCAR\_Fx_l}{dt} = k_{3f}\, Fx_l\, CAR - k_{3r}\, CAR\_Fx_l \tag{8.10}$$

$$\frac{dCAR}{dt} = k_{3r}\, CAR\_Fx_l - k_{3f}\, Fx_l\, CAR \tag{8.11}$$

The complex CAR_Fx_l stimulates the phase I and phase II enzymes activity for drug metabolism.

The enzymes of phase I in this model are the Cyp2b and Cyp3a: the activity of the first is studied through the microsomal pentoxyresorufin-O-dependentylation (PROD) and the activity of the second is studied throughthe O-debenzylation of benzyloxyquinoline (BQ). The enzyme of phase II is called uridine 5'-diphospho-glucuronosyltransferase (UDPGT) and its activity was measured using T4 as substrate.

$$\frac{dCyp3a_A}{dt} = Vmax_{4a}\, \frac{(CAR\_Fx_l)^{n_{4a}} + b_{4a}\, k_{4a}^{n_{4a}}}{(CAR\_Fx_l)^{n_{4a}} + k_{4a}^{n_{4a}}} - k_{5a}\, Cyp3a_A \tag{8.12}$$

$$\frac{dCyp2b_A}{dt} = Vmax_{4b}\, \frac{(CAR\_Fx_l)^{n_{4b}} + b_{4b}\, k_{4b}^{n_{4b}}}{(CAR\_Fx_l)^{n_{4b}} + k_{4b}^{n_{4b}}} - k_{5b}\, Cyp3a_A \tag{8.13}$$

$$\frac{dUDPGT_A}{dt} = Vmax_{4u}\, \frac{(CAR\_Fx_l)^{n_{4u}} + b_{4u}\, k_{4u}^{n_{4u}}}{(CAR\_Fx_l)^{n_{4u}} + k_{4u}^{n_{4u}}} - k_{5u}\, UDPGT_A \tag{8.14}$$

Although the enzymatic activities were measured on material extracted from a *in vivo* study, the measurement were performed *in vitro* and not directly in the organism, modifying some factors, such as the concentration of the enzyme. To take account of this, a supplementary proportionality factor was thus applied between the concentration of the enzyme in the model (the absolute one, A) and the concentration measured in the experimentation (the relative one, R): $Cyp3a_R = C_a \cdot Cyp3a_A$, $Cyp2b_R = C_b \cdot Cyp2b_A$, $UDPGT_R = C_u \cdot UDPGT_A$.

In the liver the glucuronidation of $T4$ is performed by $UDPGT$ enzyme: in this form the $T4$ degrades and can be expelled by the organism (we denote the concentration of this form with $T4G$).

$$\frac{dT4_l}{dt} = k_{12f}\, T4_b - k_{12r}\, T4_l - kcat_6\, UDPTG_A\, \frac{T4_l}{Km_6 + T4_l} \tag{8.15}$$

$$\frac{dT4G}{dt} = kcat_6\, UDPGT_A\, \frac{T4_l}{Km_6 + T4_l} - k_7\, T4G \tag{8.16}$$

8.4.3 Brain compartment

The brain compartment regroups both hypothalamus and pituitary entities in the same compartment. The exchange of $T4$ with the blood is expressed by:

$$\frac{dT4_h}{dt} = k_{13f}\, T4_b - k_{13r}\, T4_h \tag{8.17}$$

TRH is a releasing hormone, produced by the hypothalamus, that stimulates the release of TSH: the presence of T4 has an influence on TRH/TSH axis. TRH is produced at its maximum rate without the presence of $T4_h$ that inhibits its production. TRH variable denotes its concentration.

$$\frac{dTRH}{dt} = Vmax_{14}\, \frac{k_{14}^{n_{14}}}{k_{14}^{n_{14}} + T4_h^{n_{14}}} \tag{8.18}$$

The mRNA of TSH is present in the equations because it is one of the marker measured in the mechanistic experinments in [96]. The concentration of mRNA of TSH is denoted as $mRNA_{TSH}$.

$$\frac{dmRNA_{TSH}}{dt} = Vmax_{16}\, \frac{k_{16b}^{n_{16b}}}{k_{16b}^{n_{16b}} + T4_h^{n_{16b}}}\, \frac{TRH^{n_{16a}}}{k_{16a}^{n_{16a}} + TRH^{n_{16a}}} - k_{17}\, mRNA_{TSH} \tag{8.19}$$

The translation of mRNA of TSH increases with the presence of TSH and decreases with the presence of T_4. The concentration of TSH protein (TSH) is:

$$\frac{TSH_h}{dt} = k_{18}\, mRNA_{TSH} + k_{19f}\, TSH_b - k_{19r}\, TSH_h \tag{8.20}$$

8.4.4 Thyroid compartment

The thyroid hosts the follicular cell proliferation by Fluopyram, that eventually leads to the production of Adenoma. In the compartment there is also the presence of T_4 whose production is stimulated by TSH.

(8.21)

$$\frac{dTSH_t}{dt} = k_{20f}\, TSH_b - k_{20r}\, TSH_t \tag{8.22}$$

The presence of TSH in the thyroid can induce follicular proliferation (variable CP).

$$\frac{dCP}{dt} = Vmax_{24}\, \frac{TSH_t^{n_{24}}}{k_{24}^{n_{24}} + TSH_t^{n_{24}}} - k_{25}\, CP \tag{8.23}$$

TABLE 8.1: Parameter values of the model

k_{1f}	0.034	$Kcat_6$	1743.553114	k_{20r}	1.005482275
k_{1r}	Dose of Fluopyram	$Vmax_6$	65666.87588	k_{21f}	1.236398116
k_{2f}	999460	k_7	31588.85616	k_{21r}	0.2002760808
k_{2r}	0.415897	k_{12f}	80.64227305	n_{22}	0.8689751827
k_{3f}	1e-006	k_{12r}	10307.95894	$Vmax_{22}$	14.1733236
k_{3r}	116.536	k_{13f}	0.302741393	k_{22}	0.1929138594
k_{4a}	410568	k_{13r}	0.6426040942	k_{23}	2.77
b_{4a}	0.0815793	k_{14}	5	k_{24}	0.3533230358
n_{4a}	0.982814	n_{14}	1.027112014	n_{24}	1.597710749
k_{4b}	12213.9	$Vmax_{14}$	22.93543793	$Vmax_{24}$	4162.908906
b_{4b}	0.0220242	k_{15}	8.32	k_{25}	3.965767295
n_{4b}	4.49535	k_{16a}	1.219779758	a_{26}	1e-006
k_{4u}	56434.4	k_{16b}	0.2450688022	b_{26}	0.06976043718
b_{4u}	0.630659	n_{16a}	2.21198226	k_{27}	11.05080028
n_{4u}	5.45077	n_{16b}	0.06145468961	C	10.83538438
$Vmax_{4a}$	53945.1	$Vmax_{16b}$	35.76667853	C_a	9.6375
$Vmax_{4b}$	31.4309	k_{17}	1.164964739	C_b	0.01
$Vmax_{4u}$	124777	k_{18}	0.3011508688	C_u	0.0100004
k_{5a}	5646.85	k_{19f}	1.260950263	km_6	5385.524094
k_{5b}	1e-006	k_{19r}	0.1000043066		
k_{5u}	17061.2	k_{20f}	2.489579679		

8.4.5 The data

The data required to calibrate the model parameters are from Bayer CropScience and from the literature [37]. For more details about the parameter and initial condition estimation of this model, see [79]. The parameters are shown in Table 8.1 and the initial conditions in Table 8.2.

8.4.6 Different experiments in silico

We design two different types of experiments to apply PPA.

The first type of experiment is divided in two consequential time windows:

- 3 days (72 hours) of preparation, where the mouse is not exposed to Fluopyram ($k_{1r} = 0$).
- 28 days (672 hours) where the mouse is exposed to 1500 pm of Fluopyram ($k_{1r} = 75773[\frac{nM}{h}]$).

TABLE 8.2: Initial values of the model

Fx_b	0
$T4_b$	186.535714901588
TSH_b	0.0109393373647131
Fx_l	0
CAR_Fx_l	0
CAR	999953
$Cyp3a_A$	0.779337771754164
$Cyp2b_A$	692240.059366051
$UDPGT_A$	4.61232141015872
$T4_l$	1.45911386914052
$T4G$	6.89547850285214e-05
$T4_h$	87.8800534936859
TRH	0.137857583256398
$mRNA\_TSH$	0.100620544848501
TSH_h	0.440940258740178
$T4_t$	1162.44839786602
TSH_t	0.0270858598724828
C_P	17.0538586985414

The second type of experiment is divided in three consequential time windows:

- 3 days (72 hours) of preparation, where the mouse is not exposed to Fluopyram ($k_{1r} = 0$).
- 28 days (672 hours) where the mouse is exposed of 1500 pm of Fluopyram ($k_{1r} = 75773[\frac{nM}{h}]$).
- 28 days (672 hours) where the mouse recovers after the exposition of Fluopyram ($k_{1r} = 0$).

The parameter k_{1r} is represented by a step function that changes its value during the dynamics of the experiments (see Figure 8.3).

In Figure 8.4a is represented the dynamics of the cellular proliferation in the thyroid for the first type of experiment performed with model parameters of Table 8.1 (experiment 1A). If we perform the second type of experiment with the parameter listed in Table 8.1 (experiment 2A) it is possible to see in Figure 8.4b that the TFC proliferation in the recovery phase is slighty increasing, contrary to what has been observed in the mechanistic *in vivo* experiments in Figure 6 of [96] where the TFC proliferation in the same phase returned to the level of control animals after 28-day recovery. The same goes for the other system variables. We then decided to change the degradation of

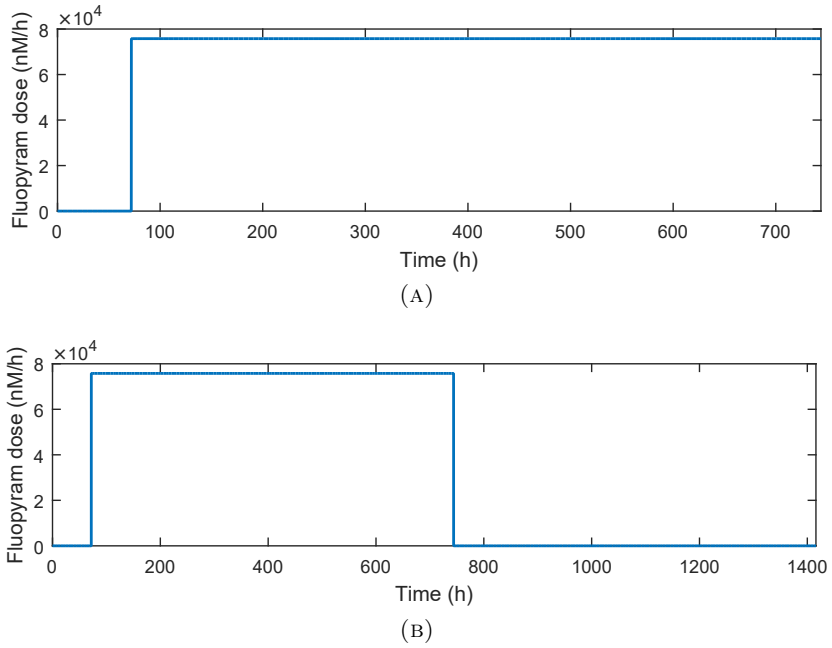
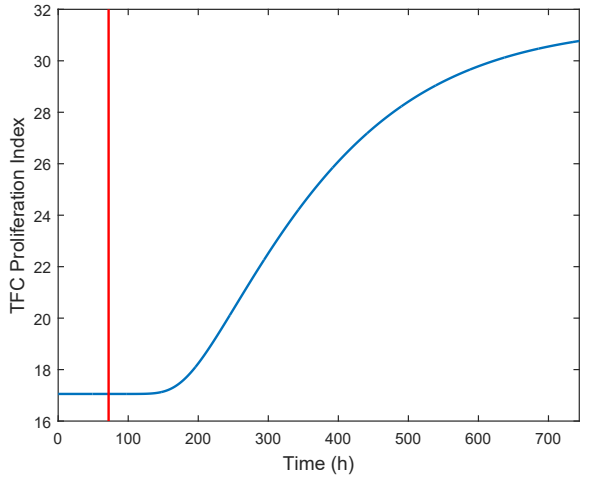


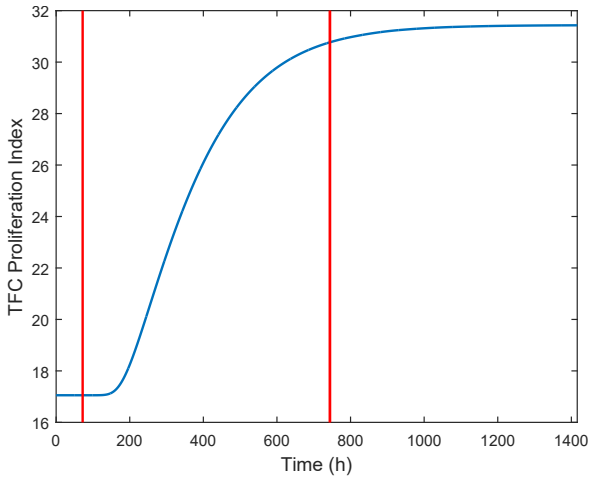
FIGURE 8.3: **Exposition of Fluopyram dose**. A. First type of experiment, B. Second type of experiment.

Fluopyram to $k_{1f} = 8000$ and the degradation of $k_{5a} = 3700$. We obtain these values performing multiple simulations with different values for k_{1f} and k_{5a} and choose the ones that give the most consistent model dynamics with the experimental results in [96], including the recovery phase. In future steps, we plan to verify the biological validity of these values using new data from Bayer CropScience database (for example the half life of Fluorpyram) and to study the influence of these parameters on the final order of the key events. Figure 8.4c shows, for the second type of the experiment, the dynamics of the TFC proliferation with the new parameters (experiment 2B). The dynamics of all the system variables for the different experiments (1A, 2A, 2B) are shown in Appendix E.1. For the sake of clarity, from now on we label the experiment with number 1 if it is of the first type, 2 if it is of the second type, with letter A if the model has the original set of parameters of Table 8.1 and with letter B if the model has the set of parameters of Table 8.1 but with $k_{1f} = 8000$ and $k_{5a} = 3700$.

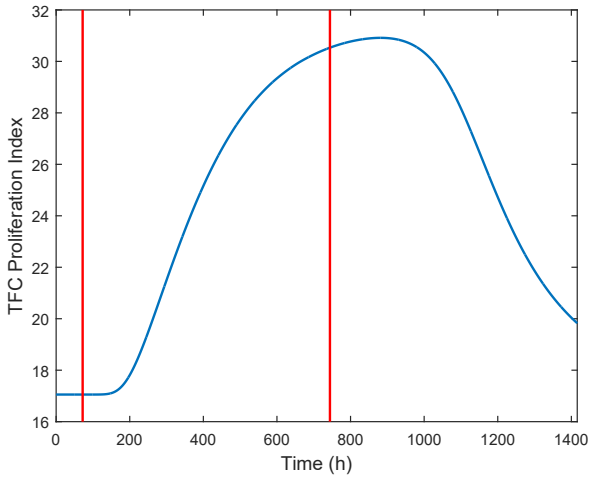
Using the hierarchical graph, it is possible to see that the change of degradation of Fluopyram (k_{1f}) affects the dynamics of the other variables because Fx_b variable is contained in the first level of the graph while the change of the degradation of $Cyp3a$ (k_{5a}) only affects the dynamics of variable $Cyp3a_A$ because the latter is contained in the third level and is not connected to any variable of lower levels (see Figure 8.5). Furthermore it is interesting to observe that the node that contains the variable $UDPGT_A$ works as a switch between the two graphs of level 3: if $UDPGT_A$ reaches a high value $T4_l$ will have low value (and consequently all the other variables contained in that graph) and



(A)



(B)



(C)

FIGURE 8.4: **Exposition to Fluopyram dose**. A. TFC profile in the first type of experiment with model parameters of Table 8.1 (experiment 1A), B . TFC profile in the second type of experiment with model parameters of Table 8.1 (experiment 2A), C . TFC profile in the second type of experiment with model parameters of Table 8.1 but where $k_{1f} = 8000$ and $k_{5a} = 3700$ (experiment 2B).

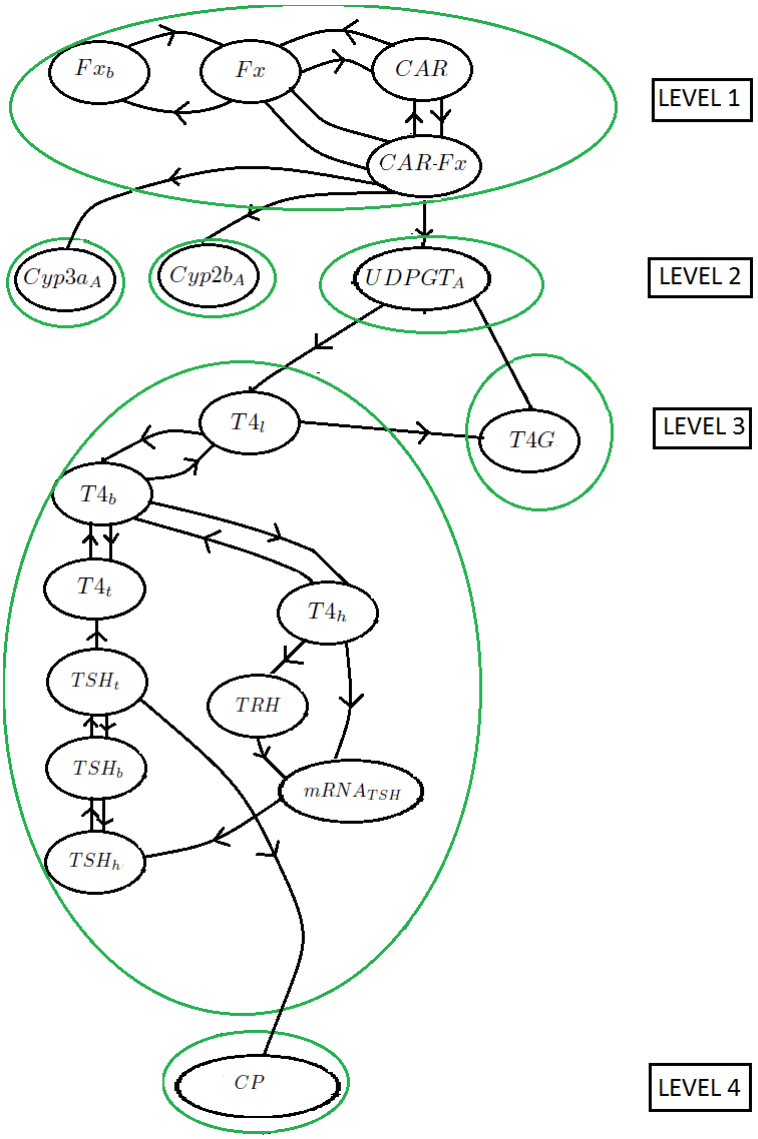


FIGURE 8.5: **Hierarchical graph of the toxicological model divided in four levels.**

$T4G$ a high value; if $UDPGT_A$ reaches a low value we will have the opposite case.

In the following sections we perform PPA on the first type of experiment with the original set of parameters (experiment 1A) to verify if the results obtained with the original parameter set of Benjamin Miraglio are in agreement with MoA hypothesis. In case experiment 1A does not give satisfactory results, we perform PPA on the second type of experiment with the new set of parameters (experiment 2B) to verify, if the modifications applied in experiment 2B give the correct concordance (including the recovery phase) or if further research are needed.

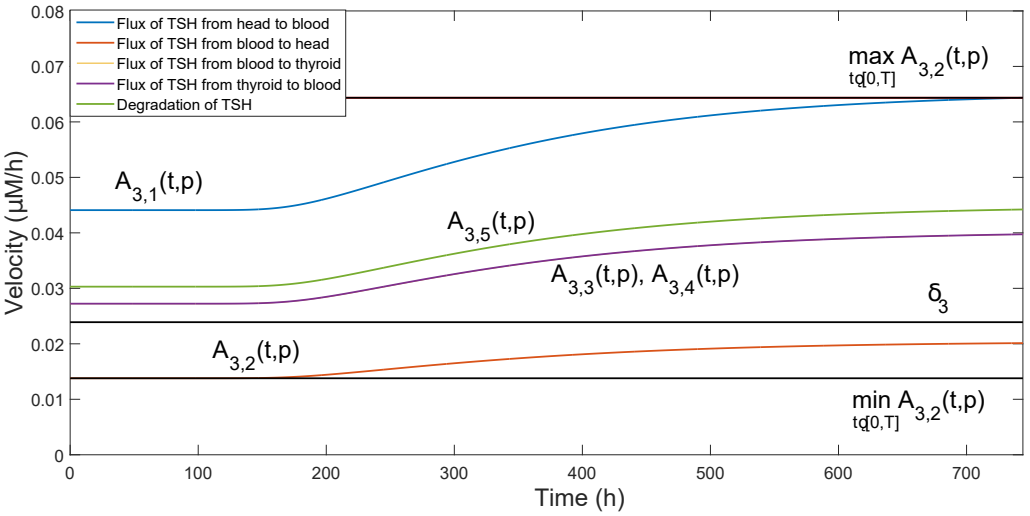


FIGURE 8.6: ***TSH$_b$* processes**. The processes $f_{3,j}(t,p)$ (where $j = 1,\ldots,5$) are compared using their absolute value over time, $A_{3,j}(t,p) = |f_{3,j}(t,p)|$. For every j $A_{3,j}(t,p)$ its maximum and minimum values are calculated ($\max_{t\epsilon[0,T]} A_{3,j}(t,p)$ and $\min_{t\epsilon[0,T]} A_{3,j}(t,p)$), then the higher of the j maximum and the lower of the j minimum are chosen (in this case $\max_{t\epsilon[0,T]} A_{3,1}(t,p)$ and $\min_{t\epsilon[0,T]} A_{3,2}(t,p)$). Using equation 8.4, δ_3 are set. The processes $f_{3,1}(t,p)$, $f_{3,3}(t,p)$, $f_{3,4}(t,p)$, $f_{3,5}(t,p)$ are *active* and $f_{3,2}(t,p)$ is always *inactive*. In this variable system, no processes crosses the threshold, so no switching times are collected in S_3.

TABLE 8.3: Switching times of the model in the experiment 1A (units [h])

$t^1_{1,3}$	209	$t^1_{5,2}$	173	$t^1_{9,2}$	144	$t^1_{14,1}$	217
$t^1_{1,4}$	206	$t^1_{6,1}$	173	$t^1_{11,1}$	141	$t^1_{14,2}$	218
$t^1_{4,1}$	297	$t^1_{6,2}$	173	$t^1_{11,2}$	141	$t^1_{17,1}$	233
$t^1_{4,2}$	299	$t^1_{7,1}$	133	$t^1_{12,1}$	423	$t^1_{17,2}$	234
$t^1_{4,3}$	173	$t^1_{7,2}$	133	$t^1_{12,2}$	425	$t^1_{18,1}$	239
$t^1_{4,4}$	173	$t^1_{8,1}$	88	$t^1_{13,1}$	212	$t^1_{18,2}$	239
$t^1_{5,1}$	173	$t^1_{9,1}$	144	$t^1_{13,2}$	212		

8.5 Absolute principal process analysis on the experiment 1A

We decompose each ordinary differential equation in processes that we considered having a precise biological meaning. Then, we calculated the absolute value of each process using 8.3. For each variable system we calculate the threshold δ_i using 8.4. As an example, Figure 8.6 shows how we compute the threshold δ_3 from the processes of the variable TSH_b.

Table 8.3 contains the switching times t^s_{ij} for all variables.

TABLE 8.4: Switching times of the model in the experiment 2B (units [h])

$t^1_{1,3}$	133	$t^1_{5,2}$	125	$t^1_{9,2}$	155	$t^1_{14,1}$	237
$t^2_{1,3}$	1236	$t^2_{5,2}$	1276	$t^2_{9,2}$	1153	$t^2_{14,1}$	1417
$t^1_{1,4}$	131	$t^1_{6,1}$	125	$t^1_{11,1}$	137	$t^2_{14,2}$	238
$t^2_{1,4}$	1234	$t^2_{6,1}$	1276	$t^2_{11,1}$	1111	$t^2_{14,2}$	
$t^1_{4,1}$	209	$t^1_{6,2}$	125	$t^1_{11,2}$	137	$t^1_{17,1}$	253
$t^2_{4,1}$	1015	$t^2_{6,2}$	1276	$t^2_{11,2}$	1111	$t^2_{17,1}$	
$t^1_{4,2}$	212	$t^1_{7,1}$	92	$t^1_{12,1}$	459	$t^1_{17,2}$	254
$t^2_{4,2}$	1018	$t^2_{7,1}$		$t^2_{12,1}$	1110	$t^2_{17,2}$	
$t^1_{4,3}$	125	$t^1_{7,2}$	92	$t^1_{12,2}$	460	$t^1_{18,1}$	260
$t^2_{4,3}$	1276	$t^2_{7,2}$		$t^2_{12,2}$	1112	$t^2_{18,1}$	
$t^1_{4,4}$	125	$t^1_{8,1}$	88	$t^1_{13,1}$	232	$t^1_{18,2}$	260
$t^2_{4,4}$	1276	$t^2_{8,1}$		$t^2_{13,1}$		$t^2_{18,2}$	
$t^1_{5,1}$	125	$t^1_{9,1}$	155	$t^1_{13,2}$	232		
$t^2_{5,1}$	1276	$t^2_{9,1}$	1153	$t^2_{13,2}$			

Once we obtain them we are able to build the *temporal process map* where we can visually see the *activity/inactivity* of each process.

We can see in Figure 8.7 that the sequence of *activation* of the processes is not following the temporal concordance of MoA schema of Figure 8.1. In Figure 8.8 a *heat process map* is applied only to the *active* processes: because in this map each process absolute value is standardized only by the process itself we do not group the processes by variable. Neither considering every process independent from its system variable and applying the *heat process map* we match the temporal concordance of MoA schema of Figure 8.1. In both cases the presence of Fluopyram in the liver and the formation of the CAR_Fx complex become *active* after the activation of the production of $Cyp3a_A$-$Cyp2b_A$-$UDPTG_A$ and the expression of $UDPTG_A$ happens slightly after the glucuronidation of the tyroxine.

8.6 Absolute principal process analysis on the experiment 2B

We apply the same steps as the first experiment. The switching times are shown in Table 8.4.

In both maps 8.9- 8.10 we can notice better results: the *activation* of Fluopyram happens considerably before with respect to the one in the maps of the experiment 1A and the *activation* time of the production of the complex CAR_Fx is getting closer to the *activation* time of the variables $Cyp3a_A$, $Cyp2b_A$ and $UDPTG_A$. The same happens

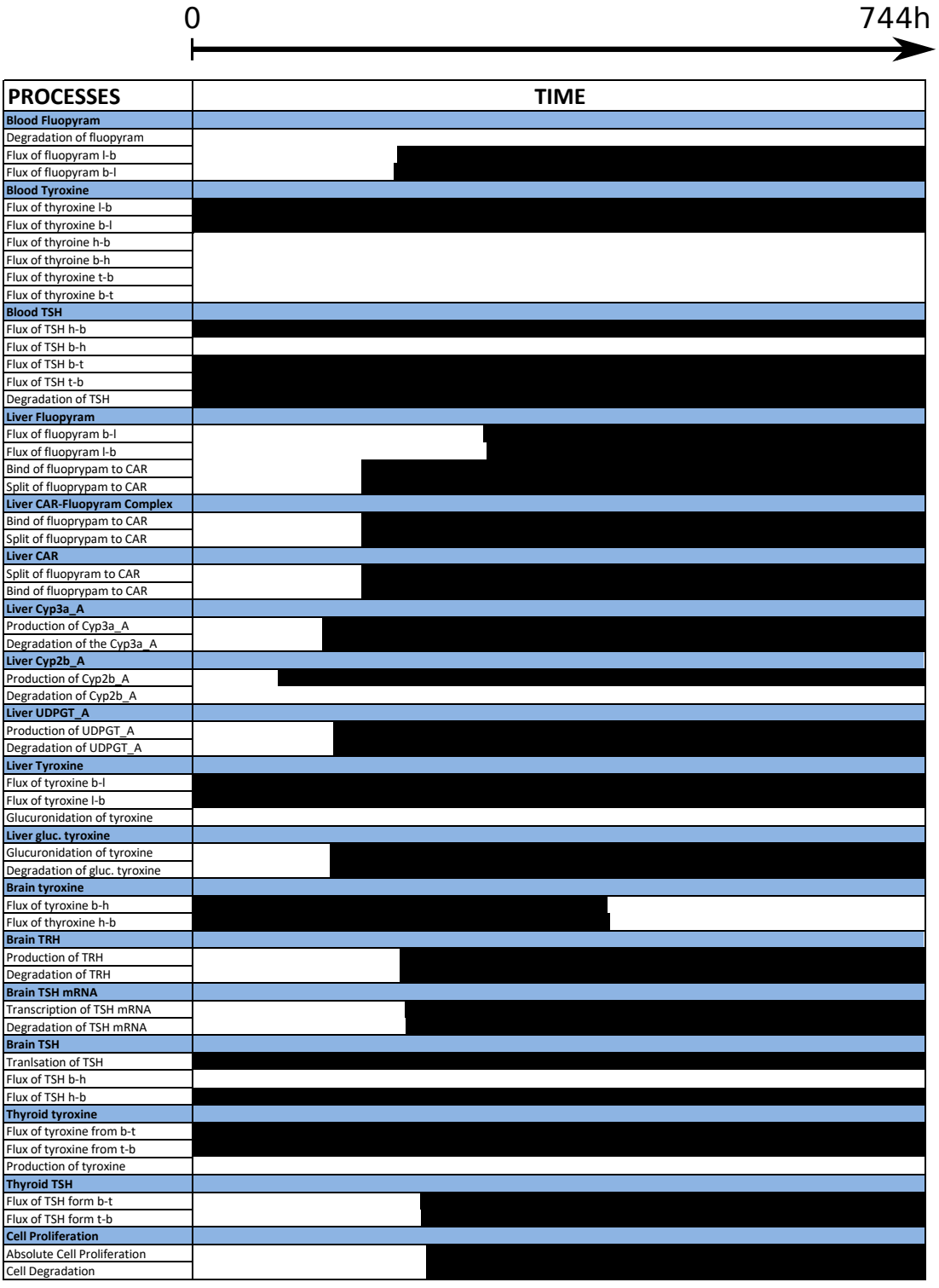


FIGURE 8.7: ***Temporal process map* of the experiment 1A**. *Activity* of the 50 model processes during a 744-hour period. Processes are listed in the first column (white background), ordered by variable (grey background). Their *activity* is depicted in the second column between 0 and 744h: a horizontal black, resp. white, bar when the process is *active*, resp. *inactive*. Values for the *switching times* are given in Table 8.3.

FIGURE 8.8: ***Heat process map* of the experiment 1A**. Because we do not consider the *inactive* processes and we consider the *active* process that appears in more variable systems as the same process (same line), we have in this map 26 processes (during a 744-hour period).

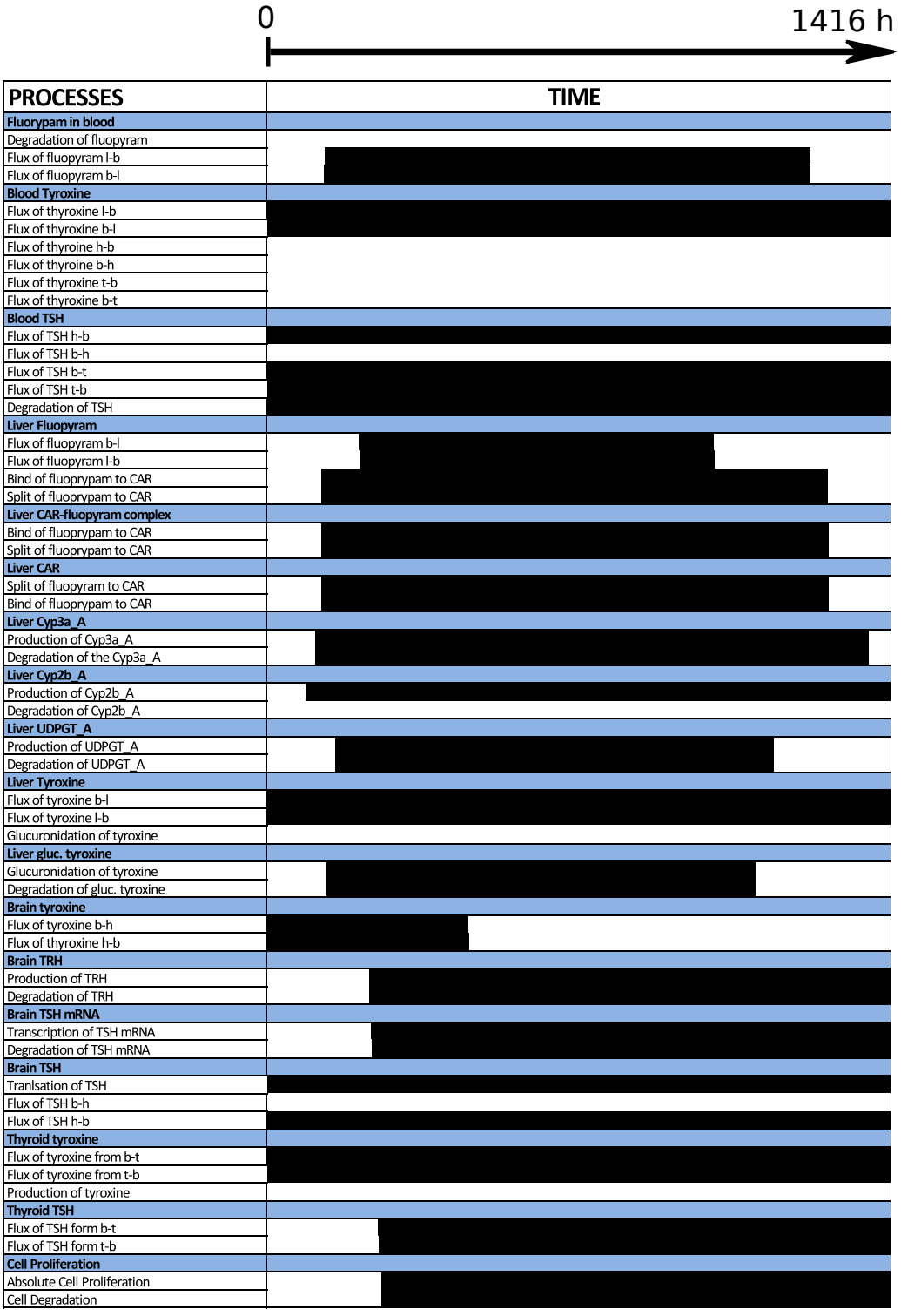


FIGURE 8.9: ***Temporal process map* of the experiment 2B**. *Activity* of the 50 model processes during a 1416-hour period (with recovery phase). Processes are listed in the first column (white background), ordered by variable (grey background). Their *activity* is depicted in the second column between 0 and 1416h: a horizontal black, resp. white, bar when the process is *active*, resp. *inactive*. Values for the *switching times* are given in Table 8.4.

FIGURE 8.10: ***Heat process map* of the experiment experiment 2B**. Because we do not consider the *inactive* processes and we consider the *active* process that appears in more variable systems as the same process (same line), we have in this map 26 processes (during a 1416-hour period).

between the *activation* of the production of $UDPGT_A$ and the *activation* of glucuronidation of tyroxine. This proves that the change of the value of the Fluopyram degradation parameter $K_{1f} = 8000$ changes considerably both the system dynamics in the recovery phase and the temporal order of *activation* of the processes. But still, we are not able to obtain the same sequence between the latter and the temporal concordance of MoA.

8.7 Conclusion and future steps

In conclusion, we have shown that APPA was really useful to understand the core mechanisms of this model and how analyzing the temporal *activation* of processes was important for checking the consistency of toxicological models. In this work we started the first steps to get to the temporal *activation* of processes (using PPA) and to the temporal concordance of the key events of the proposed MoA the same order, giving interesting results in the system dynamics: in fact, through *temporal process map* and *heath process map* it was possible to see that with the original parameter set the process were not become *active* in the order we expected. Furthermore the recovery phase of the experiment had to fit with respect to experimental data of [96]: then, with a new parameter set, we obtained a more consistent dynamic and also better results for the two maps but we were still not able to obtain the same order between the temporal concordance of key events and the temporal *activation* of processes.

The next steps of this work are many.

We will perform the APPA on experiments 1B and 2A to complete the study on all the possible cases and to verify if there are significant differences between the analyses of experiments 1A and 2B.

Furthermore, because the *temporal process map* and *heat process map* show different results for some processes (for example, the translation of TSH process in Brain TSH variable), it could be interesting to apply also the *temporal process map* to the reduced model (the model without the *inactive* processes) to see if the two maps will show a better match (and to verify if the differences are due to the different approaches or because the first map was applied to the original model and the second one to the reduced model).

After choosing the best method to compare the temporal order of activation of the processes through APPA and the temporal concordance of MoA on the Fluopyram model, we will re-calibrate the model to obtain the same temporal sequence between the two and to see if the system dynamics will further improve. The new calibration not only will involve the parameters k_{5a} and k_{1f} but also other uncertain parameters of the model.

New data from Bayer CropScience and results of the mechanistic experiments performed in [96] during the dosing phase and recovery phase can help us to achieve this goal. Once we will obtain the new parameter set, we will vary through a sensitivity analysis the half-life (tmax/2) of the system variables of interest (Fluopyram, TSH, ...) during the recovery phase to see if the ranking order (the order of activation of the processes) will also vary. This is a useful method to test the robustness of the re-calibrated model.

Chapter 9

Model and control of the gene expression machinery in *E. coli*

This chapter is confidential. The results can not be used without the express written consent by the authors.

This chapter describes my contribution to the ANR project RESET, which aims at arresting and restarting the gene expression machinery of *E. coli*. I have been involved in the development, by my co-supervisor Delphine Ropers, of a model of the gene expression machinery and its analysis by *principal process analysis*.

9.1 Introduction

As we have seen in Chapter 3, biotechnological approaches often rely on the obtention of products of interest through the growth control of *E. coli*. Arresting the growth opens the possibility to channel resources into the production of a desired metabolite, instead of wasting nutrients on biomass production [39, 110]. Different approaches are used to limit growth: for example, the use of antibiotics targeting the transcription or translation machinery or limiting nutrients essential for cell growth [77]. These methodologies however have a number of drawbacks and can be ineffective: the first solution can lead to cell death and the second to cell adjustments of their flux distribution and of their enzyme level to nutrient limitations. The aim of the project is to propose a novel strategy for improving product yield and productivity, focusing on global but reversible changes of the cellular physiology, by controlling the gene expression machinery (GEM). In fact, as we have seen in Chapter 3, RNA Polymerase (RNAP) plays a fundamental role in the synthesis of the other proteins. The idea is to arrest the GEM in a precise and controlled

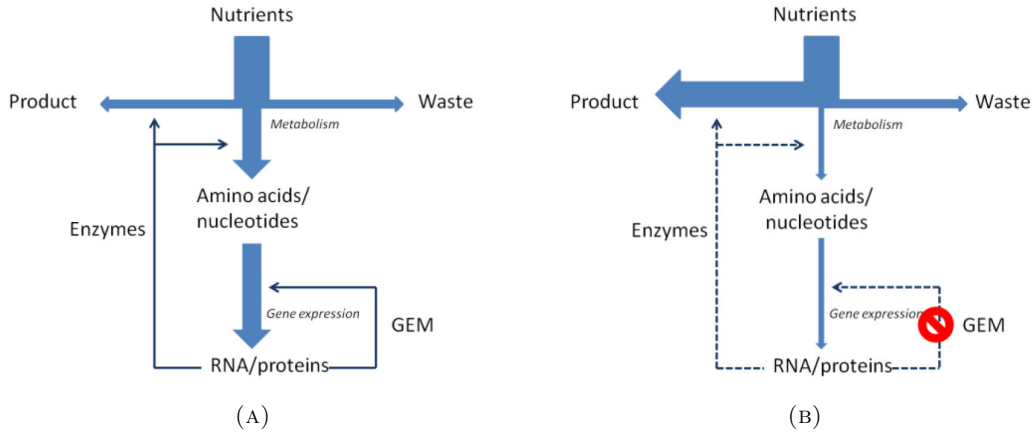


FIGURE 9.1: Outline of approach for improving product yields and productivity, based on the repeated arrest/restart of the GEM. A. Under normal operation, bacterial cells reserve the major part of the incoming nutrient fluxes for the synthesis of building blocks for macromolecules (RNAs, proteins) necessary for growth. B. The arrest of the synthesis of RNA polymerase by removing the external inducer reduces the fluxes towards the synthesis of macromolecules at the profit of fluxes in other product pathways.

way, by externally controlling the expression of two RNAP genes (*rpoB* and *rpoC*) using IPTG as an inducer (see Section 3.4 for details about gene expression control by IPTG): in this way it is possible to create non-growing cells with a functional metabolism that utilizes substrates for the synthesis of specific target compounds rather than for biomass. Contrary to other methods, in which the fluxes in one or the other pathway are favored or disfavored by over-expressing or deleting enzymes, respectively, arresting the GEM completely blocks the demand for other building blocks of protein and RNA synthesis: in this way the enzymes present at the time of growth arrest remain functional, as well as the pathways involved in the synthesis and secretion of a target product. When different factors compromise the cell survival, i.e. the degradation of enzymes and other proteins threatens the stability of metabolism, the GEM can be switched on again, thus altering phases of growth and product synthesis (see Figure 9.1). This control of the gene expression machinery proved to be effective for the production of glycerol at nearly theoretical yields [51]. Although different partners participate to this project, this chapter focus on my contributions to the mathematical modeling aspects of the project.

Mathematical models of GEM and its effect on metabolic fluxes are developed in RESET, in order to understand and optimize the effect of the externally controlled genetic circuits. I worked on two models of the gene expression machinery. The first one describes the functioning of the wild-type GEM, starting from existing studies in the literature [28, 61, 115] and information by decades of work on the cellular physiology of gene expression [29, 34, 40, 93]. This model is described in Section 9.2. In Section 9.3, we present the second model, which extends the previous one with the control of *rpoBC* transcription by IPTG.

9.2 The model

The model, shown in Figure 9.2, can be divided into four principal modules, each assuring a different task in the cell: the ribosome module for the production of stable RNAs, their maturation and assembly with ribosomal proteins into ribosomes; the bulk protein module for the synthesis of cellular proteins; the RNA polymerase module for the production of the RNA polymerase subunits and their assembly into functional RNA polymerase; and the metabolic module involved in the production of pools of amino acids and ppGpp.

ppGpp is an alarmone involved in the stringent response in bacteria, causing the inhibition of ribosome synthesis when there is a shortage of amino acids. This causes translation to decrease while biosynthesis of amino acids is stimulated [113].

The GEM model has been developed by Delphine Ropers using the mass-action law, quasi-equilibrium and quasi-steady-state approximations. It is composed of 13 ordinary differential equations and 5 algebraic expressions, described in Figure 9.3 and 9.4, and 53 parameters calibrated from literature data (e.g. [29]). The state variables correspond to the intracellular concentrations of bulk (b), *rpoBC* (o), and r-protein mRNAs (m), of ribosomal (n) and transfer (T) RNAs, of bulk proteins (B), RNA polymerase $\beta\beta'$ subunits (β), RNA polymerase (P), r-proteins (M), and ribosomes (R). The state variables G, A, and C describe the intracellular concentrations of ppGpp, amino acids, and tRNAs charged with amino-acids, respectively. Five algebraic variables denote the free intracellular concentrations of RNA polymerase (Pf), ribosomes (Rf), amino acids (Af) and ppGpp (Gf), and the specific growth rate (μ). All the variables are expressed in μM with the exception of A and A_f, expressed in M. Mass-balance equations describe the synthesis and degradation of the network components. For instance, the rate of change of bulk mRNA concentration is described as the difference between its transcription rate r_b^t and its consumption through growth dilution at a rate $\mu\, b$ and degradation by RNase E at a rate $e_b\, b$ (see also Figure 9.3):

$$\frac{d}{dt}b = r_b^t(Pf, p) - (\mu + e_b(Rf, B, p))\, b. \tag{9.1}$$

The transcription rate depends on the concentration of free RNA polymerase (see also Figure 9.4)

$$r_b^t = d_b\, k_b\, \frac{Pf}{Pf + K_b}, \tag{9.2}$$

while the expression for bulk mRNA degradation, e_b expresses the fact that RNase E and ribosomes compete for their binding to the messenger RNAs: binding of RNase leads to the mRNA degradation, while the latter is avoided by the ribosome binding

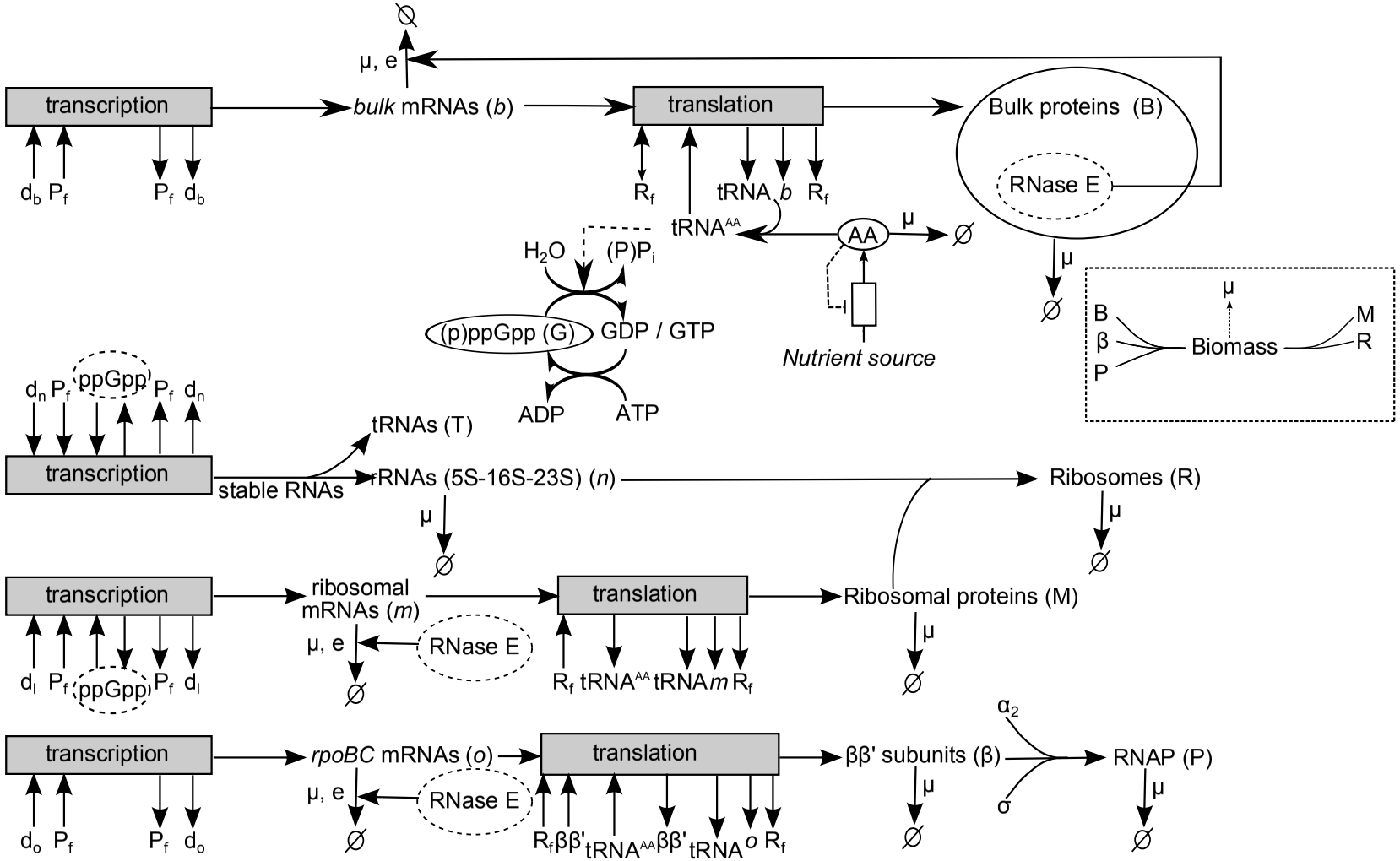


FIGURE 9.2: Simplified representation of the gene expression machinery in *E. coli*. Grey boxes represent the transcription and translation processes. Consumption of network components in various processes are denoted by black arrows, while dashed arrows represent the regulatory effect of a component in a reaction. Intracellular pools are represented by a circle. When they play a role in various parts of the network, they are represented as dashed circles for clarity. Reactions for biomass formation are represented in the box. Abbreviations and notations used: AA - amino acids, $tRNA^{AA}$ - charged tRNAs, *b* - *bulk* mRNAs, *n* - stable RNAs, *o* - *rpoBC* mRNAs, *m* - r-protein mRNAs, *Pf* - free RNA polymerase, *Rf* - free ribosome, d_b, d_n, d_m, d_o - promoters of *bulk*, *rrn*, r-protein, and *rpoBC* genes, respectively. μ: growth rate, e: degradation by ribonucleases (Ropers et al., in preparation).

and the subsequent translation of the mRNA. For the same reason, translation rates thus depend also on the concentration of free ribosomes and RNase E. For instance the translation rate of bulk proteins, r_B^T is (see also Figure 9.4):

$$r_B^T = k_B\, b\, C\, \frac{Rf}{Rf + K_B\left(1 + \frac{B}{KE_B}\right)}. \tag{9.3}$$

$$
\begin{aligned}
\frac{d}{dt}b &= r_b^t(Pf, p) - (\mu + e_b(Rf, B, p))\, b \\
\frac{d}{dt}B &= r_B^T(Rf, b, B, C, p) - \mu\, B \\
\frac{d}{dt}o &= r_o^t(Pf, p) - (\mu + e_o(Rf, \beta, B, p))\, o \\
\frac{d}{dt}\beta &= r_\beta^T(Rf, o, B, \beta, C, p) - (\mu + k_P^m)\, \beta \\
\frac{d}{dt}P &= r_P^m(\beta, p) - \mu\, P \\
\frac{d}{dt}n &= r_n^t(Pf, Gf, p) - (\mu + k_R^m\, M)\, n \\
\frac{d}{dt}m &= r_m^t(Pf, Gf, p) - (\mu + e_m(Rf, B, p))\, m \\
\frac{d}{dt}M &= r_M^T(Rf, m, B, C, p) - (\mu + k_R^m\, n)\, M \\
\frac{d}{dt}R &= r_R^m(M, n, p) - \mu\, R \\
\frac{d}{dt}T &= \rho\, r_n^t(Pf, Gf, p) - \mu\, T \\
\frac{d}{dt}C &= r_C(T, C, Af, p) - J_R - \mu\, C \\
\frac{d}{dt}G &= k_G - \left(\frac{k_{spoT}}{1 + \frac{Pf}{K_g}}\, \frac{K_{CT}^2}{K_{CT}^2 + \left(\frac{C}{T}\right)^2} + \mu\right) G \\
\frac{d}{dt}A &= r_A - \mu\, A \\
Pf &= P - \left(\frac{1}{k_b} + \frac{L_b}{c_b}\right) r_b^t - \left(\frac{1}{k_o} + \frac{L_o}{c_o}\right) r_o^t - \left(\frac{1}{k_n} + \frac{L_n}{c_n}\right) r_n^t - \left(\frac{1}{k_m} + \frac{L_m}{c_m}\right) r_m^t - \frac{Pf\, Gf}{K_g} - d_{ns}\, \frac{Pf}{Pf + K_{ns}} \\
Rf &= R - m\, \frac{Rf}{Rf + K_M\left(1 + \frac{B}{K_E}\right)} \left(1 + k_M\, \frac{L_M - 1}{c_M}\right) - o\, \frac{Rf}{Rf + K_\beta\left(1 + \frac{B}{K_E} + \frac{\beta}{KI_\beta}\right)} \left(1 + k_\beta\, \frac{L_\beta - 1}{c_\beta}\right) \\
&\quad - b\, \frac{Rf}{Rf + K_B\left(1 + \frac{B}{K_E}\right)} \left(1 + k_B\, \frac{L_B - 1}{c_B}\right) \\
Gf &= \frac{G}{1 + \frac{Pf}{K_g}} \\
Af &= A - (C + AB + AM + A\beta) \times 10^{-6} \\
\mu &= \frac{J_R}{AB + AM + A\beta}
\end{aligned}
$$

FIGURE 9.3: Model equations according to the network structure in Figure 9.2.

$$
\begin{aligned}
r_b^t &= d_b\, k_b\, \frac{Pf}{Pf + K_b} \\
r_o^t &= k_o\, d_o\, \frac{Pf}{Pf + K_o} \\
r_n^t &= d_n\, k_n\, \frac{Pf}{Pf(1 + \alpha_n\, Gf) + K_n} \\
r_m^t &= k_m\, d_m\, \frac{Pf}{Pf\,(1 + \alpha_m\, Gf) + K_m} \\
r_B^T &= k_B\, b\, C\, \frac{Rf}{Rf + K_B\, \left(1 + \frac{B}{K_E}\right)} \\
r_\beta^T &= k_\beta\, o\, C\, \frac{Rf}{Rf + K_\beta\, \left(1 + \frac{B}{K_E} + \frac{\beta}{KI_\beta}\right)} \\
r_M^T &= k_M\, m\, C\, \frac{Rf}{Rf + K_M\, \left(1 + \frac{B}{K_E}\right)} \\
r_P^m &= k_P^m\, \beta \\
r_R^m &= k_R^m\, n\, M \\
r_A &= V_m^A\, \frac{{KI_A}^2}{{KI_A}^2 + Af^2} \\
J_R &= L_B \times r_B^T + L_M \times r_M^T + L_\beta \times r_\beta^T \\
e_b &= kE_B\, \frac{B}{B + K_E\, \left(1 + \frac{Rf}{K_B}\right)} \\
e_o &= kE_\beta\, \frac{B}{B + K_E\, \left(1 + \frac{Rf}{K_\beta} + \frac{\beta}{KI_\beta}\right)} \\
e_m &= kE_M\, \frac{B}{B + K_E\, \left(1 + \frac{Rf}{K_M}\right)} \\
r_C &= V_m^C\, \frac{Af}{K_A + Af}\, \frac{T - C}{T - C + K_U} \\
AB &= L_B\, B \\
AM &= L_M\, (M + R) \\
A\beta &= L_\beta\, (\beta + P)
\end{aligned}
$$

FIGURE 9.4: Kinetic rate laws.

9.2.1 Growth rate

In this subsection, we focus on the modeling of the bacterial growth rate, which is my main contribution to the development of the GEM model. The interesting aspect of our approach is that we do not consider the growth rate as a constant or as a Michaelis-Menten equation depending on an input variable, but we deduce it from the total amount of cellular biomass.

We consider that biomass formation (schematically represented in Figure 9.5) results from the accumulation of newly synthesized proteins.

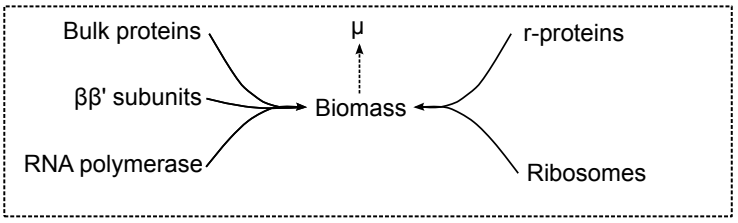


FIGURE 9.5: **Biomass formation**. The accumulation of biomass results from de novo synthesis of protein species (r-proteins alone or within ribosomes, $\beta\beta'$ subunits alone or within RNA polymerase, and Bulk proteins. The growth rate corresponds to the relative rate of biomass accumulation.

Since protein translation results from the incorporation of amino acids into proteins, describing the formation of biomass amounts to keeping track of the mass of amino acids incorporated into cellular proteins, that is:

$$\mathbf{B} = V \times MW_A \cdot 10^{-6} \times At \tag{9.4}$$

where V is the cell volume (expressed in L), MW_A is the molecular weight in g/mol of the amino acids (whose concentration are multiplied by 10^{-6} to convert them into mol·L) and the total concentrations of amino acids $At = AB + AM + A\beta$ (where $AB = L_B\, B$, $AM = L_M\,(M + R)$, $A\beta = L_\beta\,(\beta + P)$). We define the growth rate μ (expressed in min^{-1}) as the relative increase of the cell volume V:

$$\mu = \frac{dV}{dt}\,\frac{1}{V}. \tag{9.5}$$

During the cell growth, the cell density α (expressed in g/L) is considered to be constant, which implies that the biomass $\mathbf{B}$ is proportional to the cell volume V:

$$\mathbf{B} = \alpha\, V. \tag{9.6}$$

We rewrite the expression for the growth rate μ, using Equations (9.4)-(9.5)-(9.6):

$$\mu = \frac{d\mathbf{B}}{dt}\,\frac{1}{\mathbf{B}} = \frac{dAt}{dt}\,\frac{1}{At} + \frac{dV}{dt}\,\frac{1}{V} \tag{9.7}$$

Given the definition of μ in Equation (9.5), we can write the following relation :

$$\frac{dAt}{dt}\,\frac{1}{At} = 0 \Leftrightarrow \frac{dAt}{dt} = 0. \tag{9.8}$$

The time derivative of the concentration of amino acids incorporated into proteins is defined as:

$$\frac{dAt}{dt} = \frac{dAB}{dt} + \frac{dAM}{dt} + \frac{dA\beta}{dt} \tag{9.9}$$

where

$$\frac{dAB}{dt} = L_B\, r_B^T - L_B\, \mu\, AB \tag{9.10}$$

$$\frac{dAM}{dt} = L_M\, r_M^T - L_M\, \mu\, AM \tag{9.11}$$

$$\frac{dA\beta}{dt} = L_\beta\, r_\beta^T - L_\beta\, \mu\, A\beta. \tag{9.12}$$

We can rewrite Equation (9.9) as:

$$\frac{d}{dt}At = L_B\, r_B^T + L_M\, r_M^T + L_\beta\, r_\beta^T - \mu\,(AB + AM + A\beta) = J_R - \mu\,(AB + AM + A\beta). \tag{9.13}$$

Then, the final expression for the growth rate is:

$$\mu = \frac{J_R}{AB + AM + A\beta}. \tag{9.14}$$

The total concentration of amino acids incorporated into proteins (expressed in $\text{mol}\cdot\text{L}^{-1}$) is proportional to the cell density (expressed in $\text{g}\cdot\text{L}^{-1}$) which implies that

$$\alpha = (L_B\, B + L_\beta\,(\beta + P) + L_M\,(M + R)) \times MW_A \cdot 10^{-6}, \tag{9.15}$$

is constant.

Given the expression of μ in Equation (9.14), the relation expressed in Equation (9.15) implies that:

$$L_B\, B + L_\beta\,(\beta + P) + L_M\,(M + R) = constant \tag{9.16}$$

should be verified at all time t.

9.3 The effect of IPTG on *E. coli* growth

The data from the experiments in [51] suggest that the switch between cell growth and growth arrest, after the removal of IPTG from the media, is due to the highly ultrasenstive response of the growth rate to a change in concentration of the $\beta\beta'$ subunits. Different experiments were performed with different IPTG concentrations, varying from 0 to 1000 μM. Two different responses to IPTG were observed: in the first category, for IPTG concentrations of 30 μM and higher, growth is close to normal (compared to a wild-type *E. coli*) while in the second category, for IPTG concentrations of 20 μM or lower, growth stops after few hours. In the second case, the quantity of newly synthesized β and β' subunits is not enough to sustain growth, whereas growth reaches the maximum rate for higher concentrations of IPTG. To measure the intracellular

activity of $\beta\beta'$ subunits a gene coding a fluorescent protein was inserted after *rpoBC* genes. Figure 9.7c shows the concentration of β' once the bacteria have reached steady-state growth. For concentrations above a threshold between 20 and 30 μM we can notice a high cell growth, reaching a saturation effect after IPTG concentration higher than 100 μM. It is possible to notice that the dependence of the growth rate on the β' concentration appears to be highly switch like or *ultrasensitive* in the sense of [62]. The data, fitted with a curve fitted with a Hill function of order 10, indicates that growth requires a threshold level of RNA polymerase to switch from zero to maximal growth.

We want to verify if the proposed GEM system is able to reproduce the same type of response with different concentrations of IPTG as in [51]. Until now, we have considered a wild-type model of *E. coli*, in which the IPTG control on the growth rate of the bacterium is not modeled. To achieve this task, we need to control the transcriptional rate of *rpoBC* mRNA r_o^t (see Figure 9.4). We express the positive effect of IPTG on the rate as follows (for more details see [102]):

$$sr_o^t = \left(\frac{1}{1 + (\frac{KI_I}{I})^{n_I}} \right) k_o \, d_o \, \frac{Pf}{Pf + K_o} \tag{9.17}$$

where I is the concentration of IPTG, $KI_I = 40$ μM is the dissociation rate of IPTG and $n_I = 2.6$ is the Hill coefficient of the response to IPTG, based on the information in [65]. In the absence of IPTG in the culture medium ($I = 0$ μM), there is an arrest of *rpoBC* mRNA transcription ($sr_o^t = 0$). If there is a high concentration of IPTG in the medium ($I = 1000$ μM), the engineered strain has the same transcription rate as in the wild type ($sr_o^t = r_o^t$). Figure 9.6 shows the relationship between the IPTG concentration and the transcriptional activity: the engineered strain in a medium with 100 μM of IPTG has a similar transcription rate to the one in a medium with 1000 μM of IPTG (similar results have been found in [51]).

After the implementation of Equation 9.17, we perform multiple simulations with different values of the IPTG input I, between 0 and 100 μM. We then analyze the relation between the predicted growth rate and the concentration of $\beta\beta'$ when the system reaches a steady state. The plot is shown in Figure 9.7d. With the current calibration the model is able to qualitatively reproduce the hypersensitivity of bacterial growth to $\beta\beta'$ concentration observed in Figure 9.7d.

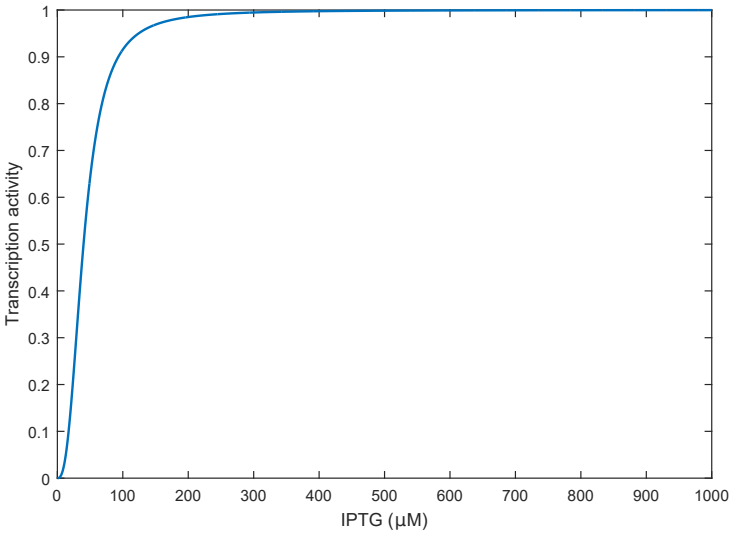


FIGURE 9.6: **IPTG effect**. The response of transcriptional activity of *rpoBC* mRNA to different concentrations of IPTG. The curve represents a Hill function with degree $n_I = 2.6$ and concentration threshold $KI_I = 40\ \mu M$.

9.4 Model analysis with three-level PPA

As we have seen in the previous sections, *E. coli*, under stress conditions as a nutrient downshift or a removal of IPTG, arrests its growth rate. We are interested in identifying the core mechanisms at play when we apply a specific stress condition to the bacterium: in this manner, we can verify if the GEM of the model responds in the expected way.

To analyze the internal mechanisms of the proposed model we apply *principal process analysis* (PPA), using two thresholds to detect not only the *inactive* processes but also to make a further distinction between processes with a moderate *activity* and a full *activity*.

Furthermore we apply PPA both to the differential equations and to the conservation equations of the system.

We refer to Section 6.2.1 and 6.2.2 for the detailed explanation about PPA: in this section we present briefly the important additions we bring to this numerical approach for analyzing the GEM model.

9.4.1 Methodology

Consider the following ODE model of biological network:

$$\dot{x} = f(x, p) \tag{9.18}$$

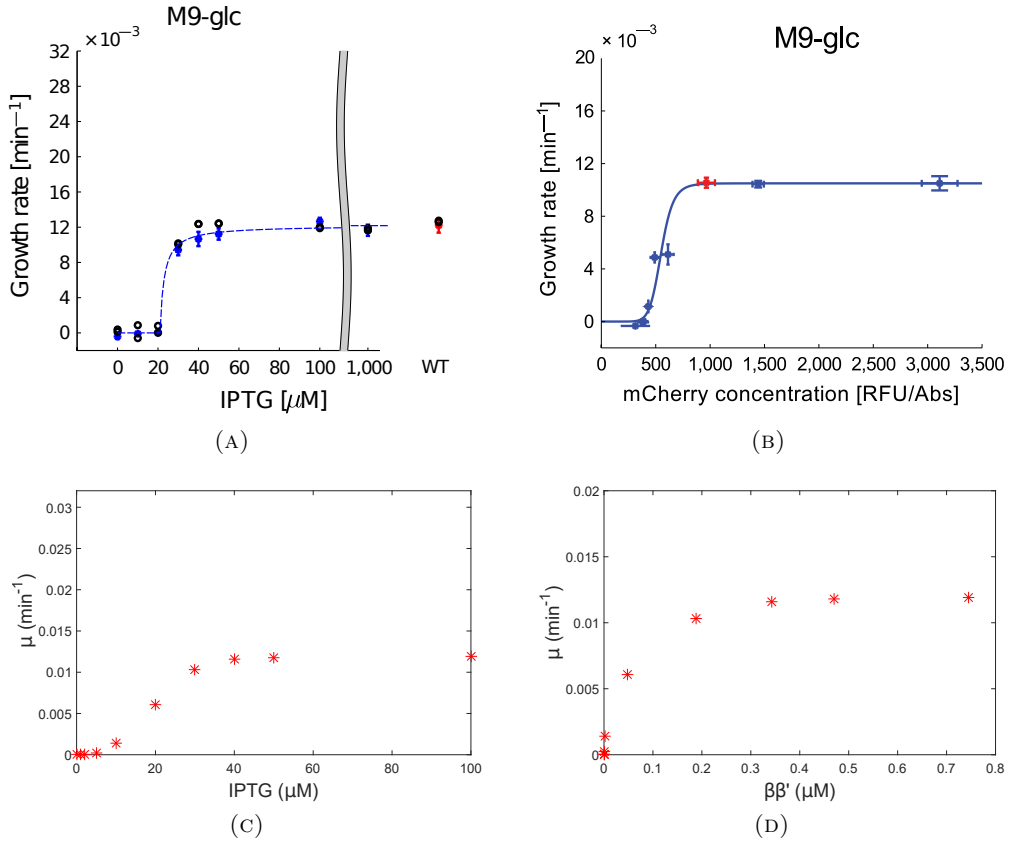


FIGURE 9.7: A. Different steady-state responses of growth rates of the engineered strain for different values of IPTG. The blue points represent the engineered strains while the red points, the wild type. Picture taken from [51]. B. Quantitative dependence of the growth rate on β' concentration in a medium with different IPTG concentrations. The blue points represent the engineered strains while the red point the wild type. Picture taken from [51]. C. Dependence of the growth rate on different IPTG concentrations in the proposed system at its steady state. Simulations performed using different values of IPTG (0, 1, 2, 5, 10, 30, 40, 50, 100 μM). D. Growth rate in function of $\beta\beta'$ subunit concentrations for the different values of IPTG listed in (C) in the proposed system at its steady state.

where $x = (x_1, x_2, ..., x_n)\ \epsilon\ \mathbb{R}^n$ is the vector of component concentrations, $x0 = (x0_1, x0_2, \ldots, x0_n)\ \epsilon\ \mathbb{R}^n$ the vector of their initial values and $p\ \epsilon\ \mathbb{R}^b$ the vector of parameters.

Each equation is decomposed into a sum of biological processes:

$$\dot{x}_i = \sum_j f_{ij}\,(x, p) \tag{9.19}$$

where f_{ij} represents the j^{th} process involved in the dynamical evolution of the i^{th} variable of the system over a period of time [0,T].

Each conservation equation is also decomposed into a sum of relevant concentration terms:

$$H_i = \sum_j h_{ij}\,(x,p) \tag{9.20}$$

where h_{ij} represents the j^{th} relevant term considered in the conservation laws with total concentration H_i over a period of time [0,T].

Comparison criteria are needed to weigh the influence of the different processes f_{ij} on the time evolution of each variable x_i and of the different terms on the dynamical contribution for the total concentration H_i.

In this work we associate a relative weight W_{ij}^d to each process to make it dimensionless:

$$W_{ij}^d(t,p) = \frac{|f_{ij}(x(t),p)|}{\sum_j |f_{ij}(x(t),p)|} \tag{9.21}$$

where $0 \leq W_{ij}^d(t,p) \leq 1$ and $\sum_j W_{ij}^d(t,p) = 1$.

Similarly, we associate a relative weight W_{ij}^a to the terms involved in the conservation laws:

$$W_{ij}^a(t,p) = \frac{|h_{ij}(x(t),p)|}{\sum_j |h_{ij}(x(t),p)|} \tag{9.22}$$

where $0 \leq W_{ij}^a(t,p) \leq 1$ and $\sum_j W_{ij}^a(t,p) = 1$.

Definition: *Let the continuous function $f_{ij}(x(t),p)$ (resp. $h_{ij}(x(t),p)$) be the j^{th} process (resp. term) of $\dot{x}_i(t)$ (resp. $H_i(t)$) in $t \in [0,T]$ and let the thresholds $\delta \in$ [0,1], $\nu \in$ [0,1] with $\delta < \nu$.*
We call a process $f_{ij}(x(t),p)$ (resp. a term $g_{ij}(x(t),p)$) always inactive when $W_{ij}(t,p) < \delta \ \forall \ t \in [0,T]$.
We call a process $f_{ij}(x(t),p)$ (resp. a term $g_{ij}(x(t),p)$) inactive at time t when $W_{ij}(t,p) < \delta$.
We call a process $f_{ij}(x(t),p)$ (resp. a term $g_{ij}(x(t),p)$) moderately active at time t when $\delta \leq W_{ij}(t,p) < \nu$.
We call a process $f_{ij}(x(t),p)$ (resp. a term $g_{ij}(x(t),p)$) fully active at time t when $W_{ij}(t,p) \geq \nu$.
Switching time for a process $f_{ij}(x(t),p)$ is the time t_{ij}^s when $W_{ij}(t,p) = \delta$ or $W_{ij}(t,p) = \nu$. A process can have $0,1,...,z$ switching times.
The switching time set S_i for the i^{th} variable contains all the switching times t_{ij}^s where $j = 1,..,k$ and $s = 1,...,z$.
The global switching time set S is the union of all S_i.

9.4.2 Different applications

We apply three-level PPA on two experiments *in silico*, in which we put *E. coli* cells under two different types of stress conditions.

In the first experiment, the system starts from a steady-state condition of the wild-type model, in which cells grow exponentially in a minimal medium. At time τ_1 we shift the bacterial population in a poorer medium permitting a moderate synthesis of amino acids (nutrient downshift). At time τ_2 we shift cells back to the initial medium (nutrient upshift).

In the second experiment, the system starts from a steady-state condition of the engineered-type model in a medium with a high concentration of IPTG. At time τ_1 we shift the bacterial population in a medium with no IPTG (growth arrest). At time τ_2 we shift it back to the initial medium (growth restart).

We apply the *Boolean Process Map* for all analyses: it shows the time-dependent activity of processes, ordered by variables, during the whole system dynamics $[t_0, T]$. *Fully Active* processes are depicted by a black bar, *moderately active* processes by a grey one and *inactive* processes by a white one. We set the threshold δ at 0.1 and the threshold ν at 0.4.

The parameter values and initial conditions for the model are listed in Appendix F.

9.4.2.1 Nutrient stress condition

The GEM network includes many interlaced feedback loops whose functioning varies with the environmental conditions. We study here their functioning in the case of changing nutritional conditions. Figures 9.8 and 9.9 shows simulation results and the *boolean process map* for this case. The first red line indicates the time (minute 200) of the transfer of bacterial cells to a poorer medium supporting a lower growth rate (nutrient downshift), while the second line indicates the time (minute 1200) at which cells are transferred back to the richer medium (nutrient upshift).

We notice that, in the first 200 minutes of the experiment, the degradation factor is the main reason for the decrease of the concentration of mRNAs and ppGpp (*fully active*) while the dilution is negligible (*inactive*). Also for rRNAs and r-proteins the dilution factor is not an important process: they are both mainly consumed through their assembly into new ribosomes. The dilution of charged tRNAs is not an important process as well since they are essentially consumed by the protein synthesis. The RNA polymerase is present equally in three forms at this stage: the free form, the form

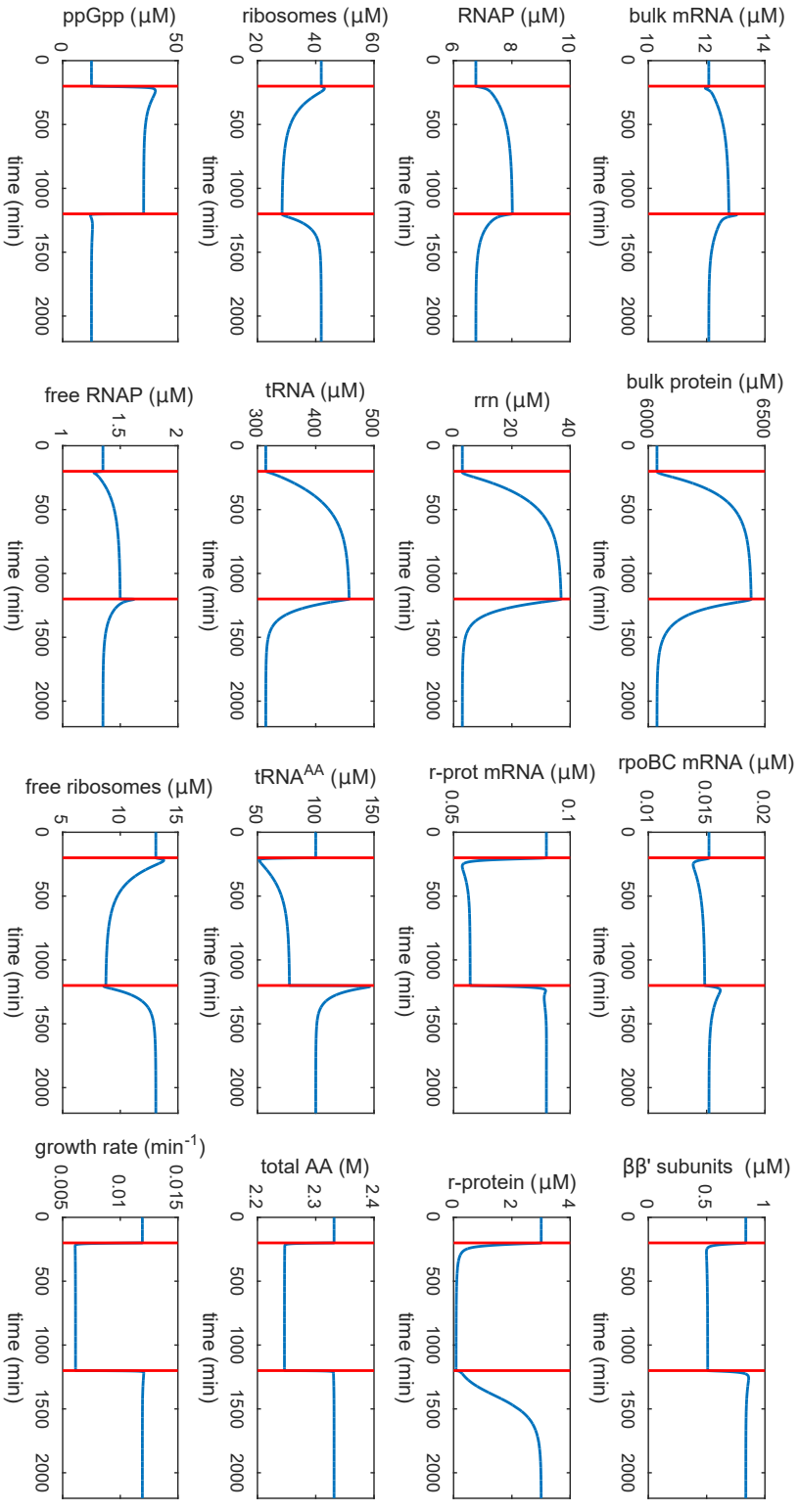


FIGURE 9.8: Simulation of growth arrest and restart following a nutrient downshift and upshift. The system starts from a steady-state condition of the wild-type model mimicking the exponential growth of cells in minimal medium supplemented with glucose. At time $\tau_1 = 200\ min$ (first red bar) the bacterial population is shifted to a medium with lower capabilities to synthesize amino acids (nutrient downshift). At time $\tau_2 = 1200\ min$ (second red bar) it is shifted again to the initial medium (nutrient upshift).

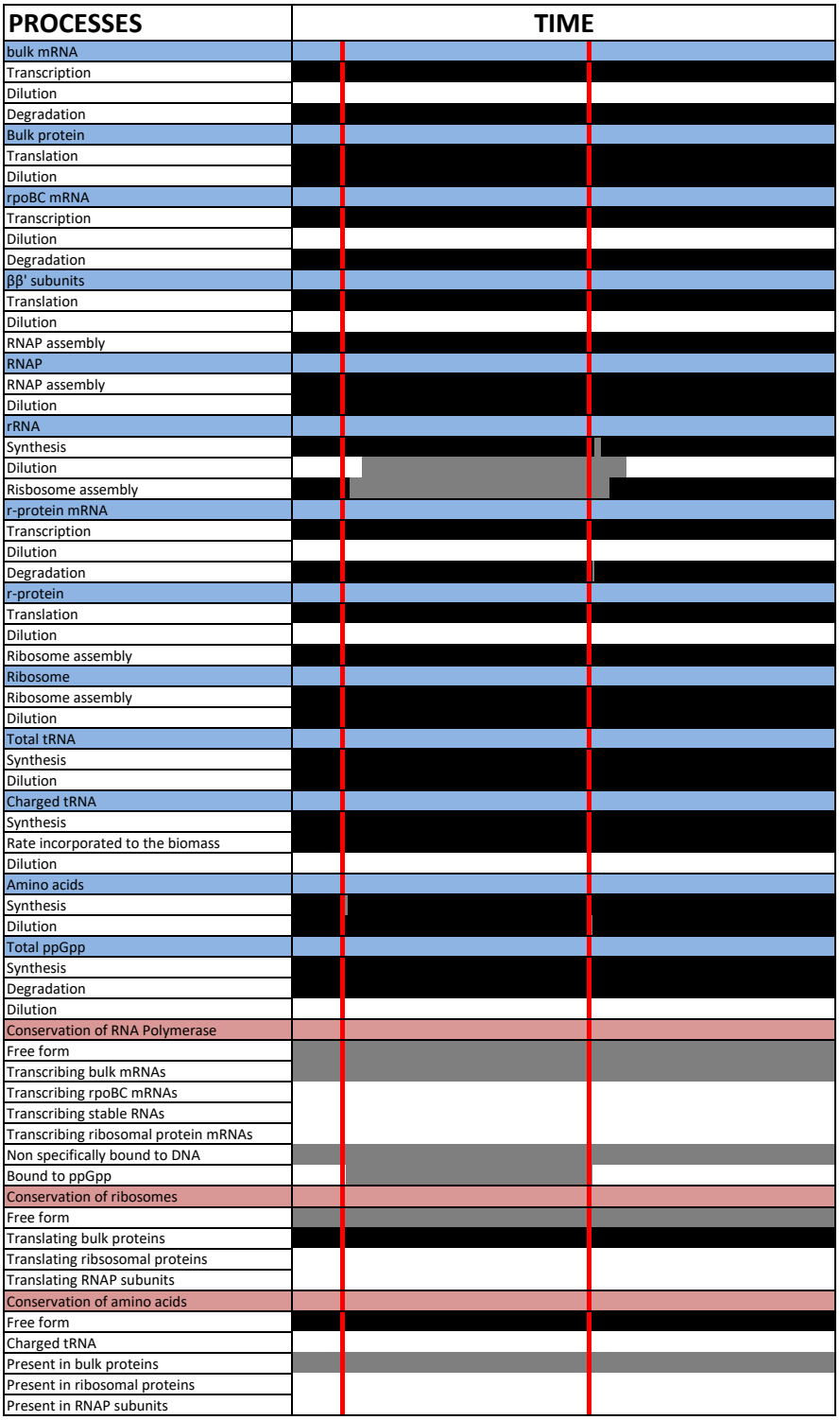


FIGURE 9.9: **Boolean Process Table for downshift/upshift experiment.** Activity of the 50 model processes during a 2200-minute period. Processes are listed in the first column (white background), ordered by variable (blue background) and by conservation equation (pink background). Their activity is depicted in the second column between 0 and 500 min: a horizontal black bar when the process is *fully active*, grey when it is *moderately active* and white when it is *inactive*.

transcribing bulk mRNAs, and the aspecifically bound form. Most of the ribosomes are used to translate bulk proteins and a non-negligible part of them is present in free form. The amino acids are most present in the free form and a significative part is used to form new proteins.

At minute 200 bacteria are shifted to a poorer medium supporting a lower growth rate. To simulate this part of the experiment we divide the maximal velocity of amino-acid synthesis (V_m^A) by 5. In these conditions, growth is not halted but the reduced growth rate restrains cells from spending their resources for growth only. As expected, the nutrient downshift, performed at minute 200, results in lower intracellular pools of amino acids and thus, charged tRNAs (see Figures 9.8 and 9.9). This immediately affects the translation rate of the proteins. To avoid that cells waste resources to synthesize new ribosomes while they cannot no longer translate proteins, ppGpp blocks *de novo* synthesis of ribsosomes through the inhibition of the transcription of stable RNAs and r-protein mRNAs. In the absence of protein translation the accumulation of biomass stops and growth is quickly arrested.

These concentrations and the growth rate are restored to their original levels following the nutrient upshift (Figure 9.8): in fact at minute 1200 the bacterial population is shifted back to the initial medium. To simulate this part of the experiment we use the initial value of the maximal velocity of amino-acid synthesis (V_m^A). Following the addition of nutrients, new amino acids and charged tRNAs are produced and translation is restored; in addition ppGpp is degraded, which relieves the inibithion of ribosome synthesis. This leads to a quick production of biomass and faster growth rate.

A number of feedback loops are active throughout the growth in these both favorable and less favorable conditions. This is for instance the case of the positive feedback loop involving the RNA polymerase, which has been studied in Chapter 5. The RNA polymerase stimulates its own expression, since it transcribes its own genes *rpoBC*. Transcription of *rpoBC* is *fully active* throughout the nutrient downshift and upshift, in the sense that it is always above the threshold value ν (see Figure 9.9 and 9.10a-9.10b). However the transcription and also the reduction of the dilution by growth are compensated for a *fully active* degradation, which is even more so following nutrient downshift, because the decreased concentration of ribosomes gives room to RNase E to bind to *rpoBC* mRNAs and to degrade them (see Figure 9.9 and 9.10a-9.10b).

The accumulation of ppGpp during the nutrient downshift inactivates the positive feedback loop involving RNA polymerase and the ribosomes. RNAP stimulates the formation of new ribosomes by transcribing ribosomal RNAs and r-protein mRNAs. Newly formed ribosomes in turn activate the synthesis of new RNAP. We observe in Figure 9.9 that a significant amount of RNA polymerase is bound to ppGpp in this phase, which

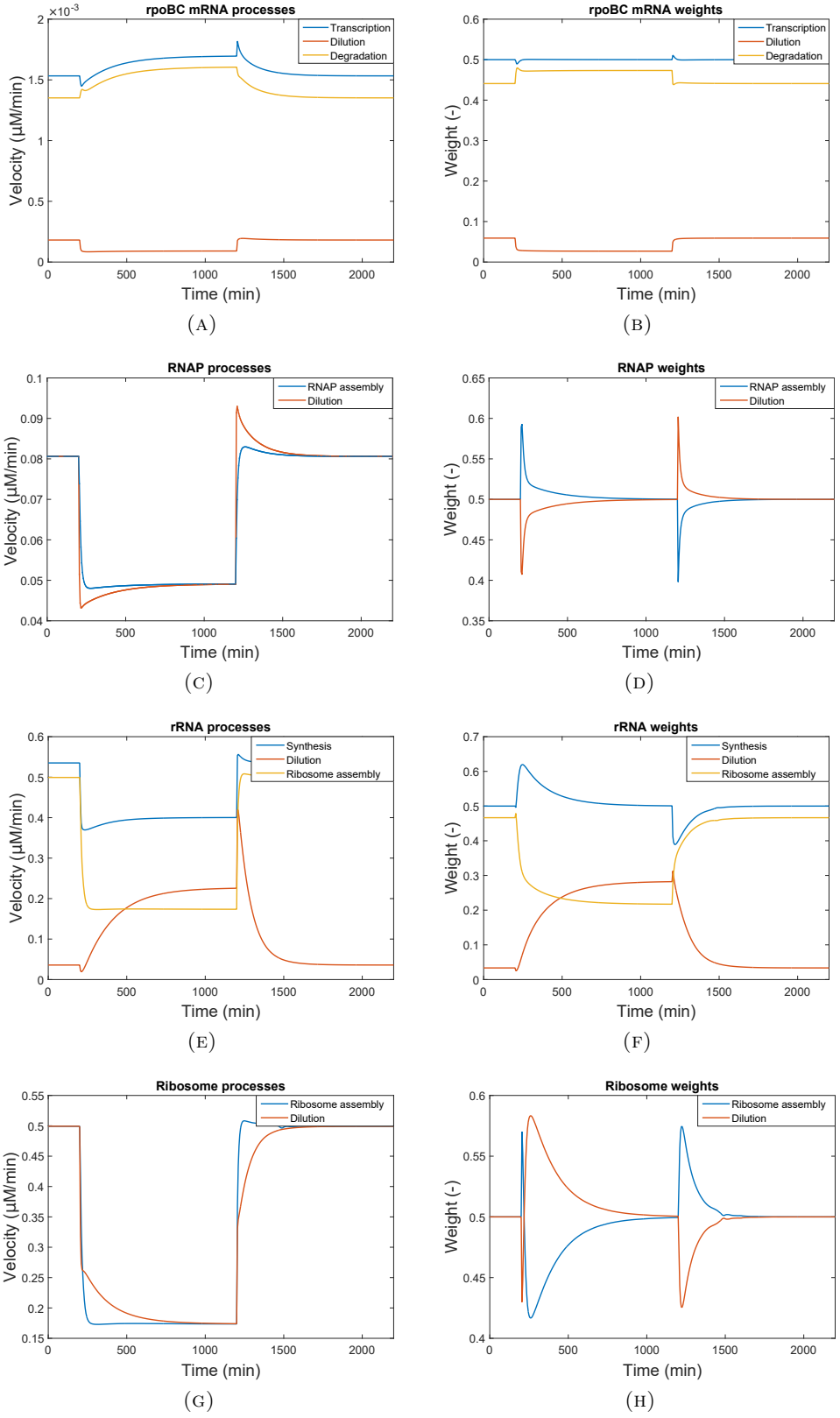


FIGURE 9.10: Absolute values of the processes (left) and their relative weights (right) for the first experiment.

ultimately impacts the assembly of stable RNAs into new ribosomes. In addition to the direct effect of ppGpp on the ribosome concentration, the positive feedback loop made of the auto-activation of ribosomes through the stimulation of r-protein translation is less important. In conditions of downshift, the degradation of r-protein mRNAs by RNase E is always *fully active*, but is strengthened by the reduced concentration of ribosomes (see Figure 9.8 ,9.9 and 9.10e-9.10h).

The *principal process analysis* provides also an explanation for model predictions inconsistent with experimental data. This is the case of the RNAP and bulk protein concentrations that are predicted to slightly accumulate during the downshift. For instance, the reduced growth rate concentrates the proteins and is not compensated for a lower maturation rate of RNAP (see Figure 9.10d) and a lower translation rate for bulk proteins. A new calibration of the model is currently under way to correct this problem, based on the use of additional experimental data that were not taken into account during the first model calibration.

9.4.2.2 IPTG stress condition

In this section, we study the effect of controlling the growth rate with IPTG. We use for that purpose the model described in Section 9.3, which relates the transcription rate of *rpoBC* to the external concentration of IPTG. Figures 9.11 and 9.12 show simulation results and the *boolean process map* for this case. Initial conditions for the simulation correspond to cells growing exponentially in glucose minimal medium (V_m^A is at its reference value) supplemented with IPTG (parameter I in Equation 9.17 equals to 1000 μM). In these conditions, as explained in Section 9.3, the engineered strain behaves similarly to the wild type ($sr_o^t = r_o^t$): the dynamics and core process *activities* are the same.

In Figures 9.11 and 9.12 the first red line indicates the time (minute 200) of the transfer of bacterial cells to a growth medium without IPTG (parameter I switched to 0 μM), while the second line indicates the time (minute 1200) at which IPTG is reintroduced into the medium (parameter I switched again to 1000 μM). As expected, removing IPTG stops the transcription of *rpoBC*, the subsequent translation of $\beta\beta'$ subunits and RNAP assembly. As a consequence, the transcription of all cell mRNAs and their translation stop also (note that this is less visible in the *boolean process map* where these processes appear as *active*. Simply they still contribute to the total processes, while the latter drop to zero in the absence of IPTG). The consequence of the reduced concentration of cell mRNAs and proteins is a growth arrest. The simulations also reproduce a behavior observed experimentally for the growth rate after removal of IPTG: a lag time of several hours is needed before we observe a growth arrest [51]. Indeed, despite the arrest of its

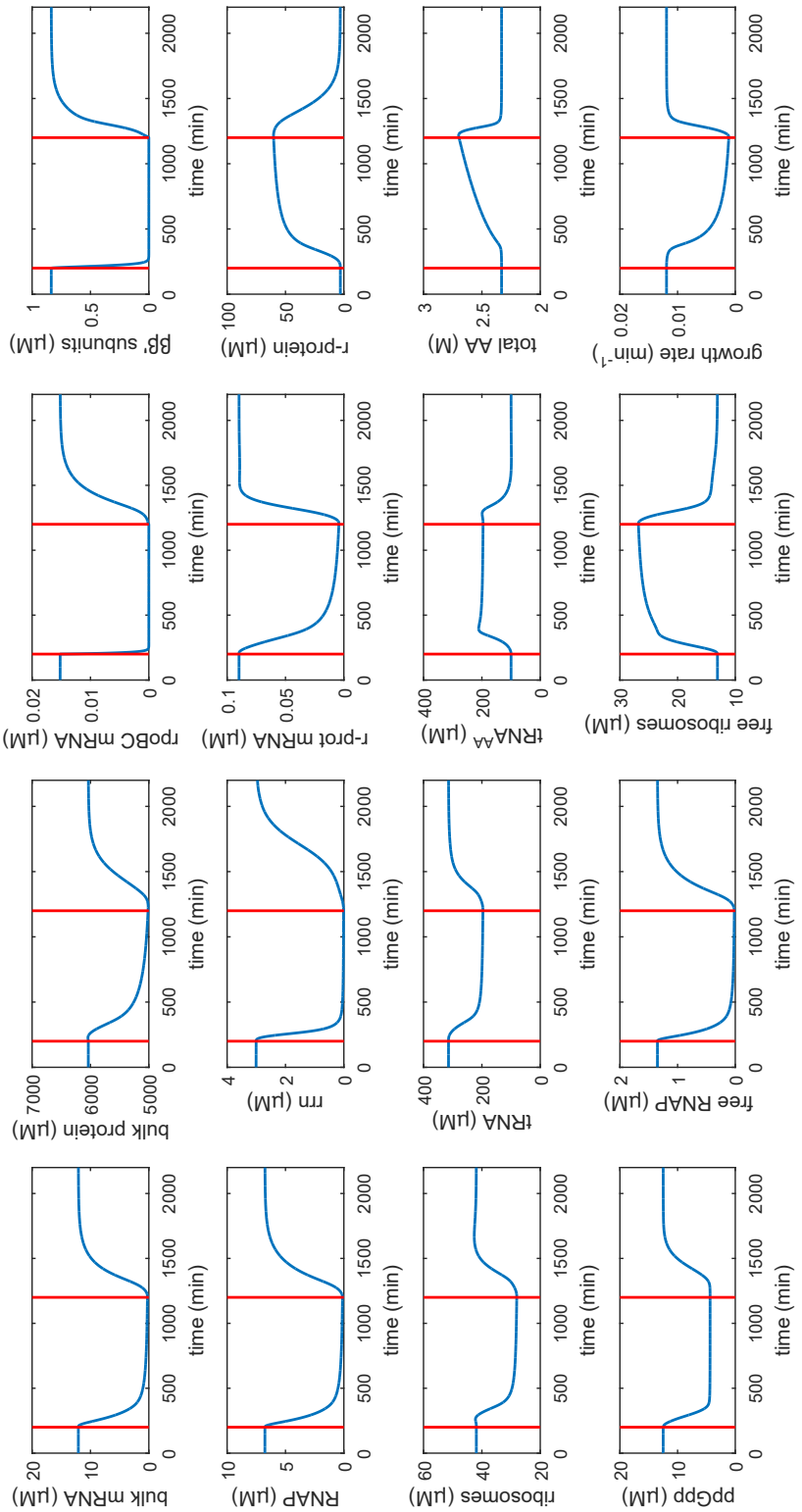


FIGURE 9.11: Simulation of the growth arrest and restart with IPTG. The system starts from a steady-state condition of the engineered-strain model in a medium with a high concentration of IPTG (1000 μM). At time $\tau_1 = 200$ min (first red bar) the bacterial population is shifted in a medium with no IPTG (0 μM). At time $\tau_2 = 1200$ min (second red bar) it is shifted again to the initial medium (IPTG=1000 μM).

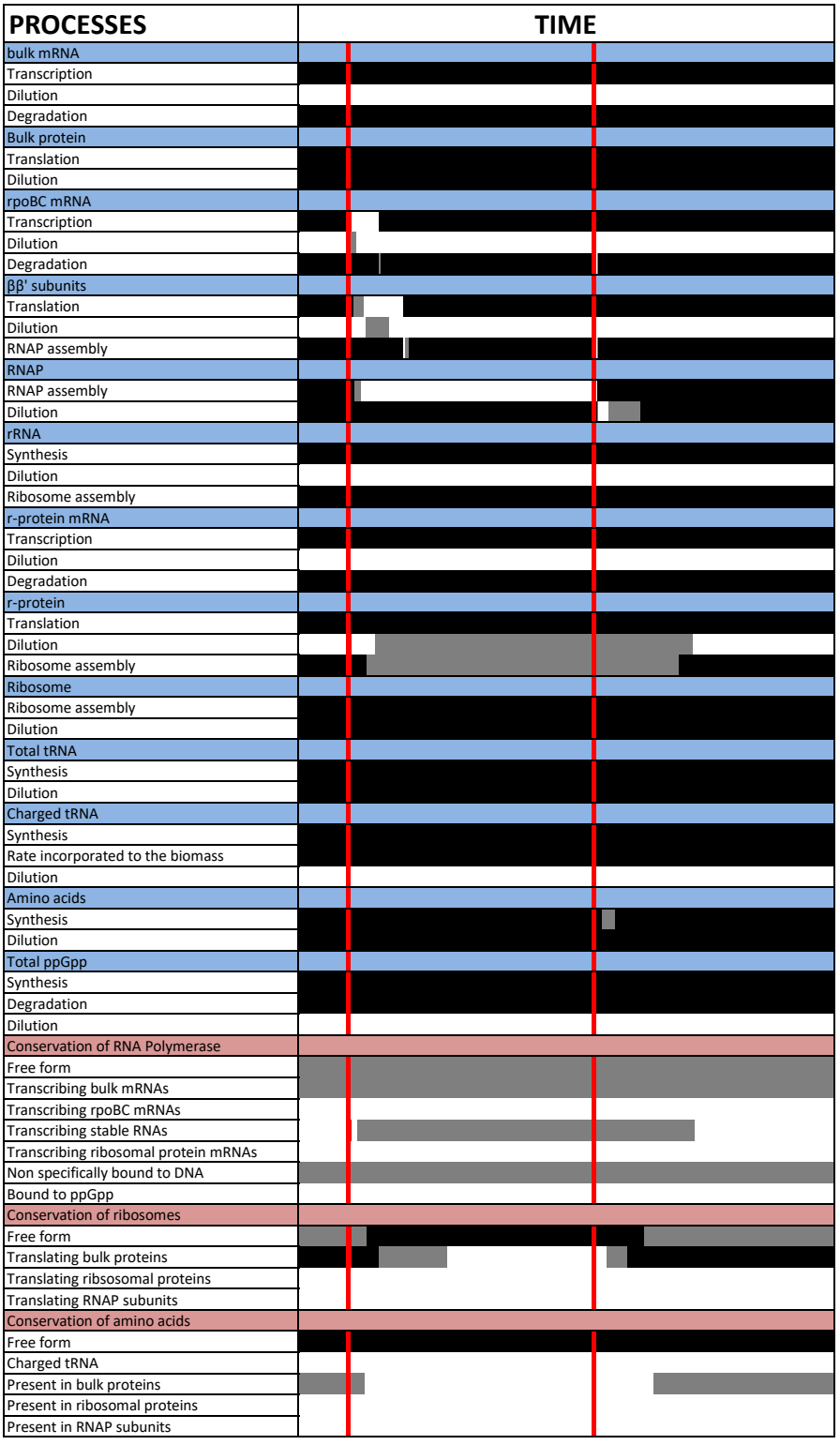


FIGURE 9.12: **Boolean Process Table growth arrest/restart experiment with IPTG.** Activity of the 50 model processes during a 2200-minute period. Processes are listed in the first column (white background), ordered by variable (blue background) and by conservation equation (pink background). Their activity is depicted in the second column between 0 and 500 min: a horizontal black bar when the process is *fully active*, grey when it is *moderately active* and white when it is *inactive*.

production, RNAP is a stable protein, whose concentration decreases through growth dilution. Several repeated cell divisions are needed before the concentration of RNAP becomes limiting and that protein synthesis can no longer support growth.

As shown in Figures 9.11 and 9.12, ppGpp does not play any role in the whole simulation: its synthesis and degradation are *fully active* both in the presence and absence of IPTG, but the resulting concentration of the alarmone remains low. The proportion of RNAP bound to ppGpp is below the threshold level δ and does not affect the transcription of stable RNAs and r-protein mRNAs. Despite the decrease of RNAP concentration when IPTG is absent, a non negligible part of RNAP is even transcribing stable RNAs. However this will not be sufficient to support the formation of ribosomes in quantity high enough to support growth.

The accumulation of bulk proteins contributes a lot to the formation of biomass. As can be seen in Figure 9.12, the ribosomes translating bulk proteins become negligible when IPTG is absent, and amino acids are no longer incorporated into bulk proteins. These unused ribosomes and amino acids accumulate as free forms within cells.

The r-proteins are predicted to accumulate within cells in the absence of IPTG. These proteins are stable and translated as long as ribosomes are in sufficient quantities. Their reduced synthesis is largely compensated by their reduced dilution by growth and assembly into new ribosomes. In the latter case indeed, stable RNAs are no longer available to make new ribosomes and the free form of r-proteins accumulates. They are no experimental data available to validate or invalidate this model prediction. It might be possible that the concentration of r-proteins decreases in the absence of IPTG, because a negative feedback mechanism is known to cause the degradation of excess of r-proteins, which are not incorporated into ribosomes [55]. Proteomics experiments are planned in the framework of the RESET project by our CEA partners from the EDYP team, to characterize the phenotype of the engineered strain. The experiments will consist of measurements of the relative content of the different cell proteins in the presence or absence of IPTG. These experimental results, notably concerning the r-proteins, will be confronted to the model predictions and will potentially lead to further model adjustments to improve its predictive capabilities.

Note that the system, after the removal of IPTG, can reach a steady state where the growth rate is null, but we prefer to restart the growth at 1200 minutes because we want to mimic the experiments of [51], in which cells restart growth more easily when there is a residual activity.

After the re-introduction of IPTG in the medium, growth rate and the system variables are restored to their values reached before the removal of the inducer at minute 200.

There is a delay before *rpoBC* mRNAs accumulate: despite their *fully active* transcription, mRNAs are still highly degraded by RNAse E in the absence of protection by ribosomes whose concentration is still low (and consequently $\beta\beta'$-subunit and RNAP concentrations).

9.5 Conclusion

This chapter uses the various methods described in the manuscript (model reduction and *principle process analysis*) to contribute to the modeling of the gene expression machinery and the analysis of its dynamical functioning. This algebro-differential model is used to analyze the functioning of GEM in wild-type cells responding to nutritional changes, as well as engineered strains, in which the growth rate is under the control of IPTG. The model predictions are consistent with experimental data in most cases. As discussed above, new calibrations are planned to improve the predicted response of RNAP concentration to a nutrient downshift. As well, potential model refinements will be considered depending on the results of the proteomics experiments obtained with the engineered strain.

The project is still on-going, but the first results of *principle process analysis* described above allow to start analyzing the regulatory mechanisms at work during the different phases of growth on a rich or poor nutrient or with or without IPTG. We have already analyzed the functioning of the feedback loops that directly affect RNAP and ribosome levels. It is puzzling to observe that, despite its important complexity, the functioning of the network can be directly affected by targeting only one component, the RNA polymerase. Living organisms are characterized by complex networks with redundant interactions, that can compensate for certain inactive interactions. In the present case, the RNA polymerase of *E. coli* is unique, contrary to its Eukaryotes couterparts. Affecting this network component breaks down the functioning of the whole network, as could be observed in the simulation and PPA results in Section 9.4.2.2, and experimentally in [51].

The RESET project relies on the idea that the control of growth rate exerted by IPTG is preferable over a control through nutrient limitations. Although, at this stage, the GEM model has not been not connected yet with a metabolic model developed by Adrien Henry and Olivier Martin from INRA in Le Moulon, we can already provide some explanations based on the simulation and PPA results in Sections 9.4.2.1 and 9.4.2.2. In the former case, the nutrient limitation imposes a stress to the cell, as shown by the accumulation of the alarmone ppGpp. In addition to its direct impact on protein synthesis through the inhibition of the formation of new ribosomes, ppGpp is also known

to elicit more general stress responses that involve the sigma factor σ^S responsible for establishing genetic program allowing the cell survival in adverse conditions [45]. Part of the cell resources are thus allocated to the stress response rather than the formation of a high value product. To the contrary, the model does not predict a situation of stress in growth-arrested cells in the absence of IPTG: the concentration of ppGpp is low, similar to the levels detected in cells grown in minimal medium [83]. We clearly see that cell resources are no longer allocated to mRNA transcription and protein translation, thus to the biomass formation. They are ready to be channeled to the synthesis of products of interest. The connection of the model to the model of metabolism will help to better analyze the conditions for the re-channeling of resources and will also allow to study how optimal control could be applied to optimize the product yield.

As seen above, the growth rate responds non linearly to the concentration of RNA polymerase in the engineered strain. Whether the hypersensitivity is due to a bistable behavior is not known at this stage. The model will be used to answer this question, together with additional experiments. However, such behaviors are known to introduce heterogeneity in the cell response and we did observe variability of growth rate and gene expression with the engineered strain in single cell experiments [51]. This is the subject of the following chapter, in which we have used a much simpler version of the GEM model to study the relation between growth and RNAP concentration at the single-cell level.

Chapter 10

Single-cell model calibration of growth control experiments in *E. coli*

In this chapter we discuss the work done in collaboration with Eugenio Cinquemani, research scientist at Inria Grenoble-Rhône-Alpes. The experimental data we used were obtained by Jérôme Izard during his PhD thesis in the Laboratoire Adaptation et Pathogénie des Micro-organismes (Univ. Grenoble - Alpes). The experiments were done in collaboration with Ariel Lindner at the Centre for Research and Interdisciplinarity in Paris.

This work will be a part of a future journal paper.

10.1 Introduction

Optimizing growth is an important topic in cell biology. As we have seen in Chapter 3, control of *E. coli* cell growth and metabolism can increase the production of high-value biotechnological products like glycerol.

We will first recapitulate the characteristics of the engineered strain and the monitoring of its growth in single-cell experiments, as these concepts will be needed later on in the chapter.

The control enables the growth of an *E. coli* population up to a certain biomass and the subsequent arrest of the growth, allowing certain enzymes to become functional and produce metabolites of interest [51]. Because cell populations cannot survive without

growing for an extended period of time (i.e because of degradation of enzymes and other proteins), growth can be switched on again, thus alternating phases of growth and product synthesis. To achieve this regulatory mechanism, an *E. coli* strain was constructed in which the transcription of the *rpoB* and *rpoC* genes, encoding the two subunits $\beta\beta'$ of RNA Polymerase (RNAP), are under the control of an IPTG-inducible promoter. The control of the production of RNA Polymerase has direct effect on cell growth and duplication. Indeed, as we have seen in Chapter 3 and 9 the synthesis of mRNA starts with transcription where RNAP plays a key role. Inhibiting the latter, mRNAs, and thus proteins, cannot be synthesized. By substituting the original promoter with an IPTG-inducible promoter and adding extra copies of the Lac repressor gene (*lacI*) makes it possible to have a synthetic control of *E. coli* growth. If there is no IPTG in the culture medium, the Lac repressor protein (that is always expressed) binds in the operator area of *rpoB* and *rpoC* genes, so that RNAP cannot bind the promoter to start the transcription of its own gene. This results in dilution of RNAP during the residual growth, followed by the arrest of synthesis of the other cellular proteins as well. If instead IPTG is added to the medium, imported IPTG molecules bind Lac repressor proteins, preventing their binding on the promoter, thus allowing RNAP to transcribe its own gene. This makes further RNAP molecules available for the expression of the different genes, thus allowing *E.coli* to grow and divide. In order to quantify protein synthesis capabilities of *E. coli* over time, a gene encoding a red fluorescent protein (RFP) was placed under the control of the promoter of a constitutive gene.

The experiment in [51] consists of monitoring and controlling the growth of such modified *E. coli* cells in a microfluidics device in minimal (M9) medium. Cells are trapped in dead-end channels, and growth leads offspring eventually leave the channel from its open end. The cell that always remains at the dead-end is monitored throughout the experiment, leading to measurements (cell growth and gene expression) taken every approximately 10 minutes for every channel. In the first 800 minutes of the experiment IPTG is provided (allowing the growth and division of *E. coli*); from minute 800 to 1150, IPTG is removed (leading to arrest of growth and cell division); then IPTG is provided again (allowing re-start of growth and cell division). As shown by removal and reinjection of IPTG, growth arrest is reversible.

The accompanying fluorescent reporter data from constitutive gene expression are used for image analysis purposes in [51]. Here, we additionally exploit RFP profiles to study the relationships between growth and RNAP expression. To do this, we develop a deterministic model of RFP expression dynamics. As customary in this type of models, the growth rate acts as the system input and fluorescence is the system output. This implicit assumption on a causal relationship is a point that we will rediscuss in the light of our modeling results at the end of the chapter.

Earlier results [67] showed that treating growth rate data as an input profile fixed a priori by data preprocessing (i.e. before the inference of model parameters) does not lead to satisfactory modeling. On the other hand, growth rate data are themselves uncertain data that need to be explained together with the corresponding gene expression data. In this spirit, we consider a modeling approach where growth rate and gene expression are jointly fitted with explicit account for their respective uncertainty (see also [101]).

Experimental single-cell data display large variability between the different cell fluorescence profiles, hinting at the fact that an average model may not be an appropriate description of the system. Therefore we focus on a calibration approach that accounts for cell-to-cell variability and that can model response of different cells. Together with the calibration of an average model, we thus perform single-cell calibration and compare the results from the two approaches.

The results of this chapter will show that fitting a mean model to the average data by considering cell-to-cell variability as bare "measurement error" leads to an unsatisfactory model of the average data itself, which is instead better explained in terms of mean of single-cell fits. In addition, single-cell calibration enables the study of the variability of single-cell dynamics across the population. Yet, some inconsistency remains that prompts further research.

10.2 Model

Figure 10.1 summarizes the dynamic of the experiment [51].

The first scheme (A) corresponds to the situation in the presence of IPTG, between minute 200 and 800, as well as from minute 1150 to the end of the experiment. The second one (B) corresponds to the the absence of IPTG, between minute 800 and 1150.

The growth rate of the bacteria was quantified by measuring the cell area of the newly-formed bacteria in successive frames of time-lapse microscopy. The concentration of the reporter protein RFP was determined by dividing the cell fluorescence by the cell area.

The model reactions are the following:

$$P_{rpoBC} + RNAP \overset{K_2}{\rightleftharpoons} P_{rpoBC} - RNAP \overset{k_1}{\rightarrow} 2RNAP \overset{\mu}{\rightarrow} \emptyset$$
$$P_{rpoBC} + LacI4 \overset{K_5}{\rightleftharpoons} P_{rpoBC} - LacI4$$
$$LacI4 + IPTG \overset{K_4}{\rightleftharpoons} LacI4 - IPTG$$
$$P_{rfp} + RNAP \overset{K_3}{\rightleftharpoons} P_{rfp} - RNAP \overset{k_p}{\rightarrow} RFP_{immature} + RNAP$$
$$RFP_{immature} \overset{k_m}{\rightarrow} RFP_{mature}$$
$$RFP_{immature} \overset{\mu}{\rightarrow} \emptyset$$

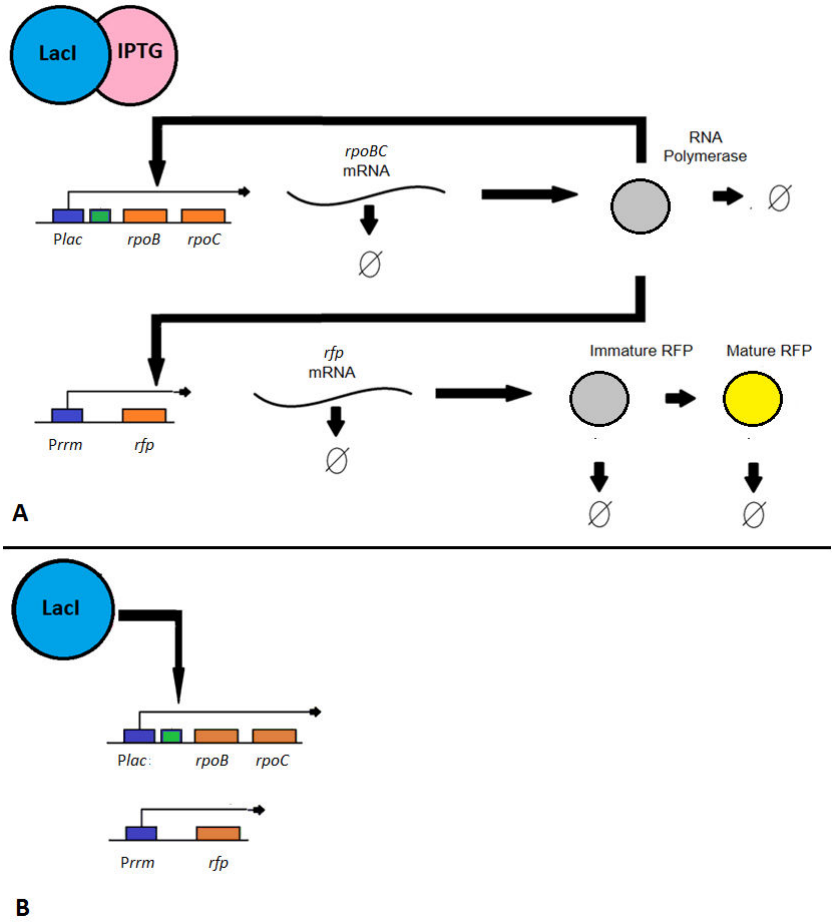


FIGURE 10.1: **Model experiment**. A- In the presence of IPTG, the protein LacI is inactivated (forming a complex with it) and then RNA Polymerase can transcribe its own genes, producing *rpoB* and *rpoC* mRNAs encoding the subunits β and β' that form, with the subunits α, the RNA Polymerase: this is a positive loop because the RNAP enhances its own expression. RNAP binds also to the promoter of the gene *rfp*, leading the transcription into *rfp* mRNA coding for the fluorescence protein RFP. Fluorescence activity of RFP in reponse to light excitation depends on post-transcriptional modifications: the protein maturation gives rise to an additional reaction step from RFP to active RFP. The synthesis of the mRNAs and proteins in the system is counterbalanced by growth dilution and degradation of the gene product. B- In the absence of IPTG LacI can bind to the operator region near the promoter of the genes *rpoB* and *rpoC*, leading to the arrest of RNAP and RFP transcription and to the subsequent arrest of the growth rate.

$$RFP_{mature} \xrightarrow{\mu} \emptyset$$
$$P_{LacI} + RNAP \overset{K_6}{\rightleftharpoons} P_{LacI} - RNAP \xrightarrow{k_l} LACI4 + RNAP$$
$$LacI4 \xrightarrow{\mu} \emptyset$$

The protein LacI in its tetrameric form ($LacI4$) controls the binding of RNAP to the promoter of the *rpoB* and *rpoC* genes (P_{rpoBC}). The presence of IPTG inhibits the binding of $LacI4$ to P_{rpoBC}, while the absence of IPTG allows the binding of $LacI4$ to P_{rpoBC}. In turn this hampers the binding of RNAP to P_{rpoBC}. The parameters of the model are: the maximal synthesis rate of RNAP (k_1), the dissociation constant of RNAP and P_{rpoBC} complex (K_2), the dissociation constant of P_{rfp} and RNAP (K_3), the maximal synthesis rate of RFP (k_p), the dissociation constant of IPTG and $LacI4$ complex (K_4), the dissociation constant of P_{rpoBC} and P_{LacI4} complex (K_5), the dissociation constant of P_{LacI4} and RNAP complex (K_6), the maturation rate of RFP (k_m) and the maximal synthesis rate of $lacI4$ (k_l).
The dynamics of the concentrations of RNAP (P), of immature RFP (F_{im}), of mature RFP (F_m) and of $LacI4$ (L) are modeled by mass-action laws with Michaelis-Menten-type reaction rates, resulting in the system of equations:

$$\frac{dP}{dt} = -\mu\, P + k_1 \frac{P}{P + K_2\left(1 + \frac{L}{K_4 + K_5\, I}\right)} \tag{10.1}$$

$$\frac{dF_{im}}{dt} = k_p \frac{P}{P + K_3} - (\mu + k_m)\, F_{im} \tag{10.2}$$

$$\frac{dF_m}{dt} = k_m\,(F_{im}) - \mu\, F_m \tag{10.3}$$

$$\frac{dL}{dt} = k_l \frac{P}{P + K_6} - \mu\, L \tag{10.4}$$

where I is the concentration of IPTG.

The input of the system is the growth rate μ and the output is the fluorescence $Y = K_f\, F_m$, where K_f is a conversion factor.

10.3 Methodology

To estimate the parameters of the model, different calibration approaches are considered: in a first approach, we consider that the model represents an average cell, and

an average growth rate profile is used to explain an average fluorescence profile. In a second approach, the same model equations are considered for the response of the single cell, and different response profiles observed in different cells are explained in terms of different parameter values for different cells [75], as well as different growth rate profiles. In both approaches growth rate is treated as an input measured with error and simultaneously fitted with the dynamical model parameters.

10.3.1 Data

From the microfluidics experiments, after a first image analysis phase, we are provided with time course measurements for C channels. For every channel, the cell that sits at the bottom is monitored over time. This is a growing cell that divides repeatedly in the course of an experiment at times t_j^d, with $j = 1, ..., m$. After every division, measurements pertain the daughter cell that takes the position of the mother cell at the channel dead-end. For simplicity we will refer to this as one cell, so that we have as many cells as channels, and keep into account the discontinuities that measurements undergo at division times. Refer to Figure 10.2, showing the raw-time course measurements for one cell. For each of C cells, at observation times $t_1 < t_2 < ... < t_n$, we are provided with measurements of cell size $A(t_i)$, total cell fluorescence $F(t_i)$ and, by straightforward division, normalized fluorescence per unit size $Y(t_i) = F(t_i)/A(t_i)$ (later in this chapter, we refer to the normalized fluorescence simply as fluorescence). For all measurements the observation times of interest t_i are between minute 400 and 1400 and they are not necessarily identical across cells. It can be appreciated that both A and F show discontinuities at division times t_j^d (which are also available as data), whereas the profile of Y is essentially continuous thanks to normalization. From this data for C cells, to perform estimation of the model parameters and the (input) growth rate profiles, we need to extract first raw growth rate data from cellular size profiles $A(t_i)$, as well as statistics on the measurement uncertainty associated with single-cell growth rate and fluorescence data $Y(t_i)$. For the fitting of the average model, we also need to compute mean fluorescence and growth rate data that will be treated as the "average-cell" data.

10.3.2 Extraction of cellular profiles

In this section, for a given cell, we discuss how we obtain growth rate data $\mu(t_i)$ from $A(t_i)$, and how, on the basis of $\mu(t_i)$ and $Y(t_i)$, we obtain the associated uncertainties. To obtain the latter we rely on a statistical procedure known as *bootstrap*, which relies on an initial fit of the data $A(t_i)$ and $Y(t_i)$. We will discuss this initial fit for fluorescence Y first, then move on to the more complex case of A, from which an initial estimate of the

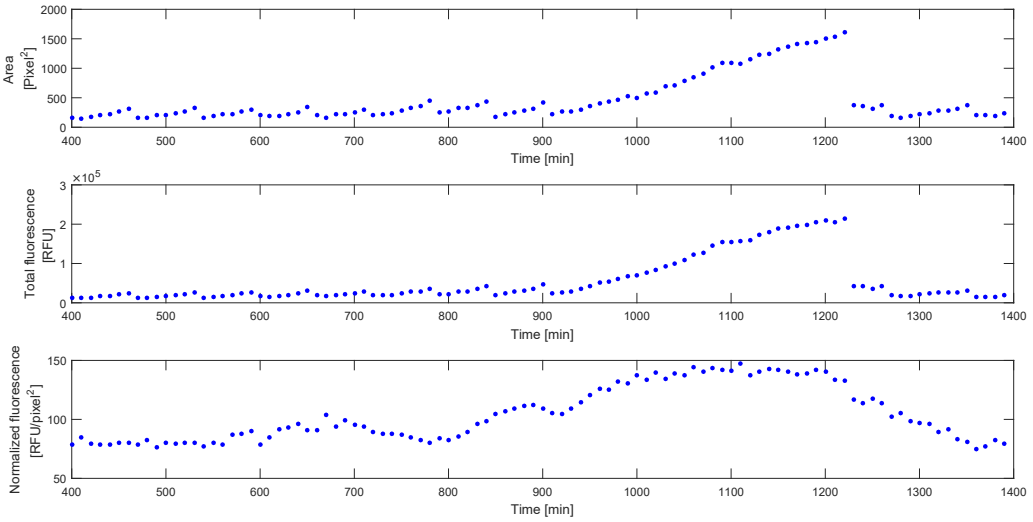


FIGURE 10.2: **Area and fluorescence measurements**. Measurements for one cell at observation times t_i, $i =, \ldots, n$, consisting of cell area $A(t_i)$ (top plot), total cell fluorescence $F(t_i)$ (center plot) and normalized fluorescence per unit size $Y(t_i)$ (bottom plot). The top plot shows that removing IPTG (at minute 800) from the microfluidics channels does not cause immediate consequences in cell division: the RNAP protein has a long half-life and the arrest of cell division starts with a delay of approximately 100 minutes. In the same way there is a delay of approximately 100 minutes after the addition of IPTG (at minute 1150) before growth-limited cells resume normal growth: this lag period is probably necessary for replenishing the pool of RNA polymerase and other cellular components necessary for cellular growth. The microfluids experiments also reveal that, in the time period in which IPTG is not in medium M9, cell division stops and elongated cells appear to restore their own division after a successive insertion of IPTG. The cause of this filamentous morphology is currently unknown. It might involve the bacterial SOS response [53], but could also be a consequence of the decrease in concentration of a protein necessary for cell division when RNA polymerase is diluted out (and transcription of this factor stops): for more details see [51]. Anyhow, during the experimental time while IPTG is not inserted we assist at a remarkable decrease of the growth rate of the single cell. The arrest of the synthesis of RNAP causes the arrest of the production of the protein RFP responsible for the total cellular fluorescence (see center plot) with a similar delay as we have seen in the first graph: during the time of the experiment while IPTG is not inserted the fluorescence increases because the cell does not divide. The decrease in fluorescent protein synthesis rate appears to occur at a slower rate than the decrease of growth rate, whence the increase of normalized fluorescence in absence of IPTG observed in the bottom plot.

profile μ follows, and eventually explain the *bootstrap* procedure for the quantification of data uncertainty.

We apply as a fitting method the *cubic smoothing spline*. Consider the relation $Y_i = f(t_i)$. The smoothing spline is an estimate $\hat{f}$ of the function f, that is defined to be the minimizer of

$$\sum_{i=1}^{n}(Y_i - \hat{f}(t_i))^2 + \lambda_Y \int_{t_1}^{t_n} \hat{f}''(t)^2 dt \tag{10.5}$$

in a space of suitably smooth functions [120], where λ_Y is a smoothing parameter, controlling the trade-off between fidelity to the data and roughness of the function

estimate. To choose the value λ_Y, we perform *leave-one-out cross-validation*. For any candidate value of λ_Y, we solve the problem above n times, each time ignoring one point (t_i, Y_i) in the fit and computing the error that the fitting result commits in predicting (t_i, Y_i). Taking the median of these n prediction errors, we quantify the predictive capability of the solution associated with that value of λ_Y. The final value of λ_Y is chosen numerically as the one that minimizes this median prediction error. We denote by $\hat{Y}(t_i)$ the associated solution of Formula (10.5) finally obtained on the basis of all data points.

Toward computation of the growth rate, in order to compute an initial fit of cell size, consider the relation $A_i = g(t_i)$. To properly account for cell division, we calculate the area profile in-between division times t_j^d , $j = 1, ..., m$: the function $g(t)$ (defined from t_1 to t_n) is decomposed in the $m-1$ functions $g_j(t)$ (defined from t_j^d to t_{j+1}^d). If the time window is large (numerous data points, as in the case of slow growth in absence of IPTG), we calculate the profile of the area $\hat{g_j}$ as we do for the fluorescence in Formula (10.5), that is, by finding the smoothing spline solution to

$$\sum_{i:t_j^d \le t_i \le t_{j+1}^d} (A_i - \hat{g_j}(t_i))^2 + \lambda_{A_j} \int_{t_j^d}^{t_{j+1}^d} \hat{g_j}''(t)^2 dt \tag{10.6}$$

with value of the smoothing parameter λ_{A_j} again fixed by *leave-one-out cross validation*. Correspondingly, we calculate growth rate estimates $\hat{\mu}_i = \hat{h}_j(t_i)$ (for all time points t_i between t_j^d and t_{j+1}^d) by means of the formula

$$\hat{h}_j(t) = \frac{1}{\hat{g}_j(t)} \frac{d\hat{g}_m(t)}{dt} \tag{10.7}$$

If instead the time window is small (few data points), fitting the curve with a *cubic smoothing spline* may lead to artifacts. Assuming in this case exponential growth $\hat{A}(t) = a \exp^{bt}$ (as expected in the presence of IPTG), we perform a *linear regression* on the log of the experimental data:

$$\min_{a,b} \sum_{i=1} (\log A(t_i) - log\hat{A}(t_i, a, b))^2 = \min_{a,b} \sum_{i=1} (\log A(t_i) - \log a - b\, t_i)^2 \tag{10.8}$$

and set $\hat{\mu} = \frac{1}{A}\dot{A} = \frac{a\, b\, e^{b\, t}}{a\, e^{b\, t}} = b$ for all times t_i between the division times considered. In this case, cross-validation is not needed since this parametric solution does not incur the risk of overfitting. Overall, the estimate $\hat{\mu}$ over the whole experiment is obtained by the juxtaposition of all estimates obtained in-between division times.

Based on these initial fits, in order to quantify the uncertainty of the data, we consider the residual between the experimental data at time t_i and the point of the fitted curve at the same times t_i: $\varepsilon(t_i) = A(t_i) - \hat{A}(t_i)$ and $\xi(t_i) = Y(t_i) - \hat{Y}(t_i)$. Then we perform *bootstrap*, a statistical technique allowing one to draw statistics from a finite data set by random sampling. We resample the residuals K times to obtain K different residual vectors ε^k and ξ^k, where $k = 1, \ldots, K$. Correspondingly we define K vectors of synthetically generated data $A^k(t_i) = \hat{A}(t_i) + \varepsilon^k(t_i)$ and $Y^k(t_i) = \hat{Y}(t_i) + \xi^k(t_i)$, with $k = 1, \ldots, K$. We then calculate for each new set of data estimates $\hat{\mu}^k$ and $\hat{Y^k}$ as we did before for $\hat{\mu}(t_i)$ and $\hat{Y}(t_i)$ and we calculate the uncertainty (standard deviation) for the fluorescence and growth rate data at the different times t_i as the standard deviation $\sigma_{\hat{Y}}$ of the K mean fluorescence profiles and $\sigma_{\hat{\mu}}$ of the K growth rate profiles at the same times t_i.

10.3.3 Calculation of average cell profiles

After the calculation of the estimates $\hat{Y}^c$ and $\hat{\mu}^c$ and related uncertainties for each cell $c = 1, \ldots, C$, it is possible to obtain the *weighted average* of the c profiles for the fluorescence:

$$\hat{\bar{Y}}(t_i) = \sum_{c=1}^{C} \hat{Y}^c(t_i) W^c_{\hat{Y}}(t_i) \tag{10.9}$$

and for the growth rate:

$$\hat{\bar{\mu}}(t_i) = \sum_{c=1}^{C} \hat{\mu}^c(t_i) W^c_{\hat{\mu}}(t_i) \tag{10.10}$$

where $W^c_{\hat{Y}}(t_i)$ is the weight associated with the fluorescence profile of the c^{th} cell at time t_i and $W^c_{\hat{\mu}}(t_i)$ is the weight associated with the growth rate profile of the c^{th} cell at time t_i, defined by:

$$W^c_{\hat{Y}}(t_i) = \frac{\frac{1}{\sigma^2_{\hat{Y}^c}(t_i)}}{\sum_{c=1}^{C} \frac{1}{\sigma^2_{\hat{Y}^c}(t_i)}} \tag{10.11}$$

and

$$W^c_{\hat{\mu}}(t_i) = \frac{\frac{1}{\sigma^2_{\hat{\mu}^c}(t_i)}}{\sum_{c=1}^{C} \frac{1}{\sigma^2_{\hat{\mu}^c}(t_i)}} \tag{10.12}$$

These mean profiles will be used for the calibration of the average cell mode. Note that, because single-cell fits are used to compute the average data, times t_i may be chosen identical across cells even if they are not in the original data sets.

10.3.4 Calibration of the model

The model equations can be summarized by

$$\dot{x} = \Phi(x, \mu, \theta), \tag{10.13}$$

with measured fluorescence

$$y = \Gamma(x, \theta). \tag{10.14}$$

Here we discuss the approach to find the unknown input μ and parameters θ to fit growth rate and fluorescence data. For single cells, these are $\hat{Y}^c$ and $\hat{\mu}^c$ (and relevant uncertainties), and calibration is repeated for $c = 1, \ldots, C$, thus getting C estimates of θ and μ. For the average model, the data to fit is $\hat{\bar{Y}}$ and $\hat{\bar{\mu}}$.

For both single cell and average cell approaches, the dynamical model parameters that we estimate are $\theta = [k_1, K_2, K_3, K_4, K_5, K_6, k_m, R_m(0), L(0)]$. The search space for the parameter values is defined by the intervals $[0.01\,\theta_0, 100\,\theta_0]$, where θ_0 is an initial guess derived from [67]. The remaining model parameters are determined by θ via stationarity assumptions of the system dynamics at the beginning of the experiment.
In all cases, the unknown input μ is modeled by a *B-spline curve* of sixth degree, defined over the domain $[t_0, t_n]$, with 30 knots, which can be expressed as a linear function of a vector of parameters η. Thus, $\mu(t) = B(t)\,\eta$, where $B(t)$ is a vector of basis functions, and the problem of input estimation becomes that of estimating η. In accordance with their definition, we estimate the coefficients η in a boundary of $[-1, 1]$. To perform simultaneous estimation of input and dynamics, i.e. of η and θ, the objective function that we minimize for the average cell is the negative log-likelihood function (under Gaussianity and mutual independence assumptions of the errors affecting the data $\hat{\bar{\mu}}(t_i)$ and $\hat{\bar{Y}}(t_i)$)

$$\sum_{t=1}^{n} \left(\frac{\hat{\bar{Y}}(t_i) - Y(t_i; \theta, \eta)}{\sigma_{\hat{\bar{Y}}(t_i)}} \right)^2 + \sum_{t=1}^{n} \left(\frac{\hat{\bar{\mu}}(t_i) - B(t_i)\,\eta}{\sigma_{\hat{\bar{\mu}}(t_i)}} \right)^2 \tag{10.15}$$

where $Y(t; \theta, \eta)$ is the solution of the system dynamics of Equations (10.13)-(10.14) for the candidate θ and input $\mu = B\,\eta$. Here, $\sigma_{\hat{\bar{\mu}}(t_i)}$ and $\sigma_{\hat{\bar{Y}}(t_i)}$ are the standard deviations of C cellular profiles $\hat{Y}^c$ and $\hat{\mu}^c$, evaluated at time t_i, that express the cell-to-cell variability. Similarly, the objective function for the single cell calibration is:

$$\sum_{t=1}^{n} \left(\frac{\hat{Y}^c(t_i) - Y(t_i; \theta, \eta)}{\sigma_{\hat{Y}^c(t_i)}} \right)^2 + \sum_{t=1}^{n} \left(\frac{\hat{\mu}^c(t_i) - B(t_i)\,\eta}{\sigma_{\hat{\mu}^c(t_i)}} \right)^2 \tag{10.16}$$

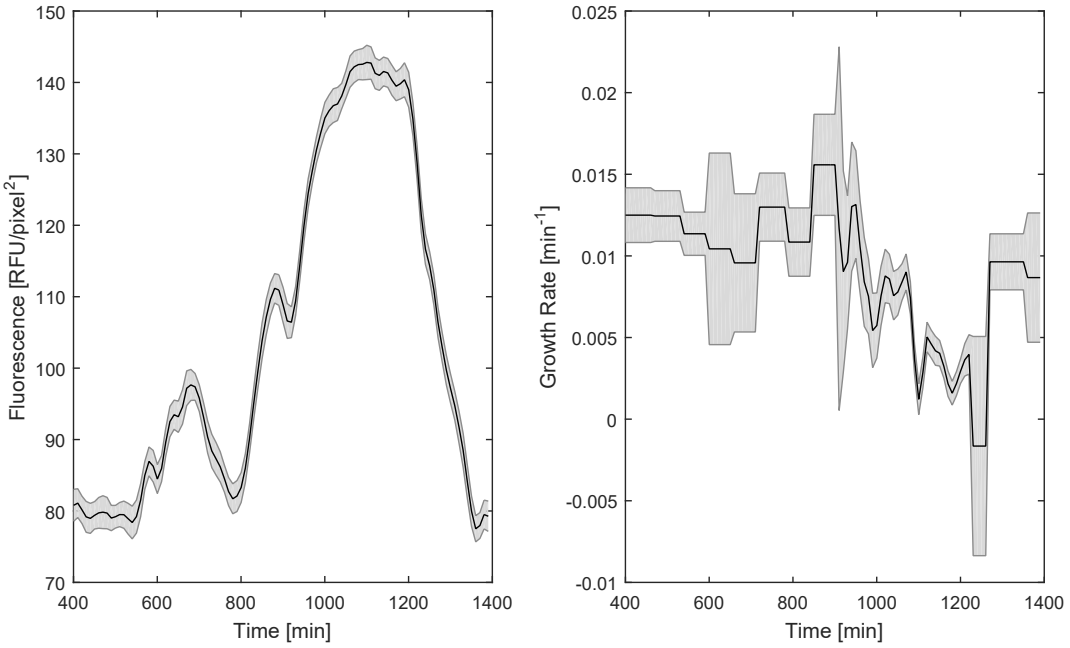


FIGURE 10.3: **Cellular profiles**. Fluorescence profile $\hat{Y}$ (left picture) and growth rate profile (right picture) $\hat{\mu}$ for the cell of Figure 10.2 with their interval of confidence $\pm 2\sigma_{\hat{Y}}$ and $\pm 2\sigma_{\hat{\mu}}$.

10.4 Results

10.4.1 Cellular profiles

We considered data from 20 cells among those experimentally observed in [51]. On this data we performed the data processing (computation of growth rates and of data uncertainties) of Sections 10.3.2-10.3.3 and the calibrations of single-cell and average-cell models of Section 10.3.4. Figure 10.3 shows, as an example, the data analysis of Section 10.3.2 for the cell of Figure 10.2. The growth rate profile shows, for small time windows in-between cellular divisions, a constant value (estimated by linear regression using Equation (10.8)) and for long time window (i.e. from minute to 875 to 1280) a curve modeled by cubic spline (through Equation (10.7)).

From the C profiles, we calculate average profiles using Equations (10.10)-(10.9). The resulting average fluorescence and growth rate profiles are shown in Figure 10.4.

10.4.2 Calibration of the average cell model

We perform the calibration of the average-cell model in Matlab, using fmincon to minimize the objective function (10.15) with respect to the vectors θ and η, in combination with ode45 to solve the dynamics of Equations (10.13)-(10.14). In Figure 10.5, fitting

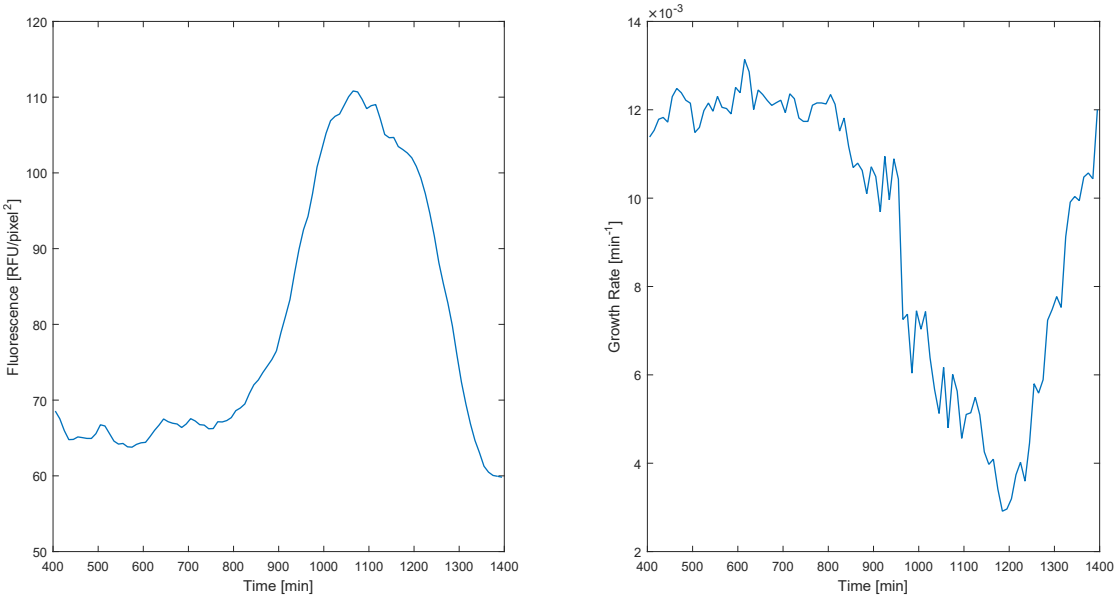


FIGURE 10.4: **Average profile**. Weighted average of fluorescence profiles (left picture) and weighted average of growth rate profiles (right picture). Average from 20 cells.

TABLE 10.1: Resulting parameters from the average-cell calibration.

Average-cell parameters									
Par.	K_1	K_2	K_3	K_4	K_5	K_6	K_{mat}	$R_m(0)$	$L(0)$
Val.	0.0896	6.2272	7.5574	0.1000	2.0002	7.1589	0.0028	0.8957	0.9795

results for the average cell model are shown. The fitted growth rate (input) profile appears to anticipate the observed growth dynamics, which is in agreement with the fact that the corresponding estimated output has to fit fluorescence transitions that appear to anticipate the observed growth-rate transitions. Yet, the model output presents a lower amplitude compared with the fluorescence profile data. Growth rate and fluorescence fits both fall within the data confidence intervals, which are large due to the large cell-to-cell variability.

The parameter values we obtain from the calibration of the average-cell model are shown in Table 10.1.

10.4.3 Calibration of the single-cell models

We perform single-cell calibration in Matlab for every cell c, using fmincon to minimize the objective function in Equation (10.16) with respect to the vectors θ and η and ode45 to solve the dynamics of Equations (10.13-10.14).

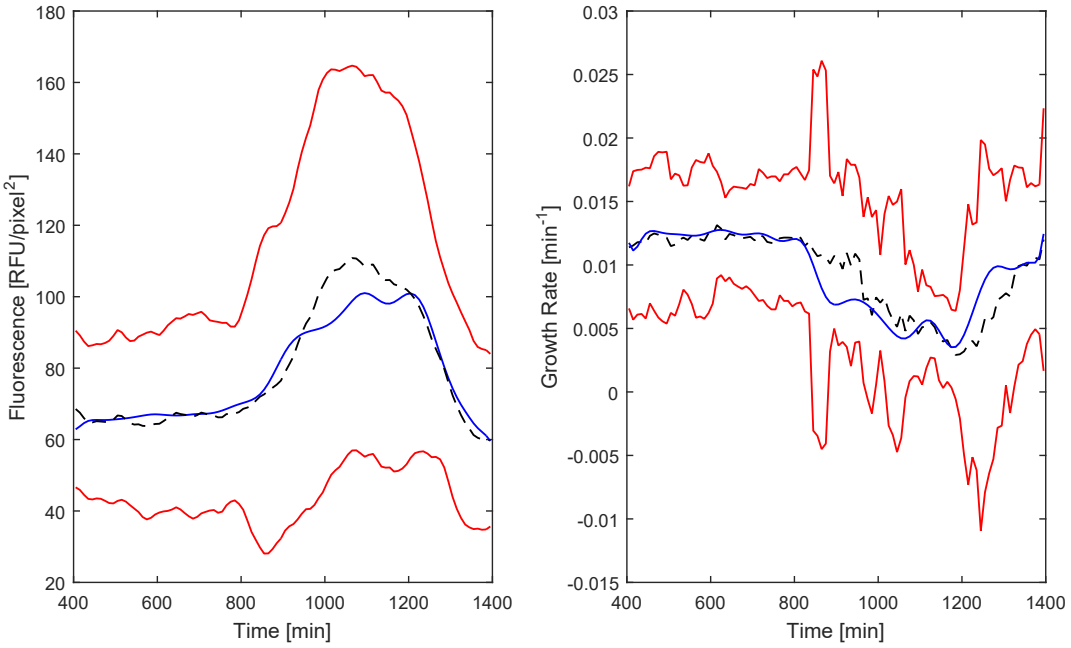


FIGURE 10.5: **Average-cell calibration**. Left (resp. right) picture shows the fluorescence (resp. growth rate) profile of the average cell (black dashed line) with its interval of confidence (red lines) and the curve obtained with the average-cell calibration (blue line).

TABLE 10.2: Statistics of the parameters from the 20 single-cell calibration.

Single-cell parameters									
Par.	K_1	K_2	K_3	K_4	K_5	K_6	K_{mat}	$R_m(0)$	$L(0)$
Mean.	0.1905	6.2578	7.6467	0.0113	0.4632	4.2821	0.0034	0.9508	1.8135
St.d.	0.0738	4.0969	6.6396	0.0247	1.5843	10.7783	0.0034	0.1564	1.6863

The mean and standard deviation of parameter values obtained from the 20 cells are shown in Table 10.2 (for the table containing the parameter values for each single-cell model, see Appendix G).

As an example, Figure 10.6 shows the calibration of the fluorescence and growth rate profiles for the cell of Figure 10.2 and 10.3. The fitted model explains the single-cell data reasonably well, yet with discrepancies due to the fact that single-cell noise (fast fluctuations of the dynamics over time) is not explicitly accounted for by the model. Yet things become interesting when comparing single-cell estimates with the average-cell model calibrated earlier on, as discussed in the next section.

10.4.4 Comparison

To compare the results from the two approaches (single-cell vs. average-cell modeling) we first of all computed the weighted mean of the 20 cell profiles from the single-cell

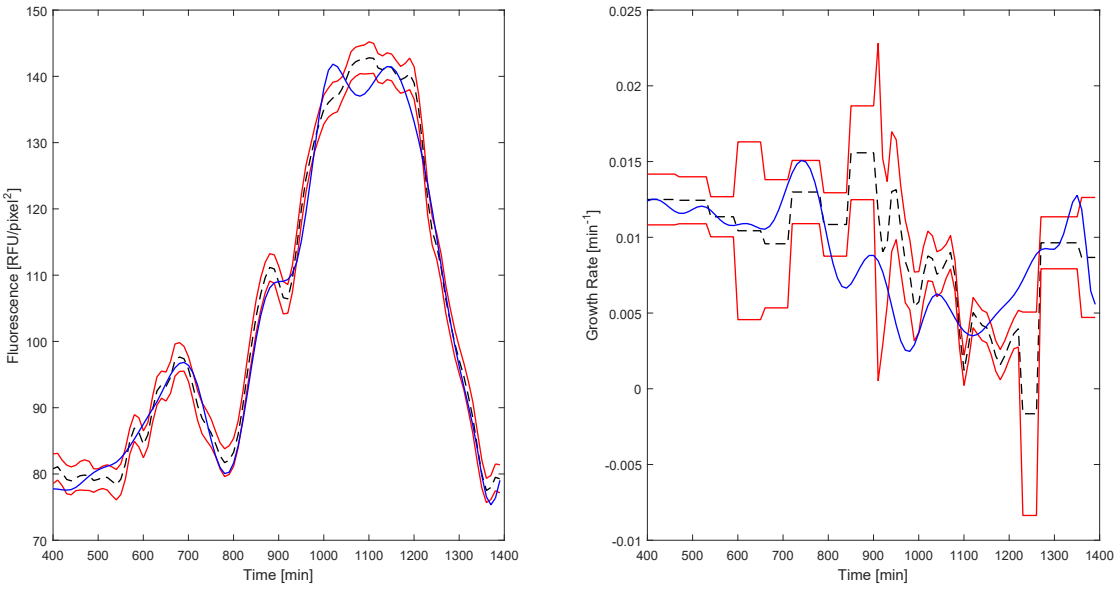


FIGURE 10.6: **Single-cell calibration**. Left (resp. right) picture shows the fluorescence (resp. growth rate) profile of the second cell (black dashed line) with its interval of confidence (red lines) and the curve obtained with single-cell calibration (blue line).

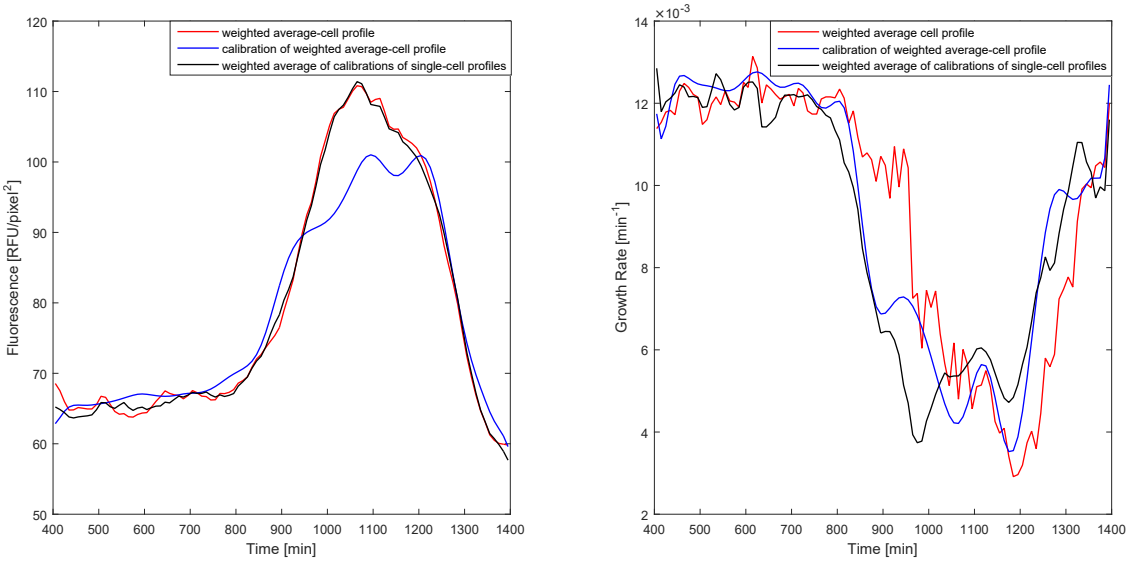


FIGURE 10.7: **Calibration of average-cell vs average of single-cell calibrations**. Left (resp. right) picture shows the weighted average cell profile of the fluorescence (resp. growth rate) in red. The profile given by the average cell model calibration is shown in blue and the weighted average of the 20 profiles given by single-cell model calibrations is shown in black.

model calibration of Section 10.4.1 with the weights computed as in Formulas (10.11)-(10.12). The results are shown in Figure 10.7. The average of single-cell fluorescence (gene expression response) profiles from the fitted single-cell models explains the average data much better than average-cell calibration, whereas both approaches show similar issues in the fit of growth rate profiles, namely, both approaches estimate that growth-rate transitions occur before they actually take place in the data. The improvement on gene

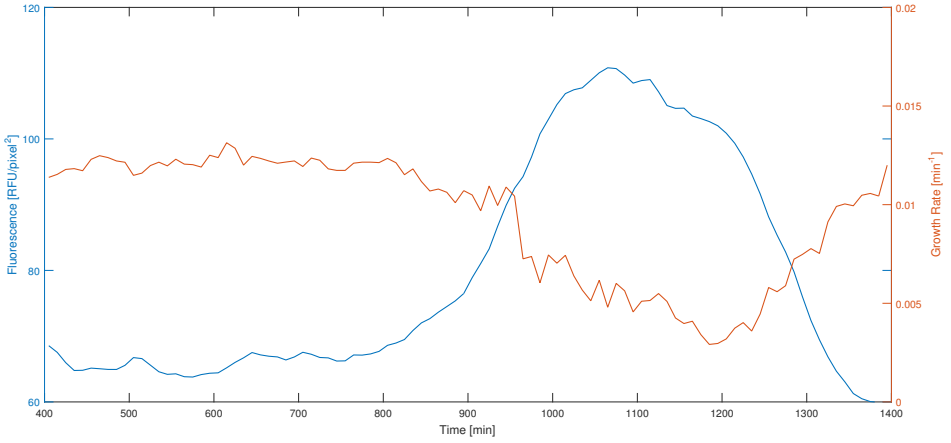


FIGURE 10.8: **Fluorescence and growth rate transitions**. After the removal of IPTG at minute 800, the fluorescence increase seems to occur before a significant decrease of growth rate. Both curves are the average-cell profiles calculated from Equations (10.9)-(10.10).

expression predictions is somewhat surprising and hints that single-cell dynamical modeling is superior. This can be explained by the large single-cell response variability, and by the fact that the average data is not an actual behavior of any of the cells of the population. In other words, the average cell approach is simply an unsuitable approximation in this case. On the other hand, the fact that growth rate estimates from model calibrations show transitions before those observed in the data is qualitatively understood in terms of the very nature of the model, where fluorescence dynamics depend causally on the growth rate profile. Indeed, in the data, fluorescence (gene expression) transitions related with IPTG removal seem to occur before significant growth-rate transitions (see Figure 10.8), which cannot be explained by a causal model where, informally speaking, fluorescence (output) changes can only follow growth rate (input) changes. This prompts for future investigation of nonstandard models where growth rate is not an input but rather, similar to gene expression, a resultant of yet-to-determine regulatory response mechanisms.

Notwithstanding the causality issue, it is interesting to also compare the estimated single-cell model parameters with the parameter calibration for the average model. Scatter plots of the parameter estimates are compared in Figure 10.9. The red dots that represent the parameter values of the average cell calibrated model fall inside the corresponding clouds of single-cell parameter estimates, but they clearly do not coincide with the average of the single-cell estimates. Moreover, for most parameters, these clouds are as large as one order of magnitude relative to the corresponding value for the average model. All this provides additional evidence of the strong variability of individual cell dynamics, and of the importance of explaining population-average data not in terms

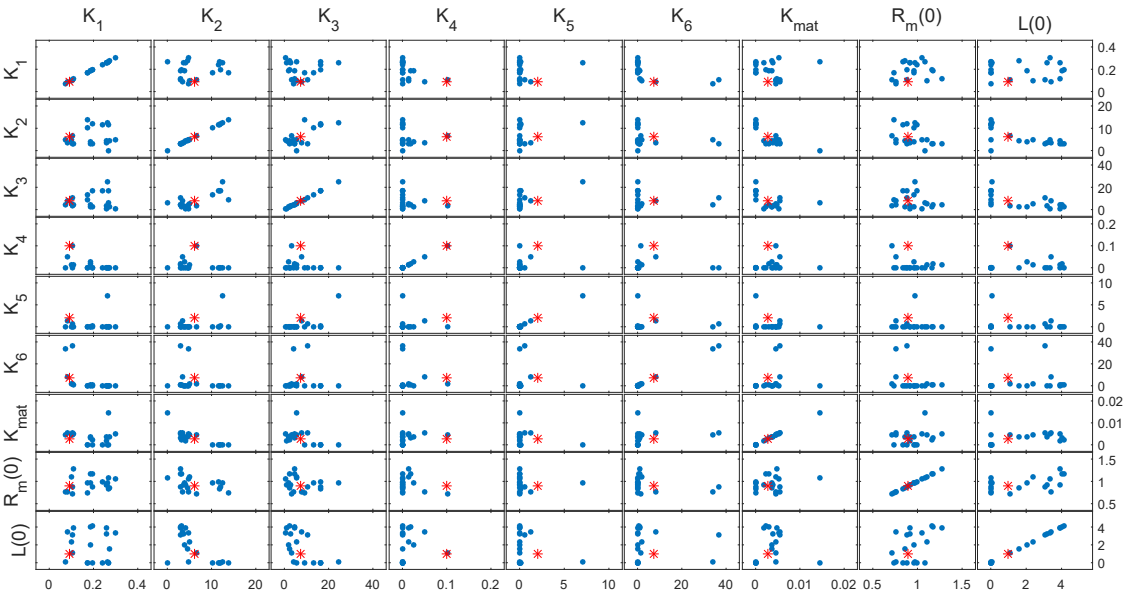


FIGURE 10.9: **Variability of parameter estimates**. Every box shows the scatter plot of the estimates of a pair coefficients from (K_1, K_2, K_3, K_4, K_5, K_6, K_{mat}, $R_m(0)$, $L(0)$). Red stars represent the parameter values of the average cell model and blue dots represent the parameter values of the 20 singe-cell models.

of response of a virtual average cell, but in terms of average response of different individual cells. On the other hand, further investigation is required to quantify practical identifiability of the model parameters [10, 21, 94].

10.5 Conclusion

In this chapter we have studied the problem of modeling growth and gene expression from single-cell growth arrest and restart experiments in *E. coli*. Using the approach from [67, 101], in a model of gene expression where growth rate appears as input profile, we have regarded growth rate data as uncertain and solved the problem of simultaneous estimation of input and dynamical model parameters. The model was first interpreted as a description of an average cell to explain the average growth and fluorescent reporter profiles. Then it was reinterpreted as a description of a single observed cell, and different model calibrations were carried out in correspondence of the different cells.

From the comparison of the two approaches in terms of population-average dynamics, we found that the single-cell calibration approach leads to far better results than the average cell calibration regarding the fitting of the fluorescent reporter output, which we explained by the ability of the single-cell approach to account for the large variability of single-cell responses to the control experiments [42, 75].

However both approaches yield unsatisfactory results regarding the fitting of the system growth rate input. We explained this in terms of the causal relationship between growth and gene expression that is hard-coded in the model.

We argued that this assumption may not be appropriate in the light of the available data and the model inference results. This consideration can lead to possible future research steps on the modeling of the relationships between growth and gene expression. On the other hand, the remaining discrepancies between single-cell model predictions and data motivate further advancements in the modeling of the intrinsic noise (i.e. gene expression randomness) and/or the extrinsic noise (i.e. parameter fluctuations over time) in single cell, and the validation and possible redesign of the models on new experiments. Exploitation of a Mixed-Effects approach to population modeling, as in [75], is another possible development avenue of this work.

Chapter 11

Conclusion and perspectives

The research work presented in this doctoral thesis proposed classical, less classical and new methods to analyze, reduce and calibrate biological models.

In particular we focused on deterministic models describing the gene regulatory network of *E. coli*, but also circadian rhythms in insects and mammals, cellular signaling pathways, and toxicological effects in mice due to pesticides to test our new numerical approach. We were able to obtain useful biological information applying the tools and methods proposed.

11.1 Classical tools for the analysis and reduction of biological models

In the first part of the thesis (Chapter 5) we focused on a system describing RNA Polymerase in *E. coli* and its positive effect on the transcription of its own gene. Based on mass-actions laws a detailed mechanistic model has been written, where every process was accounted for. Because the high dimension of the resulting deterministic model was difficult to handle, we reduced it into a much simpler system by time-scale arguments and studied the mathematical properties of the reduced model. Especially, to investigate the stability of the system we used monotone system theory and concavity properties.

We then showed how the quantity of ribosomes in the bacterium affects the stability of the reduced system: the latter was able to mimic the growth arrest of *E. coli* because of harmful environmental conditions and the restarts of its growth due to favorable conditions, according with [104, 119]. Because the loop proposed is not isolated from the rest of the cell, an interesting perspective could be to extend the model with other mechanisms considered in the GEM model.

In this work we have also modeled the growth rate as a Michaelis-Menten equation in function of the concentration of RNA polymerase: an interesting perspective can be to add IPTG in the system to control the synthesis of RNA polymerase and consequently the bacterial growth rate.

11.2 New tools for the analysis and reduction of biological systems

In the second part of the thesis we presented a new numerical approach, called *principal process analysis* (PPA), which allows to analyze the dynamics of a biological system, focusing on the contribution of each of its processes over time.

The main goal of PPA is model analysis: in Chapter 6, 7, 8, 9 and in Appendix B we applied different visualization tools to summarize information about biological processes in one figure (*boolean, dynamical process, 3-D and heat process maps*). The actual maps can be improved in future works and new maps can be designed.

In Chapter 6 and Appendix B we used PPA for model simplification purpose. In fact, neglecting the *always inactive* processes, we simplify the original model in a reduced model and then, neglecting the *inactive* processes for every time window, we decomposed it into multiple submodels. We compared their dynamics and tested the quality of the reduction by means of global relative errors. To obtain a dynamics closer to the original one, it could be interesting to perform a new parameter calibration of the reduced models. This method can lead to lower global errors and provide an interesting alternative to more classical model reduction approaches.

In Chapter 6 we performed a global sensitivity analysis on the model parameters to test the robustness of PPA that: we could do the same analysis on the initial values of the system.

In Chapter 7, after dividing the initial-condition space in rectangles we performed PPA in each of them and we studied the possible transition of the system solutions. In this way we have tested the robustness of our method with respect to initial conditions on a model of two dimensions, for simplicity reasons. A future perspective can be to verify this method on models of higher dimension and to apply it also to some specific parameters spanning several orders of magnitude. In the work we also showed different pathways that a model, contained in a rectangle, can perform toward other rectangles, based on the vector field on the edge of the rectangles. Knowing the values of each component of the vector field makes it possible to study the most probable pathway used, gaining more information about the system behavior.

In Chapter 8 we introduced a variant of our numerical approach, called *absolute principal process analysis* (APPA): this method uses the absolute values of the processes as a criteria to compare the processes during the system dynamics, contrary to PPA that uses process weights. We apply it to understand the core mechanisms of the proposed toxicological model and to analyze the temporal *activation* of processes for checking the consistencies between the model and the proposed MoA. In a future work, this analysis will help us to refine the model calibration that will be tested through a sensitivity analysis performed on the half-lives of the system variables of interest.

An interesting future application can be to use APPA as a model simplification method: for this application the method has to be tested on different models, as we did for the classical PPA.

In Chapter 9 we performed a three-level PPA, gaining more information about model processes as the usual two-level PPA. In future work we will perform PPA on more levels to have a finer method for model analysis. The choice of the threshold between *inactivity*, *moderately activity* and *fully activity* can be also improved as we did in Chapter 8 for APPA.

Another possible extension of PPA is to apply it on the full coupled system of equations instead of working on each equation separately: this would help to analyze *activities* or *inactivities* of processes shared by several equations.

Furthermore a MATLAB or PYTHON tool can be implemented to perform PPA automatically, choosing different comparison methods, thresholds and visual tools. This tool will allow fast analysis, helping mathematicians and biologists for system analysis, model calibration and model reduction.

In Chapter 6, 7, 8, 9, in Appendix B and in the works of [88, 95] PPA was applied to very different biological models showing its high applicability on the deterministic systems (both ODE and DAE systems). An interesting application could be to design a similar approach to stochastic systems, starting from simple cellular models in presence of intrinsic noise (i.e. random timing of biochemical reactions) or extrinsic noise (i.e. partitioning error in cell division).

11.3 Design and analysis of the gene expression machinery in *E. coli*

In the third part of the thesis (Chapter 9) we presented a model designed by my co-supervisor Delphine Ropers that describes the gene expression machinery in *E. coli*. We

showed my contribution on the mathematical modeling, regarding especially the growth rate and its control by IPTG.

Contrary to the phenomenological functions often used in models of biomass formation, our description is mechanistic and allows to relate the components of the gene expression machinery to the biomass.

The design of the control by IPTG gave interesting results: with the current calibration, the model proposed was able to reproduce the hypersensitivity of the bacterial growth to $\beta\beta'$ concentration observed in the work of [51], although we noticed some imprecision regarding the transition phase. In future works we will apply different designs of the control of IPTG on *E. coli*, adding also as a model parameter the concentration of Lac operon, and we will test which control will give the best results.

We also applied three-level PPA both on the wild-type and engineered strain of *E. coli* to gain knowledge on the core mechanisms of the GEM of *E. coli* under conditions of stress (nutrient downshift and IPTG removal). We will perform other experiments in different conditions to see if the core mechanisms of the system responds as expected.

New model calibrations are also under way, to refine the model predictions in the wild-type and the engineered strains. The results of the proteomics experiments to come within the RESET project will be pivotal for that purpose.

The PPA applied to the current calibrated form of the model already provided us with information on the functioning of important feedback loops. We could analyze the chain of processes that lead to growth arrest in response to nutrient deprivation or IPTG removal. An interesting result is the observation that the stress response elicited by growth arrest following a nutrient downshift does not allow the diversion of cell resources from biomass formation. On the contrary, when IPTG removal is the cause of growth arrest, there is no stress response and cell resources are no longer used for biomass formation. The extension of the model with the metabolism will allow to study this rechanneling of cell resources to the formation of high value product.

Once the parameter calibration will be finished, an interesting idea can be to reduce its structure using both classical and new tools we presented in this thesis.

11.4 Single-cell and average cell calibration of the gene expression machinery control in *E. coli*

In the fourth part of the thesis (Chapter 10) we have studied the problem of modeling growth and gene expression from single-cell growth arrest and restart experiments in

E. coli. We used the approach from [67, 101], in a model of gene expression where growth rate seems to be an input profile of the system. We then decided to generate growth rate data from cell area data with their uncertainty and solved the problem with a simultaneous estimation of input and dynamical model parameters. First we performed the calibration on average cell data that described the average growth and average fluorescence of rapporteur gene, then we performed the calibration on each single cell.

Comparing the two approaches in terms of population-average dynamics, we found the single-cell profile gave a far better fit of the fluorescence reporter output: in fact it took account of the cell-to-cell variability response to the control experiments [42, 75]. Anyway, applying both approaches we were not able to perform a satisfactory fitting of the system input, describing cellular growth rate. This is probably due to an indirect relationship between the fluorescence profile of the rapporteur gene and growth in the model considered.

This problematic can lead to a new modeling of the gene expression of the model and its growth. New perspectives concern the modeling of the intrinsic noise (i.e. gene expression randomness) and/or the extrinsic noise (i.e. parameter fluctuations over time) in single cell, and the validation and possible redesign of the models on new experiments. A possible future step for this work is to apply a Mixed-Effects approach to population modeling, as in [75].

Chapter 12

Conclusion et perspectives (en français)

Le travail de recherche présenté dans cette thèse de doctorat a proposé des méthodes classiques, moins classiques et nouvelles pour analyser, réduire et calibrer les modèles biologiques.

En particulier, nous nous sommes concentrés sur des modèles déterministes décrivant le réseau de régulation des gènes de *E. coli*, mais aussi les rythmes circadiens dans les insectes et les mammifères, les voies de signalisation cellulaire et les effets toxicologiques de pesticides chez les souris, pour tester notre nouvelle approche numérique. Nous avons pu obtenir des informations biologiques utiles en appliquant les outils et les méthodes proposés.

12.1 Outils classiques pour l'analyse et la réduction des modèles biologiques

Dans la première partie de la thèse (Chapitre 5), nous nous sommes concentrés sur un système décrivant la polymérase d'ARN dans *E. coli* et son effet d'activation sur la transcription de son propre gène. Sur la base des lois d'actions de masse, un modèle mécaniste détaillé a été écrit, où chaque processus a été pris en compte. Étant donné que la dimension élevée du modèle déterministe était difficile à gérer, nous l'avons réduit dans un système beaucoup plus simple par des arguments d'échelle de temps et étudié les propriétés mathématiques du modèle réduit. En particulier, pour étudier la stabilité du système, nous avons utilisé la théorie des systèmes monotones et les propriétés de concavité.

Nous avons ensuite montré comment la quantité de ribosomes dans la bactérie affecte la stabilité du système réduit: ce dernier a pu simuler l'arrêt de croissance de *E. coli* en raison de conditions environnementales nuisibles et des redémarrages de sa croissance en raison de conditions favorables, selon [104, 119]. Étant donné que la boucle proposée n'est pas isolée du reste de la cellule, une perspective intéressante pourrait être d'étendre le modèle avec d'autres mécanismes considérés dans le modèle GEM.

Dans ce travail, nous avons également modélisé le taux de croissance comme une équation de Michaelis-Menten en fonction de la concentration de l'ARN polymérase: une perspective intéressante peut être d'ajouter l'IPTG dans le système pour contrôler la synthèse de l'ARN polymérase et par conséquent le taux de croissance bactérienne.

12.2 Nouveaux outils pour l'analyse et la réduction des systèmes biologiques

Dans la deuxième partie de la thèse, nous avons présenté une nouvelle approche numérique, appelée *analyse de processus principaux* (PPA), qui permet d'analyser la dynamique d'un système biologique, en mettant l'accent sur la contribution de chacun des processus dans le temps.

L'objectif principal de PPA est l'analyse de modèle: dans le Chapitre 6, 7, 8, 9 et dans l'Annexe B, nous avons appliqué différents outils de visualisation pour résumer les informations sur les processus biologiques dans une figure (*carte booléenne, carte dynamique, carte 3-D et carte de chaleur de processus*). Ces cartes peuvent être améliorées dans les travaux futurs et les nouvelles cartes peuvent être conçues.

Dans le Chapitre 6 et Annexe B, nous avons utilisé la PPA pour simplifier le modèle. En fait, en négligeant les processus *toujours inactifs*, nous simplifions le modèle original e un modèle réduit et, négligeant les processus *inactive* pour chaque fenêtre de temps, nous l'avons décomposé en plusieurs sous-modèles. Nous avons comparé leur dynamique et testé la qualité de la réduction au moyen d'erreurs relatives globales. Pour obtenir une dynamique plus proche de l'original, il pourrait être intéressant d'effectuer une nouvelle calibration des paramètres des modèles réduits. Cette méthode peut entraîner des erreurs globales plus faibles et constituer une alternative intéressante aux approches de réduction de modèles plus classiques.

Dans le Chapitre 6, nous avons effectué une analyse de sensibilité globale sur les paramètres du modèle pour tester la robustesse de la PPA: nous pourrions effectuer la même analyse sur les valeurs initiales du système.

Dans le Chapitre 7, après avoir divisé l'espace de condition initiale en rectangles, nous avons effectué la PPA dans chacun d'eux et nous avons étudié la transition possible des solutions du système. Ainsi, nous avons testé la robustesse de notre méthode par rapport aux conditions initiales sur un modèle à deux dimensions, pour des raisons de simplicité. Une perspective future peut être de vérifier cette méthode sur des modèles de dimension supérieure et de l'appliquer aussi à certains paramètres spécifiques couvrant plusieurs ordres de grandeur. Dans le travail, nous avons également montré des chemins différents selon lesquels un modèle, à partir d'un rectangle, peut avoir des transitions vers d'autres rectangles, en fonction du champ de vecteurs sur le bord des rectangles. La connaissance des valeurs de chaque composante du champ de vecteurs permet d'étudier la voie la plus probable utilisée, en obtenant plus d'informations sur le comportement du système.

Dans le Chapitre 8, nous avons introduit une variante de notre approche numérique, appelée *analyse absolue de processus principaux* (APPA): cette méthode utilise les valeurs absolues des processus en tant que critère pour comparer les processus pendant le système dynamique, contrairement à PPA qui utilise des poids de processus. Nous l'appliquons pour comprendre les mécanismes fondamentaux du modèle toxicologique proposé et pour analyser les processus temporels de traitement des contraintes entre le modèle et le MoA (*mode of action*) proposé. Dans un travail futur, cette analyse nous aidera à affiner la calibration du modèle qui sera testée grâce à une analyse de sensibilité effectuée sur les demi-vies des variables du système d'intérêt.

Une application future intéressante peut être d'utiliser APPA comme méthode de simplification du modèle: pour cette application, la méthode doit être testée sur différents modèles, comme nous l'avons fait pour la PPA classique.

Dans le Chapitre 9, nous avons effectué une PPA à trois niveaux, en obtenant plus d'informations sur les processus modèles qu'avec la PPA de deux niveaux habituel. Dans un travail futur, nous effectuerons la PPA avec plus de niveaux pour avoir une méthode plus fine pour l'analyse des modèles. Le choix du seuil entre *inactivité*, *modérément activité* et *pleine activité* peut également être amélioré comme nous l'avons fait dans le Chapitre 8 pour l'APPA.

Une autre extension possible de la PPA est de l'appliquer sur le système couplé d'équations au lieu de travailler séparément sur chaque équation: cela aiderait à analyser l'*activité* ou l'*inactivité* des processus partagés par plusieurs équations.

De plus, un outil MATLAB ou PYTHON peut être implémenté pour effectuer automatiquement la PPA, en choisissant différentes méthodes de comparaison, seuils et outils

visuels. Cet outil permettra une analyse rapide, aidant les mathématiciens et les biologistes pour l'analyse du système, la calibration du modèle et la réduction du modèle.

Dans le Chapitre 6, 7, 8, 9, en Annexe B et dans les travaux de [88, 95] la PPA a été appliquée à des modèles biologiques très différents montrant sa grande utilité sur les systèmes déterministes (systèmes ODE et DAE). Une application intéressante pourrait être de concevoir une approche similaire à celle des systèmes stochastiques, à partir de modèles cellulaires simples en présence de bruit intrinsèque (c'est-à-dire du au hasard des réactions biochimiques) ou du bruit extrinsèque (c'est-à-dire l'erreur de partitionnement dans la division cellulaire).

12.3 Modélisation et analyse du mécanisme d'expression des gènes dans *E. coli*

Dans la troisième partie de la thèse (Chapitre 9), nous avons présenté un modèle conçu par ma co-encadrante Delphine Ropers qui décrit le mécanisme d'expression de gènes dans *E. coli*. Nous avons montré ma contribution sur la modélisation mathématique, en particulier le taux de croissance et son contrôle par IPTG.

Contrairement aux fonctions phénoménologiques souvent utilisées dans les modèles de formation de biomasse, notre description est mécaniste et permet de relier les composantes de la machinerie d'expression des gènes à la biomasse.

La conception du contrôle par IPTG a donné des résultats intéressants: avec la calibration actuel, le modèle proposé a été capable de reproduire l'hypersensibilité de la croissance bactérienne à la concentration $\beta\beta'$ observée dans le travail de [51], bien que nous avons constaté une certaine imprécision quant à la phase de transition. Dans les travaux futurs, nous appliquerons différentes conceptions du contrôle d'IPTG sur *E. coli*, ajoutant également comme paramètre la concentration de Operon Lac, et nous testerons quel contrôle donne les meilleurs résultats.

Nous avons également appliqué une PPA à trois niveaux à la fois sur la souche de type sauvage et modifiée de *E. coli* pour acquérir des connaissances sur les mécanismes de base du GEM de *E. coli* dans des conditions de stress (baisse des éléments nutritifs et suppression de l'IPTG). Nous effectuerons d'autres expériences dans différentes conditions pour voir si les mécanismes de base du système répondent comme prévu.

Des calibrations de nouveaux modèles sont également en cours, afin d'affiner les prédictions du modèle dans les souches de type sauvage et de génie. Les résultats des expériences de protéomique dans le cadre du projet RESET seront essentiels à cette fin.

La PPA appliqué à la forme calibrée actuelle du modèle nous a déjà fourni des informations sur le fonctionnement des boucles de rétroaction importantes. Nous pourrions analyser la chaîne de processus qui conduit à un arrêt de la croissance en réponse à la privation de nutriments ou à l'élimination de l'IPTG. Un résultat intéressant est l'observation selon laquelle la réponse au stress provoquée par l'arrêt de la croissance suite à une baisse des éléments nutritifs ne permet pas le détournement des ressources cellulaires de la formation de biomasse. Au contraire, lorsque l'élimination de l'IPTG est la cause de l'arrêt de la croissance, il n'y a pas de réponse au stress et les ressources cellulaires ne sont plus utilisées pour la formation de biomasse. L'extension du modèle avec le métabolisme permettra d'étudier cette réallocation des ressources cellulaires pour la formation d'un produit à haute valeur ajoutée. Une fois la calibration du paramètre terminé, une idée intéressante peut être de réduire sa structure à l'aide des outils (classiques et nouveaux) que nous avons présentés dans cette thèse.

12.4 Calibration d'un modèle de contrôle de la machinerie d'expression des gènes dans *E. coli* en utilisant les profils de la cellule individuelle et la moyenne

Dans la quatrième partie de la thèse (Chapitre 10), nous avons étudié le problème de la croissance de la modélisation et de l'expression des gènes à partir de l'arrêt de la croissance d'une cellule et des expériences de redémarrage dans *E. coli*. Nous avons utilisé l'approche de [67, 101], dans un modèle d'expression génique où le taux de croissance semble être un profil d'entrée du système. Nous avons ensuite décidé de générer des données de taux de croissance à partir des données de la surface cellulaire avec leur incertitude et avons résolu le problème avec une estimation simultanée des paramètres de l'entrée et du modèle dynamique. D'abord, nous avons effectué la calibration des données cellulaires moyennes qui décrivent la croissance moyenne et la fluorescence moyenne du gène rapporteur, puis nous avons effectué la calibration sur chaque cellule.

En comparant les deux approches en termes de dynamique moyenne de la population, nous avons constaté que le profil d'une seule cellule donnait une meilleure calibration de la sortie du modèle (la fluorescence du gène rapporteur): en fait, il tient compte de la réponse de la variabilité entre les cellules dans les expériences de contrôle [42, 75]. Quoi qu'il en soit, en appliquant les deux approches, nous n'avons pas été en mesure d'effectuer un ajustement satisfaisant de l'entrée du système, en décrivant le taux de croissance cellulaire. Ceci est probablement dû à une relation indirecte entre le profil de fluorescence du gène rapporteur et la croissance du modèle considéré.

Cette problématique peut conduire à une nouvelle modélisation de l'expression génique du modèle et de sa croissance. De nouveaux paramètres décrivent la modélisation du bruit intrinsèque (c'est-à-dire les fluctuations de l'expression du gène) et/ou le bruit extrinsèque (c'est-à-dire les fluctuations des paramètres au fil du temps) dans une seule cellule; la validation et la calibration de ces modèles sont possibles avec de nouvelles expériences. Une étape future possible pour ce travail est d'appliquer une approche à effets mixtes à la modélisation de la population, comme dans [75].

Appendix A

List of publications

1. S. Casagranda, D. Ropers and J.-L. Gouzé. Model reduction and process analysis of biological models. In *2015 23rd Mediterranean Conference on Control and Automation (MED), IEEE, 2015, pp.1132-1139.*

2. S. Casagranda and J.-L. Gouzé. Principal Process Analysis and reduction of biological models with order of magnitude. In *Proceedings of the 20th IFAC World Congress, 2017.*

3. I. Belgacem, S. Casagranda, E. Grac, D. Ropers and J.-L. Gouzé. Reduction and stability analysis of a transcription-translation model of RNA polymerase. *Bulletin of Mathematical Biology*, 2016. (Submitted).

4. S. Casagranda, S. Touzeau, D. Ropers and J.-L. Gouzé. Principal Process Analysis of biological models. *Journal of Theoretical Biology*, 2016. (Submitted).

The work contained in Chapters 8, 9, 10 will be a consistent part of three different future journal papers.

The works were also exposed at *The 5th International Symposium on Positive Systems (POSTA)* in Rome (Italy) in 2016 and at *Journées annuelles 2017 du GT Bioss* in Montpellier (France) in 2017.

Appendix B

First application of principal process analysis on biological models

In this chapter we discuss a work that has been presented at the 23rd *Mediterranean Conference on Control and Automation* MED, held in Torremolinos, Spain, on June 16th-19th, 2015 (with peer reviewed proceedings) and has been accepted as a conference paper in which I am first author (see Appendix A).

We do not insert these sections in a regular chapter to avoid redundancy with Chapter 6.

We present the first application of our method called *Principal Process Analysis*, presented in Chapter 6, that is able to analyze key processes for a dynamical network of high dimension and that is based on *a priori* knowledge of the system trajectory and the simplification of the mathematical model. The method consists of the model decomposition into biologically meaningful processes, whose *activity* or *inactivity* is evaluated during the time evolution of the system. The structure of the model is reduced to the core mechanisms involving active processes only. We assess the quality of the reduction by means of global relative errors and apply our method to two models of the circadian rhythm in *Drosophila* [72] and the influence of RKIP on the ERK signaling pathway [68].

B.1 Methodology

We describe below the basics of the method. We will use as a running example the 8^{th} variable of the *Drosophila* circadian clock model [72] (see Section B.2 and the model equations in Section B.5), which describes concentration changes of the double phosphorylated form of protein TIM ($T_2 = x_8$):

$$\frac{dT_2}{dt} = V_{3T}\frac{T_1}{K_{3T}+T_1} - V_{4T}\frac{T_2}{K_{4T}+T_2} - k_3P_2T_2 + k_4C - v_{dt}\frac{T_2}{K_{dT}+T_2} - k_dT_2 \quad \text{(B.1)}$$

B.1.1 Principal process analysis (PPA)

Consider the following ODE model of biological network:

$$\dot{x} = f\,(x,p) \quad \text{(B.2)}$$

where $x = (x_1, x_2, ..., x_n)\ \epsilon\ \mathbb{R}^n$ is the vector of component concentrations, $x0 = (x0_1, x0_2, ..., x0_n)\ \epsilon\ \mathbb{R}^n$ the vector of their initial values and $p\ \epsilon\ \mathbb{R}^b$ the vector of parameters. Each equation is decomposed into a sum of biological processes:

$$\dot{x}_i = \sum_j f_{ij}\,(x,p) \quad \text{(B.3)}$$

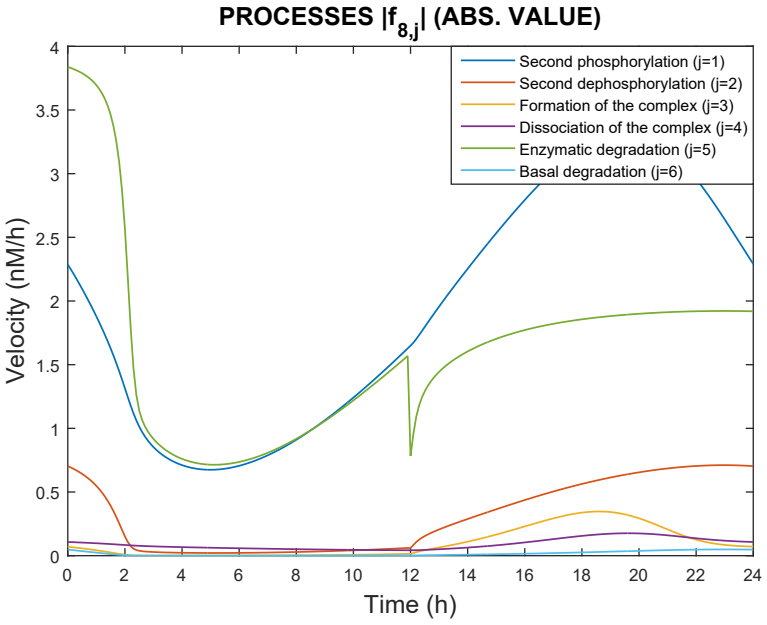
where f_{ij} represents the j^{th} process involved in the dynamical evolution of the i^{th} variable of the system over a period of time [0,T].

Example: *Equation (B.1) includes seven processes, each associated with a specific biological function. They take a positive or negative value, depending on whether they affect positively or negatively the variation of T_2 concentration. The equation of the protein is rewritten as:*

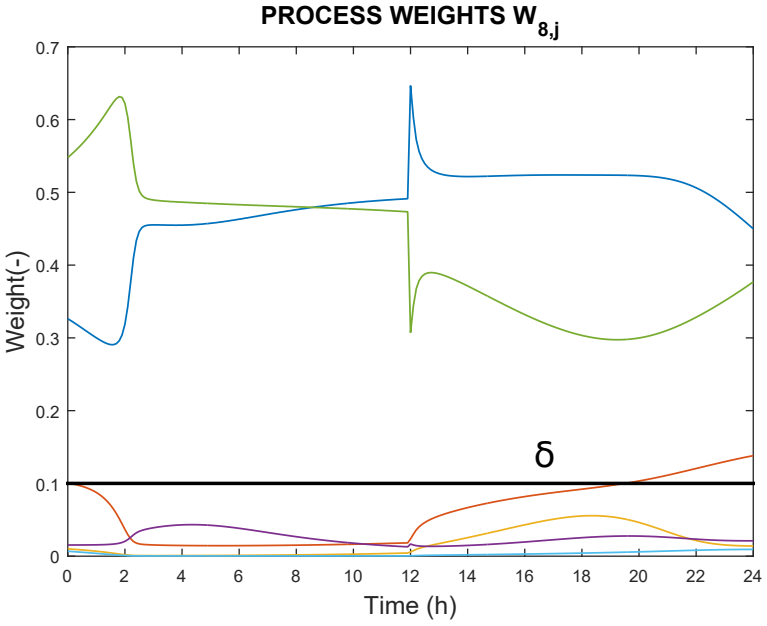
$$\dot{x}_8 = f_{8,1} + f_{8,2} + f_{8,3} + f_{8,4} + f_{8,5} + f_{8,6} \quad \text{(B.4)}$$

where $f_{8,1} = V_{3T}\frac{T_1}{K_{3T}+T_1}, ..., f_{8,6} = -k_dT_2$.

Comparison criteria are needed to weigh the influence of the different processes f_{ij} on the time evolution of each variable x_i. There are several alternatives. For instance, we can compare their absolute value $\left(|f_{ij}(x,p)|\right)$, scale it by the i^{th} initial condition $\left(\frac{|f_{ij}(x(t),p)|}{x_{0i}}\right)$, or scale it by the solution of the i^{th} ODE $\left(\frac{|f_{ij}(x(t),p)|}{x(t)_i}\right)$. In this work we



(A)



(B)

FIGURE B.1: Dynamics of processes that change the concentration of the double phosphorylated protein TIM (T_2, see Equations (B.1) and (B.4)) over a 24-hour time window. B.1a: Absolute value of the processes along time (one color per process). B.1b: Weights associated with the processes along time. The threshold δ is set at 0.1.

associate a relative weight to each process to make it dimensionless:

$$W_{ij}(t,p) = \frac{|f_{ij}(x(t),p)|}{\sum_j |f_{ij}(x(t),p)|} \tag{B.5}$$

where $0 \leq W_{ij}(t,p) \leq 1$ and $\sum_j W_{ij}(t,p) = 1$.

Definition: *Let the continuous function* $f_{ij}(x(t),p)$ *be the* j^{th} *process of* $\dot{x}_i(t)$ *in* $t \in [0,T]$ *and let the threshold* $\delta \in [0,1]$.
We call a process $f_{ij}(x(t),p)$ *always inactive when* $W_{ij}(t,p) < \delta \ \forall \ t \in [0,T]$.
We call a process $f_{ij}(x(t),p)$ *inactive at time* t *when* $W_{ij}(t,p) < \delta$.
We call a process $f_{ij}(x(t),p)$ *active at time* t *when* $W_{ij}(t,p) \geq \delta$.
Switching time for a process $f_{ij}(x(t),p)$ *is the time* t^s_{ij} *when* $W_{ij}(t,p) = \delta$. *A process can have* $0,1,...,z$ *switching times.*
The switching time set S_i *for the* i^{th} *variable contains all the switching times* t^s_{ij} *where* $j = 1,..,k$ *and* $s = 1,...,z$.
The global switching time set S *is the union of all* S_i.
Example: *We set* $\delta = 0.1$ *and apply Equation (B.5) to the six processes of Equation (B.1). We obtain their dynamical weight:*

- *the weight of processes* $W_{8,3}$, $W_{8,4}$, $W_{8,6}$ *is always below* δ *and their related processes* $f_{8,3}$, $f_{8,4}$, $f_{8,6}$ *are always inactive;*
- *the processes* $W_{8,1}$ *and* $W_{8,5}$ *are always above* δ *and* $f_{8,1}$ *and* $f_{8,5}$ *are active during the whole dynamics;*
- *the weight* $W_{8,2}$ *of the process* $f_{8,2}$ *crosses the threshold twice and the switching times* $t^1_{8,2} = 0.2h$ *and* $t^2_{8,2} = 19.8h$ *are collected in the set* S_8. *Figure B.1a shows the dynamics of the absolute values of processes involved in Equation (B.1) during a day, while relative weights defined in (B.5) are shown in Figure B.1b.*

B.1.2 Visualization of process activities

For models as complex as the mammalian circadian clock model, it is advantageous to qualitatively visualize process *activities* or *inactivities* during the system dynamics. PPA allows to visually summarize this information in one figure with the help of graphical tools. They are described below.
Boolean Process Map: shows the time-dependent activity of processes, ordered by variables, during the whole system dynamics $[t_0, T]$. *Active*, resp. *inactive*, processes are depicted by a white, resp. black, bar.
Dynamical Process Map: displays the activity of processes and their interactions with

variables. In this map, we distinguish three types of process activity to take account of system equations sharing common processes. Variables (represented by boxes) are connected by processes (arrows), which can be either *inactive* (shown in black), *active* for all the variables involved (red) or, *active* for some variables involved and *inactive* for the others (yellow).

3-D Process Map: depicts qualitatively for each process, the time-dependent evolution of its intensity. Process *activities* are averaged per hour, leading to the discretization of time. Vertical bars represent process weights for each hour. Their color code represents the intensity of process weights relatively to the other weights.

Heat Process Map: it allows to study qualitatively the evolution of the intensity of the *active* process activity using colors. Values along the rows (the absolute process values $W_{i,j}(t,p)$) are standardized

$$SW_{i,j}(t,p) = \frac{W_{i,j}(t,p) - \bar{W}_{i,j}}{\sigma_{Wi,j}}, \tag{B.6}$$

where $\bar{W}_{i,j}$ is the mean of all the values of the process weights $SW_{i,j}(t,p)$ in the time window $[0,T]$ and $\sigma_{Wi,j}$ its standard deviation. The standardized process weights assume a red color (active) if their value at a generic instant of time $t = \tau\epsilon[0,T]$ is above the mean, black if their value is equal to the mean and green (inactive) if their value is below the mean of a column across all rows (the mean of all the standardized process weights $SW_{i,j}(t,p)$ at time τ). If the red (or green) color is lighter it means that the process is more *active* (*inactive*).

B.1.3 First model reduction

The first step of PPA identifies *always inactive* processes and remove them from the original system.

The threshold value δ must be chosen in the range [0,1], preferentially at a low value to avoid neglecting important processes. Otherwise the dynamics of the new system would change significantly. The objective is to obtain $g(x^r)$, the function approximating $f(x)$ and including less processes.

We introduce the ODE system (B.7), which approximates system (B.2):

$$\dot{x}^r = g\left(x^r, p^r\right) \tag{B.7}$$

where $x^r = (x_1^r, x_2^r, ..., x_n^r)\ \epsilon\ \mathbb{R}^n$ is the vector of component concentrations, $x0 = (x0_1, x0_2, ..., x0_n)\ \epsilon\ \mathbb{R}^n$ the vector of their initial values, and $p^r\ \epsilon\ \mathbb{R}^c$, where $c \leq b$ is

the vector of parameters. The model reduction approach relies basically on the following theorem: if the vector fields of two systems are close ($f(x) \approx g(x)$), then the solutions of the original and approximated systems are close during some time interval under the assumptions on the Lipschitz conditions listed in [57, p. 96, Th. 3.4].

At this stage, dynamical weights have been assigned to every process and a value has been set for the threshold δ. We can now apply the following rule to define $g(x^r, p^r)$:
if $W_{ij}(x(t), p) < \delta \ \forall \ t \in [0,T]$ *then* $g_{ij} = 0$;
if not, $g_{ij} \equiv f_{ij}$.
We thus define x^r as an approximation of x and p^r as a subset of p.
Example: *Because* $f_{8,3}$, $f_{8,4}$, $f_{8,6}$ *are always inactive,* $g_{8,3} = 0$, $g_{8,4} = 0$, $g_{8,6} = 0$ *and* $g_{8,1} \equiv f_{8,1}$, $g_{8,2} \equiv f_{8,2}$, $g_{8,5} \equiv f_{8,5}$. *The resulting ODE for* x_8^r *is:*

$$\frac{dT_2^r}{dt} = V_{3T}\frac{T_1^r}{K_{3T} + T_1^r} - V_{4T}\frac{T_2^r}{K_{4T} + T_2^r} - v_{dt}\frac{T_2^r}{K_{dT} + T_2^r} \tag{B.8}$$

To assess the quality of the reduced model $g(x^r)$, we numerically compute the global relative error between the original and the reduced models on the six outputs of the system: the concentrations of *Per* mRNA (M_P), *Tim* mRNA (M_T), total PER protein (P_{Tot}), total TIM protein (T_{Tot}), cytosolic complex (C), nuclear complex (C_N) [1]. If y_h and y_h^r are the h^{th} output of the original and the reduced systems respectively, one possible form of global relative error is:

$$e_h = \frac{\int |y_h(t) - y_h^r(t)|dt}{\int |y_h(t)|dt} \tag{B.9}$$

B.1.4 Creation of chains of sub-models

The second step of PPA consists in defining sub-models. The time period during which the system evolves can be split into time intervals using the switching times t_b (with $b = 1, \ldots, d$) previously grouped in set S and sorted in ascending order: this allows creating a succession of sub-models for each time window, which contain the core mechanisms in that period of time. To avoid large chains of sub-models, we reduce the number of time windows by grouping closer switching times with the easy-to-compute k-means clustering [54]. Given our global switching time set $S = [t_1, t_2, ..., t_d]$, this leads us to group the d switching times into z ($\leq d$) clusters C=$\{C_1, C_2, ..., C_z\}$, so as to minimize

[1] The outputs P_{Tot} and T_{Tot} are: $P_{Tot} = P_0 + P_1 + P_2$, $T_{Tot} = T_0 + T_1 + T_2$.

the within-cluster sum of square (or within-cluster inertia):

$$\text{argmin}_C \sum_{v=1}^{z} \sum_{t \epsilon C_v} ||t - \mu_v||^2 \tag{B.10}$$

where μ_v is the mean of the switching times in C_v. We assume that processes with a switching time in cluster C_v switch together at time $t_v^r = \mu_v$, the mean switching time in cluster C_v. There is no precise rule to choose the number of clusters z, but it can be related to the difference between the maximum and minimum number of *active* processes during the time evolution of the system: if the difference is low, z should be chosen low as well. Such an approach could be:

$$z \approx \frac{\max_v(n_{act}^v) - \min_v(n_{act}^v)}{2}, \tag{B.11}$$

where n_{act}^v denotes the number of active processes in the v^{th} time window.

We eventually end up with a chain of $z+1$ sub-models in the time interval $[0, T]$, the first one being valid in $[0, t_1^r]$, while the last is valid in $[t_z^r, T]$. To test the quality of this second model reduction in each time window, we compute the error (B.9) between the original model and each sub-model. The global error can be calculated with or without the propagation error: in the first case, for each time window $[t_{v-1}^r, t_v^r]$ ($v = 1, \ldots, z+1$ with $t_0^r = 0$ and $t_{z+1}^r = T$), the initial values of the h outputs of the sub-model SM_v are equal to the final values at t_{v-1}^r of the sub-model SM_{v-1}; in the second case, they are equal to the values of the original model at t_{v-1}^r.

B.2 Model for circadian rhythms in *Drosophila*

B.2.1 Description

We consider the model in [72] given by an ODE system of dimension 10. The model describes the circadian oscillations of the proteins timeless (TIM) and period (PER), which involve a negative feedback loop: the double phosphorylated forms of these proteins can be degraded in the cytoplasm or form the PER-TIM complex which, following its transport to the nucleus, inhibits the transcription of the *Tim* and *Per* genes and the subsequent accumulation of their mRNAs and proteins. The light sets the period of the oscillations to 24 hours precisely, by increasing the velocity of the degradation of the double phosphorylated form of TIM (from 2 nM/h to 4 nM/h in the model). Since the same oscillatory behavior is repeated every day, we focus our analysis on 24 hours only and use the model parameters given in Figures 2 and 4 of [72].

B.2.2 Model reduction

After having decomposed every kinetic equation of the model into a set of processes that we consider to be biologically relevant, we simulate their relative weight as shown before for (B.1) in Figure B.1. Every time that the weight of a process crosses the threshold δ, the corresponding process changes its state. The timing of the process *activation/inactivation* in each equation and the *active/inactive* processes along time is conveniently shown in a *Boolean Process Map*, as displayed in Figure B.2.
From this analysis, we obtain a reduced model by neglecting the processes that are

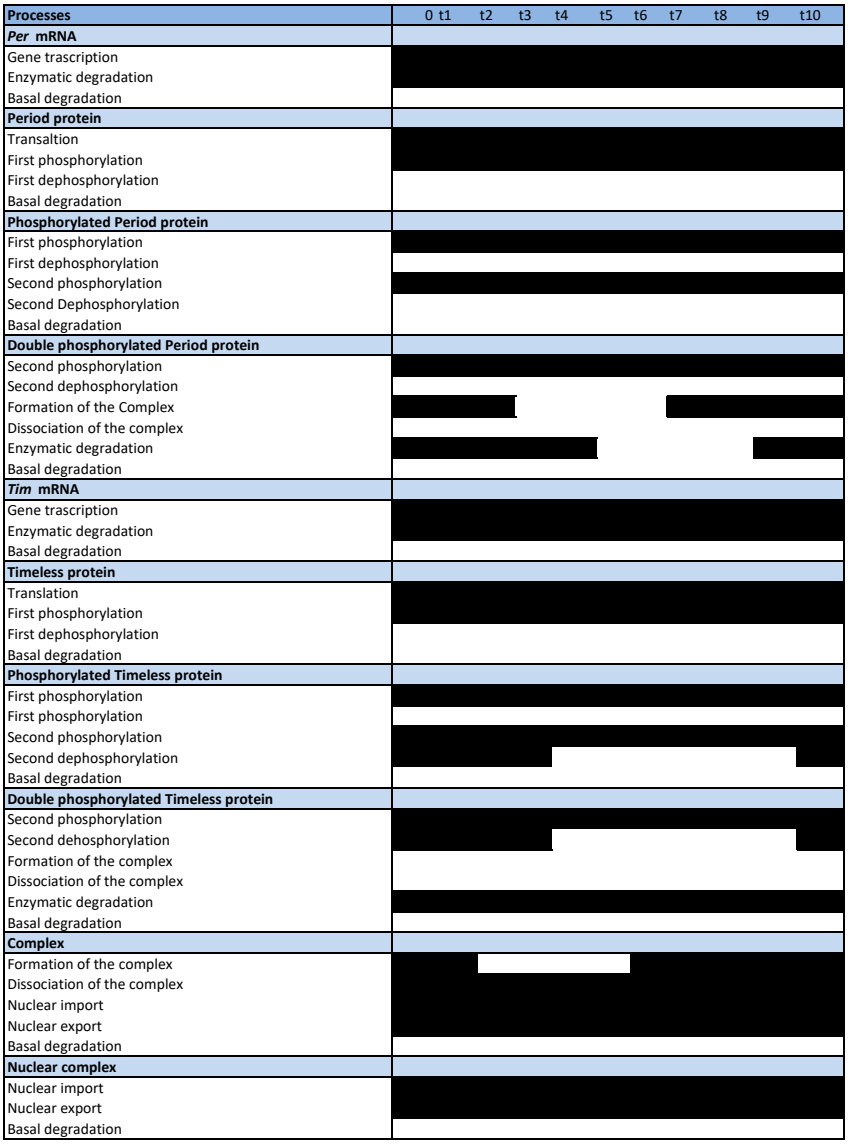


FIGURE B.2: *Boolean Process Map* of the process *activity* for the ODE model in *Drosophila* [72]. Black: *active* state. White: *inactive* state. Times t_d are, in the order, 0.2, 0.5, 1.3, 1.8, 6, 12.3, 15.8, 19.6, 19.8, and 21.5 h. The total duration of the simulation is 24 h.

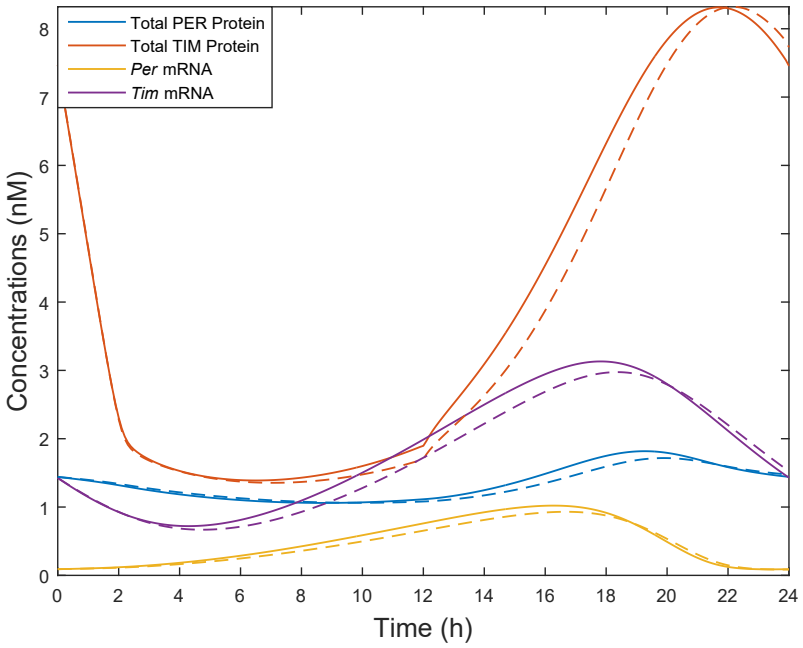


FIGURE B.3: Output dynamics for the original and reduced model. The firsts are represented with continuous lines while the seconds with dotted lines.

always inactive (under the threshold $\delta = 0.1$ during the whole time). We hence neglect 18 of 44 processes: $f_{1,3}$, $f_{2,3}$, $f_{2,4}$, $f_{3,2}$, $f_{3,4}$, $f_{3,5}$, $f_{4,2}$, $f_{4,6}$, $f_{5,3}$, $f_{6,3}$, $f_{6,4}$, $f_{7,2}$, $f_{7,5}$, $f_{8,3}$, $f_{8,4}$, $f_{8,6}$, $f_{9,5}$, $f_{10,3}$. To test the quality of the reduced model, we calculate the global relative error between the original and the reduced models for each variable using (B.9). The results in Table B.1 show a good match between the original model and the reduced one. The dynamics of the system outputs for the original and the first reduced model are shown in Figure B.3.

B.2.3 Qualitative tool: heat process map

In the previous analysis we used a Boolean approach to study the process *activity* or *inactivity* by means of the *Boolean Process Map*. We also have considered alternative graphical representations in this paper, since we also want to qualitatively understand, in one graph, the change over time of the intensity of the *activity/inactivity* of important biological processes. For the circadian model, we chose an *Heat Process Map* as a qualitative approach, where the individual values contained in a matrix are represented using colors. The *Heat Process Map* for the reduced model $g(x^r)$ is shown in Figure B.4.

TABLE B.1: **Global relative Errors for the reduced *Drosophila* model**

Output	G. Rel. Err. (%)
Period mRNA	10.82
Total Period Protein	3.70
Timeless mRNA	7.54
Total Timeless Protein	5.80
Complex	6.08
Nuclear Complex	4.56

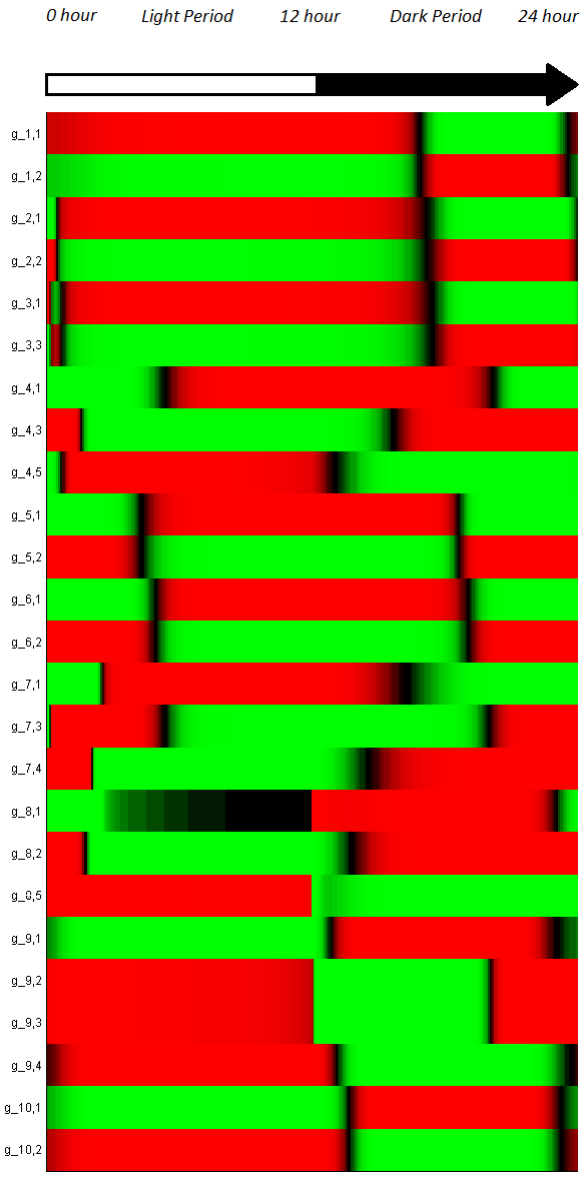


FIGURE B.4: *Heat Process Map* applied to the *Drosophila* Model.

B.2.4 Creation of sub-models based on time windows

The *Boolean Process Map* is not only useful to identify negligible processes for model reduction, it can be used also to study the evolution of important (non-negligible) processes and create a succession of sub-models containing the core mechanisms, each one being valid for a certain time window.

Figure B.6(A) shows the *based-event* grid that is built, based on the switching times contained in the global switching time Set S, while Figure B.5 represents the number of *active* processes as a function of time for the *Drosophila* model. The minimum number of *active* processes is 21 and the maximum is 26 and the number of time windows is 10.

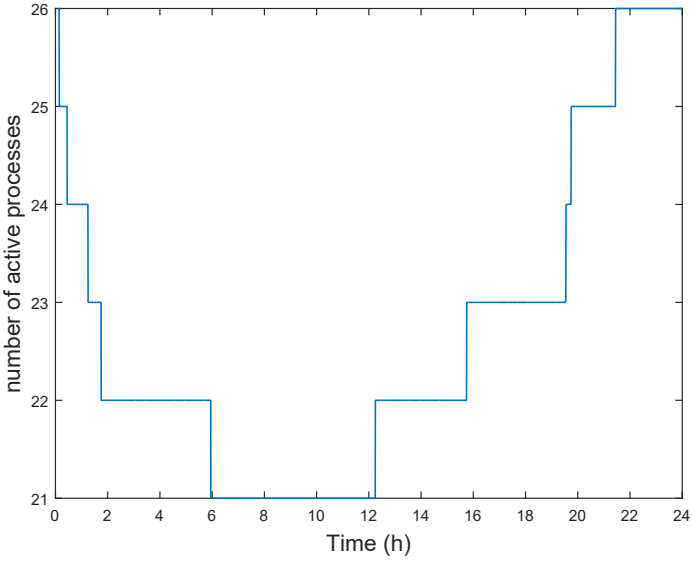


FIGURE B.5: Evolution of the number of active processes as a function of time. The function increases or decreases at the switching time t_b in the time interval of the system evolution.

We use the k-means clustering to compact together more switching times: applying Equation (B.11), we choose $z = 2$. The steps are shown in Figure B.6. We have thus created 2 sub-models: the first one (valid from 0 to 1.96 h and from 17.8 h to 24 h) coincides with the reduced model and the second one (valid from 1.96 h to 17.8 h) suppresses also the processes $f_{4,3}$, $f_{4,4}$, $f_{7,4}$, $f_{8,2}$, $f_{9,1}$.

In Figure B.7 is shown a *Dynamical Process Map* for the two sub-models, where red lines represent active processes in that time-window, while inactive ones are represented in black. The first sub-model essentially corresponds to a situation of night time and the second one, of day time. The results in Table B.2 show a good match between the

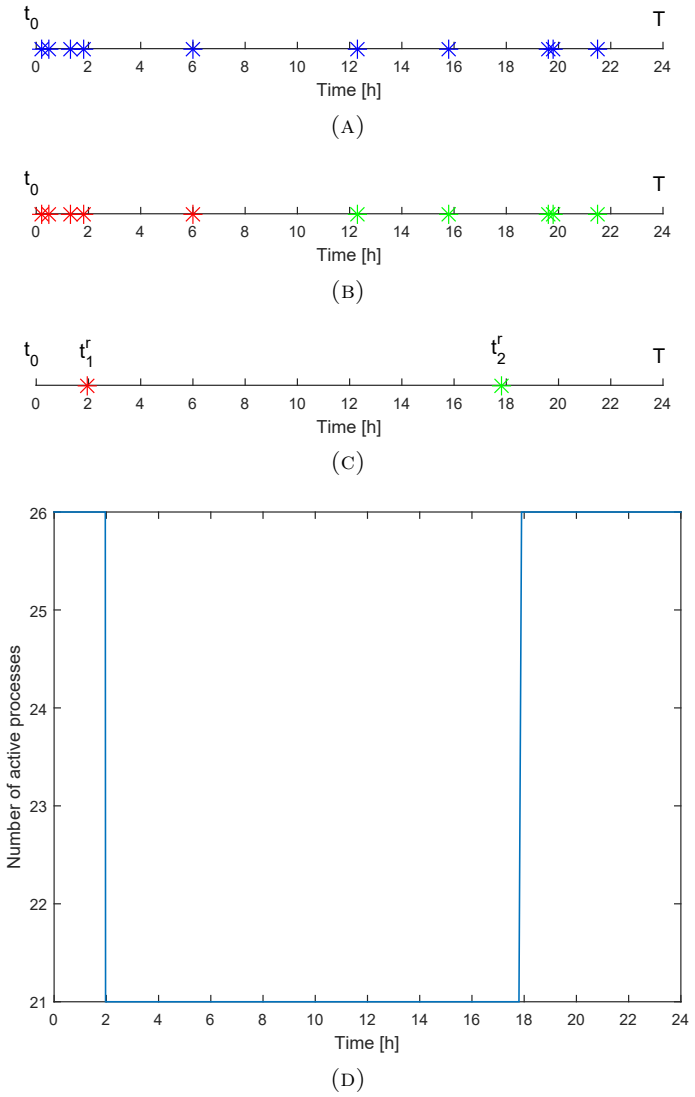


FIGURE B.6: (A) shows the *based-event* grid: every time window is delimited by the switching time t_b in the time interval of the system evolution (upper picture). In (B) the algorithm subsequently chooses the membership of every t_b in the Cluster C_1 or C_2 to minimize the WCCS expressed by B.10 (middle picture). In (C) the cluster is then replaced with its centroid (in this case the mean) that will be the new approximate switching time. (D) shows the approximation of Figure B.5. The number of time windows becomes 2.

original model and the first sub-model, and the original model and the second sub-model.

The dynamics of the system outputs for the original and the for the first sub-model ($SM1$) and second sub-model ($SM2$) with propagation error are shown in Figure B.8.

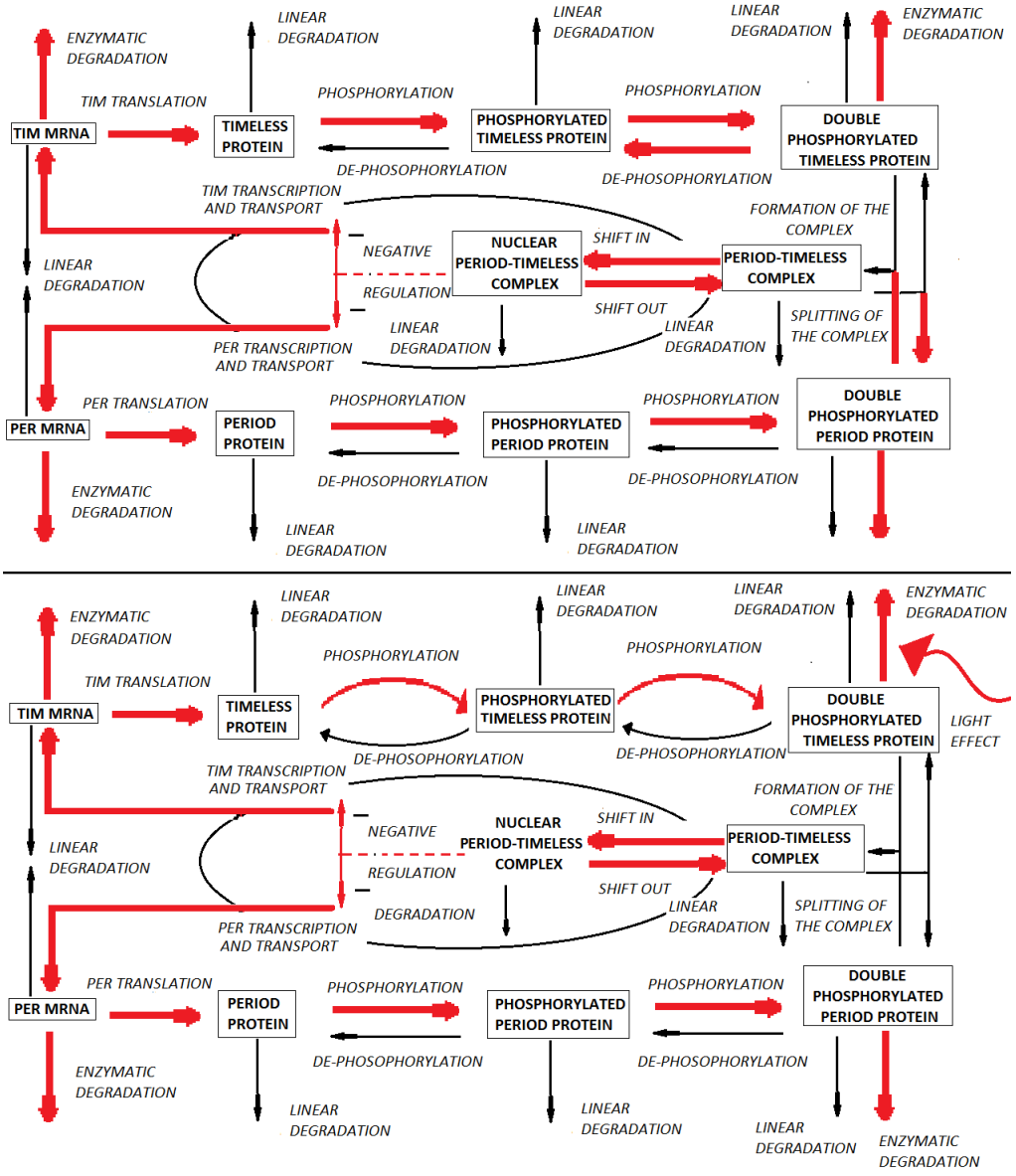


FIGURE B.7: Upper panel: *Dynamical Process Map* for the sub-model valid from 0 to 1.96 h and 17.8 h to 24 h. Lower panel: *Dynamical Process Map* for the sub-model valid from 1.96 h to 17.8 h. Black: *inactive* processes. Red: *active* processes.

TABLE B.2: **Global relative Errors for the two sub-models of the *Drosophila* model**

Output	G. Rel. Err. SM1 (%)	G. Rel. Err. SM2 (%)
Period mRNA	13.63	7.70
Total Period Protein	1.61	7.06
Timeless mRNA	9.95	5.96
Total Timeless Protein	2.64	10.96
Complex	4.74	3.97
Nuclear Complex	5.36	5.85

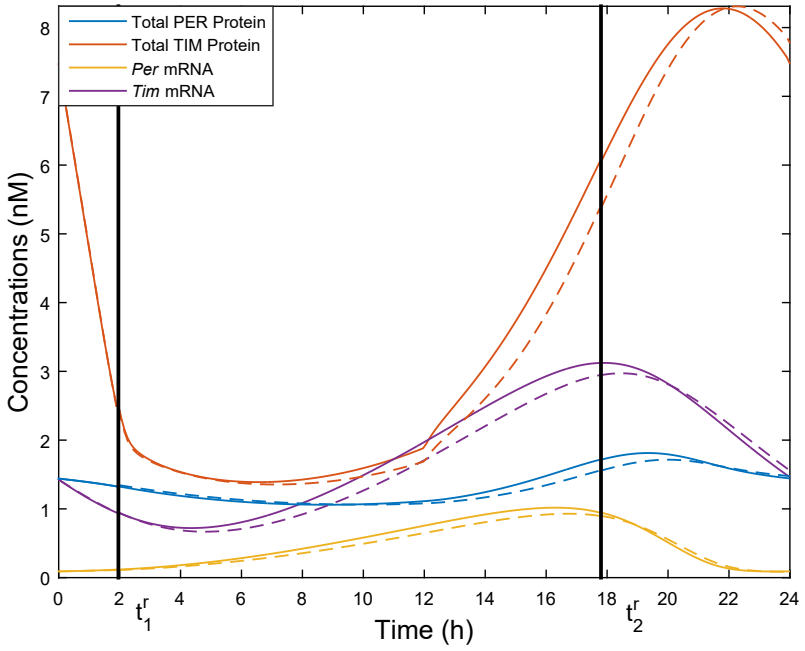


FIGURE B.8: Output dynamics for the original and sub-models with propagation errors. The firsts are represented with continuous lines while the seconds with dotted lines. The t_1^r and t_2^r are the approximate *switching times*.

B.3 Model for the influence of RKIP on the ERK signaling pathway

B.3.1 Description

We consider the model in [68] of dimension 11. The ERK signaling pathway controls important cellular phenomena like proliferation or differentiation. The model describes the inhibition of the activation of RAF by RKIP, which regulates the ERK signaling pathway. We use in our analysis the parameters of [68] and initial values in [90].

Since no switches occur during the simulation (due to the strong *activity/inactivity* boundary separation and the absence of process weights crossing the threshold δ), we will show neither the *Boolean Process Map* nor the creation of sub-models.

B.3.2 Model reduction

We simulate the relative weight of the processes for every ODE as we did for the previous model. We are able to neglect 12 of 34 processes: $f_{1,2}$, $f_{2,2}$, $f_{3,2}$, $f_{3,4}$, $f_{4,2}$, $f_{5,3}$, $f_{6,3}$, $f_{7,2}$, $f_{8,2}$, $f_{9,2}$, $f_{10,2}$, $f_{11,2}$. The results in Table B.3 show a really good match between the original model and the reduced one with a low global error. In addition to our study, other model reduction approaches with different goals have been applied to the ERK model in [68]: the quasi-steady-state-approximation used by Petrov *et al.* in [90] and the automatic complexity analysis by Lebiedz *et al.* in [70]. The work of [90] is concerned with the separation of variables with fast dynamics from those with slow dynamics with respect to a time scale and uses a mathematical scaling. The components of the ODEs related to the fast variables are expressed by algebraic equations. Analysing the resulting reduced model allowed the authors of the study to conclude that the variable m_4 (the complex Raf-1*/RKIP/ERK-PP) has the biggest influence on the system when it approaches its quasi-steady state. The work of [70] uses two different methods: the first one combines dynamic sensitivity analysis with singular value decomposition to find a minimal dimension of the model and the second one permits to reduce actually the dimension of the model and determine the variables which contribute more to the full dynamics of the system (variables m_5, m_8, m_{11}). Our analysis is different in the sense that we are not interested in which variable gives the bigger contribution to the dynamics of the full system but which mechanisms (processes) give the highest contribution to the dynamics of the variable: we study the influence of the processes. It is a reduction method that does not change in general the dimension and the biological structure of the system. For instance, the model in [68] is composed by two types of processes: *association processes*, where two or more proteins combine together to form a complex, and *dissociation processes* which correspond to the reverse mechanism. While the *association processes* are always *active* during the time evolution of the system, most of the *dissociation processes* can be considered negligible: it means that *association processes* play the bigger role in the dynamic for every variable of the system.

B.3.3 Qualitative tool: 3-D process map

We chose this time a three-dimensional bar graph as a qualitative approach to show the evolution of the intensity of the *activity* for every process (Figure B.9). The height of

TABLE B.3: **Global relative Errors for the reduced ERK model**

Variable	G. Rel. Err. (%)
Raf-1*	0.0012
RKIP	0.106
Raf-1*/RKIP	6.172
Raf-1*/RKIP/ERK-PP	0.0036
ERK-P	9.07
RKIP-P	0.138
MEK-PP	0.0056
MEK-PP/ERK	0.0148
ERK-PP	0.0013
RP	0.00049
RKIP-P	0.03451

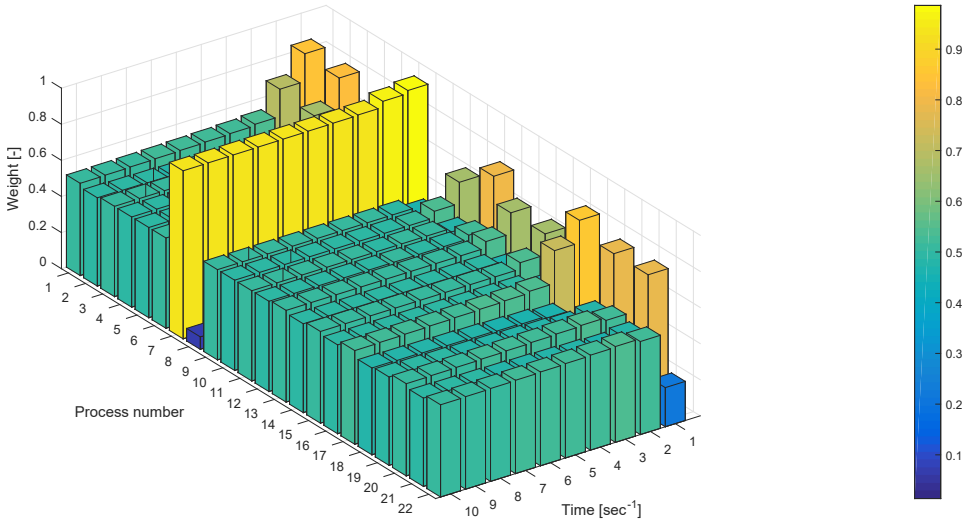


FIGURE B.9: 3D bar graph: the x-axis represents the processes, the y-axis corresponds to the time evolution and the z-axis, to the value of the weight.

the column gives the intensity of each process activity, while the color code indicates the weight of the process. In this graph we put the processes of the reduced model $g(x^r)$ during a one-second dynamics discretized in six time intervals. Figure B.9 shows that the most *active* process during the system dynamics is the seventh process, the association of Raf-1*/RKIP and ERK-PP to form the Raf-1*/RKIP/ERK-PP complex. This process strongly influences the variable m_4, which was found by Petrov *et al.* to be the most important one for the system dynamics.

B.4 Conclusions

We have presented in this appendix Principal Process Analysis, a method that allows to reduce biological models of high dimension and to analyze their core mechanisms and how these evolve dynamically, provided that their initial conditions and parameter values are known. Reduced models and sub-models obtained by this approach can be also refined, by calibrating the reduced vector of parameters so as to obtain models that better approximate the original one. In Chapter 6, Parameter Sensitivity Analysis will be applied to the original and reduced models to test the robustness of our technique. In Chapter 6, a biological model of higher dimension will be also analyzed to check further the quality of PPA.

B.5 Supplementary materials

TABLE B.4: **Variables of *Drosophila* model**

Variable	Description
M_P	Period mRNA
P_0	Period Protein
P_1	Phosphorylated Period Protein
P_2	Double Phosphorylated Period Protein
M_T	Timeless mRNA
T_0	Timeless Protein
T_1	Phosphorylated Timeless Protein
T_2	Double Phosphorylated Timeless Protein
C	Complex
C_N	Nuclear Complex

Full *Drosophila* model (see [72])

$$\frac{dM_P}{dt} = v_{sP}\frac{K_{IP}^n}{K_{IP}^n + C_N^n} - v_{mP}\frac{M_P}{K_{mP} + M_P} - k_d M_P$$

$$\frac{dP_0}{dt} = k_{sP}M_P - V_{1P}\frac{P_0}{K_{1P} + P_0} + V_{2P}\frac{P_1}{K_{2P} + P_1} - k_d P_0$$

$$\frac{dP_1}{dt} = V_{1P}\frac{P_0}{K_{1P} + P_0} - V_{2P}\frac{P_1}{K_{2P} + P_1} - V_{3P}\frac{P_1}{K_{3P} + P_1} + V_{4P}\frac{P_2}{K_{4P} + P_2} - k_d P_1$$

$$\frac{dP_2}{dt} = V_{3P}\frac{P_1}{K_{3P} + P_1} - V_{4P}\frac{P_2}{K_{4P} + P_2} - k_3 P_2 T_2 + k_4 C - V_{dP}\frac{P_2}{K_{dP} + P_2} - k_d P_2$$

$$\frac{dM_T}{dt} = v_{sP}\frac{K_{IT}^n}{K_{IT}^n + C_N^n} - v_{mT}\frac{M_T}{K_{mT} + M_T} - k_d M_T$$

$$\frac{dT_0}{dt} = k_{sT}M_T - V_{1T}\frac{T_0}{K_{1T} + T_0} + V_{2T}\frac{T_1}{K_{2T} + T_1} - k_d T_0$$

TABLE B.5: **Variables of ERK model**

Variable	Description
m_1	Raf-1*
m_2	RKIP
m_3	Raf-1*/RKIP
m_4	Raf-1*/RKIP/ERK-PP
m_5	ERK-P
m_6	RKIP-P
m_7	MEK-PP
m_8	MEK-PP/ERK
m_9	ERK-PP
m_{10}	RP
m_{11}	RKIP-P

$$\frac{dT_1}{dt} = V_{1T}\frac{T_0}{K_{1T}+T_0} - V_{2T}\frac{T_1}{K_{2T}+T_1} - V_{3T}\frac{T_1}{K_{3T}+T_1} + V_{4T}\frac{T_2}{K_{4T}+T_2} - k_d T_1$$

$$\frac{dT_2}{dt} = V_{3T}\frac{T_1}{K_{3T}+T_1} - V_{4T}\frac{T_2}{K_{4T}+T_2} - k_3 P_2 T_2 + k_4 C - V_{dT}\frac{T_2}{K_{dT}+T_2} - k_d T_2$$

$$\frac{dC}{dt} = k_3 P_2 T_2 - k_4 C - k_1 C + k_2 C_N - k_{dC} C$$

$$\frac{dC_N}{dt} = k_1 C - k_2 C_N - k_{dN} C_N$$

Reduced *Drosophila* model and first sub-model

$$\frac{dM_P}{dt} = v_{sP}\frac{K_{IP}^n}{K_{IP}^n + C_N^n} - v_{mP}\frac{M_P}{K_{mP}+M_P}$$

$$\frac{dP_0}{dt} = k_{sP} M_P - V_{1P}\frac{P_0}{K_{1P}+P_0}$$

$$\frac{dP_1}{dt} = V_{1P}\frac{P_0}{K_{1P}+P_0} - V_{3P}\frac{P_1}{K_{3P}+P_1}$$

$$\frac{dP_2}{dt} = V_{3P}\frac{P_1}{K_{3P}+P_1} - k_3 P_2 T_2 + k_4 C - V_{dP}\frac{P_2}{K_{dP}+P_2}$$

$$\frac{dM_T}{dt} = v_{sP}\frac{K_{IT}^n}{K_{IT}^n + C_N^n} - v_{mT}\frac{M_T}{K_{mT}+M_T}$$

$$\frac{dT_0}{dt} = k_{sT} M_T - V_{1T}\frac{T_0}{K_{1T}+T_0}$$

$$\frac{dT_1}{dt} = V_{1T}\frac{T_0}{K_{1T}+T_0} - V_{3T}\frac{T_1}{K_{3T}+T_1} + V_{4T}\frac{T_2}{K_{4T}+T_2}$$

$$\frac{dT_2}{dt} = V_{3T}\frac{T_1}{K_{3T}+T_1} - V_{4T}\frac{T_2}{K_{4T}+T_2} - V_{dT}\frac{T_2}{K_{dT}+T_2}$$

$$\frac{dC}{dt} = k_3 P_2 T_2 - k_4 C - k_1 C + k_2 C_N$$

$$\frac{dC_N}{dt} = k_1 C - k_2 C_N$$

Second *Drosophila* sub-model

$$\frac{dM_P}{dt} = v_{sP}\frac{K_{IP}^n}{K_{IP}^n + C_N^n} - v_{mP}\frac{M_P}{K_{mP} + M_P}$$

$$\frac{dP_0}{dt} = k_{sP}M_P - V_{1P}\frac{P_0}{K_{1P} + P_0}$$

$$\frac{dP_1}{dt} = V_{1P}\frac{P_0}{K_{1P} + P_0} - V_{3P}\frac{P_1}{K_{3P} + P_1}$$

$$\frac{dP_2}{dt} = V_{3P}\frac{P_1}{K_{3P} + P_1} - V_{dP}\frac{P_2}{K_{dP} + P_2}$$

$$\frac{dM_T}{dt} = v_{sP}\frac{K_{IT}^n}{K_{IT}^n + C_N^n} - v_{mT}\frac{M_T}{K_{mT} + M_T}$$

$$\frac{dT_0}{dt} = k_{sT}M_T - V_{1T}\frac{T_0}{K_{1T} + T_0}$$

$$\frac{dT_1}{dt} = V_{1T}\frac{T_0}{K_{1T} + T_0} - V_{3T}\frac{T_1}{K_{3T} + T_1}$$

$$\frac{dT_2}{dt} = V_{3T}\frac{T_1}{K_{3T} + T_1} - V_{dT}\frac{T_2}{K_{dT} + T_2}$$

$$\frac{dC}{dt} = -k_4C - k_1C + k_2C_N$$

$$\frac{dC_N}{dt} = k_1C - k_2C_N$$

Full model ERK model (see [68])

$$\frac{dm_1}{dt} = -k_1m_1m_2 + k_2m_3 + k_5m_4$$

$$\frac{dm_2}{dt} = -k_1m_1m_2 + k_2m_3 + k_{11}m_{11}$$

$$\frac{dm_3}{dt} = k_1m_1m_2 - k_2m_3 - k_3m_3m_9 + k_4m_4$$

$$\frac{dm_4}{dt} = k_3m_3m_9 - k_4m_4 - k_5m_4$$

$$\frac{dm_5}{dt} = k_5m_4 - k_6m_5m_7 + k_7m_8$$

$$\frac{dm_6}{dt} = k_5m_4 - k_9m_6m_{10} + k_{10}m_{11}$$

$$\frac{dm_7}{dt} = -k_6m_5m_7 + k_7m_8 + k_8m_8$$

$$\frac{dm_8}{dt} = k_6m_5m_7 - k_7m_8 - k_8m_8$$

$$\frac{dm_9}{dt} = -k_3m_3m_9 + k_4m_4 + k_8m_8$$

$$\frac{dm_{10}}{dt} = -k_9m_6m_{10} + k_{10}m_{11} + k_{11}m_{11}$$

$$\frac{dm_{11}}{dt} = k_9m_6m_{10} - k_{10}m_{11} - k_{11}m_{11}$$

Reduced ERK model

$$\frac{dm_1}{dt} = -k_1m_1m_2 + k_5m_4$$

$$\frac{dm_2}{dt} = -k_1m_1m_2 + k_{11}m_{11}$$

$$\frac{dm_3}{dt} = k_1 m_1 m_2 - k_3 m_3 m_9$$

$$\frac{dm_4}{dt} = k_3 m_3 m_9 - k_5 m_4$$

$$\frac{dm_5}{dt} = k_5 m_4 - k_6 m_5 m_7$$

$$\frac{dm_6}{dt} = k_5 m_4 - k_9 m_6 m_{10}$$

$$\frac{dm_7}{dt} = -k_6 m_5 m_7 + k_8 m_8$$

$$\frac{dm_8}{dt} = k_6 m_5 m_7 - k_8 m_8$$

$$\frac{dm_9}{dt} = -k_3 m_3 m_9 + k_8 m_8$$

$$\frac{dm_{10}}{dt} = -k_9 m_6 m_{10} + k_{11} m_{11}$$

$$\frac{dm_{11}}{dt} = k_9 m_6 m_{10} - k_{11} m_{11}$$

Appendix C

Supplementary materials of Chapter 5

C.1 Monotone systems

Monotone systems form an important class of dynamical systems, and are particularly well adapted to mathematical models in biology [112], because they are defined by conditions related to the signs of Jacobian matrix. Such a sign for one element reflects the fact that some variable will contribute positively to the variation of some other variables, a kind of qualitative dependence frequently found in biological models. The reader may consult the reference [107] for a review and an exhaustive presentation of the theory of monotone systems.

In summary, if the system is cooperative, then the flow preserves the partial order of trajectories in $\mathbb{R}^n$ (the flow is monotone). Consider an autonomous differential system:

$$\dot{x} = f(x) \tag{C.1}$$

where, $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^n$.

Therefore the system is monotone if $x_{01} \leq x_{02}$ (this inequality must be understood coordinate by coordinate: i.e. $x_{01i} \leq x_{02i}, \forall \ \ i \in [1, \ldots, n]$), implies that $x(t, x_{01}) \leq x(t, x_{02}) \ \ \forall \ \ t$, where $x(t, x_0)$ corresponds to the evolution with respect to time starting from the initial condition x_0.

Cooperativity is easy to check by looking at the signs of the elements of the Jacobian matrix, that should verify

$$\frac{\partial f_i}{\partial x_j}(t,x) \geq 0 \quad \forall i \neq j.$$

These systems have a strong tendency to converge to the set of their equilibria ([107]). It can be shown that almost any solution converges to the set of equilibria except a set of zero measure. In particular, there are no stable periodic solutions. For more precise theorems, see [107].

C.2 Tikhonov's theorem

This theorem applies to reduced systems of the form:

$$\begin{aligned} \dot{x} &= f(x,y,\epsilon) \\ \dot{y} &= \frac{1}{\epsilon} g(x,y,\epsilon). \end{aligned} \tag{C.2}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $0 < \epsilon \ll 1$ (ϵ a very small parameter), $x(0) = x_0, y(0) = y_0$. So, when ϵ tends to 0 ($\dot{y}$ evolves very rapidly compared to $\dot{x}$), the system (C.2) is equivalent to the system:

$$\begin{aligned} \dot{x} &= f(x,y,0) \\ g(x,y,0) &= 0 \end{aligned}$$

This is valid only if the fast subsystem $\dot{y} = g(x,y,0)$ satisfies some conditions which are given as follows:

- Existence and uniqueness of the steady state (there exists a unique solution, $y^* = \phi(x)$ of $g(x,y,0) = 0$).
- Stability of the steady state y^* of the fast subsystem $\dot{y} = g(x,y,0)$ for fixed x.

These conditions are given by Tikhonov's theorem (for a complete description see [57]), which ensures that y will converge rapidly to a *quasi steady state* ($y = \phi(x)$, depending only on x). Therefore the reduced system using Tikhonov's theorem is:

$$\dot{x} = f(x, \phi(x), 0), \quad x(0) = x_0.$$

Appendix D

Supplementary materials of Chapter 6

D.1 Full mammalian model

Model equations

Equations listed in [73, 74].

mRNAs of *Per* gene

$$\begin{aligned} \frac{dM_P}{dt} &= v_{sP}\frac{B_N^n}{K_{AP}^n+B_N^n} - v_{mP}\frac{M_P}{K_{mP}+M_P} - k_{dmp}M_P \\ \dot{x_1} &= f_{1,1} + f_{1,2} + f_{1,3} \end{aligned}$$

mRNAs of *Cry* gene

$$\begin{aligned} \frac{dM_C}{dt} &= v_{sC}\frac{B_N^n}{K_{AC}^n+B_N^n} - v_{mC}\frac{M_C}{K_{mC}+M_C} - k_{dmc}M_C \\ \dot{x_2} &= f_{2,1} + f_{2,2} + f_{2,3} \end{aligned}$$

mRNAs of *Bmal1* gene

$$\begin{aligned} \frac{dM_B}{dt} &= v_{sB}\frac{K_{IB}^n}{K_{IB}^n+B_N^n} - v_{mB}\frac{M_B}{K_{mB}+M_B} - k_{dmb}M_B \\ \dot{x_3} &= f_{3,1} + f_{3,2} + f_{3,3} \end{aligned}$$

Non-phosphorylated PER protein in the cytosol

$$\begin{aligned} \frac{dP_C}{dt} &= k_{sP}M_P - V_{1P}\frac{P_C}{K_P+P_C} + V_{2P}\frac{P_{CP}}{K_{dP}+P_{CP}} + k_4PC_C - k_3P_CC_C - k_{dn}P_C \\ \dot{x_4} &= f_{4,1} + f_{4,2} + f_{4,3} + f_{4,4} + f_{4,5} + f_{4,6} \end{aligned}$$

Non-phosphorylated CRY protein in the cytosol

$$\begin{aligned}\frac{dC_C}{dt} &= k_{sC}M_C - V_{1C}\frac{C_C}{K_P+C_C} + V_{2C}\frac{C_{CP}}{K_{dP}+C_{CP}} + k_4PC_C - k_3P_CC_C - k_{dnc}C_C\\ \dot{x_5} &= f_{5,1} + f_{5,2} + f_{5,3} + f_{5,4} + f_{5,5} + f_{5,6}\end{aligned}$$

Phosphorylated PER protein in the cytosol

$$\begin{aligned}\frac{dP_{CP}}{dt} &= V_{1P}\frac{P_C}{K_P+P_C} - V_{2P}\frac{P_{CP}}{K_{dP}+P_{CP}} - v_{dPC}\frac{P_{CP}}{K_d+P_{CP}} - k_{dn}P_{CP}\\ \dot{x_6} &= f_{6,1} + f_{6,2} + f_{6,3} + f_{6,4}\end{aligned}$$

Phosphorylated CRY protein in the cytosol

$$\begin{aligned}\frac{dC_{CP}}{dt} &= V_{1C}\frac{C_C}{K_P+C_C} - V_{2C}\frac{C_{CP}}{K_{dP}+C_{CP}} - v_{dCC}\frac{C_{CP}}{K_d+C_{CP}} - k_{dn}C_{CP}\\ \dot{x_7} &= f_{7,1} + f_{7,2} + f_{7,3} + f_{7,4}\end{aligned}$$

Non-phosphorylated PER-CRY complex in the cytosol

$$\begin{aligned}\frac{dPC_C}{dt} &= -V_{1PC}\frac{PC_C}{K_P+PC_C} + V_{2PC}\frac{PC_{CP}}{K_{dP}+PC_{CP}} - k_4PC_C + k_3P_CC_C + k_2PC_N - k_1PC_C - k_{dn}PC_C\\ \dot{x_8} &= f_{8,1} + f_{8,2} + f_{8,3} + f_{8,4} + f_{8,5} + f_{8,6} + f_{8,7}\end{aligned}$$

Non-phosphorylated PER-CRY complex in the nucleus

$$\begin{aligned}\frac{dPC_N}{dt} &= -V_{3PC}\frac{PC_N}{K_P+PC_N} + V_{4PC}\frac{PC_{NP}}{K_{dP}+PC_{NP}} - k_2PC_N + k_1PC_C - k_7B_NPC_N + k_8I_n - k_{dn}PC_N\\ \dot{x_9} &= f_{9,1} + f_{9,2} + f_{9,3} + f_{9,4} + f_{9,5} + f_{9,6} + f_{9,7}\end{aligned}$$

Phosphorylated PER-CRY complex in the cytosol

$$\begin{aligned}\frac{dPC_{CP}}{dt} &= V_{1PC}\frac{PC_C}{K_P+PC_C} - V_{2PC}\frac{PC_{CP}}{K_{dP}+PC_{CP}} - v_{dPCC}\frac{PC_{CP}}{K_d+PC_{CP}} - k_{dn}PC_{CP}\\ \dot{x_{10}} &= f_{10,1} + f_{10,2} + f_{10,3} + f_{10,4}\end{aligned}$$

Phosphorylated PER-CRY complex in the nucleus

$$\begin{aligned}\frac{dPC_{NP}}{dt} &= V_{3PC}\frac{PC_N}{K_P+PC_N} - V_{4PC}\frac{PC_{NP}}{K_{dP}+PC_{NP}} - v_{dPCN}\frac{PC_{NP}}{K_d+PC_{NP}} - k_{dn}PC_{NP}\\ \dot{x_{11}} &= f_{11,1} + f_{11,2} + f_{11,3} + f_{11,4}\end{aligned}$$

Non-phosphorylated BMAL1 protein in the cytosol

$$\begin{aligned}\frac{dB_C}{dt} &= k_{sB}M_B - V_{1B}\frac{B_C}{K_P+B_C} + V_{2B}\frac{B_{CP}}{K_{dP}+B_{CP}} - k_5B_C + k_6B_N - k_{dn}B_C\\ \dot{x_{12}} &= f_{12,1} + f_{12,2} + f_{12,3} + f_{12,4} + f_{12,5} + f_{12,6}\end{aligned}$$

Phosphorylated BMAL1 protein in the cytosol

$$\begin{aligned}\frac{dB_{CP}}{dt} &= V_{1B}\frac{B_C}{K_P+B_C} - V_{2B}\frac{B_{CP}}{K_{dP}+B_{CP}} - v_{dBC}\frac{B_{CP}}{K_d+B_{CP}} - k_{dn}B_{CP}\\ \dot{x_{13}} &= f_{13,1} + f_{13,2} + f_{13,3} + f_{13,4}\end{aligned}$$

Non-phosphorylated BMAL1 protein in the nucleus

$$\begin{aligned}\frac{dB_N}{dt} &= -V_{3B}\frac{B_N}{K_P+B_N} + V_{4B}\frac{B_{NP}}{K_{dP}+B_{NP}} + k_5 B_C - k_6 B_N - k_7 B_N PC_N + k_8 I_N - k_{dn} B_N \\ \dot{x_{14}} &= f_{14,1} + f_{14,2} + f_{14,3} + f_{14,4} + f_{14,5} + f_{14,6} + f_{14,7}\end{aligned}$$

Phosphorylated BMAL1 protein in the nucleus

$$\begin{aligned}\frac{dB_{NP}}{dt} &= V_{3B}\frac{B_N}{K_P+B_N} - V_{4B}\frac{B_{NP}}{K_{dP}+B_{NP}} - v_{dBN}\frac{B_{NP}}{K_d+B_{NP}} - k_{dn} B_{NP} \\ \dot{x_{15}} &= f_{15,1} + f_{15,2} + f_{15,3} + f_{15,4}\end{aligned}$$

Inactive complex between PER-CRY and CLOCK-BMAL1 in the nucleus

$$\begin{aligned}\frac{dI_N}{dt} &= -k_8 I_N + k_7 B_N PC_N - v_{dIN}\frac{I_N}{K_d+I_N} - k_{dn} I_N \\ \dot{x_{16}} &= f_{16,1} + f_{16,2} + f_{16,3} + f_{16,4}\end{aligned}$$

Model parameters

Parameters listed in [74, p.546]: set 1.

$k_1(h^{-1}) = 0.4$, $k_2(h^{-1}) = 0.2$, $k_3(nM^{-1}h^{-1}) = 0.4$, $k_4(h^{-1}) = 0.2$, $k_5(h^{-1}) = 0.4$, $k_6(h^{-1}) = 0.2$, $k_7(nM^{-1}h^{-1}) = 0.5$, $k_8(h^{-1}) = 0.1$, $K_{AP}(nM) = 0.7$, $K_{AC}(nM) = 0.6$, $K_{IB}(nM) = 2.2$, $k_{dmb}(h^{-1}) = 0.01$, $k_{dmc}(h^{-1}) = 0.01$, $k_{dmp}(h^{-1}) = 0.01$, $k_{dnc}(h^{-1}) = 0.12$, $k_{dn}(h^{-1}) = 0.01$, $K_d(nM) = 0.3$, $K_{dp}(nM) = 0.1$, $K_p(nM) = 0.1$, $K_{mB}(nM) = 0.4$, $K_{mC}(nM) = 0.4$, $K_{mP}(nM) = 0.31$, $k_{stot}(h^{-1}) = 1.0$, $k_{sB}(h^{-1}) = 0.12k_{stot}$, $k_{sC}(h^{-1}) = 1.6k_{stot}$, $k_{sP}(h^{-1}) = 0.6k_{stot}$, $n = 4$, $m = 2$, $V_{phos}(nMh^{-1}) = 0.4$, $V_{1B}(nMh^{-1}) = 0.5$, $V_{1C}(nMh^{-1}) = 0.6$, $V_{1P}(nMh^{-1}) = V_{phos}$, $V_{1PC}(nMh^{-1}) = V_{phos}$, $V_{2B}(nMh^{-1}) = 0.1$, $V_{2C}(nMh^{-1}) = 0.1$, $V_{2P}(nMh^{-1}) = 0.3$, $V_{2PC}(nMh^{-1}) = 0.1$, $V_{3B}(nMh^{-1}) = 0.5$, $V_{3PC}(nMh^{-1}) = V_{phos}$, $V_{4B}(nMh^{-1}) = 0.2$, $V_{4PC}(nMh^{-1}) = 0.1$, $v_{dBC}(nMh^{-1}) = 0.5$, $v_{dBN}(nMh^{-1}) = 0.6$, $v_{dCC}(nMh^{-1}) = 0.7$, $v_{dIN}(nMh^{-1}) = 0.8$, $v_{dIN}(nMh^{-1}) = 0.8$, $v_{dPC}(nMh^{-1}) = 0.7$, $v_{dPCC}(nMh^{-1}) = 0.7$, $v_{dPCN}(nMh^{-1}) = 0.7$, $v_{mB}(nMh^{-1}) = 0.8$, $v_{mC}(nMh^{-1}) = 1.0$, $v_{mP}(nMh^{-1}) = 1.1$, $vstot(nMh^{-1}) = 1.0$, $v_{sB}(nMh^{-1}) = v_{stot}$, $v_{sB}(nMh^{-1}) = v_{stot}$, $v_{sC}(nMh^{-1}) = 1.1v_{stot}$, $v_{sP}(nMh^{-1}) = 1.5v_{stot}$

Initial conditions

The unit of the initial conditions is nM.

$M_P(0)$ =2.188 $M_C(0)$ =1.633, $M_B(0)$ =9.498, $P_C(0)$=2.008, $C_C(0)$=1.884, $P_{CP}(0)$ =0.129, $C_{CP}(0)$ =0.473, $PC_C(0)$ =1.228, $PC_N(0)$ =0.177, $PC_{CP}(0)$ =0.203, $PC_{NP}(0)$ =0.101, $B_C(0)$ =2.523, $B_{CP}(0)$ =0.929, $B_N(0)$ =1.787, $B_{NP}(0)$ =0.318, $I_N(0)$ =0.051

D.2 Switching times

See Table B.4.

D.3 Neglected procceses

D.3.1 First reduced model

Neglected processes are:
$f_{1,3}$, $f_{2,3}$, $f_{3,3}$, $f_{4,6}$, $f_{5,3}$, $f_{5,6}$, $f_{6,4}$, $f_{7,4}$, $f_{8,2}$, $f_{8,7}$, $f_{9,6}$, $f_{9,7}$, $f_{10,4}$, $f_{11,4}$, $f_{12,3}$, $f_{12,6}$, $f_{13,2}$, $f_{13,4}$, $f_{14,2}$, $f_{14,6}$, $f_{14,7}$, $f_{15,4}$, $f_{16,1}$, $f_{16,4}$.

D.3.2 Second reduced model: sub-models

Neglected processes in SM1 are:
$f_{1,3}$, $f_{2,3}$, $f_{3,3}$, $f_{4,6}$, $f_{5,3}$, $f_{5,6}$, $f_{6,4}$, $f_{7,4}$, $f_{8,2}$, $f_{8,7}$, $f_{9,6}$, $f_{9,7}$, $f_{10,4}$, $f_{11,4}$, $f_{12,3}$, $f_{12,6}$, $f_{13,2}$, $f_{13,4}$, $f_{14,2}$, $f_{14,6}$, $f_{14,7}$, $f_{15,4}$, $f_{16,1}$, $f_{16,4}$.

In SM2, we supposed that processes switching state from $t_1 = 0.33$ until $t_6 = 1.5$ change simultaneously at time $t_1^r = 0.9$. Deleted processes are common to those removed in SM1, as well as: $f_{4,3}$, $f_{4,4}$, $f_{5,4}$, $f_{7,2}$, $f_{8,3}$,$f_{8,5}$, $f_{9,2}$, $f_{9,3}$, $f_{10,2}$, $f_{14,5}$.

TABLE D.1: *Switching times* (s.t), their values (v.) in [h] and associate reduced (cluster) *switching times* (t_1^r,t_2^r,t_3^r,t_4^r) (s.t.c.).

s.t.	v.	s.t.c.	s.t.	v.	s.t.c.	s.t.	v.	s.t.c.	s.t.	v.	s.t.c.
t_0	0		t_{12}	5.9		t_{24}	13.5		t_{36}	19.5	
t_1	0.3		t_{13}	7.9	6	t_{25}	13.6	12.5	t_{37}	20.3	
t_2	0.6		t_{14}	8.2		t_{26}	15.6		t_{38}	20.4	
t_3	0.8	0.9	t_{15}	8.6		t_{27}	17.3		t_{39}	20.45	
t_4	1		t_{16}	9.8		t_{28}	17.4		t_{40}	20.5	
t_5	1.1		t_{17}	10.4		t_{29}	18.5		t_{41}	20.7	20
t_6	1.5		t_{18}	11.2		t_{30}	18.9		t_{42}	20.8	
t_7	3.8		t_{19}	11.5	12.5	t_{31}	19.1	20	t_{43}	21.5	
t_8	4.1		t_{20}	12.4		t_{32}	19.2		t_{44}	21.6	
t_9	4.4	6	t_{21}	12.6		t_{33}	19.25		t_{45}	22.3	
t_{10}	5.4		t_{22}	13.3		t_{34}	19.3		t_{46}	22.9	
t_{11}	5.7		t_{23}	13.4		t_{35}	19.35		T	24	

In SM3, we supposed that processes switching state from $t_7 = 3.8$ until $t_6 = 1.5$ change simultaneously at time $t_2^r = 6$. Deleted processes are common to those removed in SM1, as well as: $f_{1,1}$, $f_{2,1}$, $f_{4,2}$, $f_{4,3}$, $f_{5,2}$, $f_{7,2}$, $f_{8,1}$, $f_{9,1}$, $f_{9,2}$, $f_{10,2}$, $f_{11,2}$, $f_{12,5}$, $f_{14,4}$.

In SM4, we supposed that processes switching state from $t_{16} = 9.8$ until $t_{26} = 15.6$ change simultaneously at time $t_3^r = 12.5$. Deleted processes are common to those removed in SM1, as well as: $f_{4,1}$, $f_{5,1}$, $f_{8,4}$, $f_{9,2}$, $f_{10,2}$, $f_{11,2}$,$f_{12,5}$, $f_{14,4}$.

In SM5, we supposed that processes switching state from $t_{27} = 17.3$ until $t_{46} = 22.9$ change simultaneously at time $t_4^r = 20.0$. Deleted processes are common to those removed in SM1, as well as: $f_{4,3}$, $f_{4,4}$, $f_{5,4}$, $f_{7,2}$, $f_{8,3}$, $f_{8,5}$, $f_{9,2}$, $f_{9,3}$, $f_{10,2}$, $f_{14,5}$.

D.4 Dynamical process maps

See the figures below.

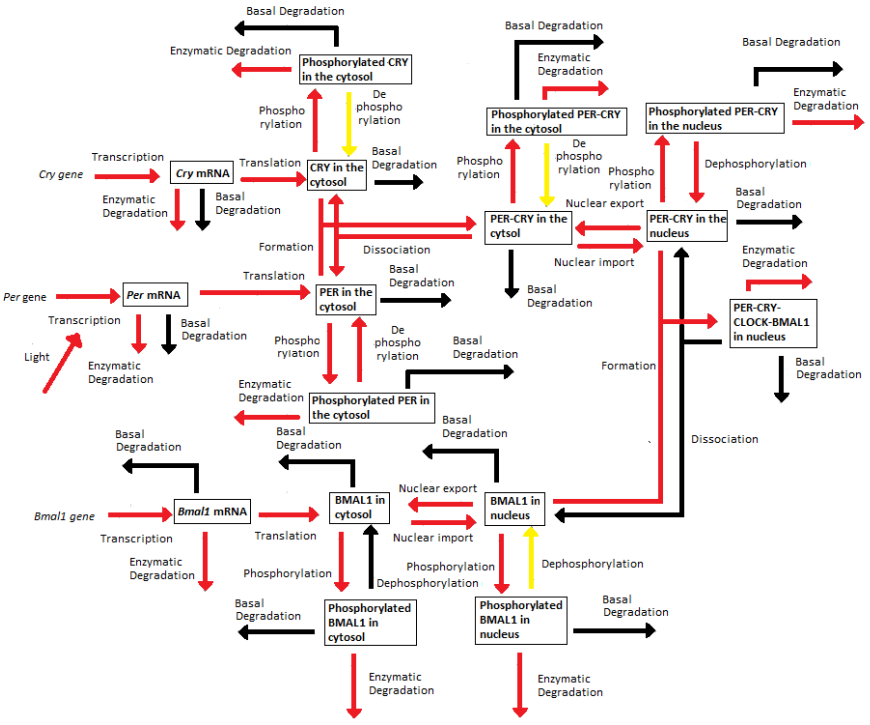


FIGURE D.1: *Dynamical process map* of SM1.

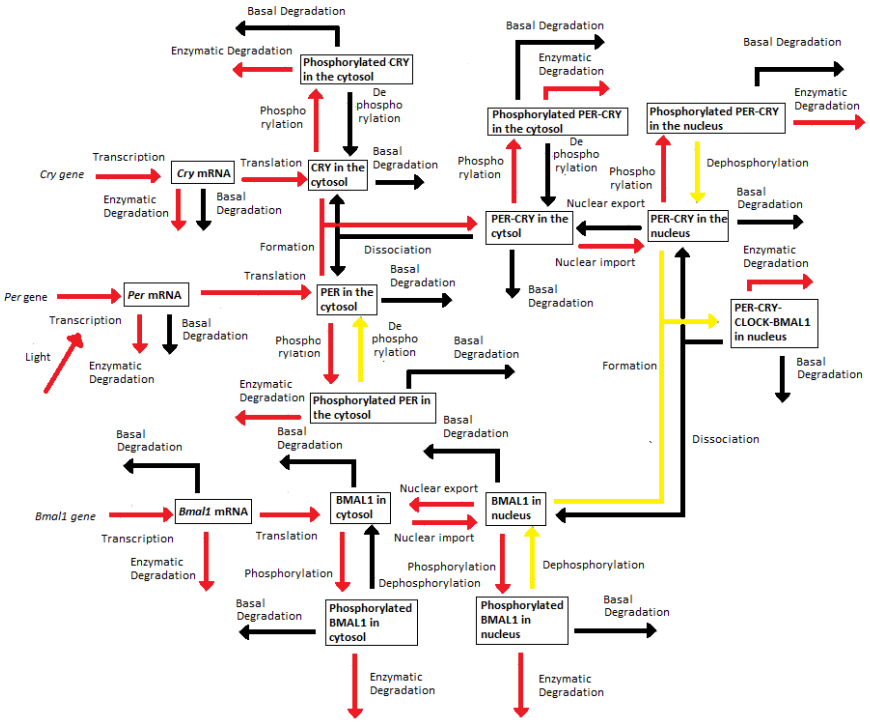


FIGURE D.2: *Dynamical process map* of SM2.

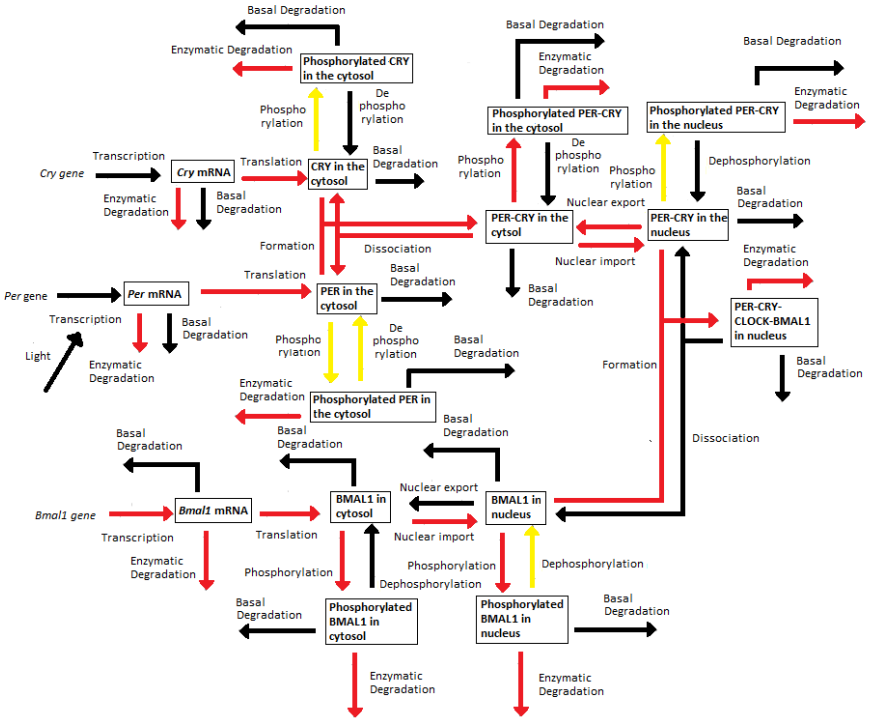


FIGURE D.3: *Dynamical process map* of SM3.

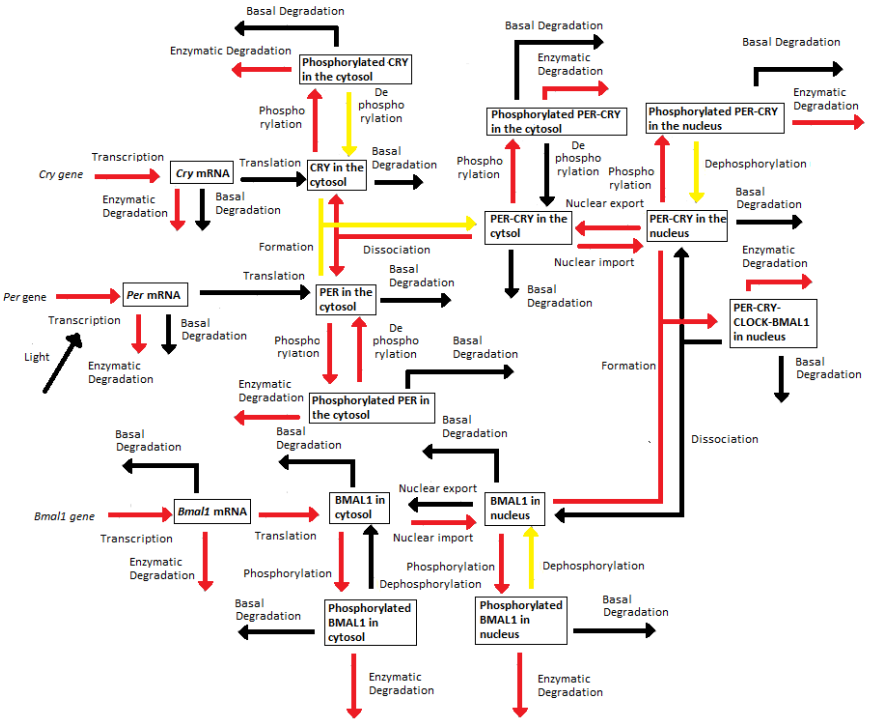


FIGURE D.4: *Dynamical process map* of SM4.

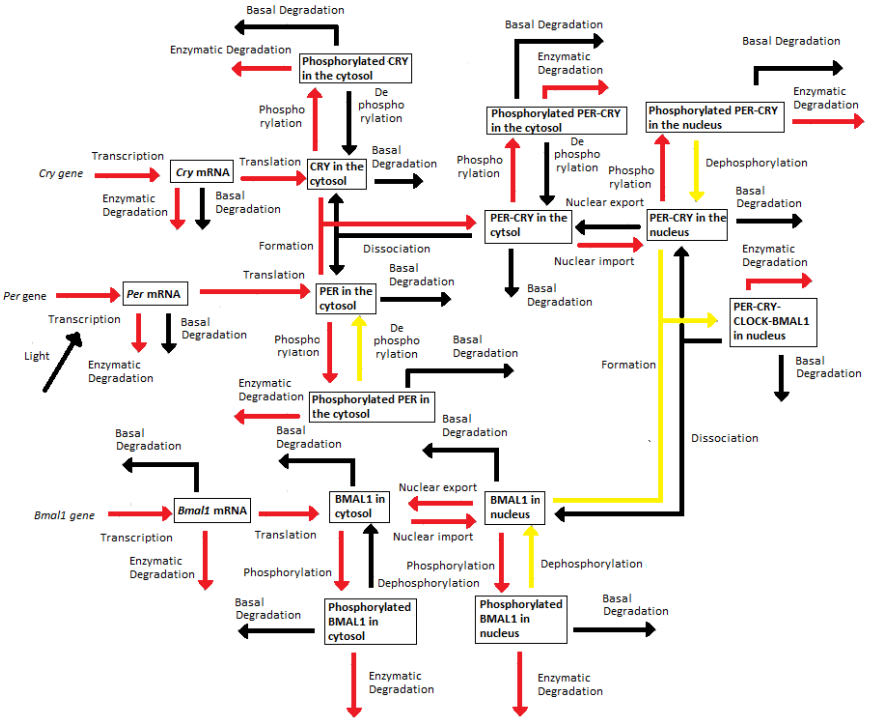


FIGURE D.5: *Dynamical process map* of SM5.

Appendix E

Supplementary materials of Chapter 8

E.1 Full dynamics of the experiments 1A, 2A, 2B

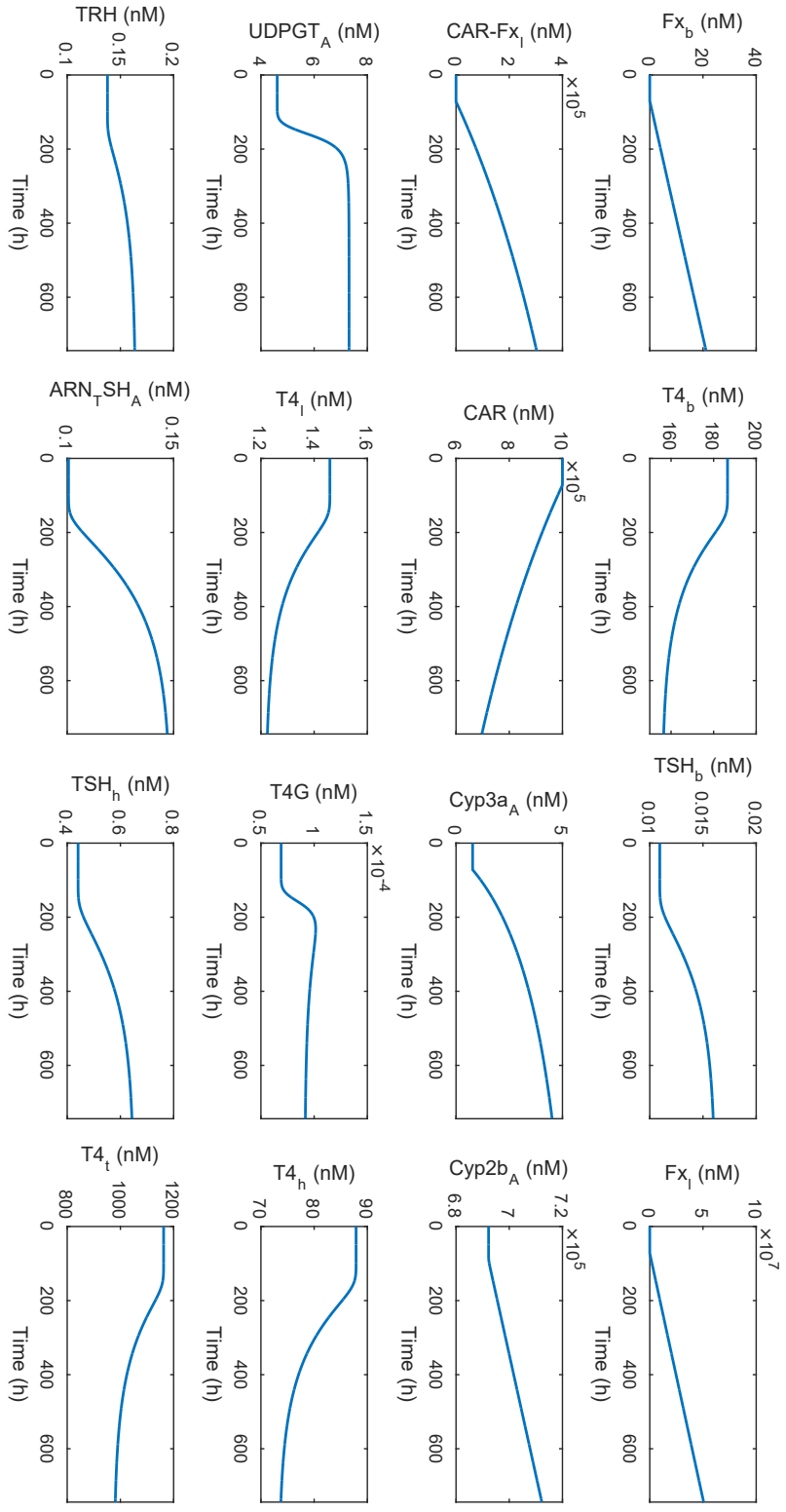


FIGURE E.1: **Full dynamics of the experiment 1A**. A - Simulation of the model parameters of Table 8.1 (744 hours).

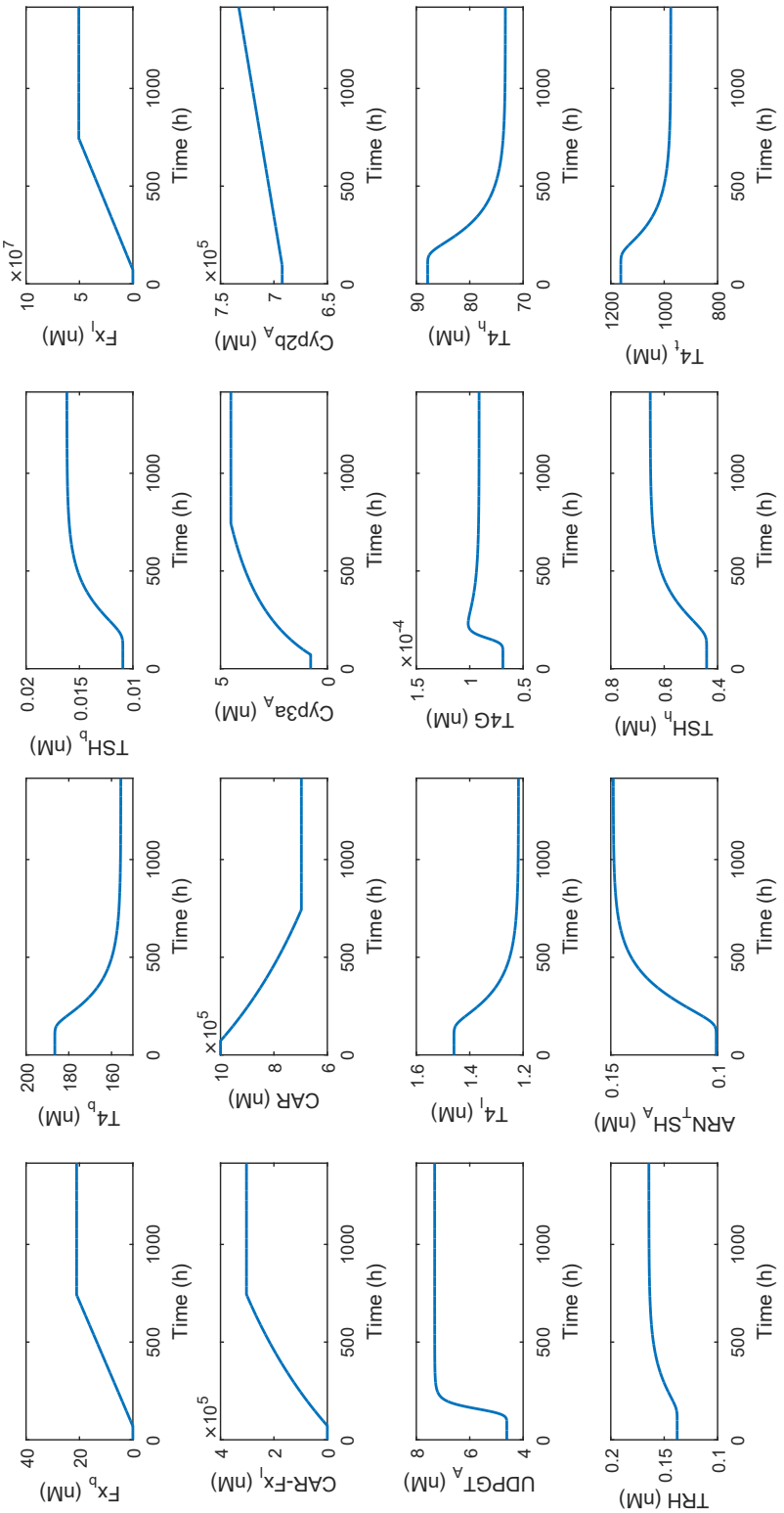


FIGURE E.2: **Full dynamics of the experiment 2A**. A - Simulation of the model parameters of Table 8.1 with recovery phase (1416 hours). The variable solutions are not returning to their values they had at the beginning of the experiment).

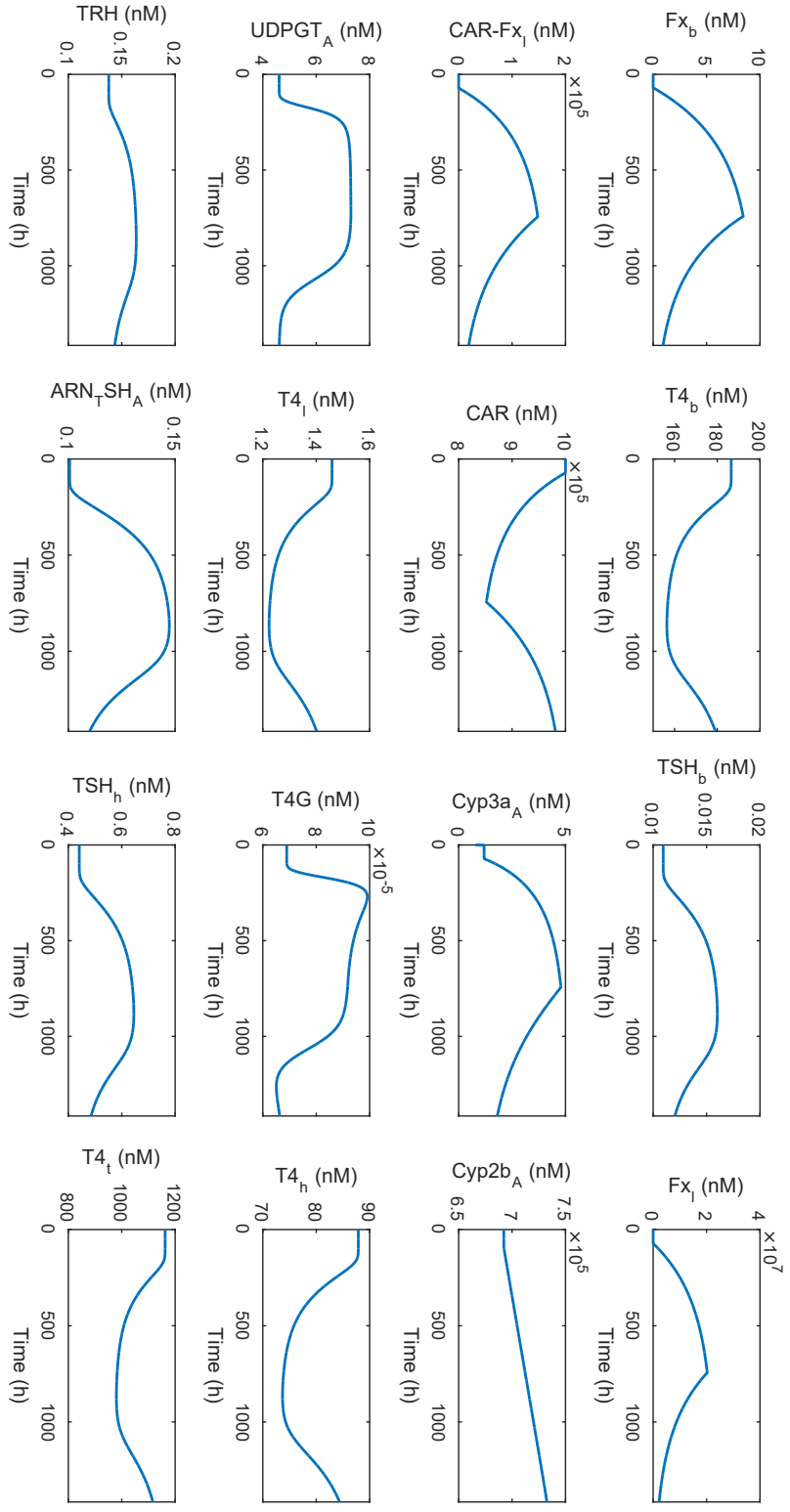


FIGURE E.3: **Full dynamics of experiment 2B with recovery phase**. A - Simulation of the model parameters of Table 8.1 but where but where $k_{1f} = 8000$ and $k_{5a} = 3700$, with recovery phase (1416 hours). The variable solutions are returning close to their values they had at the beginning of the experiment).

Appendix F

Supplementary materials of Chapter 9

F.1 Parameters and initial values of the GEM model

The parameters are shown in Table F.1.

Table F.1: Model parameters and their units.

Parameters		Units	Values
***bulk* mRNAs**			
d_b	Promoter concentration	μM	7.8867
k_b	Maximum transcription initiation rate	min^{-1}	1.5000
K_b	Binding constant	μM	6.1371
L_b	Gene length	base pairs	900
c_b	Elongation rate	min^{-1}	2572.1
Bulk proteins			
k_B	Maximum translation initiation rate	min^{-1}	0.6368
K_B	Binding constant	μM	63.3790
L_B	Protein length	amino acids	300
c_B	Elongation rate	min^{-1}	2572.1
***rpoBC* mRNAs**			
d_o	Promoter concentration	μM	0.0016
k_o	Maximum transcription initiation rate	min^{-1}	26
K_o	Binding constant	μM	34.6854
L_o	Gene length	base pairs	8253
c_o	Elongation rate	min^{-1}	2571.5
RNA Polymerase $\beta\beta'$ subunits			
k_β	Maximum translation initiation rate	min^{-1}	0.8921
K_β	Binding constant	μM	60.9419

Continued on next page

Table F.1 – continued from previous page			
Parameters		**Units**	**Values**
KI_β	Inhibition binding constant	μM	0.8360
L_β	Protein length	amino acids	2751
c_β	Elongation rate	min^{-1}	12.5831
Stable RNAs			
d_n	Promoter concentration	μM	0.0081
k_n	Maximum transcription initiation rate	min^{-1}	110
K_n	Binding constant	μM	0.4550
L_n	Gene length	base pairs	5498
c_n	Elongation rate	min^{-1}	4740
α_n	Constant	μM^{-1}	0.0274
r-protein mRNAs			
d_m	Promoter concentration	μM	0.0011
k_m	Maximum transcription initiation rate	min^{-1}	26
K_m	Binding constant	μM	1.4611
L_m	Gene length	base pairs	22680
c_m	Elongation rate	min^{-1}	2572.1
α_m	Constant	μM^{-1}	0.0908
r-proteins			
k_M	Maximum translation initiation rate	min^{-1}	1.0068
K_M	Binding constant	μM	104.004
L_M	Protein length	amino acids	7560
c_M	Elongation rate	min^{-1}	2572.1
RNase E			
kE_B	Catalytic constant	min^{-1}	0.3634
kE_β	Catalytic constant	min^{-1}	0.2859
kE_M	Catalytic constant	min^{-1}	0.189
K_E	Binding constant	μM	6036.9
RNA Polymerase			
d_{ns}	Aspecific binding site concentration	μM	4109.8e+03
K_{ns}	Non-specific binding constant	μM	2605
k_P^m	Maturation rate constant	min^{-1}	0.0964
Ribosome and tRNAs			
k_R^m	Maturation rate constant	$\text{M}^{-1}\cdot\text{min}^{-1}$	0.0552
ρ	Stoichiometry coefficient	-	7
V_m^C	Maximal velocity	$\text{M}^{-1}\cdot\text{min}^{-1}$	229380
K_A	Binding constant	M	1.1728
K_U	Binding constant	μM	13.6095
Amino-acids			
V_m^A	Maximal velocity	$\text{mM}\cdot\text{min}^{-1}$	0.1169
KI_A	Inhibition binding constant	mM	0.0893
ppGpp			
k_G	Synthesis rate constant	$\mu\text{M}\cdot\text{min}^{-1}$	3.8409
k_{spoT}	Degradation rate constant	min^{-1}	90.5822
K_{CT}	Binding constant	μM	0.5
K_g	RNA polymerase binding constant	μM	40

The initial values of the system are: $b(0) = 12.0738\ \mu M,\ B(0) = 6036.8995\ \mu M,\ o(0) = 0.0152\ \mu M,\ \beta(0) = 0.836\ \mu M,\ P(0) = 6.764\ \mu M,\ n(0) = 3.0069\ \mu M,\ m(0) = 0.0898\ \mu M,\ M(0) = 3.0069\ \mu M,\ R(0) = 41.8933\ \mu M,\ T(0) = 314.301\ \mu M,\ C(0) = 100\,\mu M,\ A(0) = 2.3313\ M,\ G(0) = 12.5\,\mu M,\ Pf(0) = 1.3496\,\mu M,\ Rf(0) = 13.0857\,\mu M,\ Af(0) = 0.1599\ M$

Appendix G

Supplementary materials of Chapter 10

G.1 Model parameters for each c calibration

TABLE G.1: Parameters from each single-cell calibration.

Single cell parameters (full)									
Par.	K_1	K_2	K_3	K_4	K_5	K_6	K_{mat}	$R_m(0)$	$L(0)$
Cell 1	0.1032	6.6938	3.2942	0.1013	0.0201	1.6232	0.0045	0.7097	1.0953
Cell 2	0.1005	4.9767	5.3085	0.0133	0.0155	2.0516	0.0045	1.1109	2.3766
Cell 3	0.1105	3.0110	4.4145	0.0134	0.0079	0.9654	0.0049	1.2765	3.9259
Cell 4	0.1843	3.0154	4.4745	0.0183	0.0013	1.1023	0.0030	1.1713	4.0180
Cell 5	0.1970	3.1915	2.2358	0.0003	0.0011	0.5013	0.0023	1.1646	4.1441
Cell 6	0.1881	3.9206	2.0934	0.0251	0.0012	0.0050	0.0037	0.9598	2.0426
Cell 7	0.2736	4.5075	2.3862	0.0003	0.0010	0.0055	0.0038	0.8617	1.5726
Cell 8	0.2583	3.6361	2.9906	0.0000	0.0011	0.0074	0.0038	0.9183	3.1788
Cell 9	0.2590	4.2018	0.8491	0.0000	0.0011	0.0074	0.0019	0.9188	3.8889
Cell 10	0.2998	4.7409	0.3906	0.0000	0.0010	0.2068	0.0052	1.0550	3.3478
Cell 11	0.2680	0.0474	5.6891	0.0000	0.0009	0.0021	0.0145	1.0858	0.0335
Cell 12	0.0839	3.5542	7.7268	0.0511	1.2942	8.0177	0.0056	0.7541	3.4210
Cell 13	0.1047	2.9187	10.6879	0.0005	0.5795	36.7303	0.0055	0.8826	3.0706
Cell 14	0.0712	4.7005	4.0511	0.0004	0.0175	33.9299	0.0046	0.7646	0.0373
Cell 15	0.2621	12.4396	24.7368	0.0010	7.0673	0.0752	0.0001	0.9724	0.0680
Cell 16	0.1940	11.9989	16.4142	0.0000	0.1331	0.0972	0.0000	0.8848	0.0138
Cell 17	0.2662	11.7642	16.4960	0.0000	0.0719	0.1087	0.0000	0.8368	0.0071
Cell 18	0.2418	11.6691	16.4580	0.0000	0.0238	0.2013	0.0000	0.9895	0.0113
Cell 19	0.1715	10.2708	13.2786	0.0000	0.0249	0.0039	0.0000	0.9582	0.0100
Cell 20	0.1725	13.8972	8.9589	0.0011	0.0002	0.0000	0.0000	0.7416	0.0063

Bibliography

[1] https://en.wikipedia.org/wiki/Prokaryote#/media/File:Prokaryote_cell.svg.

[2] http://academic.pgcc.edu/~kroberts/Lecture/Chapter%206/growth.html.

[3] https://en.wikipedia.org/wiki/Translation_(biology)#/media/File:Ribosome_mRNA_translation_en.svg.

[4] https://upload.wikimedia.org/wikipedia/commons/thumb/2/22/Lac_Operon.svg/512px-Lac_Operon.svg.png.

[5] https://it.mathworks.com/help/optim/ug/fmincon.html.

[6] 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.

[7] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, Fifth edition, 2007.

[8] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC press, 2006.

[9] M. Apri, M. de Gee, and J. Molenaar. Complexity reduction preserving dynamical behavior of biochemical networks. *Journal of Theoretical Biology*, 304:16–26, 2012.

[10] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom. Systems biology: parameter estimation for biochemical models. *Febs Journal*, 276(4):886–902, 2009.

[11] S. Bakshi, A. Siryaporn, M. Goulian, and J. C. Weisshaar. Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Molecular Microbiology*, 85(1):21–38, 2012.

[12] V. Baldazzi, D. Ropers, Y. Markowicz, D. Kahn, J. Geiselmann, and H. de Jong. The carbon assimilation network in *Escherichia coli* is densely connected and largely sign-determined by directions of metabolic fluxes. *PLoS Computational Biology*, 6(6):e1000812, 2010.

[13] G. Bastin and V. Guffens. Congestion control in compartmental network systems. *Systems & Control Letters*, 55(8):689–696, 2006.

[14] I. Belgacem and J.-L. Gouzé. Analysis and reduction of transcription translation coupled models for gene expression. *IFAC Proceedings Volumes*, 46(31):36–41, 2013.

[15] I. Belgacem and J.-L. Gouzé. Global stability of enzymatic chains of full reversible Michaelis-Menten reactions. *Acta Biotheoretica*, 61(3):425–436, 2013.

[16] I. Belgacem and J.-L. Gouzé. Stability analysis and reduction of gene transcription models. In *2013 IEEE 52nd Annual Conference on Decision and Control (CDC)*, pages 2691–2696. IEEE, 2013.

[17] I. Belgacem and J.-L. Gouzé. Mathematical study of the global dynamics of a concave gene expression model. In *2014 22nd Mediterranean Conference of Control and Automation (MED)*, pages 1341–1346. IEEE, 2014.

[18] I. Belgacem, E. Grac, D. Ropers, and J.-L. Gouzé. Stability analysis of a reduced transcription-translation model of RNA polymerase. In *2014 IEEE 53rd Annual Conference on Decision and Control (CDC)*, pages 3924–3929. IEEE, 2014.

[19] E. Benoît and J.-L. Gouzé. An algorithmic approach to orders of magnitude in a biochemical system. In *Positive Systems*, pages 233–241. Springer, 2009.

[20] G. Bernot, J.-P. Comet, A. Richard, M. Chaves, J.-L. Gouzé, and F. Dayan. Modeling and analysis of gene regulatory networks. In *Modeling in Computational Biology and Biomedicine*, pages 47–80. Springer, 2013.

[21] S. Berthoumieux, M. Brilli, D. Kahn, H. De Jong, and E. Cinquemani. On the identifiability of metabolic network models. *Journal of Mathematical Biology*, 67(6-7):1795–1832, 2013.

[22] 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(1):634, 2013.

[23] K. Bettenbrock, S. Fischer, A. Kremling, K. Jahreis, T. Sauter, and E.-D. Gilles. A quantitative approach to catabolite repression in *Escherichia coli*. *Journal of Biological Chemistry*, 281(5):2578–2584, 2006.

[24] B. Bhattacharjee, D. A. Schwer, P. I. Barton, and W. H. Green. Optimally-reduced kinetic models: reaction elimination in large-scale kinetic mechanisms. *Combustion and Flame*, 135(3):191–208, 2003.

[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 Publishing Co Inc, 2008.

[27] G. E. Box and J. S. Hunter. The 2 k—p fractional factorial designs. *Technometrics*, 3(3):311–351, 1961.

[28] H. Bremer, P. Dennis, and M. Ehrenberg. Free RNA polymerase and modeling global transcription in *Escherichia coli*. *Biochimie*, 85(6):597–609, 2003.

[29] H. Bremer and P. P. Dennis. *Escherichia coli* and *Salmonella*: cellular and molecular biology. *Washington (DC): American Society for Microbiology. Chapter, Modulation of chemical composition and other parameters of the cell by growth rate*, pages 1553–1569, 1996.

[30] R. Cukier, C. Fortuin, K. E. Shuler, A. Petschek, and J. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. *The Journal of Chemical Physics*, 59(8):3873–3878, 1973.

[31] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.

[32] H. de Jong, J.-L. 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.

[33] D. Dean, J. Reizer, H. Nikaido, and M. 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):21005–21010, 1990.

[34] 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.

[35] J.-J. Droesbeke, G. Saporta, and J. Fine. *Plans d'Expériences: Applications à l'Entreprise*. Editions Technip, 1997.

[36] L. Edelstein Keshet. *Mathematical Models in Biology*. SIAM, 2005.

[37] C. R. Elcombe, J. Odum, J. R. Foster, S. Stone, S. Hasmall, A. R. Soames, I. Kimber, and J. Ashby. Prediction of rodent nongenotoxic carcinogenesis: evaluation of biochemical and tissue changes in rodents following exposure to nine nongenotoxic NTP carcinogens. *Environmental Health Perspectives*, 110(4):363, 2002.

[38] P. Érdi and J. Tóth. *Mathematical Models of Chemical Reactions: Theory and Applications of Deterministic and Stochastic Models*. Manchester University Press, 1989.

[39] M. C. Flickinger and M. P. Rouse. Sustaining protein synthesis in the absence of rapid cell division: an investigation of plasmid-encoded protein expression in

Escherichia coli during very slow growth. *Biotechnology Progress*, 9(6):555–572, 1993.

[40] R. Fukuda, M. Taketo, and A. Ishihama. Autogenous regulation of RNA polymerase beta subunit synthesis in vitro. *Journal of Biological Chemistry*, 253(13):4501–4504, 1978.

[41] 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.

[42] A. M. González Vargas, E. Cinquemani, and G. Ferrari Trecate. Validation methods for population models of gene expression dynamics. *IFAC-PapersOnLine*, 49(26):114–119, 2016.

[43] 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.

[44] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Springer Science & Business Media, 2012.

[45] R. Hengge. Proteolysis of sigma s (RpoS) and the general stress response in escherichia coli. *Research in Microbiology*, 160(9):667–676, 2009.

[46] M. Hirsch and H. Smith. Competitive and cooperative systems: a mini-review. *Positive Systems*, pages 822–824, 2003.

[47] B. Ingalls. Mathematical modelling in systems biology: An introduction. *Internet.[cited at p. 117]*, 2013.

[48] L. Ironi and E. Lanzarone. Assigning probabilities to qualitative dynamics of gene regulatory networks. *Journal of Mathematical Biology*, 69(6-7):1661–1692, 2014.

[49] L. Ironi and L. Panzeri. A computational framework for qualitative simulation of nonlinear dynamical models of gene-regulatory networks. *BMC Bioinformatics*, 10(12):S14, 2009.

[50] L. Ironi and M. Stefanelli. A framework for building and simulating qualitative models of compartmental systems. *Computer Methods and Programs in Biomedicine*, 42(4):233–254, 1994.

[51] J. Izard, C. D. G. Balderas, D. Ropers, S. Lacour, X. Song, Y. Yang, A. B. Lindner, J. Geiselmann, and H. de Jong. A synthetic growth switch based on controlled expression of RNA polymerase. *Molecular Systems Biology*, 11(11):840, 2015.

[52] J. A. Jacquez and C. P. Simon. Qualitative theory of compartmental systems. *Siam Review*, 35(1):43–79, 1993.

[53] S. S. Justice, D. A. Hunstad, L. Cegelski, and S. J. Hultgren. Morphological plasticity as a bacterial survival strategy. *Nature Reviews Microbiology*, 6(2):162–168, 2008.

[54] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, 2002.

[55] J. Keener and M. Nomura. Regulation of ribosome synthesis. In F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter, and H. Umbarger, editors, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, pages 1417–31. ASM Press, Washington D.C., 1996.

[56] J. Keener and J. Sneyd. Mathematical Phsiology I: Cellular Physiology, 2009.

[57] H. Khalil. *Nonlinear Systems*. Prentice Hall, Third edition, 2002.

[58] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, 2002.

[59] C. D. Klaassen, M. O. Amdur, et al. *Casarett and Doull's Toxicology: the Basic Science of Poisons*, volume 5. McGraw-Hill New York, 1996.

[60] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems Biology in Practice: Concepts, Implementation and Application*. John Wiley & Sons, 2008.

[61] S. Klumpp, Z. Zhang, and T. Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7):1366–1375, 2009.

[62] D. E. Koshland, A. Goldbeter, and J. B. Stock. Amplification and adaptation in regulatory and sensory systems. *Science*, 217(4556):220–225, 1982.

[63] 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.

[64] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling. Ensemble modeling for analysis of cell signaling dynamics. *Nature Biotechnology*, 25(9):1001–1006, 2007.

[65] T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 104(14):6043–6048, 2007.

[66] B. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29(3):289–338, 1986.

[67] K. Kurma. Development and application of algorithms for the reconstruction of promoter activity statistics from reporter gene data. Technical report, INRIA Grenoble Rhône Alpes, 2015.

[68] C. Kwang Hyun, S. Sung Young, K. Hyun Woo, O. Wolkenhauer, B. McFerran, and W. Kolch. Mathematical modeling of the influence of RKIP on the ERK signaling pathway. In *Computational Methods in Systems Biology*, pages 127–141. Editors: C. Priami. Springer, 2003.

[69] M. Lamboni, H. Monod, and D. Makowski. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4):450–459, 2011.

[70] D. Lebiedz, D. Skanda, and M. Fein. Automatic complexity analysis and model reduction of nonlinear biochemical systems. In *Computational Methods in Systems Biology*, pages 123–140. Editors: M. Bernardo, P. Degano, G. Zavattano. Springer, 2008.

[71] P. D. Leenheer, D. Angeli, and E. D. Sontag. Monotone chemical reaction networks. *Journal of Mathematical Chemistry*, 41(3):295–314, 2007.

[72] J.-C. Leloup and A. Goldbeter. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *Journal of Biological Rhythms*, 13(1):70–87, 1998.

[73] J.-C. Leloup and A. Goldbeter. Toward a detailed computational model for the mammalian circadian clock. *Proceedings of the National Academy of Sciences*, 100(12):7051–7056, 2003.

[74] J.-C. Leloup and A. Goldbeter. Modeling the mammalian circadian clock: sensitivity analysis and multiplicity of oscillatory mechanisms. *Journal of Theoretical Biology*, 230(4):541–562, 2004.

[75] A. Llamosi, A. M. Gonzalez Vargas, C. Versari, E. Cinquemani, G. Ferrari Trecate, P. Hersen, and G. Batt. What population reveals about individual cell identity: Single-cell parameter estimation of models of gene expression in yeast. *PLoS Comput Biol*, 12(2):e1004706, 2016.

[76] G. A. Mackie. RNase E: at the interface of bacterial RNA processing and decay. *Nature Reviews Microbiology*, 11(1):45–57, 2013.

[77] A. Matin. Starvation promoters of *Escherichia coli*: their function, regulation, and use in bioprocessing and bioremediation. *Annals of the New York Academy of Sciences*, 721(1):277–291, 1994.

[78] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer. BioNumbers-the database of key numbers in molecular and cell biology. *Nucleic Acids Research*, 38(suppl 1):D750–D753, 2010.

[79] Miraglio B. Quantitative modeling of endocrine toxicity induced by chemical products in rodents. Technical report, Université Nice Sophia Antipolis, Polytech Nice-Sophia, 08 2014.

[80] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*, 4:55–100, 2006.

[81] J. Monod. *Recherches sur la Croissance des Cultures Bactériennes*. Hermann. 1942.

[82] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[83] K. D. Murray and H. Bremer. Control of *spoT*-dependent ppGpp synthesis and degradation in *Escherichia coli*. *J. Mol. Biol.*, 259:41–57, May 1996.

[84] A. J. Nair. *Comprehensive Biotechnology XI*. Firewall Media, 2010.

[85] A. Narang and S. S. Pilyugin. Bacterial gene regulation in diauxic and non-diauxic growth. *Journal of Theoretical Biology*, 244(2):326–348, 2007.

[86] I. Ndiaye. *Méthodes d'Analyse de Modèles de Régulation Cellulaire*. PhD thesis, Université Nice Sophia Antipolis, 2010.

[87] 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.

[88] H. Pagel, C. Poll, J. Ingwersen, E. Kandeler, and T. Streck. Modeling coupled pesticide degradation and organic matter turnover: From gene abundance to process rates. *Soil Biology and Biochemistry*, 103:349–364, 2016.

[89] D. J. Pannell. Sensitivity analysis of normative economic models: theoretical framework and practical strategies. *Agricultural Economics*, 16(2):139–152, 1997.

[90] V. Petrov, E. Nikolova, and O. Wolkenhauer. Reduction of nonlinear dynamic systems with an application to signal transduction pathways. *Systems Biology, IET*, 1(1):2–9, 2007.

[91] L. Petzold and W. Zhu. Model reduction for chemical kinetics: An optimization approach. *AIChE Journal*, 45(4):869–886, 1999.

[92] 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.

[93] K. Potrykus and M. Cashel. (p)ppGpp: still Magical? *Annu. Rev. Microbiol.*, 62:35–51, 2008.

[94] 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.

[95] C. Robles Rodriguez, C. Bideaux, S. Guillouet, N. Gorret, G. Roux, C. Molina Jouve, and C. Aceves Lara. Multi-objective particle swarm optimization (MOPSO) of lipid accumulation in Fed-batch cultures. In *2016 24th Mediterranean Conference on Control and Automation (MED*, pages 979–984. IEEE, 2016.

[96] D. Rouquié, H. Tinwell, O. Blanck, F. Schorsch, D. Geter, S. Wason, and R. Bars. Thyroid tumor formation in the male mouse induced by fluopyram is mediated by activation of hepatic CAR/PXR nuclear receptors. *Regulatory Toxicology and Pharmacology*, 70(3):673–680, 2014.

[97] A. Saltelli. Sensitivity analysis for importance assessment. *Risk Analysis*, 22(3):579–590, 2002.

[98] A. Saltelli, K. Chan, E. M. Scott, et al. *Sensitivity Analysis*, volume 1. Wiley New York, 2000.

[99] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: the Primer*. John Wiley & Sons, 2008.

[100] A. Saltelli, S. Tarantola, and K.-S. Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, 1999.

[101] M. Schelker, A. Raue, J. Timmer, and C. Kreutz. Comprehensive estimation of input signals and dynamics in biochemical reaction networks. *Bioinformatics*, 28(18):i529–i534, 2012.

[102] I. Segel. *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady State Enzyme Systems*. Wiley & Sons, 1993.

[103] L. A. Segel and M. Slemrod. The quasi-steady-state assumption: a case study in perturbation. *SIAM Review*, 31(3):446–477, 1989.

[104] Z. Shajani, M. T. Sykes, and J. R. Williamson. Assembly of bacterial ribosomes. *Annual Review of Biochemistry*, 80:501–526, 2011.

[105] H. Smith. Cooperative systems of differential equations with concave nonlinearities. *Nonlinear Analysis: Theory, Methods & Applications*, 10(10):1037–1052, 1986.

[106] H. L. Smith. On the asymptotic behavior of a class of deterministic models of cooperating species. *SIAM Journal on Applied Mathematics*, 46(3):368–375, 1986.

[107] H. L. Smith. *Monotone Dynamical Systems: an Introduction to the Theory of Competitive and Cooperative Systems*, volume 41. American Mathematical Soc. Mathematical surveys and monographs, 1995.

[108] 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.

[109] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

[110] M. Sonderegger, M. Schümperli, and U. Sauer. Selection of quiescent *Escherichia coli* with high metabolic activity. *Metabolic Engineering*, 7(1):4–9, 2005.

[111] C. Sonich Mullin, R. Fielder, J. Wiltse, K. Baetcke, J. Dempsey, P. Fenner Crisp, D. Grant, M. Hartley, A. Knaap, D. Kroese, et al. IPCS conceptual framework for evaluating a mode of action for chemical carcinogenesis. *Regulatory Toxicology and Pharmacology*, 34(2):146–152, 2001.

[112] E. D. Sontag. Some new directions in control theory inspired by systems biology. *Systems Biology*, 1(1):9–18, 2004.

[113] A. Srivatsan and J. D. Wang. Control of bacterial transcription, translation and replication by (p)ppGpp. *Current Opinion in Microbiology*, 11(2):100–105, 2008.

[114] J. Stülke and W. Hillen. Carbon catabolite repression in bacteria. *Current Opinion in Microbiology*, 2(2):195–201, 1999.

[115] 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, 2008.

[116] T. Turanyi and H. Rabitz. Local methods. *Sensitivity analysis*, pages 81–99, 2000.

[117] J. J. Tyson, C. I. Hong, C. D. Thron, and B. Novak. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophysical Journal*, 77(5):2411–2417, 1999.

[118] G. T. Van Der Horst, M. Muijtjens, K. Kobayashi, R. Takano, S.-i. Kanno, M. Takao, J. de Wit, A. Verkerk, A. P. Eker, D. van Leenen, et al. Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. *Nature*, 398(6728):627–630, 1999.

[119] A. Wada. Growth phase coupled modulation of *Escherichia coli* ribosomes. *Genes to Cells*, 3(4):203–208, 1998.

[120] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[121] E. Walter, L. Pronzato, and J. Norton. *Identification of Parametric Models from Experimental Data*, volume 1. Springer Berlin, 1997.

[122] F. Widdel. Theory and Measurement of Bacterial Growth. Technical report, Bremen: University Bremen, 2007.

[123] B. Zheng, U. Albrecht, K. Kaasik, M. Sage, W. Lu, S. Vaishnav, Q. Li, Z. S. Sun, G. Eichele, A. Bradley, et al. Nonredundant roles of the *mPer1* and *mPer2* genes in the mammalian circadian clock. *Cell*, 105(5):683–694, 2001.