Modelling, Analysis and Verification of Biological Coherent Feedforward Loop Network

Julia Smeu

Master of Science Thesis

Delft University of Technology

Delft Center for Systems and Control

Modelling, Analysis and Verification of Biological Coherent Feedforward Loop Network

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Julia Smeu

September 11, 2019

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

The world of molecular biology is composed by a complex network of interactions that are analogous to electric circuits. They govern the functions required for life, from metabolism to locomotion. In these networks, the presence of network motifs were identified, recurring elements supposedly kept by evolution. One of them is called the feedforward loop and has the function of a sign-sensitive delay element or noise-filter. Moreover, different combinations of several types of feedforward loops were identified in the transcription networks of Escherichia coli and Saccharomyces cerevisiae, called complex feedforward loops. From this finding a question arises: do different types of combined feedforward loops have a specific function? Would this identified function be useful in synthetic biology applications? Answering these questions is the ultimate goal of a research direction in systems biology, studied at the Institute of Complex Molecular Systems (ICMS) at Eindhoven University of Technology. However, biological experiments are difficult to setup and conduct in a suitable manner to generate relevant results. Therefore, it would be highly effective to be able to predict the nonlinear dynamical behaviour of these (combined) feedforward loops. Nevertheless, in order to be able to achieve this, first a single feedforward loop must be fully modelled, calibrated and analysed. This master thesis focuses on this goal and is composed of three main elements: modelling, parameter estimation and structural analysis. The modelling section comprises of the methodology derived in order to transpose the biochemical reactions into equations and perform model reduction on the feedforward loop built at ICMS. Then, a hybrid parameter estimation method was applied successfully and made it possible to perform numerical simulations of the system. Lastly, the focus was directed to structural analysis and obtaining insights about the behaviour of the network without knowledge of the parameters. This included the adaptation of metabolic network analysis tools, elementary flux mode analysis and flux balance analysis to be used on gene expression networks. As a result, it was possible to link the nonlinearity of the steady-states observed in the experimental data with the accumulation of certain compounds.

Contents

List of Figures

List of Tables

Acknowledgements

I could describe these past two years in many ways: difficult and mentally challenging, inspiring and helping me to develop tremendously. I have learned a lot, both about the field I chose and about myself during this master's degree. However, all of this could not have been possible without the following people.

First of all, I would like to thank my supervisor dr.ir. Erik Steur for all the invaluable help I was given during the thesis work and for all the reassurance I got when I was full of self-doubt.

I am wholeheartedly grateful to dr. Carlos Robles Rodriguez whom I was lucky enough to meet halfway through my thesis. I have really enjoyed working together with you and I would like to thank you for managing to change my negative thoughts about my project into positive ones.

Lastly, I would like to thank all the people that kept me going in difficult times and supported me all the steps of the way. My parents, who were and will always be my strong pillars. The friends with whom I have sweated together during this master's degree and formed my family in Delft: Ola, Karol, Leo, Patrick, Mátě, Daniel, Bart, Barbara and Nirmal. I have learnt so much from all of you.

Delft, University of Technology Julia Smeu
September 11, 2019

Chapter 1

Introduction

"Every object that biology studies is a system of systems."

— *Francois Jacob (1974)*

This chapter provides a brief introduction to the field of systems biology. More precisely, the origin of the research area, the theory and applications related to it and the connection with the field of systems and control. Subsequently, the focus will be put on the project work that was completed during the master thesis and an overview of it is presented in continuation. Moreover, concepts central to the project will be introduced, jointly with the main components of the work completed. In addition, an overview of the experimental process is described together with the analysis of the experimental data provided by the Institute of Complex Molecular Systems (ICMS) at Eindhoven University of Technology.

1-1 The Field of Systems Biology

Systems biology is considered to have two historical roots [6], both originating from molecular biology. One of them is represented by the discovery and study of the genetic material and the second one is made up of nonequilibrium thermodynamics theory. Both of them emerged from the first half of the 19th century and prior to their development biology was an isolated field of research, not linked to mathematics, systems theory and engineering. However, as a result of the pioneering work of Alan Turing published in the paper "The Chemical Basis of Morphogenesis" [7], mathematical modelling was introduced to the world of molecular biology. Another important figure to mention is Ludwig von Bertalanffy who is considered to be the father of general systems theory [8]

and focused on studying organisms as a 'whole' [9], also applying modelling principles on self-regulating systems. In addition, another notable milestone in the history of systems biology is the discovery of the lac operon's regulation in bacteria Escherichia coli (E.Coli) by Jacob and Monod in 1961 [10]. This was revolutionary as it showed that the gene expression process can be seen as a dynamical system with inputs and outputs [11].
Systems biology has come a long way since then, encompassing a research field on its own and uniting researchers from multiple disciplines: biology, systems and control, chemistry, computer science and engineering. An accurate definition of the research direction was formulated in the paper [12] in the following way:

"Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations."

Therefore, one of the significant aims of this field is to describe the behaviour of biological systems using mathematical laws. Having achieved this, some researchers started a new direction of pioneering work resulting from the theory developed from systems biology. In 1994, Leonard M. Adleman put all the theory in practice and developed the first DNA-based computer. This triggered the construction of de novo synthetic gene circuits like the oscillator [13] and the toggle switch [14]. These constitute one of the first big milestones in synthetic biology, an engineering field that focuses on the design and manipulation of artificial biological systems with specific application purposes. As a result, the first real-world applications were developed which were environmental biosensors and biofuel production pathways [11].

The two research fields of systems biology and synthetic biology have a 'symbiotic relationship': advances in one help the development of the second one and vice-versa [15]. The products of synthetic biology are considered to be analogous to electrical circuits, hence the term genetic circuits. The development of these has enormous potential in energy, environmental and medical applications. However, in order to get major breakthroughs, challenges outside the sphere of molecular biology have to be tackled. One category of challenges are 'system-level' problems and represent the necessity of application of concepts from control theory. In paper [11] there were three main challenges identified that are encountered during the development of synthetic biological systems: lack of modularity and compositionality, emergent behaviours from stochasticity and interactions between spatially distributed dynamics. In response to this challenges, research opportunities were identified for systems and control field. These include directions like developing a proper system identification methodology for biomolecular systems, exploiting time-scale separation for simplifying dynamics and increasing modularity and development of modelling frameworks.
The work completed and presented in the master thesis report has as its base the application of systems and control theory tools to the modelling and analysis of a biological system. More precisely, it tackles the above mentioned challenges by deriving a mod-

elling method for gene expression, implementing system identification using different parameter estimation methods and performing structural analysis using mathematical tools.

1-2 The Coherent Feedforward Loop

Before presenting an overview of the master thesis, the biological network motif of coherent feedforward loop (CFFL) has to be introduced. In the following its function and general structure will be presented. This bio-chemical reaction network constitutes the central element of the project. It is defined to be a network motif as its appearance number is higher in the transcription networks of bacteria E.Coli [16] and Saccharomyces cerevisiae [17] than in randomised networks. The main identified function of the CFFL is of a sign-sensitive delay element or noise-filtering device. This function is considered to represent one possible explanation for the reason why evolution kept this motif in biological networks [1].

The feedforward loop is usually illustrated as a 3-node network. The exact chemical complexes that represent these nodes usually differ, alternative variations of it can be found in transcription networks or built in laboratories. In literature [1] [18] usually one of the nodes is represented by transcription factor X, that regulates a second node, transcription factor Y. Both X and Y regulate the third element which is represented by gene Z (fig. 1-1). As a consequence, the network contains two regulation paths, a direct one from X to Z and an indirect one through Y. Similarly to the nodes variation, the way this combined regulation by X and Y is implemented also can differ from one feedforward loop to another. Moreover, according to the nature of the regulation paths (activating or inhibiting), the feedforward loop can be of different types. The project focuses only on type 1 CFFL in which all paths are of activating nature. A more detailed description of these types and the breakthrough experiment that was conducted to prove the noise-filtering behaviour of the biological system can be found in the literature study written prior to this thesis report [19]. In addition, the functioning of the device is illustrated in fig. 1-1. The input to the system is the signal S_x that activates transcription factor X. For a short perturbation of transcription factor X, only a limited amount of transcription factor Y is produced, which is not enough to reach the threshold in order to start the production of Z. However, if the impulse for transcription factor X has longer time period, then Y is produced in higher quantities and surpasses the threshold. Therefore, gene Z production starts as well.

The present master thesis project is based on the CFFL built at ICMS at Eindhoven University of Technology. In order to have a better understanding of the modelling and analysis of this specific genetic circuit, in the following subsection a thorough description of it will be presented.

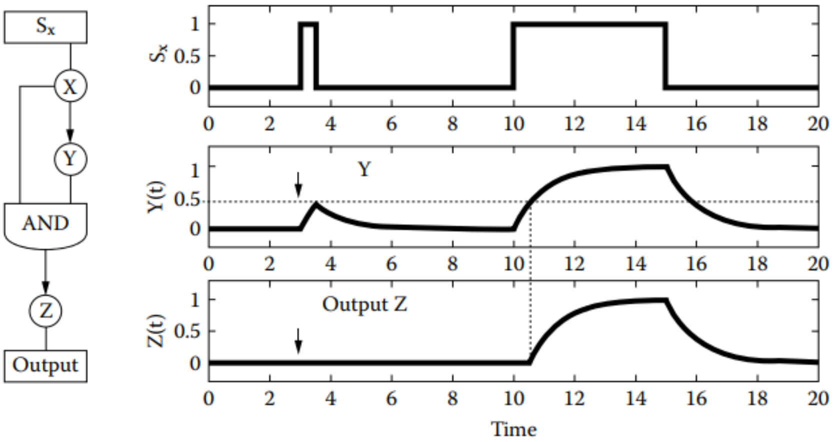


Figure 1-1: The feedforward loop network motif structure and its noise-filtering behaviour visualised [1]

1-2-1 Coherent Feedforward Loop built at ICMS

The type of feedforward loops found in literature [1] [18] are structurally the same as the CFFL analysed in the current report. However, component-wise they differ. More specifically, this can be observed by the composition of the nodes. The three nodes are represented σ-factor 70 (S70) and σ-factor 28 (S28) and the green fluorescent protein (eGFP). This biological system has one input, specifically the DNA template called, $DNA_{trigger}$ and one output, eGFP. These elements can be seen in fig. 1-2 and fig. 1-3.

The core reactions making up the CFFL are based on the process of protein synthesis. All of the three mentioned nodes are proteins, two of them (S28 and eGFP) are produced during the functioning of the network. The protein synthesis process can be divided into two main reaction sets: transcription and translation. During transcription, RNA is produced and then it is used as an input to translation which has the corresponding protein as its end-product. Therefore, the schematic from fig. 1-2 is extended to fig. 1-3 according to the processes of protein synthesis. From this it can be seen that there is one input to the network, $\mathrm{DNA_{trigger}}$ and it is varied during the experimental process. In order for the network to start functioning there are also five other chemical species added to the mixture: RNAP, S70, $\mathrm{DNA_{S28}}$, $\mathrm{DNA_{eGFP}}$ and Ribo. However the concentration of these species stays constant. Therefore, $\mathrm{DNA_{trigger}}$ is considered as the single input to the network. In the first stage of transcription reactions, $\mathrm{RNA_{trigger}}$ and $\mathrm{RNA_{S28}}$ are produced. These two species are the inputs to the translation reaction to output S28, the intermediate protein and also second node of the CFFL. Then, in the next step S28 is used in the third transcription reaction to produce $\mathrm{RNA_{eGFP}}$. This in turn is used as input to the second translation reaction to output eGFP.

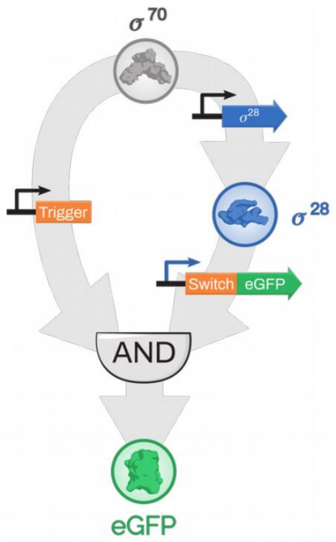


Figure 1-2: CFFL used in the thesis project - schematic made by Pascal Pieters from Eindhoven University of Technology

Moreover, there is a second significant difference in the composition of the CFFL from the project compared to the feedforward loops found in literature. This is represented by the structure of the AND gate combining the direct and indirect path from the network. In order to make sure that both the first two nodes regulate the output protein, a toehold switch is used. This is an RNA-based AND gate that requires the binding of the different RNA elements in order to start translation. A more detailed description of this mechanism can be found in [19].

1-2-2 Experimental Procedure

In the following section a short overview and the significant aspects of the experimental procedure will be presented. Firstly, the experiments are completed in vitro. Therefore, the reactions taking place in the CFFL are not influenced by any external processes that would be happening in a cell. Secondly, at the ICMS at Eindhoven University of Technology there are two types of experimental procedures used: batch and flow experiments.

Batch experiments represented the first stage of the research work. The required chemical species for the gene expression process were loaded into a microfluidic reactor that did not use the flow inlet/outlets. Then the trigger or input DNA template was added to the mixture and the reactions were let to take place without adding any inflow or outflow. The transcription and translation processes use up all the input chemical species and the reactions halt after a specific time. During the process, the fluorescence level of the output eGFP protein was measured. This is proportional to the concentration of the

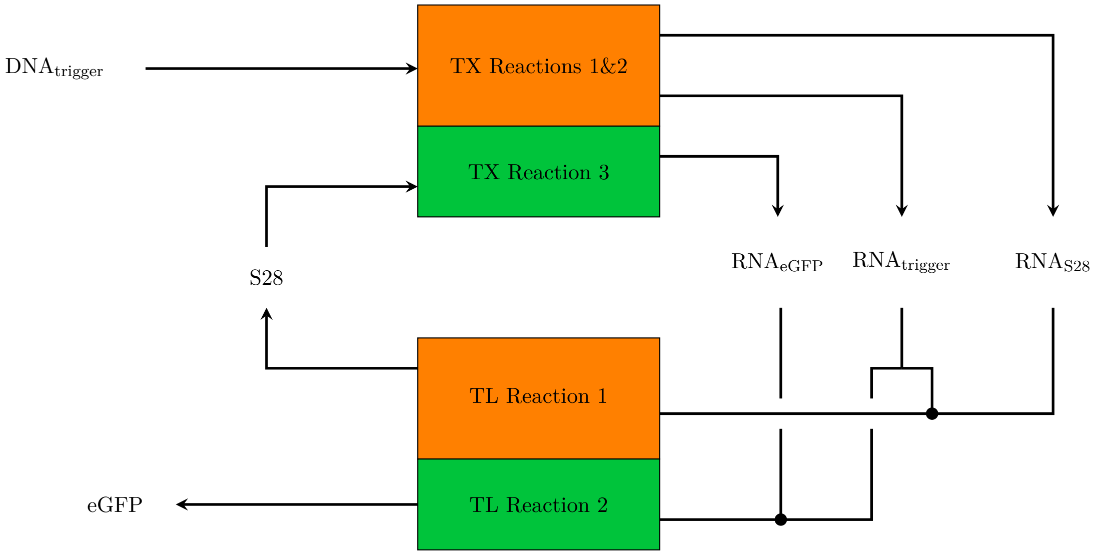


Figure 1-3: Extended schematic of CFFL

compound, therefore fluorescence is translated into concentration values which build up the experimental data. This is used for parameter estimation later on and it is visualized in fig. 1-4. The five different datasets were acquired from five different experiments. Therefore, it is likely that the initial conditions of the experiments differed from one to another. Also, it is possible to have slightly different reaction rate constants because of changes in temperature or any other discrepancy in the setup of the experiment. The initial concentrations of three chemical species are not known precisely, a range of values were given and can be found in table 1-1.

Parameters	Cell-free Reaction
RNAP [nM]	60-75 nM
S70 [nM]	<35 nM
Ribosomes [nM]	<2300 nM

Table 1-1: Ranges of initial concentrations - batch experiment

So far, successful batch experiments were completed and yielded the results that can be seen in fig. 1-4. This shows the production of the output protein, eGFP, according to the amount of input DNA$_{\text{trigger}}$ added to the experimental process. The behaviour that corresponds to the sign-sensitive delay element can be recognised: the start time of the production of eGFP is the largest when the input concentration of DNA$_{\text{trigger}}$ is the lowest. In addition, an important observation to make is the non-linearity in the amplitudes of the datasets. For the first three inputs, the amplitude increases as the input value increases as well. However, for the fourth and fifth dataset, this is not valid any more and the output settles at lower values. The noise-filter function was not verified yet as it is dynamic in character which can only be validated using flow experiments.

Therefore, the second stage comprises of adding an inflow and outflow to the reaction mixture. In this way it is made sure that the reactions don't halt and the noise-filtering behaviour of the CFFL can be experimentally tested. The researchers at the ICMS are currently working on successfully conducting flow experiments of the CFFL. More details about the technical setup of the flow experiments can be found in [19] and a schematic of the microfluidic reactor with the inlets and outlets can be seen in fig. 2-2.

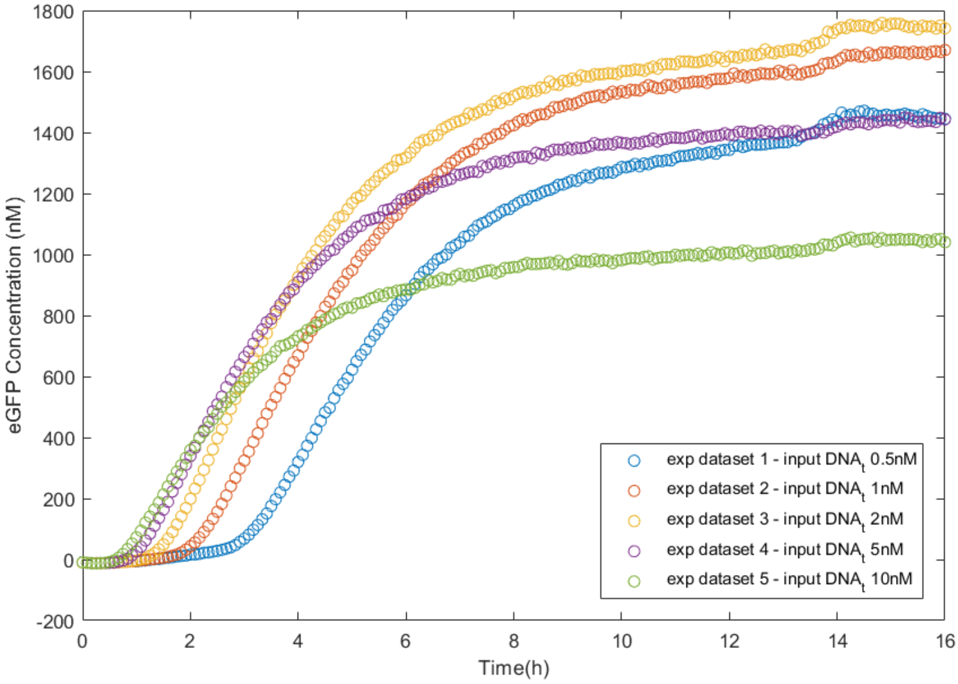


Figure 1-4: Data from Batch Experiments

1-3 Master Thesis Project Description

So far the background research field of the project and its centre element, the CFFL, was described. Moreover, the experimental procedure followed at the ICMS was also presented. In the following the research direction that guided this project will be discussed. In addition, the research objective of the master thesis will be presented, together with an overview of thesis work that was completed during the project.

1-3-1 Research Direction at the Institute for Complex Molecular Systems at Eindhoven University of Technology

As mentioned in section 1-1, the main research objective of synthetic biology is to develop de novo genetic circuits for specific application purposes. At ICMS at Eindhoven University of Technology, researchers are achieving this by focusing their work on developing combined feedforward loops. There is no knowledge available on how will these genetic circuits behave and which one of them will have potential for synthetic biology applications. Nevertheless, developing these biological circuits and conducting successful experiments in order to observe its behaviour is a cumbersome process that requires a lot of time and meticulous work in setting up the experiments. Therefore, finding a way to model and simulate these genetic circuits in advance would be a tremendous help in conducting the research work more efficiently.

1-3-2 Master Thesis Research Objective

As described in the previous subsection, there are several challenges that molecular biologists from Eindhoven University of Technology face during conducting the batch and flow experiments for a single biological network. However, their aim is to develop multiple of these genetic circuits. Accordingly, a range of solutions have to be found in order to make this entire procedure more efficient and to be able to conduct in silico experiments that already provide insights about biological networks.

Therefore, the master thesis titled 'Modelling, Analysis and Verification of Biological Coherent Feedforward Loop Network' provides a collection of tools that are implemented in different stages of the process of studying a biological network motif.

Firstly, a modelling methodology was developed to apply on the CFFL. This was specifically tailored to the type of chemical complexes (holoenzymes) used in the configuration of the network motif built at the ICMS. Nevertheless, in case of different structure, the modelling strategy could still be applied with a few modifications.

Secondly, a parameter estimation process was designed to conduct system identification on the CFFL. A hybrid method was applied which combined the performance of several optimization algorithms. Moreover, a complete calibration procedure is shown together with the analysis of the results.

Thirdly, theory and tools were studied and implemented from structural analysis of biochemical reaction networks. These are represented by methods that are applied in case knowledge about the parameters of the model is not available at all. All insights are generated based on the structure and stoichiometry of the network. Moreover, two tools used in the analysis of metabolic networks were implemented on transcription reactions successfully. These generated results about the static dynamics of the CFFL.

As a consequence, combining all these elements the aim was to develop a framework for modelling and analysing biological systems. Moreover, the goal was to make it possible for it to be applied in order to perform functional analysis not only on a single network motif but also on combined ones.

In summary, the following are the contributions of the master thesis:

1. **Development of the mathematical model of the CFFL**. The result is a set of ordinary differential equations describing the dynamics of the studied network motif. With the help of it, not only the behaviour of the output can be observed but also the dynamics of the other state variables, the other chemical species taking part in the reaction network. During this process, model reduction was also implemented.

2. **Implementation of a hybrid parameter estimation method to calibrate the CFFL**. There is only a limited amount of insights that can be gained by analysing the bio-chemical reaction network without the knowledge of the parameters. In order to be able to simulate the model, it is necessary to estimate the parameters which are represented by reaction rates and Michaelis-Menten constants. The system identification procedure was completed by using a particle swarm optimization algorithm in combination with patternsearch method.

3. **Development of a set of structural analysis tools and implementation on the CFFL**. It was desired to study what information can be deducted from the structural characteristics of the biological network. Therefore, the capacity to admit multiple equilibria was confirmed by two software packages developed for chemical reaction network analysis. Moreover, software tools used for analysis of metabolic networks were implemented and provided insights about the static dynamics of the network. The results were confirmed by the numerical simulations resulting from the parameter estimation procedure

4. **Implementation of metabolic network analysis tools on gene expression networks**. Elementary flux mode analysis and flux balance analysis are two commonly used methods in the analysis of metabolic networks. However, the reactions composing gene expression networks differ from the reactions constituting metabolic networks. As a result, a way had to be found to input the transcription/translation reactions to generate suitable results.

1-3-3 Organization of the Thesis Report

The remainder of this master thesis report comprises of the project work completed that resulted in the contributions described above. It was structured based on the three main components that build up the thesis: mathematical modelling, parameter estimation and structural analysis. In the following, the subsequent chapters of the report will be summarised.

The first part, encompassed by the second chapter of this master thesis report, comprises of the derivation of the mathematical model of the CFFL. Modelling of biological networks was identified as a challenge in systems biology that requires systems and control theory. One relevant reason for this is that biological systems are highly complex therefore in most of the cases even the biochemical reactions representing the network are an approximation of the actual system. As a consequence, mathematical modelling in this case significantly differs from the usual process found in physical sciences and engineering. In these fields models play a primary role and is possible to use them to perform full analysis and make hypothesis based on them. In biology however, mathematical models are merely used to 'document' experimental results and are actually derived based on the data available. In contrast, during this project a mathematical model is derived based on the chemical reaction network given by the researchers from the ICMS from Eindhoven University of Technology. The derivation is split into several parts and the details of how the reactions are transposed into ordinary differential equations are presented.

The second part comprises of completing system identification and it is the main topic of the third chapter. The biological model contains unknown reaction rates and Michaelis-Menten constants. However, having an estimation of the parameters makes it possible to perform numerical simulation of the CFFL to predict its behaviour. It also aids in input-output relationship analysis and gives an insight into the dynamics of the other chemical species represented by the other states of the model. These can not be observed during the experimental procedure. Therefore, in this way there is information about the underlying dynamics.

The fourth chapter encompasses the dynamics analysis of the network motif. In the previous chapter, one of the goals was to analyse the system dynamics for specific sets of parameter values. However, in this part the aim is to gain insights about the behaviour of the CFFL without the knowledge of the parameters. This will be achieved by using structural analysis methods to confirm the capacity for multiple equilibria. Moreover, a novel way of applying metabolic network analysis tools (elementary flux modes analysis and flux balance analysis) on transcription network will be demonstrated. In addition, the results from it will be discussed.

Lastly, the results and findings of the project work will be summarised and discussed in the final chapter of the master thesis report. Moreover, suggestions for further work will be given.

Chapter 2

Modelling of the Coherent Feedforward Loop

A relevant challenge in research fields that comprises of multidisciplinary work is understanding the discrepancies present between researchers coming from different backgrounds. A good example of this is the concept of model. For molecular biologists, the model of a biological network is represented by a set of biochemical reactions. On the other hand, systems and control researchers associate the term model with a set of ordinary differential equations that describe the system's dynamics. However, in order to get to this mathematical model, the biochemical reaction model has to be transposed into mathematical equations. There is no set methodology to do this, there are different directions that can result in different results. However, what is important is to understand the assumptions that can be formulated and identify the required level of complexity in order to get a model that is suitable for further work.

The following chapter encompasses the first part of the project work that corresponds to the mathematical modelling of the network motif of coherent feedforward loop (CFFL). This is a fundamental part of the thesis as the subsequent parameter estimation and dynamics analysis will have the derived model as their basis. Firstly, a few choices prior to the derivation of the equations will be motivated. Taking these into account, the method of modelling the transcription and translation processes will be presented. Lastly, an overview of the model will be given.

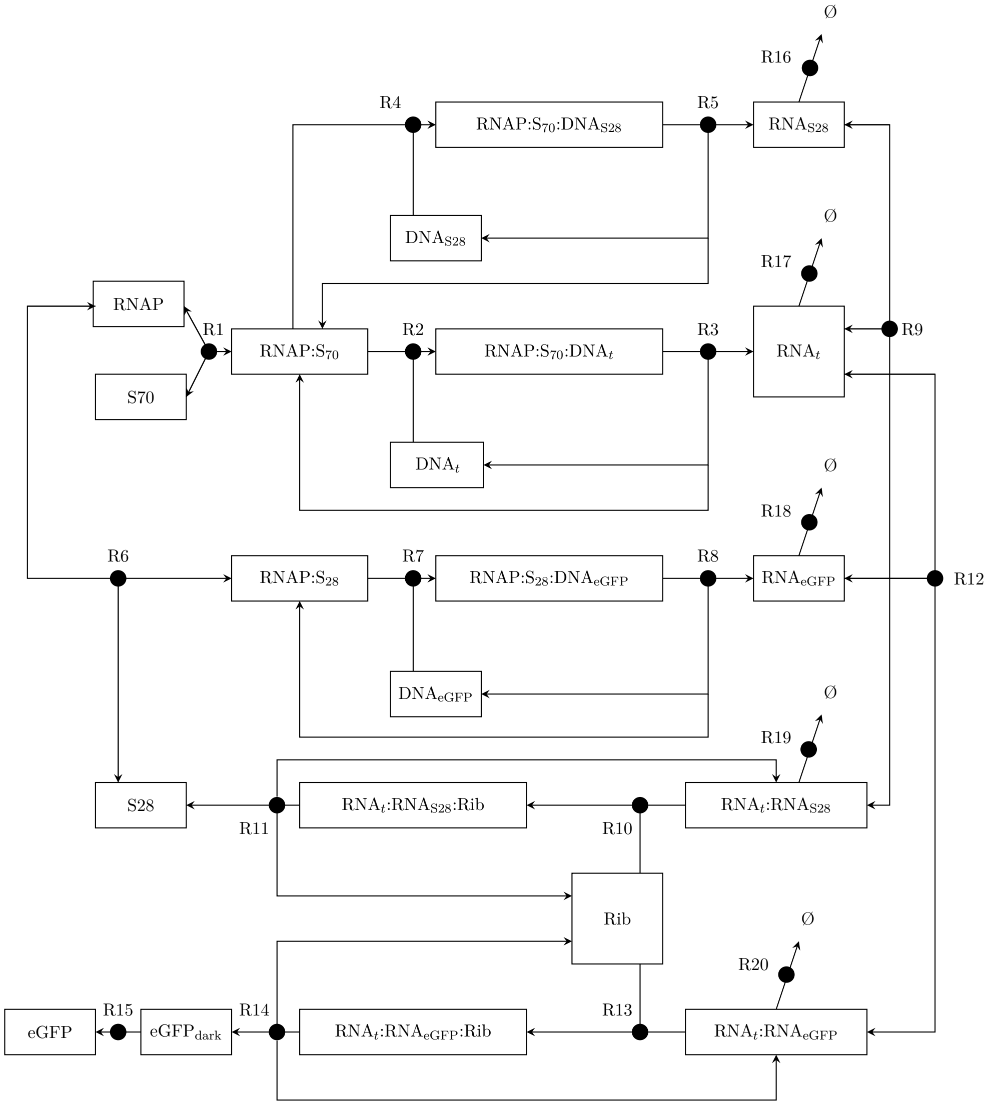


Figure 2-1: The biochemical reaction network of the CFFL

2-1 Deterministic Modelling

There are two separate starting directions when it comes to modelling of biochemical networks. One is called stochastic modelling which takes into account different possibilities to include the different microstates of the system [3]. In addition to this, there is the choice of deterministic modelling which is less complex however also less accurate. Prior to the actual process of modelling, a choice had to be made between these two frameworks. During the project deterministic modelling was chosen. The motivation for this is that there is no need for a higher-dimensional representation of the dynamics of the system. Even on the contrary, one of the aim is to have as simple representation of the dynamics as possible without compromising too much on accuracy by applying model reduction. Moreover, the given reaction network from the Institute of Complex Molecular Systems (ICMS) is an approximation of the highly complex chemical reaction network that corresponds to the coherent feedforward loop. For example, at a higher detail level perspective, transcription consists of the conversion of nucleotides to messenger RNA [20]. However, the several types of nucleotides are not included into the model. Therefore, choosing deterministic modelling was a suitable choice in order to transpose the bio-chemical reactions into ordinary differential equations.

2-2 From Reactions to Ordinary Differential Equations

The following section will focus on the methodology developed that transposes the chemical reactions forming the bio-chemical network into ordinary differential equations. A set of reactions was made available by the ICMS. Using these, mass-action modelling is applied. Subsequently, model reduction is achieved by using conservation laws derived from the stoichiometry matrix and by applying temporal differentiation of the reactions. The biological model provided by ICMS can be found in appendix A-1.

The first step is to convert the reactions composing the biological model into ordinary differential equations. This will be achieved by applying deterministic modelling, more specifically mass-action law. More details about this can be found in the literature study report accompanying the master thesis [19]. Therefore, the following system of ordinary differential equations is derived:

$$
\begin{aligned}
\dot{[RNAP]} =& k_{-1}[RNAP:S70] - k_1[RNAP][S70] + k_{-6}[RNAP:S28]- \\
&- k_6[RNAP][S28] \\
\dot{[S70]} =& k_{-1}[RNAP:S70] - k_1[RNAP][S70] \\
\dot{[DNA_t]} =& - k_2[RNAP:S70][DNA_t] + k_3[RNAP:S70:DNA_t] \\
\dot{[DNA_{S28}]} =& - k_4[RNAP:S70][DNA_{S28}] + k_5[RNAP:S70:DNA_{S28}] \\
\dot{[DNA_{eGFP}]} =& - k_7[RNAP:S28][DNA_{eGFP}] + k_8[RNAP:S28:DNA_{eGFP}] \\
\dot{[RNAP:S70]} =& - k_{-1}[RNAP:S70] + k_1[RNAP][S70]- \\
&- k_2[RNAP:S70][DNA_t] + k_3[RNAP:S70:DNA_t]- \\
&- k_4[RNAP:S70][DNA_{S28}] + k_5[RNAP:S70:DNA_{S28}] \\
\dot{[RNAP:S70:DNA_t]} =& k_2[RNAP:S70][DNA_t] - k_3[RNAP:S70:DNA_t] \\
\dot{[RNA_t]} =& k_3[RNAP:S70:DNA_t] - k_9[RNA_t][RNA_{S28}]+ \\
&+ k_{-9}[RNA_t:RNA_{S28}] - k_{12}[RNA_t][RNA_{eGFP}]+ \\
&+ k_{-12}[RNA_t:RNA_{eGFP}]
\end{aligned}
\tag{2-1}
$$

$$
\begin{aligned}
\dot{[RNAP : S70 : DNA_{S28}]} =& k_4[RNAP : S70][DNA_{S28}] - k_5[RNAP : S70 : DNA_{S28}] \\
\dot{[RNA_{S28}]} =& k_5[RNAP : S70 : DNA_{S28}] - k_9[RNA_t][RNA_{S28}] + \\
& + k_{-9}[RNA_t : RNA_{S28}] \\
\dot{[RNAP : S28]} =& - k_{-6}[RNAP : S28] + k_6[RNAP][S28] - \\
& - k_7[RNAP : S28][DNA_{eGFP}] + k_8[RNAP : S28 : DNA_{eGFP}] \\
\dot{[RNAP : S28 : DNA_t]} =& k_7[RNAP : S28][DNA_{eGFP}] - k_8[RNAP : S28 : DNA_{eGFP}] \\
\dot{[RNA_{eGFP}]} =& k_8[RNAP : S28 : DNA_{eGFP}] - k_{12}[RNA_t][RNA_{eGFP}] + \\
& + k_{-12}[RNA_t : RNA_{eGFP}] \\
\dot{[Ribo]} =& -k_{10}[RNA_t : RNA_{S28}][Ribo] + k_{11}[RNA_t : RNA_{S28} : Ribo] - \\
& - k_{13}[RNA_t : RNA_{eGFP}][Ribo] + k_{14}[RNA_t : RNA_{eGFP} : Ribo] \\
\dot{[RNA_t : RNA_{S28}]} =& k_9[RNA_t][RNA_{S28}] - k_{-9}[RNA_t : RNA_{S28}] - \\
& - k_{10}[RNA_t : RNA_{S28}][Ribo] + k_{11}[RNA_t : RNA_{S28} : Ribo] \\
\dot{[RNA_t : RNA_{S28} : Ribo]} =& k_{10}[RNA_t : RNA_{S28}][Ribo] - k_{11}[RNA_t : RNA_{S28} : Ribo] \\
\dot{[S28]} =& k_{11}[RNA_t : RNA_{S28} : Ribo] + k_{-6}[RNAP : S28] - \\
& - k_6[RNAP][S28] \\
\dot{[RNA_t : RNA_{eGFP}]} =& k_{12}[RNA_t][RNA_{eGFP}] - k_{-12}[RNA_t : RNA_{eGFP}] - \\
& - k_{13}[RNA_t : RNA_{eGFP}][Ribo] + k_{14}[RNA_t : RNA_{eGFP} : Ribo] \\
\dot{[RNA_t : RNA_{eGFP} : Ribo]} =& k_{13}[RNA_t : RNA_{eGFP}][Ribo] - k_{14}[RNA_t : RNA_{eGFP} : Ribo] \\
\dot{[eGFP_{dark}]} =& k_{14}[RNA_t : RNA_{eGFP} : Ribo] - mat[eGFP_{dark}] \\
\dot{[eGFP]} =& mat[eGFP_{dark}]
\end{aligned}
\tag{2-2}
$$

The resulting system is composed of 21 states and 19 unknown parameters. In the following, model reduction techniques will be applied to eliminate some of the states and simplify the system. However, a priori to this, conservations laws of the CFFL will be derived.

2-3 Conservation Laws from Stoichiometry

A step necessary for model reduction is analysing the stoichiometric matrix. More specifically, it can be used to identify the conservation laws of the CFFL. Prior to this, a concise way of representing the biochemical network is introduced in the following form:

$$\dot{x} = Sv(x) \tag{2-3}$$

where $S_{n \times m}$ is the stoichiometric matrix, $x \in \mathbb{R}^n_+$ is the non-negative vector of concentrations of n chemical species, and $v(x) \in \mathbb{R}^m$ represents the reaction flux vector of m

reactions. Without considering the nature of the dynamics found in $v(x)$ yet, the stoichiometric matrix is built by using the stoichiometric coefficients of the reactions and it is organised such that every column corresponds to a reaction and every row corresponds to a chemical species. In order to obtain a better understanding of the composition of the stoichiometric matrix the following example is introduced:

$$\mathrm{A} + \mathrm{B} \underset{\mathrm{k}_-}{\overset{\mathrm{k}_+}{\rightleftharpoons}} \mathrm{C} \tag{2-4}$$

This represents two reactions, a forward and reverse one and contains 3 species. Therefore, the stoichiometric matrix will be 3 × 2. The mass-action model corresponding to it is the following:

$$\begin{aligned}\frac{d[A]}{dt} &= k_-[C] - k_+[A][B]\\ \frac{d[B]}{dt} &= k_-[C] - k_+[A][B]\\ \frac{d[C]}{dt} &= k_+[A][B] - k_-[C]\end{aligned} \tag{2-5}$$

Therefore, the sample system can be written up in the following way:

$$Sv(x) = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} k_+[A][B] \\ k_-[C] \end{bmatrix} \tag{2-6}$$

As mentioned, each row corresponds to a compound. In this example first two rows represent concentrations of A and B, while the third row corresponds to the product C. Therefore, matrix entry -1 represents the corresponding chemical species being consumed. On the other hand, matrix entry 1 corresponds to the production of the compound.

The CFFL is initially built by considering every chemical compound from the biochemical reaction network that is visualised in fig. 2-1. The resulting stoichiometric matrix can be found in appendix A-2.
Computing the right null space of the matrix results in finding the steady-state flux distributions through the network while the left null space results in the conservation laws. In order to achieve this, a MATLAB toolbox called METATOOL [21] was used. The computed right-null space and left-null space can be seen in appendix A-3.

The left null-space of the stoichiometry matrix gives the following conservation laws:

$$[\mathrm{DNA}_t] + [\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_t] = C_1 \tag{2-7}$$

$$[\mathrm{DNA}_{\mathrm{S28}}] + [\mathrm{DNA}_t] - [\mathrm{S}_{70}] - [\mathrm{RNAP}:\mathrm{S}_{70}] = C_2 \tag{2-8}$$

$$[\mathrm{S}_{70}] + [\mathrm{RNAP}:\mathrm{S}_{70}] - [\mathrm{DNA}_t] + [\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_{\mathrm{S28}}] = C_3 \tag{2-9}$$

$$[\mathrm{RNAP}] - [\mathrm{S}_{70}] + [\mathrm{RNAP}:\mathrm{S}_{28}] + [\mathrm{RNAP}:\mathrm{S}_{28}:\mathrm{DNA}_{\mathrm{eGFP}}] = C_4 \tag{2-10}$$

$$[\mathrm{S}_{70}] - [\mathrm{RNAP}] - [\mathrm{RNAP}:\mathrm{S}_{28}] + [\mathrm{DNA}_{\mathrm{eGFP}}] = C_5 \tag{2-11}$$

$$[\mathrm{Ribo}] + [\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}}:\mathrm{Ribo}] + [\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}}:\mathrm{Ribo}] = C_6 \tag{2-12}$$

The last conservation law eq. (2-12) is straightforward, it contains all chemical compounds that include Ribo in it. This was expected as the concentration of ribosomes that is added to the reaction mixture is fixed, therefore the sum of concentrations of the chemical complexes containing it is constant. The other conservation laws have to be grouped in a way that it reflects the same logic as in the case of the ribosomes. This was achieved by adding up eq. (2-7), eq. (2-9) and eq. (2-10). Therefore, this results in a concentration law that comprises of every complex that contains RNAP:

$$\begin{aligned} &[\mathrm{RNAP}] + [\mathrm{RNAP}:\mathrm{S}_{70}] + [\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_t] + [\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_{\mathrm{S28}}] + \\ &\quad + [\mathrm{RNAP}:\mathrm{S}_{28}] + [\mathrm{RNAP}:\mathrm{S}_{28}:\mathrm{DNA}_{\mathrm{eGFP}}] = C \end{aligned} \tag{2-13}$$

The two conservations laws formed around the concentration of *RNAP* and *Ribo* are used in the subsequent modelling process to apply model reduction.

2-3-1 Transcription and Translation Modelling - Model Reduction

The CFFL is a 3-node biological motif, with the three nodes being represented by σ-factor 70 (S70), σ-factor 28 (S28) and green fluorescent protein (eGFP) (fig. 1-2). All of these chemical compounds are proteins, therefore the process of protein synthesis lies at the base of the CFFL's functioning. Therefore, the large reaction set can be grouped in reaction subsets that take part or in the transcription, or in the translation process (fig. 1-3). From the three proteins, S70 is added to the reaction mixture. The other two, S28 and eGFP are produced. As a consequence, there are two protein synthesis processes happening, which means two sets of transcription and translation reactions. In addition, there is one additional transcription reaction that has the product of an RNA element, $RNA_{trigger}$, that is needed for the functioning of the toe-hold switch. In summary, there are three sets of transcription reactions and two sets of translation reactions.

The first step was to understand the exact processes that are undergoing during transcription. More specifically, it had to be identified which chemical species bind with each other, which are the end-products and on what time-scale is the process completed. In addition, the kinetics of the reactions had to be identified as well, for example if it contains enzyme kinetics or just mass-action kinetics suffices. A more detailed description of the transcription process can be found in [19].

In literature there were several ways used to model transcription [20] [22] [23]. However, no source was found that specifically dealt with the the initialization of transcription using σ-factors. Therefore this represented a challenge in the modelling process. The next issue to consider was the identification of the type of dynamics: mass-action or Michaelis-Menten kinetics. The way to approach this, was to identify if the reactions are catalysed by an enzyme and have enzymatic kinetics. In [19] it was identified that transcription in the CFFL is initiated by the holoenzyme RNAP:S_{70} or RNAP:S_{28}, using DNA templates as substrates and having RNA species for end-products. In addition, the publications used for the literature study of the master thesis, all implemented enzymatic kinetics in order to model transcription.

In order to simplify the derivation process the reactions found in appendix A-1 and parts of the mass action model found in eq. (2-1) and eq. (2-2) are rewritten. In the following, the different chemical species will be denoted by letters in order to simplify the derivation. Capital letters denote chemical species and the matching small case letters represent the corresponding concentrations. The simplified reactions depicting the transcription from the CFFL, are written in the following way:

$$\begin{aligned}
&\mathrm{A}+\mathrm{B}_1 \underset{\mathrm{k}_{-1}}{\overset{\mathrm{k}_1}{\rightleftharpoons}} \mathrm{C}_1\\
&\mathrm{C}_1+\mathrm{D}_1 \xrightarrow{\mathrm{k}_2} \mathrm{E}_1 \xrightarrow{\mathrm{k}_3} \mathrm{M}_1+\mathrm{D}_1+\mathrm{C}_1\\
&\mathrm{C}_1+\mathrm{D}_2 \xrightarrow{\mathrm{k}_4} \mathrm{E}_2 \xrightarrow{\mathrm{k}_5} \mathrm{M}_2+\mathrm{D}_2+\mathrm{C}_1
\end{aligned} \tag{2-14}$$

where A is RNAP, B_1 is S70, B_2 is S28, C_1 is $\mathrm{RNAP:S_{70}}$, D_1 is $\mathrm{DNA_{trigger}}$, D_2 is $\mathrm{DNA_{S28}}$, E_1 is $\mathrm{RNAP:S_{70}:DNA_{trigger}}$, E_2 is $\mathrm{RNAP:S_{70}:DNA_{S28}}$, M_1 is $\mathrm{RNA_{trigger}}$ and M_2 is $\mathrm{RNA_{S28}}$. The above reaction set represents the first two transcription processes grouped as they use the same enzyme-sigma-factor complex for catalysis. Subsequently, the third transcription reaction set is the following:

$$\begin{aligned}
&\mathrm{A}+\mathrm{B}_2 \underset{\mathrm{k}_{-6}}{\overset{\mathrm{k}_6}{\rightleftharpoons}} \mathrm{C}_2\\
&\mathrm{C}_2+\mathrm{D}_3 \xrightarrow{\mathrm{k}_7} \mathrm{E}_3 \xrightarrow{\mathrm{k}_8} \mathrm{M}_3+\mathrm{D}_3+\mathrm{C}_2
\end{aligned} \tag{2-15}$$

where A is RNAP, B_2 is S28, C_2 is $\mathrm{RNAP:S_{28}}$, D_3 is $\mathrm{DNA_{eGFP}}$, E_3 is $\mathrm{RNAP:S_{28}:DNA_{eGFP}}$ and M_3 is $\mathrm{RNA_{eGFP}}$.

The next step is to write up the partial mass-action model that corresponds to the reactions from eq. (2-14) and eq. (2-15):

$$\begin{aligned}
\dot{a} &= k_{-1}c - k_1 ab_1 + k_{-6}c_2 - k_6 ab_2 \\
\dot{b}_1 &= k_{-1}c_1 - k_1 ab_1 \\
\dot{c}_1 &= k_1 ab_1 - k_{-1}c_1 + k_{-2}e_1 - k_2 c_1 d_1 + k_3 e_1 - k_4 c_1 d_2 + k_5 e_2 \\
\dot{d}_1 &= k_{-2}e_1 - k_2 c_1 d_1 + k_3 e_1 \\
\dot{e}_1 &= k_2 c_1 d_1 - k_{-2}e_1 - k_3 e_1 \\
\dot{m}_1 &= k_3 e_1 \\
\dot{d}_2 &= -k_4 c_1 d_2 + k_5 e_2 \\
\dot{e}_2 &= k_4 c_1 d_2 - k_5 e_2 \\
\dot{m}_2 &= k_5 e_2 \\
\dot{b}_2 &= k_{-6}c_2 - k_6 ab_2 \\
\dot{c}_2 &= k_6 ab_2 - k_{-6}c_2 - k_7 c_2 d_3 + k_8 e_3 \\
\dot{d}_3 &= -k_7 c_2 d_3 + k_8 e_3 \\
\dot{e}_3 &= k_7 c_2 d_3 - k_8 e_3 \\
\dot{m}_3 &= k_8 e_3
\end{aligned} \tag{2-16}$$

In the next step there are two important details to consider: difference in velocity (timescales) of different reactions and competitive binding. Regarding the temporal differentiation of the different reactions, the modelling strategy used in [3] was used: the reactions that contain the binding of the RNAP to the σ-factor and the binding of this complex to the DNA template are considered to be much faster than the production of RNA. Therefore the concentrations of RNAP:S$_{70}$, RNAP:S$_{28}$, RNAP:S$_{70}$:DNAt, RNAP:S$_{70}$:DNA$_{s28}$ and RNAP:S$_{28}$:DNA$_{\text{eGFP}}$ are approximated at their quasi-steady state. This translates into setting the ordinary differential equations corresponding to these states to 0. The next step is to find an expression for the mentioned concentrations and replace them into the set of equations eq. (2-16). In order to do this, a conservation law is required. Therefore, the next step is to go back to the earlier found conservation laws using the stoichiometric matrix. The Equation (2-13) is applied in the subsequent derivation and it is written up in the simplified way:

$$A_{tot} = a + c_1 + e_1 + c_2 + e_2 + e_3 \tag{2-17}$$

In addition, it has to be mentioned that using this conservation law makes it possible to include competitive binding of RNAP with S70 and S28 respectively. The next step is to express concentration of RNAP (a) from eq. (2-17):

$$a = A_{tot} - c_1 - e_1 - c_2 - e_2 - e_3 \tag{2-18}$$

In addition, the concentrations of E$_1$, E$_2$ and E$_3$ are expressed with the concentrations of C$_1$ and C$_2$:

$$\begin{aligned}
e_1 &= (d_1/K_{e1})c_1 \\
e_2 &= (d_3/K_{e2})c_1 \\
e_3 &= (d_3/K_{e3})c_2
\end{aligned} \tag{2-19}$$

Where $K_{e1} = \frac{k_3}{k_2}$, $K_{e2} = \frac{k_5}{k_4}$ and $K_{e3} = \frac{k_8}{k_7}$. The aim was achieved, everything is expressed in terms of c_1 and c_2, the inclusion in equations from eq. (2-16) can be completed. After arranging the terms in the desired way the following expressions are found:

$$\begin{aligned} c_1 &= \frac{A_{tot}(b_1/K_1)}{1 + (1 + d_1/K_{e1} + d_2/K_{e2})(b_1/K_1) + (1 + d_3/K_{e3})(b_2/K_6)} \\ c_2 &= \frac{A_{tot}(b_2/K_6)}{1 + (1 + d_1/K_{e1} + d_2/K_{e2})(b_1/K_1) + (1 + d_3/K_{e3})(b_2/K_6)} \end{aligned} \tag{2-20}$$

where $K_1 = \frac{k_{-1}}{k_1}$ and $K_6 = \frac{k_{-6}}{k_6}$

Therefore, the expression for concentrations of E_1, E_2 and E_3 are written up in the following way:

$$\begin{aligned} e_1 &= \frac{A_{tot}(b_1/K_1)(d_1/K_{e1})}{1 + (1 + d_1/K_{e1} + d_2/K_{e2})(b_1/K_1) + (1 + d_3/K_{e3})(b_2/K_6)} \\ e_2 &= \frac{A_{tot}(b_1/K_1)(d_2/K_{e2})}{1 + (1 + d_1/K_{e1} + d_2/K_{e2})(b_1/K_1) + (1 + d_3/K_{e3})(b_2/K_6)} \\ e_3 &= \frac{A_{tot}(b_2/K_6)(d_3/K_{e3})}{1 + (1 + d_1/K_{e1} + d_2/K_{e2})(b_1/K_1) + (1 + d_3/K_{e3})(b_2/K_6)} \end{aligned} \tag{2-21}$$

Having the concentrations expressed in eq. (2-21), these are replaced in the equations that depict the production rate of the three RNA species produced during transcription: M_1, M_2 and M_3 ($\mathrm{RNA}_{\mathrm{trigger}}$, $\mathrm{RNA}_{\mathrm{S28}}$ and $\mathrm{RNA}_{\mathrm{eGFP}}$) from eq. (2-16). By applying this quasi-steady-state approximation the model was simplified by eliminating 5 chemical complexes as states from the final model. This represents the end of the transcription modelling.

In the following the second part of the modelling process is presented: generating the equations for the translation reactions, expressing the production rates of the proteins S28 and eGFP. In the case of transcription, RNAP was catalysing the RNA production process. Similarly, during the translation process, ribosomes catalyse the protein production process. Therefore, it should be modelled with Michaelis-Menten kinetics. However, in literature, two modelling strategies were found to be used: some sources chose Michaelis-Menten kinetics [24] [20] [22] while others resorted to applying simple translation rate with mass-action [3] [23]. In this master thesis, it was chosen to use Michaelis-Menten kinetics in order to be able to include the concentration of ribosomes as a state and for keeping the consistency of the modelling strategy applied so far. Moreover, this also makes it possible to include competitive binding of $\mathrm{RNA}_{\mathrm{trigger}}$ with the other two RNA species.

Similarly to transcription, the translation reactions are also rewritten by denoting the different chemical species with a capital letter and the corresponding concentrations with small case letters:

$$\begin{aligned}
&\mathrm{M_1 + M_2 \xrightleftharpoons[k_{-9}]{k_9} N_1}\\
&\mathrm{N_1 + R \xrightarrow{k_{10}} S_1}\\
&\mathrm{S_1 \xrightarrow{k_{11}} P_1 + N_1 + R}
\end{aligned} \tag{2-22}$$

$$\begin{aligned}
&\mathrm{M_1 + M_3 \xrightleftharpoons[k_{-12}]{k_12} N_2}\\
&\mathrm{N_2 + R \xrightarrow{k_{13}} S_2}\\
&\mathrm{S_2 \xrightarrow{k_{14}} P_2 + N_2 + R}
\end{aligned} \tag{2-23}$$

Where $\mathrm{M_1}$, $\mathrm{M_2}$ and $\mathrm{M_3}$ are the three RNA species that were produced during transcription, $\mathrm{RNA_{trigger}}$, $\mathrm{RNA_{S28}}$ and $\mathrm{RNA_{eGFP}}$ respectively. $\mathrm{N_1}$ is $\mathrm{RNA_{trigger}:RNA_{S28}}$, $\mathrm{N_2}$ is $\mathrm{RNA_{trigger}:RNA_{eGFP}}$, $\mathrm{S_1}$ is $\mathrm{RNA_{trigger}:RNA_{S28}:Ribo}$, $\mathrm{S_2}$ is $\mathrm{RNA_{trigger}:RNA_{eGFP}:Ribo}$. $\mathrm{P_1}$ and $\mathrm{P_2}$ are the protein produced during translation, S28 and $\mathrm{eGFP_{dark}}$ respectively.

It can be observed that the structure of the reactions is similar to the transcription reactions from eq. (2-14) and eq. (2-15). The same modelling strategy is applied in this case as well. Previously mentioned conservation law is used from eq. (2-12) and the quasi-steady state approximation applied in modelling of the transcription reactions. Firstly, the partial symbolic mass-action model is written up:

$$\begin{aligned}
\dot{m}_1 &= -k_9 m_1 m_2 - k_{12} m_1 m_3 + k_{-9} n_1 + k_{-12} n_2\\
\dot{m}_2 &= -k_9 m_1 m_2 + k_{-9} n_1\\
\dot{n}_1 &= k_9 m_1 m_2 - k_{-9} n_1 - k_{10} n_1 r + k_{11} s_1\\
\dot{s}_1 &= k_{10} n_1 r - k_{11} s_1\\
\dot{p}_1 &= k_{11} s_1\\
\dot{m}_2 &= -k_{12} m_1 m_3 + k_{-12} n_2\\
\dot{n}_2 &= k_{12} m_1 m_3 - k_{-12} n_2 - k_{13} n_2 r + k_{14} s_2\\
\dot{s}_2 &= k_{13} n_2 r - k_{14} s_2\\
\dot{p}_2 &= k_{14} s_2\\
\dot{r} &= -k_{10} n_1 r + k_{11} s_1 - k_1 3 n_2 r + k_{14} s_2
\end{aligned} \tag{2-24}$$

The reactions that contain the binding of Ribo to the RNA complex pairs are considered to be faster than the production of the proteins. Therefore, the concentrations of $\mathrm{S_1}$ and $\mathrm{S_2}$ are approximated at their quasi-steady-state and the ordinary differential equations corresponding to these complexes are set to 0. As a result the following expressions for the two concentrations is written up:

$$\begin{aligned}
s_1 &= \frac{k_{10}}{k_{11}} n_1 r\\
s_2 &= \frac{k_{13}}{k_{14}} n_2 r
\end{aligned} \tag{2-25}$$

In the next step, the relevant conservation law, eq. (2-12), is presented in the following form:

$$r = R_{tot} - s_1 - s_2 \tag{2-26}$$

Where R_{tot} is the total ribosome concentration. Equation (2-26) is used to express the concentrations of S_1 and S_2 and implement the quasi-steady state approximation. It has to be mentioned, that this will also make it possible to include the competitive binding of Ribo with the two RNA complex pairs. In the next step the substitution of eq. (2-26) into eq. (2-25) is completed. Subsequently, the system of equations is solved for deriving an expression for concentrations s_1 and s_2:

$$\begin{aligned} s_1 &= \frac{R_{tot}(n_1/K_{TL1})}{1 + (n_1/K_{TL1}) + (n_2/K_{TL2})} \\ s_2 &= \frac{R_{tot}(n_2/K_{TL2})}{1 + (n_1/K_{TL1}) + (n_2/K_{TL2})} \end{aligned} \tag{2-27}$$

Where $K_{TL1} = \frac{k_{11}}{k_{10}}$ and $K_{TL2} = \frac{k_{14}}{k_{13}}$.The expressions from eq. (2-27) are substituted into the ordinary differential equations that correspond to the production rate of the two proteins, S28 and eGFP. By applying this modelling strategy, 2 states were eliminated from the model and the translation process has been mathematically modelled. Apart from transcription and translation reactions, there were two other chemical processes that are required to be modelled as well: maturation of $eGFP_{dark}$ and degradations of specific chemical complexes. The modelling of these will be presented in the subsequent subsection.

2-3-2 Maturation of $eGFP_{dark}$ and degradations

Two processes were not mentioned in the modelling work presented so far in this chapter: the maturation of $\text{eGFP}_{\text{dark}}$ in order to get the final output, eGFP and the degradations of some of the species. Maturation of $\text{eGFP}_{\text{dark}}$ is described with the reaction below:

$$\text{eGFP}_{\text{dark}} \xrightarrow{\text{mat}} \text{eGFP} \tag{2-28}$$

It suffices to model the maturation with mass-action kinetics in the following way:

$$\begin{aligned} \frac{d[eGFP_{dark}]}{dt} &= \frac{k_{14}R_{tot}(n_2/K_{TL2})}{1 + (n_1/K_{TL1}) + (n_2/K_{TL2})} - mat \cdot eGFP_{dark} \\ \frac{d[eGFP]}{dt} &= mat \cdot eGFP_{dark} \end{aligned} \tag{2-29}$$

The Michaelis-Menten term from eq. (2-29) is the one resulting from the previously derived equations. The focus in this subsection is on the term $mat \cdot eGFP_{dark}$. It is subtracted from the production rate equation of $\text{eGFP}_{\text{dark}}$, as it represents being consumed. Moreover, it is added to the production of eGFP.

Degradations are modelled similarly to maturation, with mass-action kinetics. The only difference is that instead of a maturation rate *mat*, a degradation rate *deg* is used. There were in total 5 chemical complexes identified by ICMS that required to include degradation reactions: the three RNA species and the two RNA binding pairs. The rest of the compounds' degradations are considered to be insignificant. An example of degradation term inclusion from the model is given below:

$$\dot{x}_{10} = k_{12}x_6x_8 - k_{-12}x_{10} - k_{13}x_{10}x_{14} + \frac{k_{14}R_{tot}(x_{10}/K_{TL2})}{1 + (x_9/K_{TL1}) + (x_{10}/K_{TL2})} - deg \cdot x_{10} \quad \text{(2-30)}$$

2-4 The Full Mathematical Model

The modelling framework presented in this chapter so far was applied on the CFFL and a full mathematical model was derived for it. This is composed of 14 ordinary differential equations containing 14 state variables and 24 parameters. Four of the state variables maintain a constant concentration value during the batch experiment, the reason why their production rate equals 0. The model is presented below:

$$\begin{aligned}
\dot{x_1} =& + \frac{k_{-1}RNAP_{tot}(x_2/K_1)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_1x_1x_2 + \\
& + \frac{k_{-6}RNAP_{tot}(x_{11}/K_6)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_6x_1x_{11} \\
\dot{x_2} =& \frac{k_{-1}RNAP_{tot}(x_2/K_1)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_1x_1x_2 \\
\dot{x_3} =& 0 \\
\dot{x_4} =& 0 \\
\dot{x_5} =& 0 \\
\dot{x_6} =& \frac{k_3RNAP_{tot}(x_2/K_1)([DNA_t]/K_{e1})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_9x_6x_7 - k_{12}x_6x_8 + k_{-9}x_9 + k_{-12}x_{10} - deg_6 \cdot x_6 \\
\dot{x_7} =& \frac{k_5RNAP_{tot}(x_2/K_1)(x_4/K_{e_2})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_9x_6x_7 + \\
& + k_{-9}x_9 - deg_7 \cdot x_7 \\
\dot{x_8} =& \frac{k_8RNAP_{tot}(x_{11}/K_6)(x_5/K_{e3})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_{12}x_6x_8 + k_{-12}x_{10} - deg_8 \cdot x_8
\end{aligned}$$

$$
\begin{aligned}
\dot{x_9} =& k_9 x_6 x_7 - k_{-9} x_9 - k_{10} x_9 x_{14} + \frac{k_{11} R_{tot}(x_9/K_{TL1})}{1 + (x_9/K_{TL1}) + (x_{10}/K_{TL2})} - deg_9 \cdot x_9 \\
\dot{x_{10}} =& k_{12} x_6 x_8 - k_{-12} x_{10} - k_{13} x_{10} x_{14} + \frac{k_{14} R_{tot}(x_{10}/K_{TL2})}{1 + (x_9/K_{TL1}) + (x_{10}/K_{TL2})} - deg_{10} \cdot x_{10} \\
\dot{x_{11}} =& \frac{k_{-6} RNAP_{tot}(x_{11}/K_6)}{1 + (1 + x_5/K_{e3})(x_{11}/K_6) + (1 + x_3/K_{e1} + x_4/K_{e2})(x_2/K_1)} - \\
& - k_6 x_1 x_{11} + \frac{k_{11} R_{tot}(x_9/K_{TL1})}{1 + (x_9/K_{TL1}) + (x_{10}/K_{TL2})} \\
\dot{x_{12}} =& \frac{k_{14} R_{tot}(x_{10}/K_{TL2})}{1 + (x_9/K_{TL1}) + (x_{10}/K_{TL2})} - mat \cdot x_{12} \\
\dot{x_{13}} =& mat \cdot x_{12} \\
\dot{x_{14}} =& 0
\end{aligned}
$$

Where the states correspond to the following chemical species' concentration:

$$
\begin{aligned}
&x_1 - \text{RNAP} \\
&x_2 - \text{S70} \\
&x_3 - \text{DNA}_t \\
&x_4 - \text{DNA}_{\text{S28}} \\
&x_5 - \text{DNA}_{\text{eGFP}} \\
&x_6 - \text{RNA}_t \\
&x_7 - \text{RNA}_{\text{S28}} \\
&x_8 - \text{RNA}_{\text{eGFP}} \\
&x_9 - \text{RNA}_t\text{:RNA}_{\text{S28}} \\
&x_{10} - \text{RNA}_t\text{:RNA}_{\text{eGFP}} \\
&x_{11} - \text{S28} \\
&x_{12} - \text{eGFP}_{\text{dark}} \\
&x_{13} - \text{eGFP} \\
&x_{14} - \text{Ribo}
\end{aligned}
$$

2-5 Adding inflow and outflow term

The full modelling theory presented in this chapter was focused on developing a mathematical model for simulating the batch experiment. However, in order to be able to simulate the dynamic characteristic of the CFFL which is the noise-filtering function, the experimental process has to include an inflow and outflow. This way the resources are not depleted and it is possible to apply an input that has the shape of a square-wave and observe the production of eGFP. This is achieved experimentally by using a microfluidic reactor that has multiple inlets and outlets in order to be able to load chemical species and flush out a fraction of the reaction mixture. The structure of the microfluidic chip can be seen in fig. 2-2. Another important piece of information is the time periods at which the inflow and outflow is initiated. The experiments at the ICMS are based on experimental methodology derived in [24], where the loading and flushing is completed at every 15 minutes.

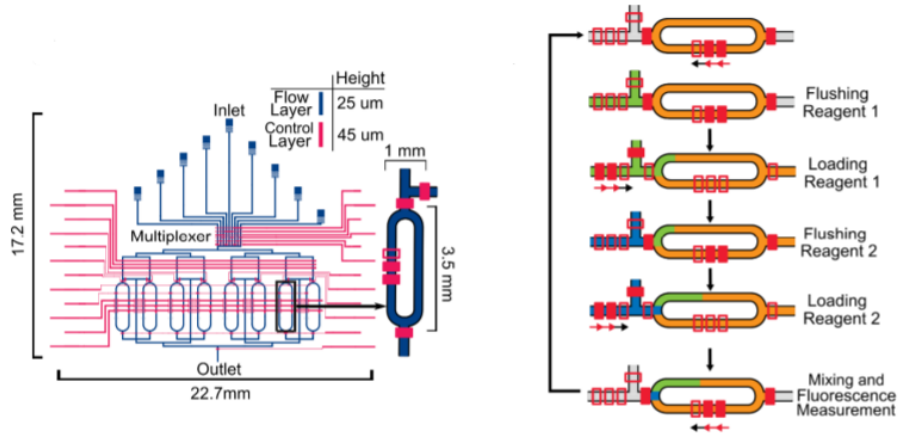


Figure 2-2: Top-view diagram and schematic of flushing, loading and mixing of reagents inside a reactor [2]

A generalized way to include inflow and outflow to the model is by expanding the expression from eq. (2-3) in the following way:

$$\dot{x} = dil(x_{in} - x) + Sv(x) \tag{2-31}$$

Where dil is the dilution fraction which corresponds to the rate the specific chemical species are flown in or respectively flown out, x_{in} is the vector of species that are flown in and x is the vector of all the species as the whole reaction mixture is flushed out. In the next step, the structure of vector x_{in} is discussed. There are in total six species that are flown in. Five of them, more precisely RNAP, S70, $\mathrm{DNA}_{\mathrm{S28}}$, $\mathrm{DNA}_{\mathrm{eGFP}}$ and Ribo are constant inputs: a specific concentration of each of them is inputted to the experiment. On the other hand, $DNA_{trigger}$ represents the input to the system and it is represented by a square-wave function to test the behaviour of the CFFL. In the following, a few equations from the mathematical model will be included to show how the dilution fraction term *dil* got incorporated into the ODE system:

$$\begin{aligned}\dot{x}_1 =& + \frac{k_{-1}RNAP_{tot}(x_2/K_1)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_1x_1x_2 + \\ &+ \frac{k_{-6}RNAP_{tot}(x_{11}/K_6)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_6x_1x_{11} \\ &+ dil\cdot RNAP_0 - dil\cdot x_1\end{aligned}$$

$$\dot{x}_3 = dil\cdot c\cdot g(x_3) - dil\cdot x_3$$

$$\dot{x}_{13} = mat\cdot x_{12} - dil\cdot x_{13}$$

The first equation represents the dynamics of RNAP. This is a complex needed in order for the reactions in the CFFL to occur, therefore there is an inflow of a constant concentration denoted by the term $dil\cdot \mathrm{RNAP}_0$. The second equation corresponds to $\mathrm{DNA}_{\mathrm{trigger}}$, the input of the system. The inflow term in this case is composed by the dilution fraction, the constant c that is used for scaling the input value and a function $g(x_3)$ that stands for the square-wave input. The last equation shows the dynamics of the output of the system, eGFP, and only contains an outflow which is represented by the term $dil\cdot x_{13}$. The same outflow term can be also observed in the first two equations. The full model with inflow-outflow modelling is included in appendix A-4.

2-6 Summary

In summary, this chapter encompassed the full description of a modelling strategy that was applied on the CFFL. First, the choice of deterministic modelling was presented and motivated. Then prior to the model reduction, the conservation laws were identified from the left-null space of the stoichiometric matrix. Subsequently, the modelling methodology for the transcription and translation reactions was presented. Both of them are modelled using Michaelis-Menten kinetics and include the competitive binding present in the network implemented with the help of the conservation laws. Moreover, the modelling of maturation and degradation was described. Lastly, the extension to a flow model was presented by incorporating the dilution fraction into the batch model. The ordinary differential equations resulting from this chapter are used in the following parts of the thesis work. Specifically, it is one of the main components of the parameter estimation process. With the correct estimated parameters, the mathematical model can be used to do in silico experiments and observe the dynamics of the other chemical species composing the network.

Chapter 3

Parameter Estimation

The noise-filtering property of the coherent feedforward loop (CFFL) is dynamic in nature. Therefore, one of the aims is to gain knowledge about the CFFL's dynamics. For this, it is necessary to know the values of the parameters of the mathematical model. There are a number of tools that can be applied to gain insights about biological systems without knowledge of these parameters, however they are mainly static analysis tools. In other word, they provide information about the steady-state operation of the system. These tools will be presented in Chapter 4. However, in order to be able to predict how the CFFL handles a pulse input, the parameters of the mathematical model have to be estimated from experimental data.

The following chapter will present the challenge of estimating the parameters of the mathematical model, more precisely the reaction rate constants, Michaelis Menten kinetics and initial concentrations used for the experiments. A priori of diving into the details of this stage of the project work, the notion of structural identifiability is introduced. Subsequently, the main components of the parameter estimation process are presented. Firstly, it is required to have experimental data that is suitable for system identification. Secondly, a mathematical model containing the unknown parameters has to be developed. This step was completed and presented in Chapter 2. The third component is a cost function or distance between the experimental data and the model prediction. And the last element is a specific optimization algorithm and procedure of implementation. The last part of the chapter comprises of the specific parameter estimation strategy applied and the discussion of the results.

3-1 Structural Identifiability

Before the tedious process of parameter estimation is initiated, there is an important issue that must be raised regarding the derived mathematical model of the biochemical reaction network. This issue is represented by the question if there is a unique combination of parameter values that generate an appropriate fit for the experimental data. If the answer is yes, then the mathematical model is said to be structurally identifiable. This is an important detail that is often overlooked in papers presenting the model and parameter estimation of a biological model [25]. The reason for this might be that the mathematical theory behind the tools used for verifying identifiability are out of the scope of the knowledge possessed by biology researchers. However, structural identifiability is commonly used in systems and control theory [26], another reason why researchers from this field should help advancing the multidisciplinary research area of systems biology. Ideally the structural identifiability or a priori identifiability is performed on a mathematical model before the experiments are conducted. Consequently, it is made sure that the way the experimental data is collected is optimal for parameter estimation procedure. For example, in some cases this would mean observing the production of an intermediate protein, therefore having another state as an output. This was not possible in the case of the master thesis project, therefore the tools for verifying structural identifiability were implemented and the parameters were identified that need to be fixed in order to obtain the desired uniqueness.

Even though it is not a strict requirement to prove structural identifiability of a biological model, there are strong arguments in favour of it presented in [25] and correspond to the aims set for this master thesis. Firstly, one of the goals is to predict the behaviour of the other chemical species taking part in the reaction network. These are not observable from the experimental data therefore their dynamics are presented by the derived model. Secondly, the modelling of the single CFFL comprises as a base for determining the behaviour of combined feedforward loops. This means that the aim is to use the model to test hypotheses that will not be tested experimentally, only in case the function of the combined feedforward loop is predicted to be interesting for synthetic biology applications. In summary, in order to ensure accurate results, structural identifiability should be proven.

In the following the definition of structural identifiability will be presented from [25]. Let $\theta = (\theta_1, \theta_2, \dots, \theta_{n_p})$ be a set of parameters and $M(\theta)$ the corresponding set of ordinary differential equations containing the parameters. A parameter θ_i is defined to be structurally identifiable if the following is true:

$$M(\theta) = M(\theta^*) \Rightarrow \theta_i = \theta_i^* \tag{3-1}$$

Subsequently, an example is given that helps demonstrate the concept of structural identifiability. The following model is considered:

$$y = a \cdot b \cdot x \tag{3-2}$$

Where x is the input, y is the output, a and b the unknown parameters. Both x and y are available from experimental data. It can be observed that parameters a and b individually are non-identifiable. Only, their product $a \cdot b$ taken as one parameter is considered to be identifiable. Therefore, as a solution these can be combined to represent one parameter.

In order to determine structural identifiability there are several software tools that are available. The three main ones are DAISY (Differential Algebra for Identifiability of SYstems) [27], GenSSI (Generating Series for testing Structural Identifiability) [28] and IdentifiabilityAnalysis [29]. All of them use different mathematical methods and were tested on the mathematical model derived for the CFFL. However, the first two mentioned, DAISY and GenSSI, required an extensive time to run and were not successfully applied. The third tool, IdentifiabilityAnalysis was successfully implemented as it was developed to be computationally efficient for complex and large models. The method used is implemented in Mathematica and comprises of a probabilistic semi-numerical algorithm described in [30].

In the following, the application and results from the analysis of the CFFL model's parameters structural identifiability will be described. The algorithm was implemented in Mathematica and it requires the following inputs: a list of differential equations that build up the model, a list of initial conditions (in case there are), the outputs of the model and separate lists of the inputs, parameters, state variables. The user has to determine which values are unknown and input them as symbolic variables to the algorithm. These can include both parameters and initial values. In the case of the CFFL, the initial conditions were assumed to be known and also the parameters that are related to the initial concentrations of RNAP and ribosomes. The rest of the parameters were all set to symbolic variables.
The initial run of the algorithm confirmed the expected results that the model is structurally unidentifiable. Moreover, it returned a list of parameters that were the ones identified to be the reason for this result: k_5, k_8, K_{e1}, K_{e2}, K_{e3}, k_{-1} and k_{-6}. All of these are a part of the transcription reactions. This result was expected as one of the big disadvantages of biological models is that they comprise of a large number of parameters and in many cases they prove to be structurally unidentifiable [25]. However, the CFFL model was further tested and two parameters, k_5 and k_8, were set to constant values. The reason for this is that these two parameters are only present in two of the ordinary differential equations building up the model. Therefore, they do not influence the trajectories of the other states and it will be possible to find a way to fix them to a value during parameter estimation. This resulted in a positive result, meaning the model became structurally identifiable. Therefore, focus was put on parameters k_5 and k_8 in the subsequent parameter estimation process in order to not allow large change in the estimated values of these parameters and have them fixed for simultaneous calibration of the datasets.

3-2 Components of the Parameter Estimation

Parameter estimation applied on biological models is a cumbersome process. First of all, most of the times the model is structurally unidentifiable. Secondly, the ODE system representing the CFFL is nonlinear which imposes additional challenges. Thirdly, the input species like RNAP or S70 are given as ranges, not exact values. In addition, the kinetic parameters can vary from one experiment to the other. In conclusion, there are a multitude of hurdles in estimating the reaction rate and Michaelis Menten constants. Hence. it is really important to approach this part of the project strategically, and find the best choice for each element of the model calibration process. As a result, the following section will present the main components of the parameter estimation and motivate the choices that were made during this process.

3-2-1 Experimental Data and Mathematical Model

Firstly, the starting point of the calibration is presented, more specifically the experimental data and the derived mathematical model used for estimation. The experimental data is the one that was plotted in fig. 1-4 and described briefly in Chapter 1. There are in total 5 output datasets, each corresponding to one constant value of input concentration of $\mathrm{DNA_{trigger}}$ (0.5, 1, 2, 5 and 10 nM). Each experiment was completed individually. Therefore, the initial conditions most probably vary and also the kinetic parameters as well. The length of the experiments were 16 hours and measurements were taken every 5 minutes. As a consequence there are 193 data points per dataset. The output is measured in μM which during the estimation and simulation algorithms is converted into nM. There is one main input that was mentioned above, the concentration of $\mathrm{DNA_{trigger}}$. Apart from this, there are 5 chemical species that are constant inputs to the batch experiment: RNAP, S70, $\mathrm{DNA_{S28}}$, $\mathrm{DNA_{eGFP}}$ and Ribo. The two DNA templates have a specified input concentration of 10 nM each. The rest of the species' input value is not known exactly, however a range of values were given that can be found in [2] and presented in table 1-1.

Even though the three species present in the table were not given with a fixed concentration, during most of the parameter estimation process they were set to have a specific fixed value. The reason for this is to combat the problem of structural identifiability.

The mathematical model required for system identification is the batch model derived and presented in section 2-4. It is made up of 14 ordinary differential equations with the corresponding 14 states and 24 unknown parameters. It contains both mass-action and Michaelis-Menten kinetics.

3-2-2 Cost function

In order to apply an optimization algorithm, it is required to calculate a cost that has to be minimised. In this case, this is represented by the difference between the experimental data and the estimated output data. In order to compute this difference, the normalized Mean of the Squared Errors (MSE) is used. The formula used for computing is the following:

$$MSE_y = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_{exp,i} - \hat{y}_i}{max(y_{exp,i})})^2 \tag{3-3}$$

Where $y_{exp,i}$ represents a point in the experimental data, $\hat{y}_i$ is the corresponding estimated point and $max(y_{exp,i})$ is the maximum value taken from the experimental data. The formula was implemented in a script as a function where a single or multiple datasets could be used in order to calculate the value of MSE. The cost function could be used by any of the optimization algorithms presented below.

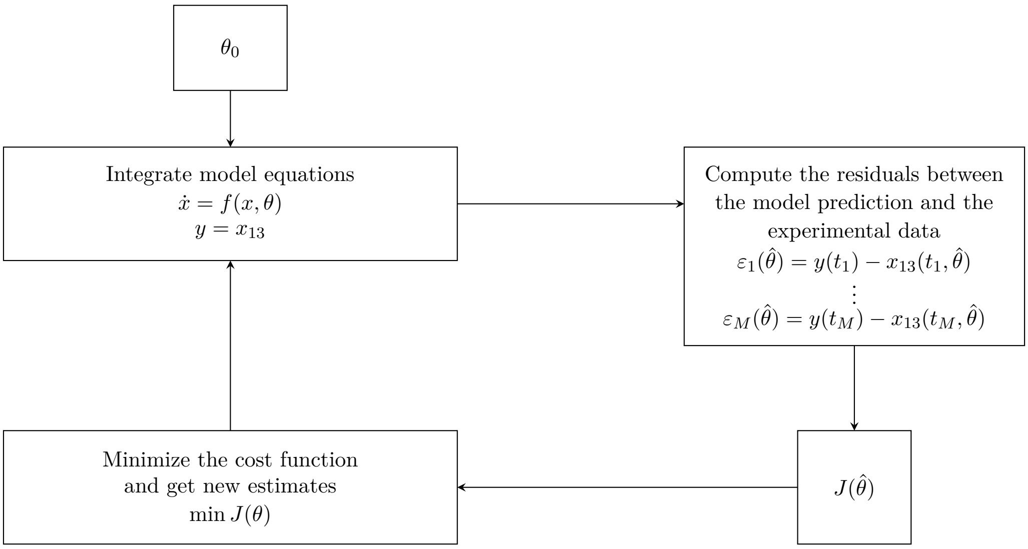

θ_0
Integrate model equations
$\dot{x} = f(x,\theta)$
$y = x_{13}$
Compute the residuals between the model prediction and the experimental data
$\varepsilon_1(\hat{\theta}) = y(t_1) - x_{13}(t_1,\hat{\theta})$
$\vdots$
$\varepsilon_M(\hat{\theta}) = y(t_M) - x_{13}(t_M,\hat{\theta})$
$J(\hat{\theta})$
Minimize the cost function and get new estimates
$\min J(\theta)$

Figure 3-1: General Structure of Optimization Code

3-2-3 Parameter Estimation Algorithms

Literature presents many different ways of implementing parameter estimation of biological models using several different algorithms [31][32][33]. However, all of them followed a similar line of thought, which was presented in [34] where the most frequent approaches to calibration of biological models are described. Two classes of parameter estimation algorithms are presented: global and local methods. Global optimization includes algorithms like simulated annealing, evolutionary algorithms, particle swarm optimization and pattern search. Local optimization is mainly composed of gradient-based methods. In [34], another category of optimization strategy is mentioned, called hybrid methods.

This comprises of combining at least two optimization algorithms in a certain way to reach global minimum. A hybrid method was chosen to be applied in the parameter estimation of the CFFL. Specifically, the particle swarm optimization algorithm is used with pattern search method. In the following there will be brief description of global and local methods. Subsequently, the algorithms chosen for system identification will be presented together with their implementation.

The starting point in the application of parameter estimation of biological models in literature [31][33][35] is the motivation of using global optimization or hybrid methods that include global methods. The reason for this is that biochemical reaction networks usually are described by a set of nonlinear ordinary differential equations. As a consequence stochastic or derivative-free optimization strategies are more desired. This is explained by the fact that gradient-based methods converge to the local minima which is to be avoided. This is especially the case for biological models where there are a high number of parameters and limited experimental data [32]. Taken this into account, there were three different optimization methods used during the parameter estimation. Two of them are derivative-free: particle swarm optimization and pattern search; the third one on the other hand is a gradient-based method called interior point algorithm. This was only used for refinement of the estimated results or in some cases helped in speeding up the estimation process. However, the two main optimization algorithms in this project are considered to be Particle Swarm Optimization (PSO) and pattern search. The choice for this hybrid method to perform model calibration was based on the successful application of it in [33][35]. In the following the different algorithms will be briefly described.

Particle Swarm Optimization (PSO)

Particle swarm optimization was used as a first step in the parameter estimation process. Initially there is little knowledge about the dimensions of the constants and also the ranges they could vary in. PSO does not require starting estimation points of the parameter values. Therefore it is useful to narrow down initially the ranges of the parameters and see an initial direction of each one, meaning which ones are really close to 0 while which ones are much larger than expected.

In the following, the implementation of PSO is briefly presented. It is a population-based stochastic algorithm and it was developed specifically for optimization of nonlinear functions [36]. A numerical vector of the dimension of the number of the parameters to be optimized is randomly initialized in the first step. This can be seen as a point in a higher-dimensional space which during the optimization moves around to test new parameter values. More precisely, a defined number of these points are initialized and moved around at the same time. As a result, they tend to cluster together in regions of the space corresponding to optimal values, hence the name particle swarm. In addition to the moving around operation, each particle is connected bidirectionally to its assigned neighbours. The next question is how do these particles move around? There are two

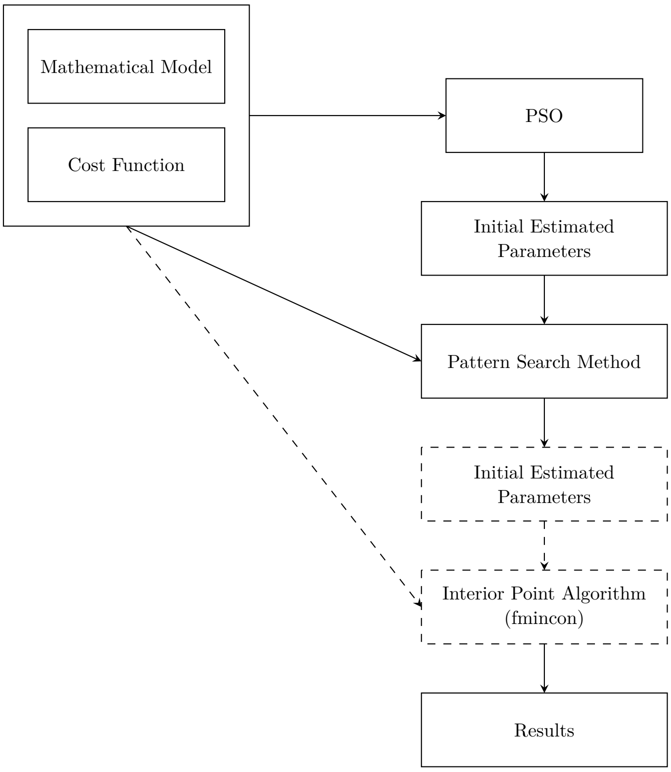

Mathematical Model
Cost Function
PSO
Initial Estimated Parameters
Pattern Search Method
Initial Estimated Parameters
Interior Point Algorithm (fmincon)
Results

Figure 3-2: Configuration of Parameter Estimation Process

stages to it: first velocity or step size is chosen, then the particle is moved according to the following formula [37]:

$$\begin{aligned} v_n = wv_n + c_1 rand()(p_{best,n} - x_n) + c_2 rand()(g_{best,n} - x_n) \\ x_n = x_n + \Delta t \cdot v_n \end{aligned} \tag{3-4}$$

where v_n is the velocity of the particle in the nth dimension, x_n is the particle's coordinate, w is the inertial weight and has an influence on to what extent does the particle maintains its original course. Scaling factors c_1 and c_2 control the relative pull of p_{best} and g_{best}, the first one being the best location found so far in the parameter space for a specific agent and the latter the best location found so far for the entire swarm. The $rand()$ function returns a random number between 0 and 1. The specific PSO algorithm used was adapted from [38].

Pattern Search Algorithm

The second algorithm used was a pattern search method, more specifically the Nelder-Mead simplex algorithm and it was implemented in MATLAB using the patternsearch function. It represents the final estimation stage and it is used to get the closest to the

optimal values as possible.
First, it requires a vector of initial points, x_j, for each parameter to be estimated that are within the set limits of the optimization. In the next step the algorithm takes into consideration a pattern P_j and a step length Δ_j. The product of these two, $s_j = \Delta_j P_j$ is added to the vector of initial points, $x_{j+1} = x_j + s_j$. The resulting parameter vectors are used to calculate the cost function. Subsequently, the values are compared and if one of them is lower than the initial value evaluated at the initial starting point $(f(x_{j+1}) < f(x_j))$, then the considered vector of parameters is replaced by the more optimal one and the polling is considered to be successful. As a result the step size is doubled and the procedure is repeated. In case the polling fails and none of the values return a lower cost then the step size is decreased [33] [39] [40].

Interior-Point Algorithm

An additional element to parameter estimation consisted of using an interior-point algorithm. This was implemented in MATLAB by using the fmincon function.
As it is gradient-based and prone to get stuck in local minima, it is used just as an additional and optional step to refine the final results or to use between PSO and pattern search to speed up the estimation process.

3-3 Procedure Description and Results

So far, all of the main components of the parameter estimation process were presented. Nevertheless, a good choice of cost function or optimization algorithm is not enough to have successful results calibrating the model. Parameter estimation is an iterative process that requires often rerunning the optimization, varying limits if it is admitted and finding ways to use the available dataset in the most efficient way possible. Therefore, in the following section an overview of the parameter estimation process of the CFFL is described combined with the analysis of the results

There were several approaches identified in literature corresponding to biological model calibration [31] [32] [34] [35]. The first step is, most of the times, finding initial values for the parameters from literature. This was not possible for the CFFL, however this issue was dealt with by implementing PSO. Subsequently, there were two directions: either parameter sensitivity analysis or start of parameter estimation (individual or simultaneous). The subsequent steps were usually chosen based on the results given from the previous stage and on how accurate were the initial parameters found from biological databases.

As values for the parameters were not found in literature for the reactions building up the CFFL, the first step in the system identification was to apply PSO to all datasets

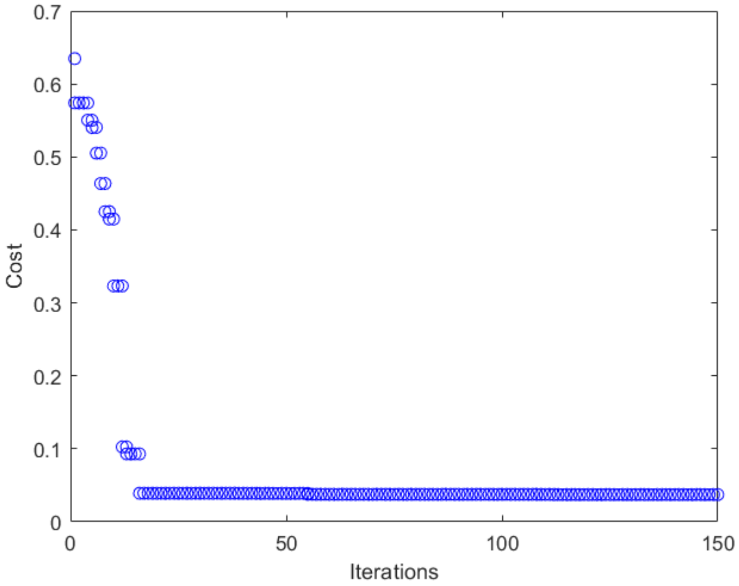


Figure 3-3: Resulting Decrease in Cost from PSO - applied on dataset 4

individually. However, the next challenge comprised of determining the appropriate bounds for the estimated parameters. This proved to be a very important detail in the entire process, as many times appropriate results were not obtained due to the fact that some parameters' values were too limited. More specifically, the mistake of having a too small range for the Michaelis-Menten constants limited the performance of the optimization in the first trials. Therefore, the first step was to consult [2] and use the limits implemented in the system identification of a genetic oscillator tuned with the help of a σ-factor. The limits used in the parameter estimation procedure of the CFFL can be found in appendix B-1. Moreover, an overall understanding of the parameters' dimension could be formulated: all reaction rate constants involved in the reversible reactions were identified to have a significantly wider range (0-200) than the reaction rate constants involved in transcription or translation (0-3.5). The Michaelis-Menten constants were allowed to have a wider range of 0-500. Moreover, the degradation constants were identified to have an even narrower range (0-1). In addition, the values corresponding to the initial concentrations of RNAP, S70 and Ribo were fixed to 70, 30 and 2000 nM.

After implementing the previously described limits and constant initial conditions, the PSO was implemented on the 5 datasets individually. The resulting parameters can be found in appendix B-2. Even though PSO provides a fast computation time, it does not perform well at a finer grain of search [41]. In other words, the algorithm is let to run until it reaches a point where a decrease in the value of the cost function stalls. This is visually presented in fig. 3-3.

In the following step of the calibration process, the estimated parameters resulting from PSO are used as initial estimation values for pattern search method. This as well was applied individually on the datasets. The results can be seen in table table 3-1.

		PS dataset 1	PS dataset 2	PS dataset 3	PS dataset 4	PS dataset 5
1	k1	0.066370612	99.999	99.999	69.90039289	99.9873748
2	k_1	0.026688265	4.656803792	45.51060126	27.32026335	0.019434147
3	Ke2	1908.486237	6287.536001	3890.92503	6613.277272	1054.662851
4	k5	3.499	3.462434054	3.389904116	3.255340061	1.428108283
5	Ke1	1.25419814	0.886635586	0.447372727	5.089642913	17.23004853
6	k3	2.881959969	3.139670488	3.406222474	3.390549044	3.415741412
7	k6	99.99871118	7.61679787	11.69312901	1.902723311	2.249834567
8	k_6	1.015590016	36.04443243	14.99521812	7.087545522	27.30217389
9	Ke3	123.4546645	30.34966799	30.19231883	27.55400473	22.24803405
10	k8	3.499	2.719444233	3.499	3.477040022	3.499
11	k9	99.99304062	99.98663769	99.98455995	89.21678093	99.99768327
12	k_9	0.013428897	0.023271717	0.020472816	6.924737122	30.08427429
13	k10	0.13730839	0.033070718	0.025783267	3.493206563	3.499
14	k11	0.214078859	0.055870859	0.039151565	0.68931565	2.742449472
15	k12	8.972694708	46.58067232	5.002268227	9.939995079	22.35169632
16	k_12	99.98574112	99.99885659	99.98770952	57.11375626	99.94757621
17	k13	0.219808053	0.126792238	0.768194229	3.492017895	1.957840303
18	k14	3.499	3.499	3.499	3.49829486	3.499
19	mat	0.999	0.999	0.999	0.996611215	0.999
20	deg6	0.369359321	0.187104414	0.098016699	0.030302341	0.092096917
21	deg7	0.013713422	0.004867992	0.037933202	0.025808642	0.039780184
22	deg8	0.994916383	0.997974695	0.998335219	0.945001413	0.905668818
23	deg9	0.019111305	0.144120237	0.07542438	0.9994226	0.175746768
24	deg10	0.999	0.999	0.999	0.00417736	0.094421948
25	atot	70	70	70	70	70
26	Rtot	2000	2000	2000	2000	2000
27	S70_0	30	30	30	30	30

Table 3-1: Estimated Parameters From Pattern Search

An important observation can be made from the results above. The values for the parameters k_3, k_5 and k_8 across all 5 datasets are similar, with the exception of one value having a lower value compared to the rest in the case of each parameter. As a consequence these will be fixed for the next step of the parameter estimation process in the following way:

$$k_3 = 3.1441$$
$$k_5 = 3.4445$$
$$k_8 = 3.499$$

This step was desired because by fixing k_5 and k_8 the structural identifiability of the model is ensured and the final set of estimated parameters will represent the only combination of values that generate the appropriate fit.

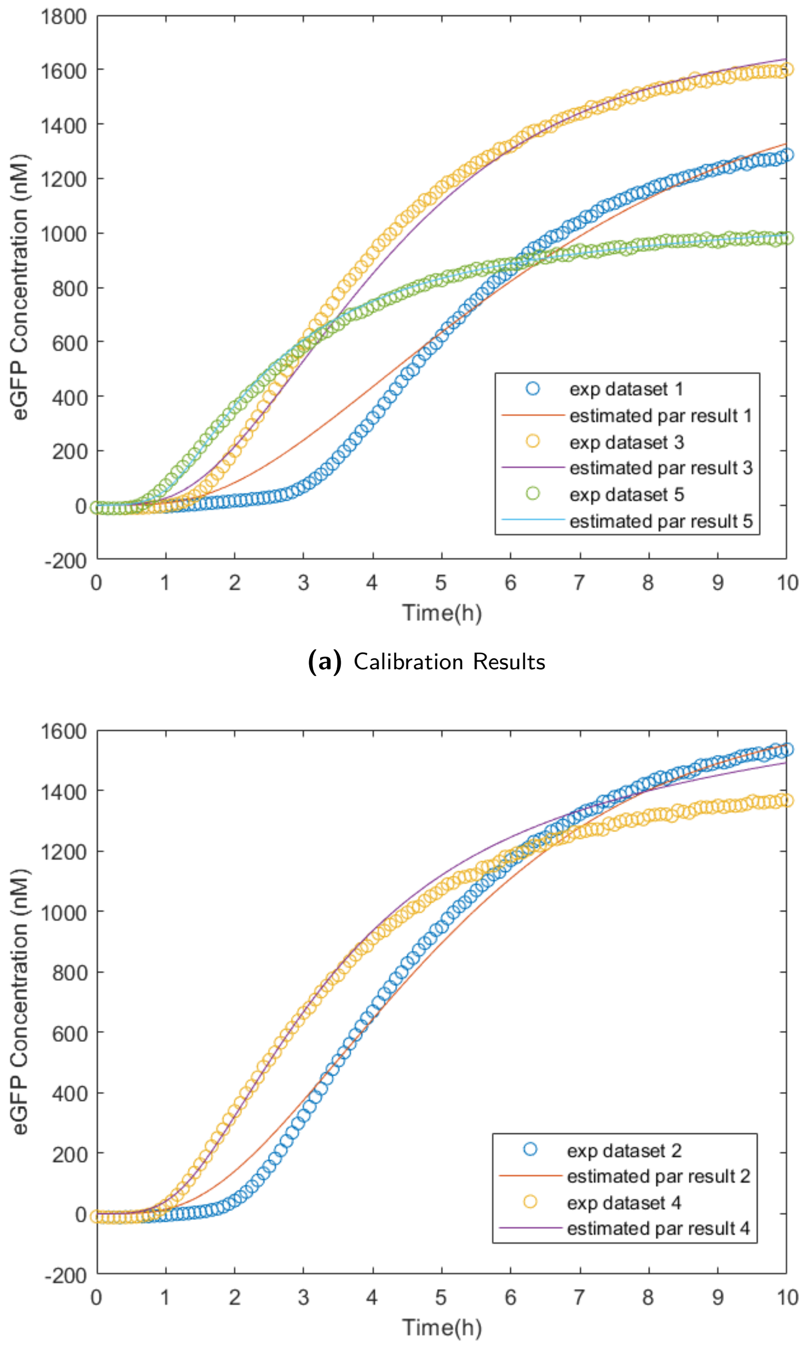
So far there are 5 datasets that have few parameters that differ significantly like $k6$ and k_6. Therefore, cross-validation with simultaneous parameter estimation will be implemented in continuation. In other words, datasets 1, 3 and 5 will be included in the cost function jointly to estimate parameters that result in an output that fits all of them. Then, the results will be used to validate the parameters on dataset 2 and 4. The simultaneous parameter estimation is completed by using the pattern search method. The initial values of the parameters used are given by the average of the parameters estimated from datasets 1, 3 and 5 from table 3-1. Another important detail that has to be mentioned that the data is cut up until 10 hours. The reason for this is that the aim of the calibration is to fit the initial dynamics as well as possible, to capture the delay function. When the entire data length is used (16 hours), it was observed that the fitting then was heavily focused on the 'plateau' region of the data. As a consequence, the start of the production and shape of the slope was neglected.
The estimated parameters can be seen in appendix B-3 and the results of the fitting and validation can be observed in fig. 3-4. The output of the system with the estimated parameters corresponding to datasets 3 and 5 is appropriately calibrated, it follows the slope of the experimental data properly. However, a compromise can be seen in the calibration curve corresponding to dataset 1. The production of the output starts faster, therefore exhibiting a shorter delay period. Subsequently, the validation results are discussed. This is really interesting to observe as it gives an indication on how accurate the estimated parameters are. In the CFFL's case the most significant detail to observe was the nonlinearity of the steady states of the experimental data. The validation on dataset 2 and 4 reflected this property. Even though the estimated output of dataset 4 did not perform well in the 'plateau' region, it still had a steady state that was less than the one corresponding to dataset 2. Therefore, validation was identified to be successful.

The last part of the parameter estimation process is composed of the final refinement of the results. This is achieved by estimating the initial parameters for each dataset individually. This is possible, as the batch experiments were conducted separately, therefore there is a high chance that the initial conditions differed for each experiment. However, the estimation bounds were set to be strict, as the variation should not be large. Moreover, only the initial concentrations of RNAP and ribosomes were estimated. These compounds were set to have the following bounds:

Parameters	Ranges
RNAP [nM]	66-75 nM
Ribosomes [nM]	1900 - 2100 nM

Table 3-2: Bounds on the estimation of initial conditions

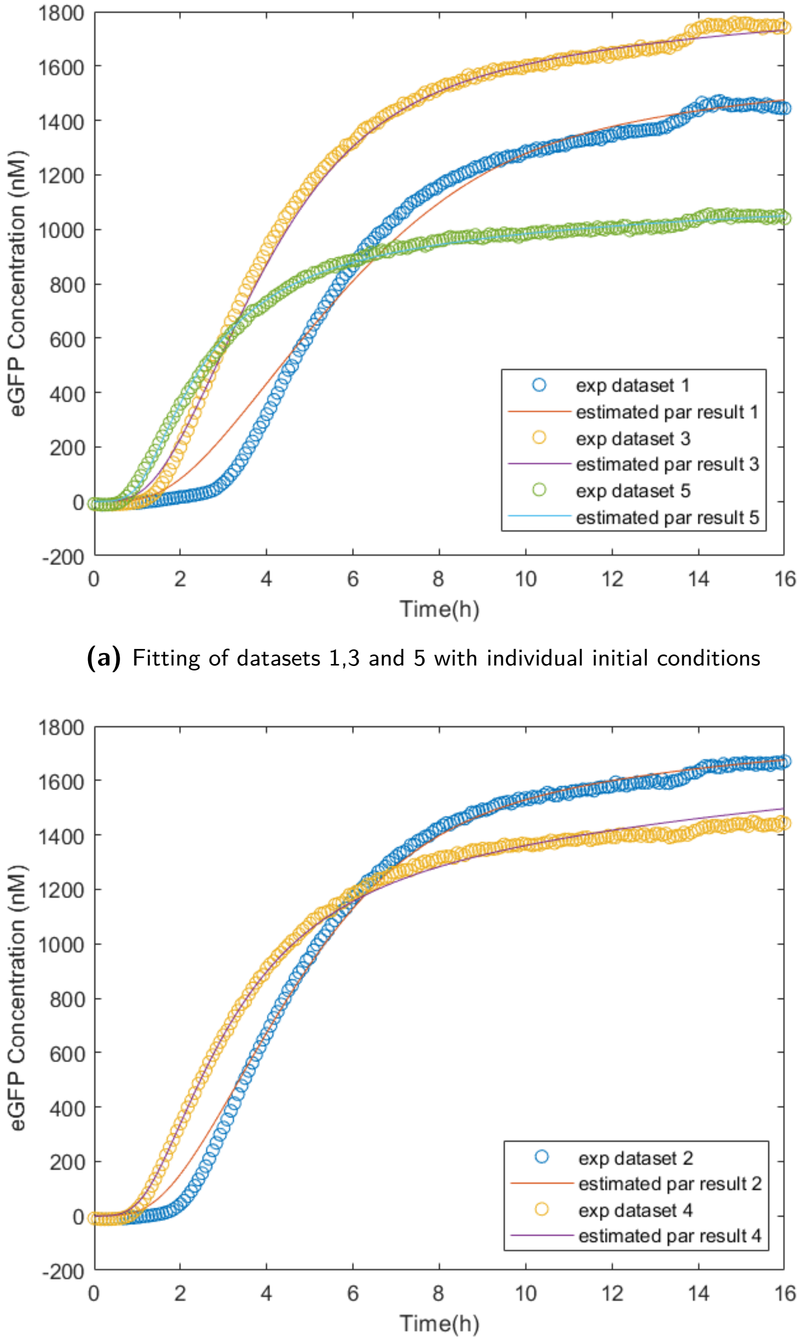
The results of the refinement can be seen in fig. 3-5. The steady-state of the curve corresponding to dataset 4 was lowered in value and resulted in a better fit after the refinement of the results.



(a) Calibration Results

(b) Validation Results

Figure 3-4: Results of simultaneous parameter estimation



(a) Fitting of datasets 1,3 and 5 with individual initial conditions

(b) Fitting of datasets 2 and 4 with individual initial condition

Figure 3-5: Results of initial conditions refinement

3-4 Summary

In summary, this chapter comprised of the system identification performed on the CFFL. In the case of biological models this was identified as a process that comprises of many challenges. The first one was found by performing identifiability analysis on the mathematical model of the network motif. This resulted in the diagnostic that the model is unidentifiable. As a consequence, this was taken into account in the subsequent parameter estimation process. However, first the individual elements of the calibration were presented: experimental data, mathematical model, cost function and the chosen parameter estimation algorithms. The latter included also the motivation for the choice of a hybrid method that combined the performance of two global optimization algorithms, PSO and pattern search method. Additionally, a gradient-based method was used to refine the results. Lastly, the calibration process was presented jointly with the results. The parameter estimation was successfully completed and a set of parameters was identified, that can be used to perform numerical analysis of the CFFL.

Chapter 4

Structural Analysis

Previous chapter heavily focused on the process of finding the suitable parameters for the derived mathematical model representing the coherent feedforward loop (CFFL). However, it is a difficult and lengthy procedure. Moreover, it requires additional experiments in many cases to have a better fitting of the data or to ensure structural identifiability. Therefore, another research direction in systems biology is to consider only the biochemical reaction network's structure and try to derive insights about the behaviour of the genetic circuit from it. In the following, some tools from structural analysis of biochemical networks will be presented and applied to gain insights about the behaviour of the CFFL. Specifically, it will be proven that the network has the capacity for multiple equilibria. Secondly, a method usually used for metabolic network analysis will be applied to the CFFL. This is called flux balance analysis and will be used to study the steady state fluxes corresponding to different objectives. All of these will help explain the source of the nonlinearity arising in the steady states of the experimental data. Moreover, it represents a set of tools that can be used to perform structural analysis on other biochemical reaction networks.

4-1 Numerical Simulations

Before discussing the insights available from the structure of the biochemical network, the CFFL is simulated with the estimated parameters from Chapter 3. The reason for this is to find the characteristics of the network that are both seen from the numerical simulations and the structural analysis of the network and find the connecting bridge between the two.
In the following section, the analysis of the dynamics of the other states of the model are discussed. As these are not available to have as experimental output, the only way

to observe them is by simulating the biological model with the estimated parameters. This way it is possible to see which chemical species end up being accumulated, from which node of the network the delay kicks in, what is getting consumed rapidly etc. The evolution of the intermediate compounds that build the network of the CFFL can be seen in fig. 4-1.

The first observations that can be made based on the numerical simulations is that two species accumulate: $\mathrm{RNA_{eGFP}}$ and S28. In addition RNAP is not fully consumed either and as expected the concentration of green fluorescent protein (eGFP) increases until reaches steady-state.. The rest of the states are all converging to 0. The delay in the output is mirrored in several of the chemical compounds' behaviour. Lower input concentrations of DNA_t results in smaller concentrations of RNA_t but larger production of $\mathrm{RNA_{S28}}$. Moreover, it also means slower production of $\mathrm{RNA_{eGFP}}$ and S28. As the latter two are directly responsible for the translation of eGFP, it is straightforward that the delay will be exhibited in the output as well.
In the following the focus is put on the nonlinearity of the final concentration values that are exhibited in the experimental data. As this is a characteristic observed at steady-state, the attention is directed to the species that accumulate during the functioning of the CFFL (fig. 4-2). Moreover, the timespan of the simulation is extended up to 50 hours in order to capture the steady-state. Both have different order corresponding to the respective input, therefore it is not possible to conclude that one of them specifically affects directly the steady-state values of eGFP. In fig. 4-2a it can be observed that the steady-state corresponding to the highest input has by far the highest value. In the output steady-state concentrations this is reflected by eGFP having a significant lower value than the rest. As a consequence, it was determined that the higher concentration of S28 and $\mathrm{RNA_{eGFP}}$ is produced, the lower value will eGFP reach. It is consistent with the logic governing the functioning of the bio-chemical reaction network as both species are needed in order for the output to be produced. However RNA_t is depleted too fast and the production of eGFP halts.

In the subsequent sections of this chapter the focus will be put on the applications of structural analysis tools that will result in specific insights about the behaviour of the CFFL without using the knowledge of the parameter values. These will be connected with the dynamics presented from the numerical analysis of the network motif.

4-2 Steady State Analysis

A significant research branch is dedicated to infer insights about the number and types of steady-states a biochemical reaction network possesses. As the biological processes are inherently nonlinear and complex, it is a challenge to determine the type and number of steady-states as the initial conditions of the system can vary. Moreover, the only

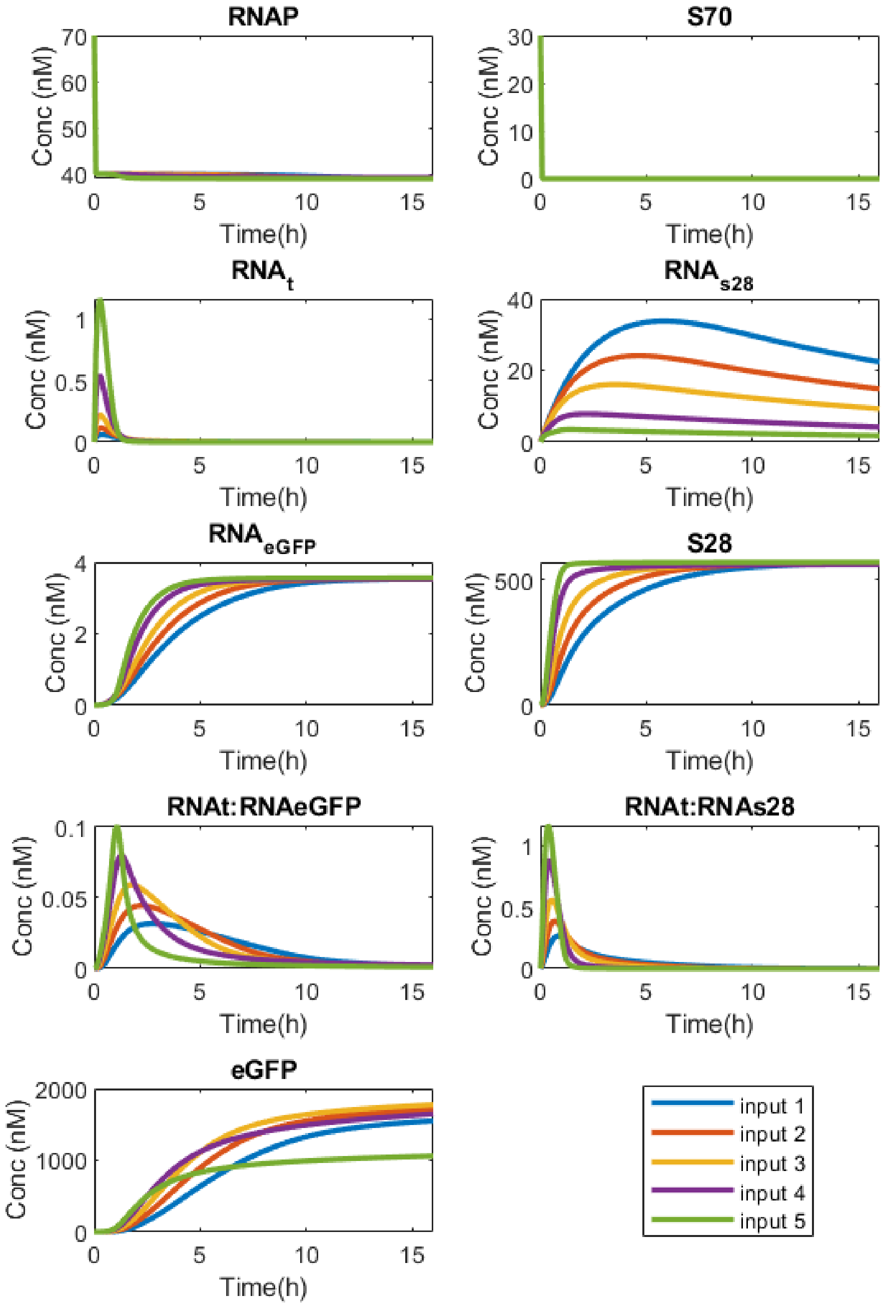
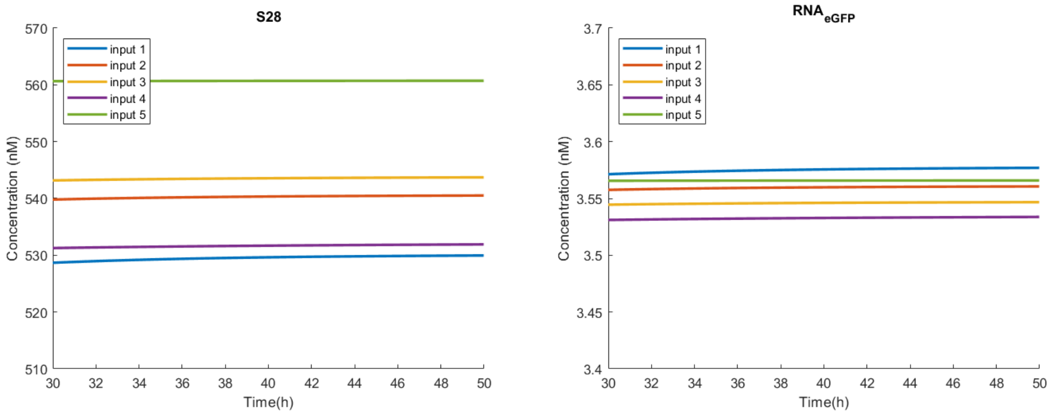


Figure 4-1: Dynamics of the intermediate chemical complexes



(a) Accumulation of S28

(b) Accumulation of RNA$_{eGFP}$

Figure 4-2: Steady-state values of S28 and RNA$_{eGFP}$

steady-states that are possible to observe experimentally, are the stable ones. This line of theory was developed and applied on the reaction networks' structure, without the additional step of mathematical modelling and specifications of the kinetics. This is possible as the equilibria are relative to the stoichiometry class. In some cases, little knowledge about the type of kinetics is required, for example if it is only mass action or it includes also Michaelis-Menten kinetics. The reason for this is that large part of the chemical reaction network theory was developed first on biological models only comprising of mass action dynamics. However, in many cases the theoretical tools were extended to include Michaelis-Menten kinetics as well.

An overview of the background theory and methods were presented in the literature study report [19]. Deficiency theory was the starting point and first direction in chemical reaction theory and made it possible to prove that certain networks' structure have deficiency zero and consequently have a locally stable equilibrium [42] [43]. In addition to this, graph theoretic approaches were developed [44] and also methods that are based on P matrix properties of the system [45]. Some of these mathematical proofs are implemented into algorithms that run tests on a specific network the user inputs. Three of these were applied to the CFFL: CoNtRol, chemical reaction network analysis tool [46]; ERNEST, a toolbox for chemical reaction network theory [47]; and CRNT, chemical reaction network toolbox [48].

Before presenting the results it must be mentioned that the CFFL network is considered to be non-autocatalytic (N1C). A system with stoichiometric matrix S and flux vector $v(x)$ has a corresponding matrix $V(x)$ defined by $V_{ij}(x) \equiv \frac{\delta v_i}{\delta x_j}$, that represents the dependence of the reaction rates on the concentrations [45]. The product SV corresponds to the Jacobian of the system. In order for a system to be N1C, it must fulfil the following condition:

Condition 4-2.1. A reaction system is N1C if $S_{ij}V_{ij} \leq 0$ for all i and j, and $S_{ij} = 0 \Rightarrow V_{ij} = 0$

Put simply, it means that the reactants do not occur on both sides of the reactions. This is true in the case of the CFFL if the network includes the intermediate complexes of RNAP:S_{70}:DNA_t, RNAP:S_{70}:DNA_{s28}, RNAP:S_{28}:DNA_{eGFP}, RNA_t:RNA_{S28}:Rib and RNA_t:RNA_{eGFP}:Rib. Therefore, in this way it is avoided to have the holoenzymes, DNA templates or ribosomes on both sides of the reactions. For the subsequent analysis it is important for the network to be N1C as it is a condition for the proofs to be valid [44] [49]. In the following the results from the mentioned algorithms will be presented combined with the mathematical theory that accompanies them. Main objective of this analysis is to gain knowledge about the steady-states of the CFFL.

4-2-1 Graph-theoretic Approach

Firstly, CoNtRol was used and the main finding from the analysis resulting from it that the CFFL network does not exclude multiple steady-states. The algorithm implemented in the software tool uses graph-theoretic approaches to analyse chemical reaction networks and the background theory behind it is presented in [50] [44]. The first step is the derivation of a directed species-reaction (DSR) graph representing the network, this can be seen in fig. 4-3. A species-reaction (SR) graph is a bipartite undirected graph with nodes being represented by both species and reactions. A DSR graph is a SR graph with directed edges, depending on the reversibility or irreversibility of the reactions. Based on this, the term cycle is defined as the minimal directed path starting and ending at the same vertex [50]. Cycles can be split into two categories: e-cycles and o-cycles. This is done by checking whether the parity of the cycle is negative or positive. The parity is defined in the following way:

$$P(C) = (-1)^{|C|/2} sign(C) \tag{4-1}$$

Where $|C|$ is the length of the cycle (number of edges) and $sign(C)$ represents the sign of the cycle which is the product of the signs of the individual edges. Negative edges correspond to dashed lines while positive edges correspond to continuous lines in fig. 4-3. In case the parity of a cycle is positive, then it is classified as en e-cycle. However, if the parity results to be negative, then it is an o-cycle. Moreover, two cycles have an S-to-R intersection if each shared path corresponds to a S-to-R path.
Having all these important elements defined, the following condition is presented which can be easily verified on the DSR graph of any network:

Condition 4-2.2. All e-cycles are s-cycles, and no two e-cycles have S-to-R intersection.

The first part of condition 4-2.2 is fulfilled by the CFFL as the DSR graph corresponding to it has edges labelled with 1 only. The second part has to be verified in order to be able to exclude multiple equilibria.

From condition 4-2.2, the following proposition was introduced in [44]i:

Proposition 4-2.1. An N1C reaction system with DSR graph satisfying condition 4-2.2 is injective.

From injectivity it follows that the chemical reaction network admits only one steady-state. The application CoNtRol was not able to run on the full CFFL network to verify if condition 4-2.2 is fulfilled by it. However, it did run on a subset of the graph, more specifically until the production of S28 and confirmed that the graph fulfils condition 4-2.2. From this point, the remaining part of the full CFFL graph was verified manually. Nevertheless, the result was negative, the possibility of multiple equilibria was not excluded. The reason for this that an S-to-R intersection was found. There were in total 4 cycles that had to be verified additionally in order to check condition 4-2.2. Two of them are illustrated by 2 different colours: cycle 1 - blue and cycle 2 - green. Both of them are e-cycles. As mentioned, all e-cycles are s-cycles in the CFFL network. However, an S-to-R intersection can be attributed to them. The location of it can be seen in fig. 4-3 where the two highlighted cycles overlap between reaction number 14 and Ribo. Therefore, the admittance of multiple steady states is not excluded.

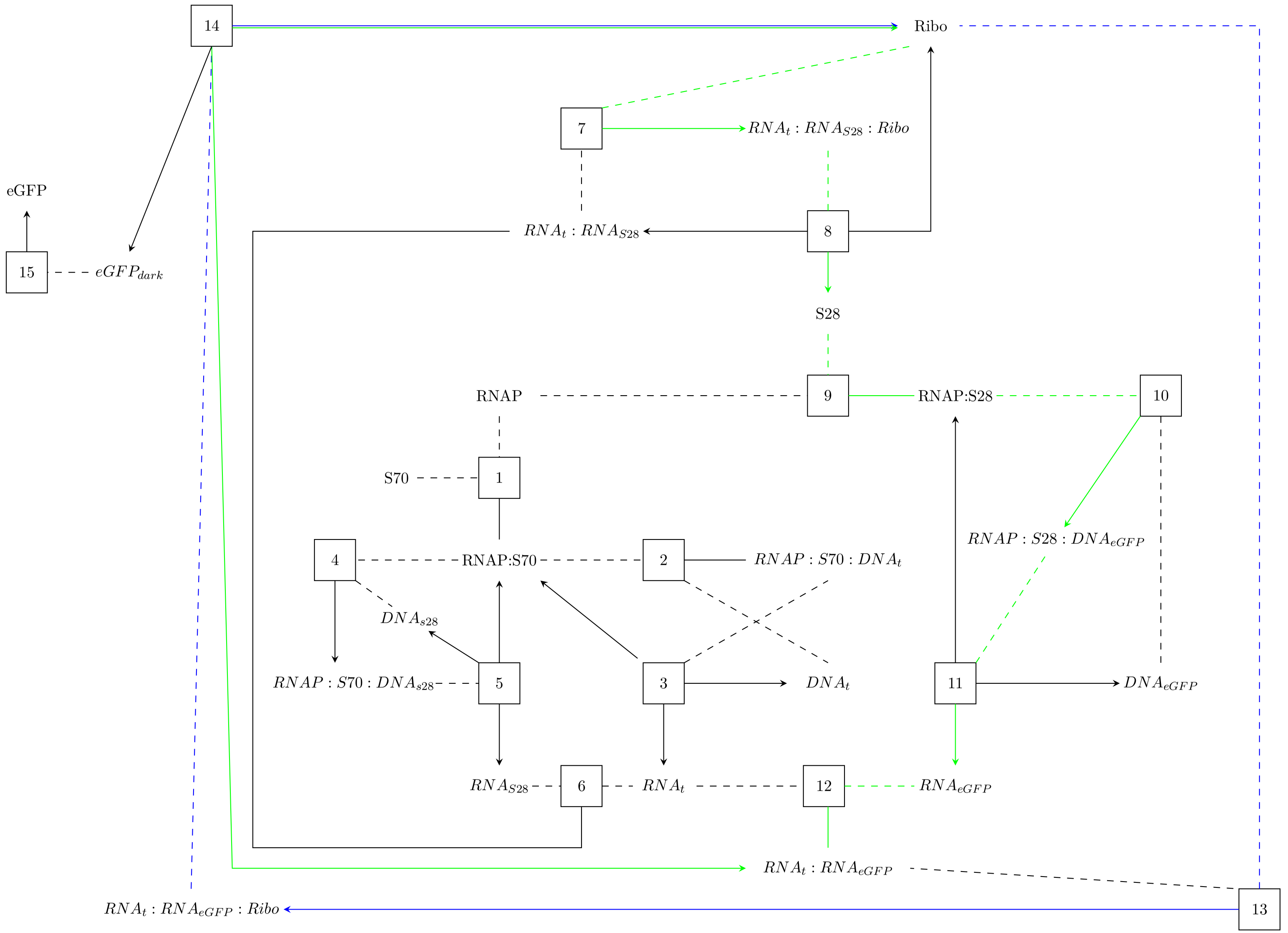


Figure 4-3: DSR graph of CFFL

4-2-2 P Matrix Properties

Another way of attempting to exclude the admittance of multiple steady-states is to verify if the stoichiometric matrix is strongly-sign-determined (SSD). The mathematical theory that connects injectivity to properties of the stoichiometric matrix can be found in [45] and will be presented briefly in this subsection. The condition is implemented as an algorithm in ERNEST toolbox and was verified on the CFFL network.

The main analysis starting point is the Jacobian of the biological system described by eq. (2-3). However, first the class of P matrices has to be introduced. P matrices are square matrices that have the property that all of the principal minors are positive. If all the principal minors are nonnegative, then it is said to be a P_0 matrix. Consequently, P matrices are nonsingular and the corresponding eigenvalues are restricted to a range [51]. In case a matrix -A is a P matrix, then A is said to be a $P^{(-)}$ matrix. The next step is to connect the previously described class of matrices and the model representing a bio-chemical reaction network. This is achieved by analysing the stoichiometry matrix and verify if it is SSD. This property is fulfilled, if all of the square submatrices are either sign-nonsingular or singular. As a consequence, the Jacobian of the system is a $P_0^{(-)}$ matrix (Theorem 3.1 from [45]). From this, injectivity of the function results and the admittance of multiple equilibria is excluded.

The same results as in section 4-2-2 were obtained by implementing ERNEST toolbox on the CFFL. The stoichiometry matrix was identified as not SSD. Therefore, the network has the capacity for multiple steady states, i.e. multiple steady states might exist.

4-2-3 Identification of Equilibria from Mathematical Model

The previous two subsections attempted to gain insights about the steady states of the network without knowledge of the type of kinetics present in the network and kinetic constants. However, it was not possible to derive a significant amount of information, apart from the fact that the network has capacity of multiple equilibria. In the next step, the type of kinetics will be assumed to be known. This means that the derived model from Chapter 2 will be used but without knowledge of the parameters.

The first step was to set the ordinary differential equation (ODE) system to be at steady-state. The equations corresponding to the DNA templates' and ribosomes' concentration were taken out as their value stays constant. Moreover, the ODE corresponding to the output, eGFP is taken out as well. Therefore, there are 9 ODEs left for analysis. From these two types of steady states were identified.

1. $[x_{1,ss} \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$
2. $[x_{1,ss} \;\; 0 \;\; 0 \;\; 0 \;\; x_{8,ss} \;\; 0 \;\; 0 \;\; x_{11,ss} \;\; 0]$

The state variables present above are the ones corresponding to the species shown in Section 2-4. The steady state of $x_{1,ss}$ can have any value (equal or less than its initial value, $RNAP_{tot}$). In the second type of steady state $x_{8,ss}$ and $x_{11,ss}$ can have any value as long as the following conditions are fulfilled:

$$\begin{aligned}&\frac{k_{-6}RNAP_{tot}(x_{11,ss}/K_6)}{1+(1+x_{5,ss}/K_{e3})(x_{11,ss}/K_6)+(1+x_{3,ss}/K_{e1}+x_{4,ss}/K_{e2})(x_{2,ss}/K_1)}-\\&-k_6x_{1,ss}x_{11,ss}=0\\&\frac{k_8RNAP_{tot}(x_{11,ss}/K_6)(x_{5,ss}/K_{e3})}{1+(1+x_{5,ss}/K_{e3})(x_{11,ss}/K_6)+(1+x_{3,ss}/K_{e1}+x_{4,ss}/K_{e2})(x_{2,ss}/K_1)}-deg_8x_{8,ss}=0\end{aligned} \tag{4-2}$$

As a conclusion, it can be seen that there is an infinite number of steady-states. However, it would be useful to investigate the stability of them, specifically what conditions have to be fulfilled in order to have stable equilibria. This is achieved by computing the Jacobian of the mathematical model, linearising it and deriving the corresponding eigenvalues. The Jacobian was computed symbolically using Mathematica. It was observed that by linearising the Jacobian with the two steady-states mentioned above, the eigenvalues can be identified being located on the diagonal of the matrix. Therefore, it was identified that by linearising with the first type of steady-state enlisted above, one of the eigenvalues will always be positive. Therefore, it is an unstable steady-state. Using the second type of steady state, it results in negative eigenvalues as long as the inequality from eq. (4-3) is fulfilled, but it also has one zero eigenvalue. Therefore, the stability of the steady-state cannot be determined.

$$RNAP_{tot}-x_{1,ss}(0.0014x_{11,ss}+1)<0 \tag{4-3}$$

4-3 Tools Applied on Metabolic Networks

The processes occurring in the cells can be categorised into four groups: sensing, signalling, regulation and metabolism [3]. All of them are constituted by biochemical reactions, however the structure of the networks and products can be different, and also distinct modelling is usually applied.

A set of tools were developed with the aim to be used on metabolic networks in order to determine ways to maximize production of certain biochemicals like amino acids, lactic acid, ethanol etc. [52]. These tools include elementary modes analysis and flux balance analysis. Both of the approaches have the starting point the network dynamics being described by the product of the stoichiometry matrix and the vector of reaction rates from eq. (2-3). The CFFL network can be also written up in this way even though it represents a transcription network. Therefore, tools developed for metabolic networks can be applied on to the CFFL in order to gain insights about the system at steady state.

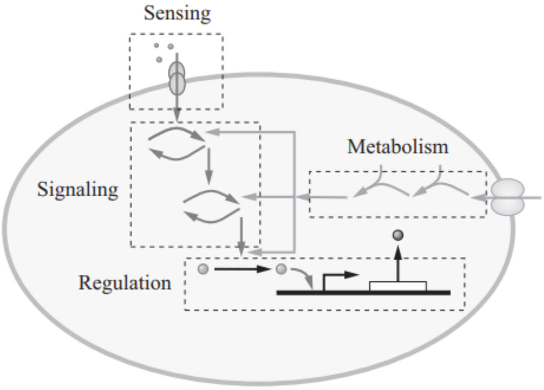


Figure 4-4: Processes in a Cell [3]

4-3-1 Elementary Flux Modes

The first analysis tool described is based on elementary modes analysis. This process starts by computing the right null-space of the stoichiometry matrix. As mentioned in Chapter 2, this results in finding the steady-state flux distributions. Therefore, the way chemical species are produced in a biochemical reaction network, specifically which reactions are activated in order to get the output, can be determined by a set of basis vectors in the null-space. A difficulty with these is that they are not unique. A way of reducing this space of solutions is to consider the irreversibility of the relevant reactions. As a result, the flux vectors that are admitted as steady-state occupy only a subset of the null-space. Therefore, in case of a stoichiometry matrix S and vector of reaction rates v, there are two conditions that have to be fulfilled:

Condition 4-3.1. $Sv = 0$

Condition 4-3.2. $v_{irr} \geq 0$

where v_{irr} is the subvector of v and represents the fluxes of the irreversible reactions [53]. As a consequence of applying the above presented conditions, the region of admitted solutions will be represented by a convex polyhedral cone, called flux cone. This is illustrated in fig. 4-5.

In the following, the mathematical background will be succinctly summarised from paper [53]. In case the reactions of the network are all irreversible, the flux cone is pointed. As a result, proven in [54], pointed polyhedral cones are made up of a finite set of unique vectors. Moreover, these can be found on the edges of the cone. However, in most cases there are reversible reactions, meaning the flux cone is not pointed. Nevertheless, in [53] it is shown that is still possible to determine an unique set of generating vectors that spans the polyhedral cone. The vectors that cannot be decomposed into two other vectors spanning the cone are defined as elementary flux modes (EFM). These can be seen as the minimal pathways in the system in order 'to get from' the external input

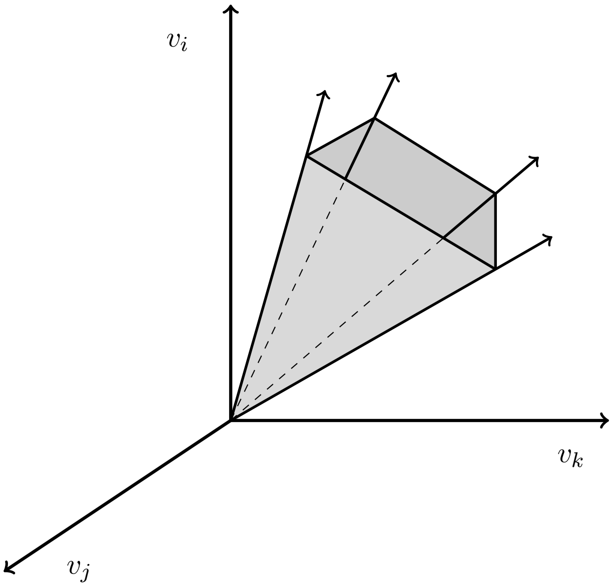


Figure 4-5: Flux cone in 3 dimensional visualisation

chemical species to the external output protein. By combining EFMs it is possible to generate all possible pathways in the network.

Several algorithms were developed to identify EFMs [55]. The software package used for the CFFL is called METATOOL, one of the first tools that was developed for elementary modes analysis [21]. It is ran in the MATLAB environment and requires a specific input file to complete the analysis. As elementary mode analysis was developed for metabolic networks, a challenge is encountered in using the METATOOL on transcription networks. Therefore, the structure of metabolic reactions have to be compared to the transcription and translation reactions in order to find a way to use METATOOL on the CFFL. Firstly, the external and internal species are chosen. The external ones are represented by RNAP, S70, $\mathrm{RNA_{trigger}}$, $\mathrm{RNA_{S28}}$, $\mathrm{RNA_{eGFP}}$, S28 and eGFP. The reason why the RNA species are considered to be external complexes is that they represent the intermediate output of the transcription reactions. It is identified that without the presence of any of the RNA species, the output protein cannot be produced. Moreover, in order for this metabolic network analysis tool to be applied on transcription networks, they must be defined as external species. The rest of the complexes taking part in the CFFL are declared as internal 'metabolites'. The next step is to write up the reactions. A significant difference was identified between the structure of the metabolic reactions and gene expression processes. The metabolic reactions are represented by the conventional enzymatic reactions that constitute the base of biochemical reactions modelling. Hence, their structure is the following:

$$\mathrm{S} + \mathrm{E} \longrightarrow \mathrm{P} + \mathrm{E} \tag{4-4}$$

Where S is the substrate, E is the enzyme and P is the product. In words, a substrate in the presence of an enzyme is transformed into a product. However, in the case of the translation or transcription reactions, the following structure is present:

$$\mathrm{A} + \mathrm{B} \longrightarrow \mathrm{P} + \mathrm{A} + \mathrm{B} \tag{4-5}$$

Where in case of transcription A is the holoenzyme, B is the DNA template and P is the RNA species. In case of translation, A is the RNA species, B is ribosome and P is the output protein. Therefore, in order to have the reactions in a similar manner as in eq. (4-4), it was decided to modify eq. (4-5) in the following way:

$$\begin{aligned} \mathrm{A} + \mathrm{B} &\longrightarrow \mathrm{P} + \mathrm{B} \\ \mathrm{A} + \mathrm{B} &\longrightarrow \mathrm{A} + \mathrm{B} \end{aligned} \tag{4-6}$$

This method of defining the reactions and with the correct specification of the external metabolites the algorithm used by METATOOL can be run correctly on the CFFL network.

In the following, the results of the algorithm are presented. There were ten elementary flux modes identified and they can be found in appendix C-1. The output file of METATOOL presented them in the following way:

$$\begin{aligned} &1 : \mathrm{RNAP} + \mathrm{S}_{70} \rightleftharpoons \mathrm{RNA}_{\mathrm{S28}} \\ &2 : \mathrm{RNAP} + \mathrm{S}_{70} \rightleftharpoons \mathrm{RNA}_t \\ &3 : \text{no net transformation of external metabolites} \\ &4 : \mathrm{RNA}_t + \mathrm{RNA}_{\mathrm{eGFP}} \rightleftharpoons \mathrm{eGFP} \\ &5 : \text{no net transformation of external metabolites} \\ &6 : \mathrm{RNAP} + \mathrm{S}_{28} \rightleftharpoons \mathrm{RNA}_{\mathrm{eGFP}} \\ &7 : \text{no net transformation of external metabolites} \\ &8 : \mathrm{RNA}_t + \mathrm{RNA}_{\mathrm{S28}} \rightleftharpoons \mathrm{S}_{28} \\ &9 : \text{no net transformation of external metabolites} \\ &10 : \text{no net transformation of external metabolites} \end{aligned} \tag{4-7}$$

Each row above corresponds to a column in the elementary modes matrix from appendix C-1. As the RNA species were identified as intermediate 'external metabolites', the total pathway from S70 to eGFP is segmented. However, it is observed that the combination of the identified elementary flux modes gives the total pathway to the production of eGFP. Therefore, there is only one way of getting from S70 to eGFP as expected from the chain of reactions building up the CFFL. Therefore, the METATOOL was applied successfully to a transcription network. In the following, to be able to gain more useful insights from the structure of the network, flux balance analysis (FBA) is applied. The knowledge acquired from application of METATOOL is transferred subsequently, as FBA is usually applied on metabolic networks and not gene expression models.

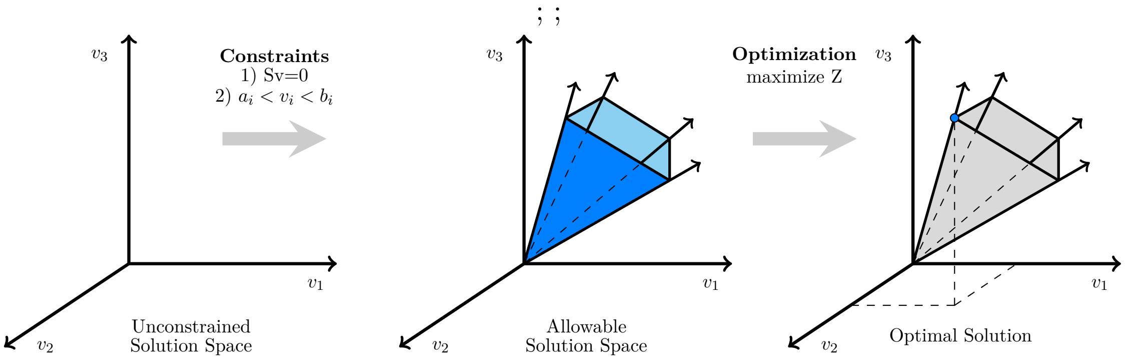


Figure 4-6: Visualisation of Constraints Based Modelling [4]

4-3-2 Flux Balance Analysis (FBA)

In the following, the static modelling tool called FBA will be briefly presented, in combination with the implementation of it. Similarly to the elementary modes analysis, it is a method applied to metabolic networks and it is used widely for applications like optimization of bio-processes in industries [56]. Its advantage is that it does not need any kinetic parameter nor species concentration in order to apply it. However, this means that the system is only studied at steady state. Moreover, it concentrates on the fluxes distributed in the network which cannot be directly connected to the concentrations of the chemical species involved in the biochemical reaction network. Nevertheless, FBA offers the possibility to infer some conclusions about the system at steady state and also explain the nonlinearity of the batch data regarding the steady-state concentrations of the output, eGFP.

FBA is made up of four parts: system definition, deriving the stoichiometric matrix, determining the relevant objective function and choosing the correct constraints and optimization [56]. The first step is constituted by determining the chemical reaction set that builds up the CFFL. These were provided by the Institute of Complex Molecular Systems (ICMS) at Eindhoven University of Technology and are presented in Chapter 2. The second step is to generate the stoichiometric matrix. This was also described in Chapter 2. The next step is to determine the relevant objective function. In the case of the CFFL several different objective functions will be applied, in other words different species' production will be chosen to be maximised. This way it will be possible to study the resulting effect on the output flux of eGFP. The constraints will be applied by taking into account the speed of the different reactions. And lastly, the optimization will be implemented using a software package.

In the following, it will be briefly presented how the optimization problem is formulated. The starting point is the derivation of the stoichiometric matrix. Steady-state is achieved by setting the product of the stoichiometric matrix with the reaction rate vector to equal

0. Therefore, all the solutions lie in the null-space of the stoichiometric matrix S. This is visualized in fig. 4-6. In order to reduce the number of solutions, constraints are defined. The first set of constraints is represented by the conservation laws originating from the stoichiometry. This makes sure that the amount produced in the network is also consumed in order for the system to be at steady state. The second set of constraints arise from setting limits to the fluxes of each reaction. As a consequence, the solution space 'shrinks' to a flux cone, the same one mentioned in section 4-3-1. In order to get to a single vector solution, the implementation of an objective function c is introduced and the following optimization problem is formulated [56]:

$$\max_{v} c^T v \;\; s.t. \;\; Sv = 0 \tag{4-8}$$

This represents a linear programming problem and solving it results in a single optimal distribution of fluxes.
There are several software tools available to implement FBA. The one used in this project is a MATLAB toolbox called COBRA [57]. The first step in the implementation of FBA using the COBRA toolbox is to define the reactions building up the CFFL. The same strategy is applied as in Section 4-3-1. First of all, like before, the external species are chosen to be represented by RNAP, S70, RNA_t, RNA_{S28}, RNA_{eGFP}, S28 and eGFP. Secondly, the transcription and translation reactions are separated into two reactions, eq. (4-6), having the structure of metabolic process. As a result, in total there are 20 reactions. In addition to this, the external species are required to have a corresponding inflow/outflow reaction. Therefore, the model which serves as an input to the FBA is comprised of 27 reactions in total. The next step is to set the bounds on the fluxes corresponding to each reaction. In this case the attention is directed to the reactions corresponding to the external species. Values that are negative represent fluxes that are flown into the network while values that are positive correspond to fluxes of species consumed/flown out. RNAP and S70 are only flown into the network, therefore their upper bound must be set to 0. The other chosen external species are only consumed, accumulated or flown out. Therefore, their lower bound are set to 0.

In the following the resulting flux distributions from FBA will be analysed. First the biochemical reaction network is visualized in fig. 4-7. The 20 modified reactions composing the CFFL can be seen noted on the edges connecting the chemical species. The 7 inflow/outflow reactions are also illustrated: R21, R22 correspond to the inflow of RNAP and S70 respectively and from R23 to R27 the outflow of RNA_t, RNA_{S28}, RNA_{eGFP}, S28 and eGFP are denoted. In the first round of FBA implementation, 5 different objective functions were tested, meaning 5 different species were chosen to be maximised. The results are shown in appendix C-2 jointly with the objective function vectors used for each optimization in appendix C-3. It is important to mention that the flux values do not represent concentrations. In the first case, the production of the output protein, eGFP is maximised. This results in an output flux of 0.333 for R27 reaction for a value of 1 for input flux of R22. However, in case the objective function is set to maximise any of the RNA species' production or the intermediate protein's (S28) production, then the output flux of R27 is close to 0, which means that no output is produced. The reason for this is that the maximised species are accumulating, however not consumed. Hence, no output protein will be produced.

In the next step the objective function will be more complex. Two species' production will be chosen to be maximised at the same time, the production of eGFP will be set to 1, and the production of a chosen second species will be set to 0.5. This way, the priority is still to produce as much eGFP as possible but it is simulated the case when one of the intermediate species is accumulating. The results of the second round of FBA can be seen in appendix C-4 jointly with the objective function vectors used for each optimization in appendix C-5. In case the production of RNA_t or $\mathrm{RNA}_{\mathrm{S28}}$ is maximised next to eGFP the results show that no output will be produced. On the other hand, if $\mathrm{RNA}_{\mathrm{eGFP}}$ is chosen then the resulting output flux for eGFP is the maximum. These three cases are not really of interest as the behaviour of the CFFL from the batch experiment can not be linked to any of them. However, in case S28 is chosen to be included in the objective function, the output flux drops from 0.333 to 0.315. The flux distribution corresponding to this case is visualised in fig. 4-8. The drop in the output flux shows that the nonlinearity of the produced eGFP concentrations can be linked to the accumulation of S28. In other words, a higher input concentration of DNA_t results in an accumulation of S28 protein and consequently a lower output concentration of eGFP. This behaviour can be linked to the results of the numerical simulations of the CFFL. The intermediate protein, S28, was shown to accumulate which is the reason for the resulting nonlinearity of the steady-states. Therefore, insights provided by a structural analysis tool were reflected in the simulations of the model with the estimated parameters.

4-4 Summary

This chapter focused mainly on deriving insights about the behaviour of the CFFL strictly from the biochemical network structure, without any knowledge about the parameters. Firstly, the numerical analysis of the network motif was presented. This way it was possible to observe the dynamics of the other chemical species, not only the output. Then, the steady-states of the system were analysed, only based on stoichiometry. However, the only result that could be derived was the fact that multiple steady-states cannot be excluded, Therefore, the mathematical model without parameters was studied and it was found that admits an infinite number of steady-states. The last two sections of this chapter encompassed the application of two static analysis tools, usually applied on metabolic networks: EFM analysis and FBA. The latter was able to explain the nonlinearity of the steady-states by connecting the accumulation of S28 to lower concentrations of eGFP produced. This was also confirmed by the numerical analysis where the same behaviour was observed.

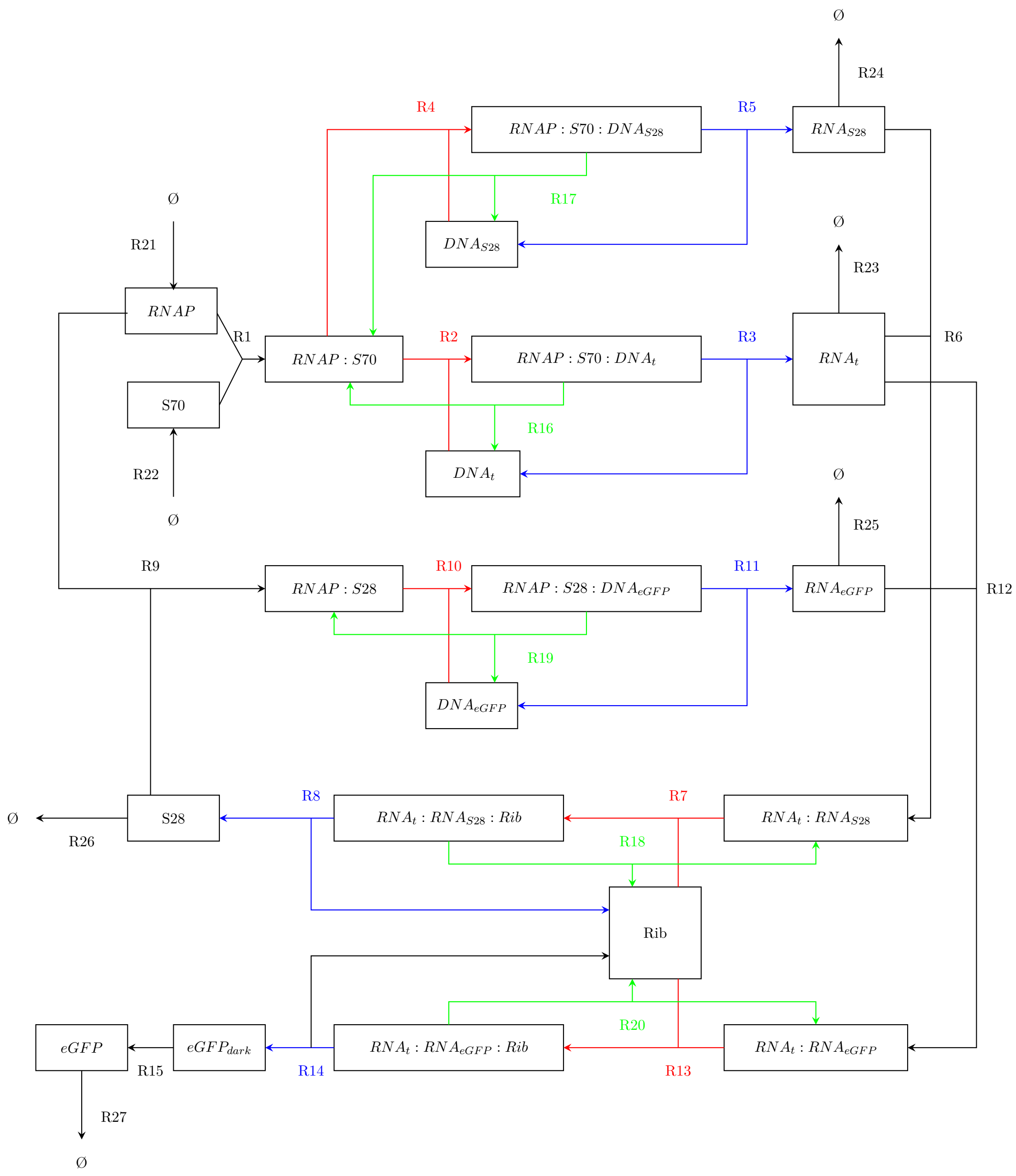


Figure 4-7: CFFL network structure for FBA

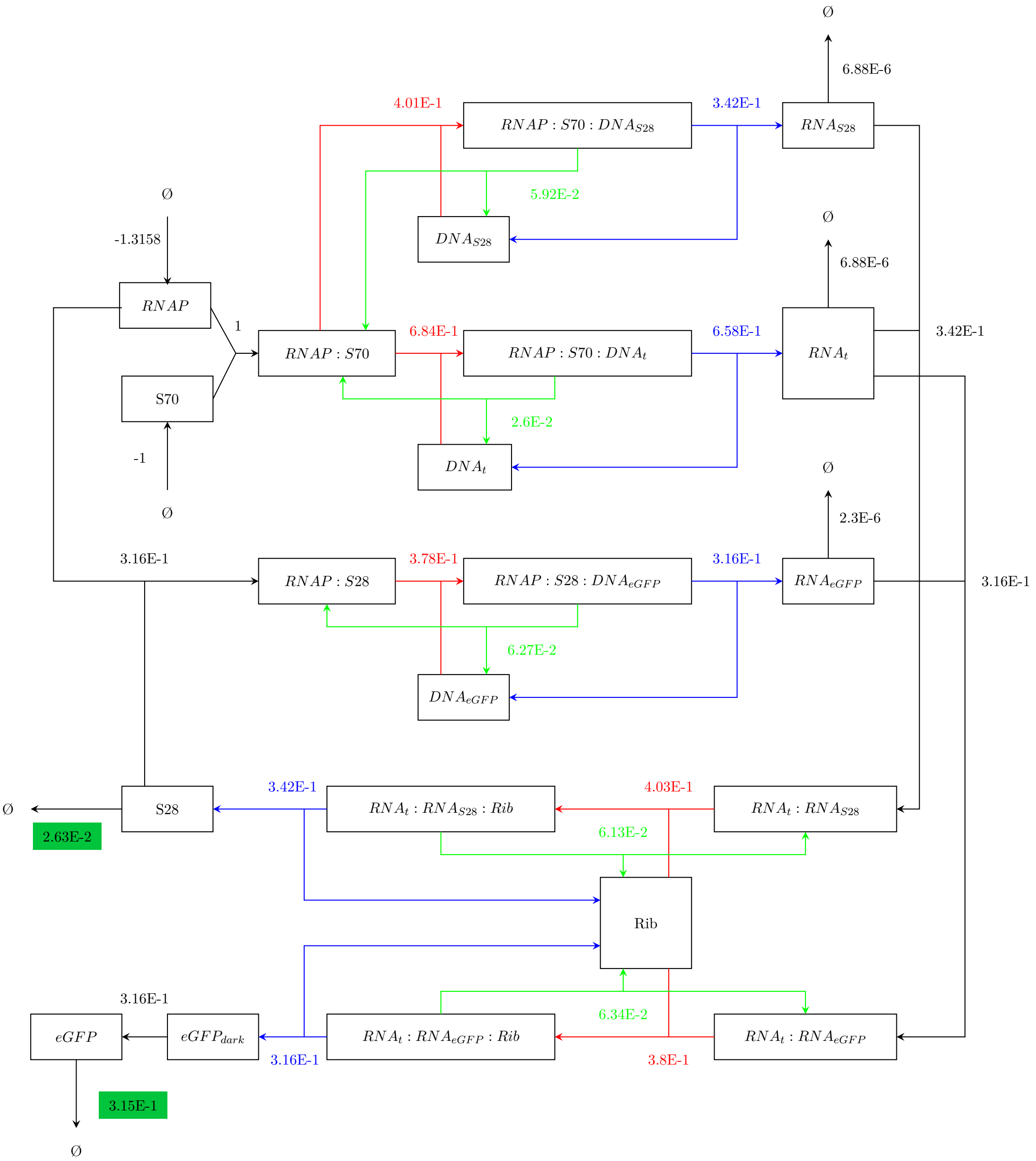


Figure 4-8: FBA - Flux distribution resulting from maximising the production of eGFP and S28

Chapter 5

Conclusion and Future Work

5-1 Conclusion

This thesis represents a multidisciplinary approach to a research problem that emerged from the field of systems biology. It combines knowledge from biology, optimization, mathematics and systems and control theory in order to fully analyse the biological network motif of coherent feedforward loop (CFFL). Moreover, during the project work, a complete modelling and analysis framework was developed that can be applied on more complex networks. More specifically, on the combined feedforward loops that are the centre of one of the research directions at the Institute of Complex Molecular Systems (ICMS) from Eindhoven University of Technology (fig. 5-1).

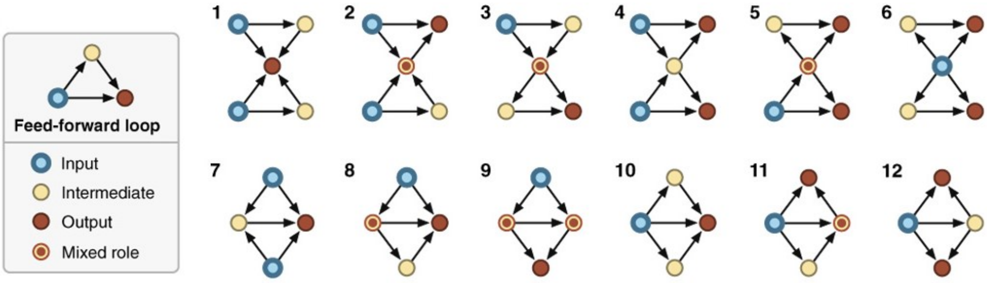


Figure 5-1: The 12 unique motif clustering types for two feedforward loops [5]

In the following the thesis work is presented together with the contributions achieved. The thesis comprises of three main parts: modelling, parameter estimation and structural analysis of the CFFL. Firstly, a modelling framework was developed that can be applied on transcription networks using σ-factors to initiate protein synthesis. It is a deterministic modelling method that includes mass-action and Michaelis-Menten kinetics.

Moreover, it presents how can the competitive binding of certain species be incorporated into the model. In addition, model reduction was also applied by implementing quasi-steady state approximation on the transcription and translation reactions. This resulted in a nonlinear ordinary differential equations system with 14 states and 27 unknown parameters.

The next part of the thesis is comprised of the parameter estimation process. The parameters of the model are unknown, hence this represents an obstacle in analysing the dynamics of the CFFL and simulate it. First an a priori identifiability test was conducted to verify if the model is identifiable or not. The result was negative, however it was identified that in case two parameters are fixed, the model becomes structurally identifiable. The next step consisted of choosing the right optimization algorithms for calibration of the model. A hybrid optimization strategy composed of particle swarm optimization and pattern search method was implemented. This made it possible to obtain suitable estimated values of the parameters of the model. Subsequently, numerical simulations were conducted to analyse the dynamics of not only the output but also the other states of the model that represent the other compounds of the CFFL.

The last part of the thesis work includes the structural analysis applied on the CFFL. These is comprised of tools that can be applied on bio-chemical reaction networks that offer insights about its static dynamics. First, the capacity for multiple equilibria was confirmed by using two different methods, one base don graph theory and the other on stoichiometry. In addition, a novel way of applying metabolic network analysis tool on gene expression networks was presented. First the reactions' structure had to be slightly changed in order to be able to implement elementary flux modes analysis and flux balance analysis on the network. Moreover, the results originating from the latter method confirm the accumulation of some specific compounds that were identified during the numerical simulations completed.

5-2 Future Work

In the following section opportunities for future work are presented, in order to extend and improve the modelling and analysis framework for biological models developed in this master thesis.

Firstly, the modelling methodology presented was derived based on the CFFL. Other types of feedforward loops have paths that are not activating but inhibiting, for example type 1 incoherent feedforward loop. In addition, these network motifs also are present in the composition of the combined feedforward loops. Therefore, the modelling tools have to be extended to include kinetics that represent inhibition.

An alternative way of improvement would be comprised of further model reduction. One way this could be achieved by observing the estimated parameter values to decide which rate constants are close to 0. As a result, these kinetic parameters are insignificant and the terms corresponding to them could be eliminated from the model. However, this has to be verified subsequently with further simulations.

The parameter estimation part of the project is significantly dependent of the availability of experimental data provided by ICMS. As a consequence, calibration was only conducted based on the batch experiment results. However, when successful flow experiments will be completed at Eindhoven University of Technology, then it will be possible to perform system identification on the flow model. This way it would be really interesting to observe the differences of the parameter values. Moreover, in order to have a more efficient parameter estimation process, the intermediate's protein output could be measured as well, in our case this would be σ-factor 28 (S28). In addition, a posteriori identifiability could be performed on the model with the estimated parameters in order to calculate confidence intervals of the estimated values [25]. Lastly, different optimization algorithms could be implemented in order to observe the difference in performance. In literature the successful application of optimization methods like simulated annealing and evolutionary algorithm could be identified. Implementing these methods on the CFFL could help in comparing the performance of different optimization algorithms applied on gene expression models.

Appendix A

Modelling

A-1 CFFL Biological Model

$$\begin{aligned}
&\mathrm{RNAP} + \mathrm{S}_{70} \xrightleftharpoons[\mathrm{k}_{-1}]{\mathrm{k}_1} \mathrm{RNAP}:\mathrm{S}_{70}\\
&\mathrm{RNAP}:\mathrm{S}_{70} + \mathrm{DNA}_t \xrightarrow{\mathrm{k}_2} \mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_t\\
&\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_t \xrightarrow{\mathrm{k}_3} \mathrm{DNA}_t + \mathrm{RNAP}:\mathrm{S}_{70} + \mathrm{RNA}_t\\
&\mathrm{RNAP}:\mathrm{S}_{70} + \mathrm{DNA}_{\mathrm{S28}} \xrightarrow{\mathrm{k}_4} \mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_{\mathrm{S28}}\\
&\mathrm{RNAP}:\mathrm{S}_{70}:\mathrm{DNA}_{\mathrm{S28}} \xrightarrow{\mathrm{k}_5} \mathrm{DNA}_t + \mathrm{RNAP}:\mathrm{S}_{70} + \mathrm{RNA}_{\mathrm{S28}}\\
&\mathrm{RNAP} + \mathrm{S}_{28} \xrightleftharpoons[\mathrm{k}_{-6}]{\mathrm{k}_6} \mathrm{RNAP}:\mathrm{S}_{28}\\
&\mathrm{RNAP}:\mathrm{S}_{28} + \mathrm{DNA}_{\mathrm{eGFP}} \xrightarrow{\mathrm{k}_7} \mathrm{RNAP}:\mathrm{S}_{28}:\mathrm{DNA}_{\mathrm{eGFP}}\\
&\mathrm{RNAP}:\mathrm{S}_{28}:\mathrm{DNA}_{\mathrm{eGFP}} \xrightarrow{\mathrm{k}_8} \mathrm{DNA}_{\mathrm{eGFP}} + \mathrm{RNAP}:\mathrm{S}_{28} + \mathrm{RNA}_{\mathrm{eGFP}}\\
&\mathrm{RNA}_t + \mathrm{RNA}_{\mathrm{S28}} \xrightleftharpoons[\mathrm{k}_{-9}]{\mathrm{k}_9} (\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}})\\
&(\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}}) + \mathrm{Ribo} \xrightarrow{\mathrm{k}_{10}} \mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}}:\mathrm{Ribo}\\
&\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}}:\mathrm{Ribo} \xrightarrow{\mathrm{k}_{11}} (\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{S28}}) + \mathrm{Ribo} + \mathrm{S}_{28}\\
&\mathrm{RNA}_t + \mathrm{RNA}_{\mathrm{eGFP}} \xrightleftharpoons[\mathrm{k}_{-12}]{\mathrm{k}_{12}} (\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}})\\
&(\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}}) + \mathrm{Ribo} \xrightarrow{\mathrm{k}_{13}} \mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}}:\mathrm{Ribo}\\
&\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}}:\mathrm{Ribo} \xrightarrow{\mathrm{k}_{14}} (\mathrm{RNA}_t:\mathrm{RNA}_{\mathrm{eGFP}}) + \mathrm{Ribo} + \mathrm{eGFP}_{\mathrm{dark}}\\
&\mathrm{eGFP}_{\mathrm{dark}} \xrightarrow{\mathrm{mat}} \mathrm{eGFP}
\end{aligned}$$

A-2 Stoichiometric Matrix of CFFL

$$S = \begin{pmatrix}
-1 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & -1 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1
\end{pmatrix}$$

A-3 Right and Left Null space of Stoichiometric Matrix

The right null-space from METATOOL:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1
\end{pmatrix}$$

The left null-space from METATOOL:

$$\begin{pmatrix}
0 & 0 & 0 & 1 & -1 & 0 \\
0 & -1 & 1 & -1 & 1 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 \\
1 & 1 & -1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

A-4 Mathematical Model - Flow Experiment

$$
\begin{aligned}
\dot{x_1} =& + \frac{k_{-1} RNAP_{tot}(x_2/K_1)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_1 x_1 x_2 + \\
& + \frac{k_{-6} RNAP_{tot}(x_{11}/K_6)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_6 x_1 x_{11} - dil \cdot x_1 + dil \cdot RNAP_0 \\
\dot{x_2} =& \frac{k_{-1} RNAP_{tot}(x_2/K_1)}{1+(1+x_4/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_1 x_1 x_2 - dil \cdot x_2 + dil \cdot S70_0 \\
\dot{x_3} =& - dil \cdot x_3 + dil \cdot c \cdot input \\
\dot{x_4} =& - dil \cdot x_4 + dil \cdot DNA_{S28_0} \\
\dot{x_5} =& - dil \cdot x_5 + dil \cdot DNA_{eGFP_0} \\
\dot{x_6} =& \frac{k_3 RNAP_{tot}(x_2/K_1)([DNA_t]/K_{e1})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_9 x_6 x_7 - k_{12} x_6 x_8 + k_{-9} x_9 + k_{-12} x_{10} - deg \cdot x_6 - dil \cdot x_6 \\
\dot{x_7} =& \frac{k_5 RNAP_{tot}(x_2/K_1)(x_4/K_{e_2})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_9 x_6 x_7 + k_{-9} x_9 - deg \cdot x_7 - dil \cdot x_7 \\
\dot{x_8} =& \frac{k_8 RNAP_{tot}(x_{11}/K_6)(x_5/K_{e3})}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - k_{12} x_6 x_8 + k_{-12} x_{10} - deg \cdot x_8 - dil \cdot x_8 \\
\dot{x_9} =& k_9 x_6 x_7 - k_{-9} x_9 - k_{10} x_9 x_{14} + \frac{k_{11} R_{tot}(x_9/K_{TL1})}{1+(x_9/K_{TL1})+(x_{10}/K_{TL2})} - deg \cdot x_9 - dil \cdot x_9 \\
\dot{x_{10}} =& k_{12} x_6 x_8 - k_{-12} x_{10} - k_{13} x_{10} x_{14} + \frac{k_{14} R_{tot}(x_{10}/K_{TL2})}{1+(x_9/K_{TL1})+(x_{10}/K_{TL2})} - deg \cdot x_{10} - dil \cdot x_{10} \\
\dot{x_{11}} =& \frac{k_{-6} RNAP_{tot}(x_{11}/K_6)}{1+(1+x_5/K_{e3})(x_{11}/K_6)+(1+x_3/K_{e1}+x_4/K_{e2})(x_2/K_1)} - \\
& - k_6 x_1 x_{11} + \frac{k_{11} R_{tot}(x_9/K_{TL1})}{1+(x_9/K_{TL1})+(x_{10}/K_{TL2})} - dil \cdot x_{11} \\
\dot{x_{12}} =& \frac{k_{14} R_{tot}(x_{10}/K_{TL2})}{1+(x_9/K_{TL1})+(x_{10}/K_{TL2})} - mat \cdot x_{12} - dil \cdot x_{12} \\
\dot{x_{13}} =& mat \cdot x_{12} - dil \cdot x_{13} \\
\dot{x_{14}} =& -dil \cdot x_{14} + dil \cdot Rib_0
\end{aligned}
$$

Appendix B

Parameter Estimation

B-1 Limits used during Parameter Estimation

Parameters	Lower Limit	Upper Limit
k1	0	250
k_1	0	150
Ke2	50	200
k5	0	3.5
Ke1	50	200
k3	0	3.5
k6	0	150
k_6	0	150
Ke3	0	500
k8	0	3.5
k9	0	250
k_9	0	150
k10	0	3.5
k11	0	1000
k12	0	100
k_12	0	100
k13	0	3.5
k14	0	1000
mat	0	1
deg6	0	1
deg7	0	1
deg8	0	1
deg9	0	1
deg10	0	1

B-2 Results from PSO on the individual datasets

		fmincon dataset 1	fmincon dataset 2	fmincon dataset 3	fmincon dataset 4	fmincon dataset 5
1	k1	2.377091663	56.89062535	93.42564825	69.90039289	60.68751711
2	k_1	0.030179715	42.65073702	32.52319288	42.32026335	42.17442705
3	Ke2	482.1283974	9243.403956	4564.749632	6296.855397	866.0741366
4	k5	3.449507318	3.078333711	3.182228945	3.255340061	3.373299644
5	Ke1	1.054560743	0.833691887	0.420900878	5.089642913	4.253037114
6	k3	3.172677645	3.113154326	3.439839393	3.390549044	3.376080365
7	k6	57.16810011	39.25227358	31.26867937	1.902723311	3.00474019
8	k_6	28.8452723	36.52567982	14.98584182	7.087545522	22.0136488
9	Ke3	542.885534	44.32291526	79.78831746	40.55400473	117.6732338
10	k8	3.495229516	3.080363819	3.481440969	3.445790022	3.484116422
11	k9	55.74351265	55.95171746	62.99053733	43.46678093	49.6020874
12	k_9	46.82445467	38.3350346	38.86385426	26.04973712	6.204767939
13	k10	3.09E12	2.74E12	1.03E12	3.27E12	2.85E12
14	k11	1.88202834	0.342970599	0.042076435	0.81431565	2.99E13
15	k12	3.46E15	3.06E15	2.69E15	9.94E14	3.71E15
16	k_12	36.28785705	38.82960206	27.40308225	32.61375626	25.08359961
17	k13	5.17E-01	1.58E-01	2.37E-01	1.12E16	3.62E-01
18	k14	3.495638009	3.15198976	3.485886002	3.46704486	3.487017788
19	mat	0.996033308	0.877163322	0.994321573	0.988798715	0.995921691
20	deg6	0.00176538	0.108776194	0.006118213	0.030302341	0.013127707
21	deg7	0.014120863	0.059759212	0.014597562	0.025808642	0.012325379
22	deg8	0.92884704	0.761815024	0.927130383	0.898126413	0.951347827
23	deg9	0.44293286	0.481031014	0.310286019	0.3978601	0.20552089
24	deg10	0.060172346	0.694276437	0.155585861	0.36355236	0.726064799

B-3 Final Estimated parameters

k1	k_1	Ke2	k5	Ke1	k3	k6	k_6	Ke3	k8	k9
249.995	1.88	51.264	3.4445	55.7946	3.1441	0.0390	28.837	296.2375	3.499	72.3162

k_9	k10	k11	k12	k_12	k13	k14	mat	deg6	deg7	deg8
149.94	0.361	49.944	99.951	25.414	3.4906	494.256	0.999	0.997	0.0477	0.997

deg9	deg10
0.0037	2.86E-5

Appendix C

Structural Analysis

C-1 The identified EFMs from METATOOL

$$\begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}$$

C-2 FBA Results - Single Objective Function

Reactions	max eGFP	max RNA_t	max RNA_{S28}	max RNA_{eGFP}	max S28
R1	1.0000E0	1.0000E0	1.0000E0	1.0000E0	1.0000E0
R2	6.9094E-01	1.0000E0	1.5934E-01	6.1619E-01	5.5547E-01
R3	6.6666E-01	9.9999E-01	7.8966E-06	5.0000E-01	5.0000E-01
R4	3.8977E-01	1.3761E-01	1.0000E3	6.1901E-01	5.5750E-01
R5	3.3334E-01	7.4484E-06	9.9999E-01	5.0000E-01	5.0000E-01
R6	3.3333E-01	4.6169E-06	3.8878E-06	4.9998E-01	4.9999E-01
R7	3.9204E-01	1.3694E-01	1.5779E-01	6.2347E-01	5.5951E-01
R8	3.3333E-01	4.6174E-06	3.8883E-06	4.9998E-01	4.9999E-01
R9	3.3333E-01	3.4005E-06	2.9500E-06	4.9998E-01	7.3422E-06
R10	3.9334E-01	1.3673E-01	1.5751E-01	6.2576E-01	1.9303E-01
R11	3.3333E-01	3.4011E-06	2.9505E-06	4.9998E-01	7.3426E-06
R12	3.3333E-01	2.0799E-06	1.9760E-06	8.4204E-06	4.8481E-06
R13	3.9410E-01	1.3652E-01	1.5723E-01	2.8825E-01	1.9270E-01
R14	3.3333E-01	2.0806E-06	1.9766E-06	8.4211E-06	4.8487E-06
R15	3.3333E-01	2.0809E-06	1.9769E-06	8.4215E-06	4.8489E-06
R16	2.4275E-02	7.5993E-06	1.5933E-01	1.1619E-01	5.5467E-02
R17	5.6436E-02	1.3760E-01	7.0846E-06	1.1902E-01	5.7498E-02
R18	5.8713E-02	1.3694E-01	1.5779E-01	1.2348E-01	5.9521E-02
R19	6.0012E-02	1.3673E-01	1.5751E-01	1.2578E-01	1.9302E-01
R20	6.0775E-02	1.3652E-01	1.5723E-01	2.8825E-01	1.9270E-01
R21	-1.3333	-1.0000	-1.0000	-1.5000	-1.0000
R22	-1.0000	-1.0000	-1.0000	-1.0000	-1.0000
R23	6.4783E-06	9.9999E-01	2.0330E-06	1.1229E-05	4.5883E-06
R24	6.4802E-06	2.8316E-06	9.9999E-01	1.1231E-05	4.5897E-06
R25	2.1600E-06	1.3215E-06	9.7484E-07	4.9997E-01	2.4947E-06
R26	3.2382E-06	1.2171E-06	9.3862E-07	5.6123E-06	4.9999E-01
R27	3.3333E-01	2.0811E-06	1.9770E-06	8.4216E-06	4.8490E-06

Table C-1: FBA Results - Single Objective Function

C-3 Single Objective Function Vectors

$$c_{eGFP} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad c_{RNA_t} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad c_{RNA_{S28}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad c_{RNA_{eGFP}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad c_{S28} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

C-4 FBA Results - Complex Objective Function

Reactions	max RNA_t	max RNA_{S28}	max RNA_{eGFP}	max S28
R1	1	1	1	1
R2	1.0000E1	1.3083E-01	6.8901E-01	6.8390E-01
R3	9.9999E-01	3.7616E-05	6.6666E-01	6.5791E-01
R4	1.0316E-01	9.9999E-01	3.8604E-01	4.0125E-01
R5	1.4958E-05	9.9996E-01	3.3334E-01	3.4209E-01
R6	1.1741E-05	1.8528E-05	3.3333E-01	3.4208E-01
R7	1.0268E-01	1.2907E-01	3.8810E-01	4.0333E-01
R8	1.1741E-05	1.8528E-05	3.3333E-01	3.4208E-01
R9	1.0324E-05	1.6551E-05	3.3333E-01	3.1583E-01
R10	1.0251E-01	1.2886E-01	3.8927E-01	3.7851E-01
R11	1.0324E-05	1.6551E-05	3.3333E-01	3.1583E-01
R12	9.1453E-06	1.4944E-05	3.3332E-01	3.1583E-01
R13	1.0235E-01	1.2867E-01	3.9060E-01	3.7918E-01
R14	9.1458E-06	1.4945E-05	3.3332E-01	3.1583E-01
R15	9.1461E-06	1.4945E-05	3.3332E-01	3.1583E-01
R16	1.4706E-05	1.3079E-01	2.2347E-02	2.5992E-02
R17	1.0315E-01	2.9800E-05	5.2701E-02	5.9164E-02
R18	1.0266E-01	1.2905E-01	5.4773E-02	6.1254E-02
R19	1.0250E-01	1.2885E-01	5.5942E-02	6.2683E-02
R20	1.0234E-01	1.2866E-01	5.7276E-02	6.3350E-02
R21	-1.0000	-1.0000	-1.3333	-1.3158
R22	-1.0000	-1.0000	-1.0000	-1.0000
R23	9.9996E-01	4.1432E-06	5.8029E-06	6.8783E-06
R24	3.2171E-06	9.9994E-01	5.8049E-06	6.8807E-06
R25	1.1791E-06	1.6068E-06	5.8098E-06	2.2932E-06
R26	1.4173E-06	1.9777E-06	2.9007E-06	2.6251E-02
R27	9.1464E-06	1.4946E-05	3.3332E-01	3.1583E-01

Table C-2: FBA Results - Complex Objective Function

C-5 Complex Objective Function Vectors

$$c_{RNA_t} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad c_{RNA_{S28}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad c_{RNA_{eGFP}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 1 \end{bmatrix} \quad c_{S28} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.5 \\ 1 \end{bmatrix}$$

Bibliography

[1] U. Alon, *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2007.

[2] M. Yelleswarapu, A. J. van der Linden, B. van Sluijs, P. A. Pieters, E. Dubuc, T. F. De Greef, and W. T. Huck, "Sigma Factor-Mediated Tuning of Bacterial Cell-Free Synthetic Genetic Oscillators," *ACS Synthetic Biology*, vol. 7, no. 12, pp. 2879–2887, 2018.

[3] D. Del Vecchio and R. M. Murray, *Biomolecular Feedback Systems*. Princeton University Press, 2015.

[4] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nature Publishing Group*, vol. 28, no. 3, pp. 245–248, 2010.

[5] T. E. Gorochowski, C. S. Grierson, and M. di Bernardo, "Organization of feed-forward loop motifs reveals architectural principles in natural and engineered networks," *Science Advances*, vol. 4, no. 3, 2018.

[6] H. V. Westerhoff and B. O. Palsson, "The evolution of molecular biology into systems biology," *Nature Biotechnology*, vol. 22, no. 10, pp. 1249–1252, 2004.

[7] A. M. Turing, "The chemical basis of morphogenesis," *Bulletin of Mathematical Biology*, vol. 52, no. 1-2, pp. 153–197, 1990.

[8] M. Drack, "Ludwig von Bertalanffy's Early System Approach," *Systems Research and Behavioral Science*, vol. 573, no. 3, pp. 549–573, 2008.

[9] S. Kesić, "Systems biology, emergence and antireductionism," *Saudi Journal of Biological Sciences*, vol. 23, no. 5, pp. 584–591, 2016.

[10] F. Jacob and J. Monod, “Genetic regulatory mechanisms in the synthesis of proteins,” *Journal of Molecular Biology*, vol. 3, no. 3, pp. 318–356, 1961.

[11] D. Del Vecchio, Y. Qian, R. M. Murray, and E. D. Sontag, “Future systems and control research in synthetic biology,” *Annual Reviews in Control*, vol. 45, pp. 5–17, 2018.

[12] T. Ideker, T. Galitski, and L. Hood, “A new approach to decoding life: Systems Biology,” *Annual Review of Genomics and Human Genetics*, vol. 2, pp. 343–372, 2001.

[13] M. Elowitz and S. Leibler, “A synthetic oscillatory network repressilator,” *Nature*, vol. 403, pp. 335–338, 2000.

[14] C. R. Cantor, J. J. Collins, and T. S. Gardner, “Construction of a genetic toggle switch in Escherichia coli,” *Nature*, vol. 403, no. 6767, pp. 339–342, 2000.

[15] D. Liu, A. Hoynes-O’Connor, and F. Zhang, “Bridging the gap between systems biology and synthetic biology,” *Frontiers in Microbiology*, vol. 4, pp. 1–8, 2013.

[16] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, “Network motifs in the transcriptional regulation network of Escherichia coli,” *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[17] T. Lee, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, and R. Young, “Transcriptional Regulatory Networks in Saccharomyces cerevisiae,” *Science*, vol. 298, pp. 799–804, 2002.

[18] L. Roselius, D. Langemann, J. Müller, B. A. Hense, S. Filges, D. Jahn, and R. Münch, “Modelling and analysis of a gene-regulatory feed-forward loop with basal expression of the second regulator,” *Journal of Theoretical Biology*, vol. 363, pp. 290–299, 2014.

[19] J. Smeu, “Input and Output Dynamics Analysis of Complex Feedforward Loops -Literature Study Report,” tech. rep., Delft University of Technology, Delft, 2019.

[20] F. Mavelli, R. Marangoni, and P. Stano, “A Simple Protein Synthesis Model for the PURE System Operation,” *Bulletin of Mathematical Biology*, vol. 77, no. 6, pp. 1185–1212, 2015.

[21] A. von Kamp and S. Schuster, “Metatool 5.0: fast and flexible elementary modes analysis,” *Bioinformatics*, vol. 22, no. 15, pp. 1930–1931, 2006.

[22] T. Stögbauer, *Experiment and quantitative modeling of cell-free gene expression dynamics*. PhD thesis, Ludwig–Maximilians–Universitat, 2012.

[23] E. Karzbrun, J. Shin, R. H. Bar-Ziv, and V. Noireaux, "Coarse-grained dynamics of protein synthesis in a cell-free system," *Physical Review Letters*, vol. 106, no. 4, pp. 1–4, 2011.

[24] H. Niederholtmeyer, V. Stepanova, and S. J. Maerkl, "Implementation of cell-free biological networks at steady state," *Proceedings of the National Academy of Sciences*, vol. 110, no. 40, pp. 15985–15990, 2013.

[25] R. Muñoz-Tamayo, L. Puillet, J. B. Daniel, D. Sauvant, O. Martin, M. Taghipoor, and P. Blavy, "Review: To be or not to be an identifiable model. Is this a relevant question in animal science modelling?," *Animal*, vol. 12, no. 4, pp. 701–712, 2018.

[26] E. Walter and L. Pronzato, *Identification of Parametric Models from Experimental Data.* Springer, 1997.

[27] G. Bellu, M. P. Saccomani, S. Audoly, and L. D'Angiò, "DAISY: A new software tool to test global identifiability of biological and physiological systems," *Computer Methods and Programs in Biomedicine*, vol. 88, no. 1, pp. 52–61, 2007.

[28] O.-T. Chis, J. R. Banga, and E. Balsa-Canto, "Sloppy models can be identifiable," *arXiv e-prints*, 2014.

[29] J. Karlsson, M. Anguelova, and M. Jirstrand, "An Efficient Method for Structural Identifiability Analysis of Large Dynamic Systems," *IFAC Proceedings Volumes*, vol. 45, no. 16, pp. 941–946, 2012.

[30] A. Sedoglavic, "A Probabilistic Algorithm to Test Local Algebraic Observability in Polynomial Time," *Journal of Symbolic Computation*, vol. 33, no. 5, pp. 735–755, 2002.

[31] F. Lei and S. B. Jorgensen, "Estimation of kinetic parameters in a structured yeast model using regularisation," *Journal of Biotechnology*, vol. 88, no. 3, pp. 223–237, 2001.

[32] G. Wozny, S. Ochoa, D. Yang, J.-U. Repke, and A. Yoo, "Modeling and Parameter Identification of the Simultaneous Saccharification-Fermentation Process for Ethanol Production," *Biotechnology Progress*, vol. 23, no. 6, pp. 1454–1462, 2007.

[33] C. E. Robles Rodriguez, *Modeling and optimization of the production of lipids by oleaginous yeasts.* PhD thesis, Université Fédérale Toulouse Midi-Pyrénées, 2016.

[34] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom, "Systems biology : parameter estimation for biochemical models," *FEBS Journal*, vol. 276, pp. 886–902, 2009.

[35] S. Katare, A. Bhan, J. M. Caruthers, and W. N. Delgass, "A hybrid genetic algorithm for efficient parameter estimation of large kinetic models," *Computers and Chemical Engineering*, vol. 28, pp. 2569–2581, 2004.

[36] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995.

[37] J. Robinson and Y. Rahmat-Samii, "Particle Swarm Optimization in Electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 2, pp. 397–407, 2004.

[38] S. N. Gaul, "Optimization using Particle Swarm," 2013.

[39] J. W. Chinneck, "Pattern Search for Unconstrained NLP," in *Practical Optimization: A Gentle Introduction*, ch. 17, 2015.

[40] R. M. Lewis and V. Torczon, "Patterns Search Algorithms for Bound Constrained Minimization," *SIAM Journal on Optimization*, vol. 9, no. 4, pp. 1082–1099, 1999.

[41] P. J. Angeline, "Evolutionary Optimization Versus Particle Swarm Optimization : Philosophy and Performance Differences," in *Evolutionary Programming VII* (V. W. Porto, , N. Saravanan, , D. Waagen, , and A. E. Eiben, eds.), pp. 601–610, Springer, 1998.

[42] F. Horn, "Necessary and Sufficient Conditions for Complex Balancing in Chemical Kinetics," *Archive for Rational Mechanics and Analysis*, vol. 49, no. 3, pp. 172–186, 1972.

[43] M. Feinberg, "Complex Balancing in General Kinetic Systems," *Archive for Rational Mechanics and Analysis*, vol. 49, no. 3, pp. 187–194, 1972.

[44] M. Banaji, "Cycle structure in SR and DSR graphs: implications for multiple equilibria and stable oscillation in chemical reaction networks," in *Transactions on Petri Nets and Other Models of Concurrency V* (K. Jensen, , S. Donatelli, , and J. Kleijn, eds.), pp. 1–21, Springer Berlin Heidelberg, 2012.

[45] M. Banaji, P. Donnell, and S. Baigent, "P matrix properties, injectivity and stability in chemical reaction systems," *SIAM Journal on Applied Mathematics*, vol. 67, no. 6, pp. 1523–1547, 2007.

[46] P. Donnell, M. Banaji, A. Marginean, and C. Pantea, "CoNtRol : an open source framework for the analysis of chemical reaction networks," *Bioinformatics*, vol. 30, no. 11, pp. 1633–1634, 2014.

[47] N. Soranzo and C. Altafini, "ERNEST: a toolbox for chemical reaction network theory," *Bioinformatics*, vol. 25, no. 21, pp. 2853–2854, 2009.

[48] P. Ellison, M. Feinberg, H. Ji, and D. Knight, "The Chemical Reaction Network Toolbox, Version 2.3," 2014.

[49] G. Shinar and M. Feinberg, "Concordant Chemical Reaction Networks," *Mathematical Biosciences*, vol. 240, no. 2, pp. 92–113, 2018.

[50] M. Banaji and G. Craciun, "Graph-theoretic approaches to injectivity and multiple equilibria in systems of interacting elements," *Communications in Mathematical Sciences*, vol. 7, no. 4, pp. 867–900, 2009.

[51] R. B. Kellogg, "On Complex Eigenvalues of M and P Matrices," *Numerische Mathematik*, vol. 19, pp. 170–175, 1972.

[52] S. Ranganathan, P. F. Suthers, and C. D. Maranas, "OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions," *PLOS Computational Biology*, vol. 6, no. 4, 2010.

[53] S. Schuster, C. Hilgetag, J. Woods, and D. Fell, "Reaction routes in biochemical reaction systems: Algebraic properties, validated calculation procedure and example from nucleotide metabolism," *Journal of Mathematical Biology*, vol. 45, pp. 153–181, 2002.

[54] R. T. Rockafellar, "Convex analysis," in *Princeton Landmarks in Mathematics and Physics*, Princeton University Press, 1996.

[55] L. Wang, S. Dash, C. Y. Ng, and C. D. Maranas, "A review of computational tools for design and reconstruction of metabolic pathways," *Synthetic and Systems Biotechnology*, vol. 2, no. 4, pp. 243–252, 2017.

[56] K. Raman and N. Chandra, "Flux balance analysis of biological systems: applications and challenges," *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 435–449, 2009.

[57] L. Heirendt, S. Arreckx, T. Pfau, and S. N. Mendoza, "Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0," *Nature Protocol*, vol. 14, pp. 639–702, 2019.

Glossary

List of Acronyms

CFFL	coherent feedforward loop
DSR	directed species-reaction
E.Coli	Escherichia coli
EFM	elementary flux modes
eGFP	green fluorescent protein
FBA	flux balance analysis
ICMS	Institute of Complex Molecular Systems
MSE	Mean of the Squared Errors
ODE	ordinary differential equation
PSO	Particle Swarm Optimization
S28	σ-factor 28
S70	σ-factor 70
SR	species-reaction
SSD	strongly-sign-determined