DOI:10.15774/PPKE.ITK.2015.008

Pázmány Péter Catholic University

Doctoral Thesis

Structural Analysis of Kinetic Systems with Application to Cell-free Expression Systems

Author:
Zoltán András Tuza

Thesis Advisor:
Dr. Gábor Szederkényi, D.Sc.

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Roska Tamás Doctoral School of Sciences and Technology
Faculty of Information Technology and Bionics

August 3, 2015

DOI:10.15774/PPKE.ITK.2015.008

DOI:10.15774/PPKE.ITK.2015.008

Abstract

Dynamical models play an important role in many fields of science and engineering. The aim of applying these models is to solve real-life problems, using that they are able to reproduce the important observed phenomena as accurately as required. Nonnegative systems form a special class of dynamical systems where all the state variables remain in the positive orthant, if the states start there. Kinetic dynamical models originate from chemistry as descriptors of chemical processes, but their range of applicability reaches far beyond (bio)chemical models as they are suitable to describe all important dynamical phenomena.

It is possible to associate a directed graph structure to a kinetic system which enables us to investigate not only the graph theoretic properties, but the dynamical properties of the kinetic system as well. The theory to investigate these properties has existed for decades but a set of optimization based approaches—exploiting that the graph structure corresponding to a given kinetic dynamics is non-unique—have been developed relatively recently. Building upon these, this thesis presents two new algorithms utilizing mathematical optimization.

It has been known that the sparse directed graph structure is not necessarily unique which is in contrast with the unique dense structure. This non-uniqueness may hamper the successful identification of a kinetic system because a unique sparse structure is often implicitly assumed. The first algorithm shows an efficient way to calculate all sparse directed graph structures of a kinetic system.

The second algorithm is developed in the newly introduced class of uncertain kinetic systems. This procedure computes the core reactions of the uncertain kinetic system and it has polynomial time complexity. The uncertainty is represented in the form of parameter intervals which makes it suitable to accommodate uncertainties ranging from temperature change to different operating regimes.

The applicability of the algorithms is illustrated by a series of examples. First, the well-known Lorenz system is transformed into kinetic form to calculate all the sparse structures. Second, a network reconstruction benchmark is used to show the computation of the core reaction set of an uncertain kinetic system. Third, it is shown that the sparse structure of a kinetic system with predetermined uncertainty may be non-unique, too.

The last part of this thesis is focusing on the modeling process of an *in vitro* system. The work starts with a list of molecular laboratory protocols to prepare the different molecular probes. Then, a series of experiments is designed to collect data about the *in vitro* system. A first principle model is developed to capture the transient behavior of the gene expression and it is shown that the model is structurally identifiable using the applied measurement setup. Finally, the quality of the time series data enables us to estimate and validate the parameters of the developed kinetic model.

DOI:10.15774/PPKE.ITK.2015.008

Acknowledgements

First of all, I would like to thank my thesis advisor Gábor Szederkényi for his guidance and endless support. I'm also grateful for Katalin Hangos who advised me during Gábor's sabbatical.

Learning and doing engineering is always a teamwork, I was and still I am really lucky to have János Rudan as my colleague.

Bernadett Ács started her PhD in my last year of grade school, but her suggestions and meticulous work made a huge impact on my Thesis. I am really thankful for her countless hours of review on this thesis.

It was also great to have so many wonderful officemates over the years who made our graduate student office a welcoming environment. Therefore, I would like to thank Dóra, Bence, Zsolt, András, Balazs, Csaba, Istvan, Norbert, Antal and Endre for that.

During my time at SZTAKI, I worked with Attila Gábor and Ralf Hannemann. I really enjoyed our talks and I thank them for all the support and feedback they gave.

In 2012, I had the honor to spend 13 months at California Institute of Technology as a Fulbright fellow. During that time, I worked with Richard M. Murray. I am really grateful for that time for Richard, he welcomed me as one of his graduate students and guided my work throughout the year. Also I was lucky enough to work with many wonderful people at Caltech. I'm personally in huge debt to Jongmin Kim and Dan Siegel, they spent many hours to guide me. I'm also really glad to work with Vipul Sighal, I really miss endless hours of debate and planning in front of the white board. Finally, I'm really grateful for rest of the biocircuit group: Anandh, Anu, Chris, Clare, Emzo, Enoch, Joe, Marcella, Shaobin, Victoria, Yong and Zach .

All of these would not have been possible without my family who supported me in many ways throughout my PhD studies.

Finally, I would like to acknowledge the financial support of the following grants: TÁMOP-4.2.1.B-11/2/KMR-2011-0002 and TÁMOP-4.2.2/B-10/1-2010-0014, the DARPA Living Foundries Program under Contract HR0011-12-C-0065, OTKA NF 104706, KAP-1.1-14/029 and the Fulbright program.

DOI:10.15774/PPKE.ITK.2015.008

Contents

Abstract ii

Acknowledgements iii

Contents iv

List of Figures vii

List of Tables x

Abbreviations xi

Notations and Symbols xii

1 Introduction 1
- 1.1 The notion and significance of dynamical and nonnegative models 1
- 1.2 Kinetic systems 2
- 1.3 Challenges of structure and parameter estimation of kinetic systems . . . 6
- 1.4 Structure and objectives of the thesis 8

2 Background and Basic Notations 10
- 2.1 Convex Optimization 10
  - 2.1.1 Linear Programming 11
  - 2.1.2 Mixed Integer Linear Programming 13
  - 2.1.3 Least Squares Optimization 14
- 2.2 Introduction to modeling of Kinetic Systems 15
  - 2.2.1 ODE description 16
  - 2.2.2 Directed graph structure 17
  - 2.2.3 Assigning the canonical CRN to a kinetic system 20
  - 2.2.4 Dynamical equivalence of kinetic systems 21
- 2.3 Known optimization methods for computing certain CRN realizations . . 22
  - 2.3.1 Computing Sparse and Dense Realizations 22
  - 2.3.2 Computing Constrained Realizations 24
  - 2.3.3 Computing Core Reactions and Core Complexes 25
  - 2.3.4 Introductory example 26
- 2.4 Estimating Parameters of Kinetic Systems 28
- 2.5 Structural Identifiability Analysis of Mathematical Models 31

DOI:10.15774/PPKE.ITK.2015.008

3 Computing All Dynamically Equivalent Sparse Chemical Reaction Network Structures **34**
3.1 Computation of all possible sparse structures . . . 35
3.2 Transformation of polynomial models into kinetic form . . . 37
3.2.1 State-dependent time-rescaling . . . 39
3.2.2 X-factorable transformation . . . 40
3.3 Results: sparse kinetic realizations of the Lorenz system . . . 41
3.3.1 State-dependent time-rescaling . . . 42
3.3.2 X-factorable transformation . . . 45
3.4 Summary . . . 48

4 Computing Structural Properties of Uncertain Kinetic Polynomial Systems **52**
4.1 Uncertain polynomial kinetic systems . . . 53
4.2 Computing Core reactions of Uncertain Polynomial Kinetic systems . . . 54
4.2.1 Algorithm for computing core reactions . . . 54
4.3 Examples . . . 57
4.3.1 Example of an Uncertain Kinetic System . . . 57
4.3.2 Network reconstruction example . . . 58
4.3.2.1 Parameter Estimation Procedure . . . 60
4.3.2.2 Estimation of matrix A_k . . . 61
4.3.2.3 Results . . . 62
4.4 Computing Sparse and Dense Realizations of Uncertain Kinetic Systems . . . 64
4.5 Summary . . . 67

5 Modeling and Parameter Estimation of a Cell-free *in vitro* System **68**
5.1 Experimental background . . . 70
5.1.1 Cell-free system . . . 70
5.1.2 Measurements . . . 72
5.2 Process Model . . . 74
5.2.1 Initial modeling steps . . . 75
5.2.2 Transcription . . . 76
5.2.3 Translation . . . 77
5.2.4 Resource degradation . . . 78
5.2.5 State-space model . . . 79
5.2.6 Measured outputs . . . 81
5.3 Analysis of the Process Model . . . 82
5.3.1 mRNA dynamics . . . 82
5.3.2 Steady state assumption . . . 82
5.3.3 Structural Identifiability . . . 83
5.4 Parameter Estimation . . . 84
5.4.1 Prediction error minimization . . . 85
5.4.2 Statistical Analysis of the Parameter Estimation . . . 87
5.5 Summary . . . 89

6 Conclusions **91**
6.1 New scientific results . . . 92
6.2 Possible directions of future work . . . 96

DOI:10.15774/PPKE.ITK.2015.008

A **97**
A.1 Special measurements 97
A.2 TXTL Software toolbox 99
A.2.1 Implementation considerations 99

References **100**

DOI:10.15774/PPKE.ITK.2015.008

List of Figures

2.1 Two dynamically equivalent realizations of the positive feedback motif. The blue edges represent the core reactions which are structurally invariant under dynamical equivalence. The black edges are the non-core reactions. The reader can notice that the graph on panel (a) is a subgraph of the graph on panel (b), this property was proven in [61]. 28

3.1 Phase-plane plot of the kinetic Lorenz system obtained via state-dependent time-rescaling. 39
3.2 Phase-plane plot of the Kinetic Lorenz system obtained using the X-factorable transformation. The inset shows the local behavior of the Lorenz attractor. 41
3.3 The canonical realization of the Lorenz system with state-dependent time-scaling. In this case, the canonical realization is also a sparse one. The parameters are: $k_{3,1} = \sigma$, $k_{4,1} = 1$, $k_{7,1} = \beta$, $k_{1,2} = \sigma$, $k_{3,5} = \rho + w_3$, $k_{4,5} = \sigma$, $k_{1,6} = |w_2 - w_1\rho - w_1w_3|$, $k_{1,7} = w_1w_2 + \beta w_3$, $k_{7,8} = w_1$, $k_{9,10} = \sigma$, $k_{3,11} = w_2$, $k_{4,12} = w_1$, $k_{5,13} = \sigma$. 43
3.4 Sparse realization of the model in Equation (3.11) that contains only the core complexes. The reaction rate coefficients are: $k_{1,2} = 1882.7$, $k_{7,4} = 2.6667$, $k_{4,7} = 1$, $k_{7,8} = 21.333$, $k_{1,10} = 1271$, $k_{9,10} = 1$, $k_{1,11} = 601.67$, $k_{3,11} = 59$, $k_{3,12} = 10$, $k_{4,12} = 35$, $k_{3,13} = 44$, $k_{4,13} = 10$, $k_{5,13} = 1$. 44
3.5 Canonical realization of the Lorenz system transformed with X-factorable transformation. This realization is sparse as well. The values of the rate coefficients are: $k_{1,2} = \sigma$, $k_{3,1} = \sigma$, $k_{4,7} = \rho + w_3$, $k_{4,5} = \sigma$, $k_{6,2} = w_1\rho + (w_1w_3 - w_2)$, $k_{7,6} = 1$, $k_{9,7} = 1$, $k_{11,15} = 1$, $k_{11,12} = 1$, $k_{12,1} = w_2$, $k_{13,14} = w_1w_2 + \beta w_3$, $k_{14,13} = \beta$. 46
3.6 One of the 48 possible sparse realizations of Equation (3.14) that contains 13 complexes. The parameters are: $k_{12,1} = 101$, $k_{1,2} = 10$, $k_{3,2} = 5$, $k_{6,2} = 2799$, $k_{4,3} = 10$, $k_{7,6} = 1$, $k_{9,7} = 100$, $k_{4,8} = 39$, $k_{11,12} = 1$, $k_{14,13} = 2.6667$, $k_{13,14} = 10103$, $k_{11,15} = 1$. 47
3.7 Sparse realization of the model in Equation (3.11) with 3 linkage classes. The reaction rate coefficients are: $k_{7,1} = 2.6667$, $k_{1,2} = 10$, $k_{4,3} = 1$, $k_{3,4} = 10$, $k_{1,6} = 1271$, $k_{1,7} = 699.33$, $k_{7,8} = 24$, $k_{9,10} = 1$, $k_{3,11} = 69$, $k_{4,12} = 33$, $k_{3,13} = 44$, $k_{4,13} = 9$, $k_{5,13} = 1$. 48

DOI:10.15774/PPKE.ITK.2015.008

3.8 Dense realization of Equation (3.11) containing 51 reactions. The reaction rate coefficients are $k_{1,2} = 679.63, k_{1,3} = 0.1, k_{1,4} = 0.1, k_{1,5} = 0.1, k_{1,6} = 602.37, k_{1,7} = 0.1, k_{1,8} = 0.1, k_{1,9} = 0.1, k_{1,10} = 669.13, k_{1,11} = 0.1, k_{1,12} = 0.1, k_{1,13} = 0.1, k_{3,1} = 0.1, k_{3,2} = 0.1, k_{3,4} = 0.1, k_{3,5} = 44.6, k_{3,6} = 0.1, k_{3,7} = 0.1, k_{3,8} = 0.1, k_{3,9} = 0.1, k_{3,10} = 0.1, k_{3,11} = 16.2, k_{3,12} = 9.3, k_{3,13} = 0.1, k_{4,1} = 0.1, k_{4,2} = 0.1, k_{4,3} = 0.1, k_{4,5} = 9.6, k_{4,6} = 0.1, k_{4,7} = 0.1, k_{4,8} = 0.1, k_{4,9} = 0.1, k_{4,10} = 0.1, k_{4,11} = 0.1, k_{4,12} = 24.4, k_{4,13} = 0.1, k_{5,13} = 1, k_{7,1} = 0.1, k_{7,2} = 1.1833, k_{7,3} = 0.1, k_{7,4} = 0.1, k_{7,5} = 0.68335, k_{7,6} = 0.1, k_{7,8} = 23.217, k_{7,9} = 0.1, k_{7,10} = 0.1, k_{7,11} = 0.1, k_{7,12} = 0.1, k_{7,13} = 0.1, k_{9,10} = 1.1, k_{9,13} = 0.1.$ 50

3.9 Dense realization of Equation (3.14) with 44 reactions. The parameter values are the following: $k_{1,2} = 10.1, k_{1,3} = 0.1, k_{3,1} = 0.1, k_{3,2} = 4.95, k_{4,1} = 0.1, k_{4,2} = 0.1, k_{4,3} = 0.1, k_{4,5} = 10.2, k_{4,6} = 0.1, k_{4,7} = 0.1, k_{4,8} = 29.2, k_{6,2} = 2799.1, k_{6,7} = 0.1, k_{7,2} = 0.45, k_{7,6} = 0.1, k_{9,2} = 0.1, k_{9,6} = 0.1, k_{9,7} = 99.9, k_{9,10} = 0.4, k_{9,13} = 0.1, k_{9,14} = 0.1, k_{11,1} = 0.1, k_{11,2} = 0.1, k_{11,3} = 0.7, k_{11,4} = 0.1, k_{11,5} = 0.1, k_{11,6} = 0.1, k_{11,7} = 0.1, k_{11,8} = 0.1, k_{11,9} = 0.1, k_{11,10} = 0.1, k_{11,12} = 0.1, k_{11,13} = 0.1, k_{11,14} = 0.2, k_{11,15} = 2.2, k_{12,1} = 100.7, k_{12,2} = 0.1, k_{12,3} = 0.3, k_{12,13} = 0.1, k_{12,14} = 0.1, k_{13,2} = 0.1, k_{13,14} = 10103, k_{14,2} = 1.2833, k_{14,13} = 0.1.$ 51

4.1 Figure shows how the increasing intervals around M is affecting the number of core reactions inside the interval. The vertical axis lists the elements of M and the horizontal axis shows the accumulation of the interval size around the values of M. The calculation start at the top left corner and goes down along the horizontal axis, then current interval gets increased and the calculation start at the top in the next column along the vertical axis. Each color represents the size of the core reaction set within the interval. In each iteration, the applied step size was 0.1. 58

4.2 Comparison of the original network from [12] (left) and the network given by parameter estimation (right). The core reactions in each case are shown with blue dashed edges. 59

4.3 Two sparse realizations of the positive feedback motif with $[M_l]_{ij} = [M]_{ij} - 0.1$ and $[M_u]_{ij} = [M]_{ij} + 0.1$. 66

5.1 The cell-free *in vitro* system consists of three main components. Tube 1 has the extracted content of *E. coli* cells. Tube 2 contains all the amino acids and nucleotides, among other chemicals. Our biocircuit is placed in Tube 3, hence different DNA fragments. Finally, when each components are mixed together the gene expression is initiated. 72

5.2 The plasmid DNA contains a constitutive promoter, then an untranslated region with aptamer and finally the GFP gene. After transcription, mRNA dynamics were measured by utilizing an RNA aptamer for the fluorescent dye malachite green [48]. Finally, translation creates a fluorescent GFP protein. Readers should note that the MGApt emission is in the far red region whereas the GFP emission is in the green spectra. . . 74

DOI:10.15774/PPKE.ITK.2015.008

5.3 Overview of the process model. The model is built around the central dogma of molecular biology with additional step accounting for resource consumptions and degradations. Forward and reverse reaction rate coefficients are denoted with green and orange colors, respectively. The $x_1, \ldots, x_{14}$ are the species concentrations. The parameters p_{11} and p_{15} are not shown, see Appendix A.2.1 for details. 80
5.4 Simulation of reaction rates for the mRNA dynamics and transcription dynamics. The reaction rate for transcription ($p_{12}x_{13}$) decays over time and eventually crosses the rate of mRNA degradation ($p_{17}x_9$) causing a peak in mRNA production. 83
5.5 Result of the structural identifiability which is called the identifiability tableau. The Jacobian of generating series coefficient were calculated and the non-zero elements of the Jacobian are shown. This helps to reduce and eventually solve the underlying algebraic equations; see [26] for details. The GenSSI software does not show the parameters for initial values of the system on the identifiability tableau. 84
5.6 Simulations with the estimated parameter set is shown in red. The Figure contains time series measurement for both channels with different initial concentration DNA (1nM, 2nM, 5 nM and 10 nM plasmid DNA concentration, shown in green, red, cyan and black respectively). The left panel shows the dynamics of MGApt, which is proportional to the mRNA concentration. The GFP dynamics is shown on the right panel. The fluorescent counts for each channel have been converted by applying Equations (5.8) and (5.9) to nM and μM, respectively. 86
5.7 Validation of the estimated parameter set over another set of data (0.1 nM, 0.2 nM, 0.5 nM of plasmid DNA). The red curves show the corresponding simulations with the parameters from Table 5.3. On average there is 20% error in the final value of GFP, the simulation qualitatively matches the time series data. The 1 nM data (black curve) was used for estimation, shown here only for comparison. 88
5.8 Samples for the joint posterior distribution of the parameters were drawn with different confidence intervals (99%, 95%, 90%, and 50%), we simulated the dynamics with these parameter sets. 90

A.1 Measurement of mRNA degradation in the gene expression system. The figure shows that, within the measured rage, mRNA degradation follows first order kinetics. 97
A.2 The experiment compares the two constructs with MGApt and without it. The endpoint measurements (final concentrations level) show that the presence of MGApt increases the overall protein production. 98

DOI:10.15774/PPKE.ITK.2015.008

List of Tables

2.1 Truth table of the possible connectives 14
2.2 List of equivalent compound statements and linear equalities or inequalities, taken from [127] 14

3.1 Comparing the computational time required to compute all different sparse realizations for the kinetic Lorenz systems given by the two transformations. The time required to solve one LP and one MILP problem with the gives network size on our desktop PC is 0.05 sec and 6 sec, respectively. The 'SD-TS' and 'X-factorable' denote state-dependent time-scaling and X-factorable transformation, respectively. 47
3.2 Comparison table of the two approaches for transforming the Lorenz system into kinetic form. In the header line, 'SD-TS' and 'X-factorable' denote state-dependent time-scaling and X-factorable transformation, respectively. 49

4.1 Comparison of the main properties of two algorithms for core reaction set computation. Keywords: $offdiag(A_k)$—number of off-diagonal entries of A_k, $diag(A_k)$—number of diagonal entries of A_k, $Kirchhoff$—Equation (4.3), $DynEq$—Equation (4.5). 56

5.1 Genotype of the plasmid used in this study. 73
5.2 The table lists the species with non-zero initial concentrations in the model. Resource (R) type species are established by the crude-cell extract protocol [110]. The values in case of enzyme (E) type species are taken from the literature. We took the average concentration of the nucleotides (ATP, CTP, GTP, UTP) and denoted the average value as NTP. (The same goes for amino acids.) 81
5.3 Numerical result of the parameter estimation. The table has 13 reaction rate coefficients and two initial concentrations. 87
5.4 The table contains parameter pairs with the strongest cross-correlation. . 89

A.1 Species - state variables 98

DOI:10.15774/PPKE.ITK.2015.008

Abbreviations

CRN	**C**hemical **R**eaction **N**etwork
LP	**L**inear **P**rogramming
MILP	**M**ixed **I**nteger **L**inear **P**rogramming
ODE	**O**rdinary **D**ifferential **E**quations
PDE	**P**artial **D**ifferential **E**quations
TXTL	**T**ranscription-**T**rans**l**ation
DNA	**D**eoxyribo**n**ucleic **a**cid
RNA	**R**ibo**n**ucleic acid
mRNA	**m**essenger RNA
GFP	**G**reen **F**luorescent **P**rotein
MGApt	**M**alachite **G**reen **Apt**amer
NTP	RNA Nucleotides (ATP,GTP,UTP,CTP)
AA	**A**mino **A**cid
RNAP	RNA **P**olymerase
PCR	**P**olymerase **C**hain **R**eaction

DOI:10.15774/PPKE.ITK.2015.008

Notations and Symbols

$\mathbb{R}_+$	non-negative real numbers
$\mathbb{N}$	natural numbers (including zero)
Q^T	transpose of the vector/matrix Q
A_k	Kirchhoff Matrix of a Kinetic System
Y	Complex composition matrix of a kinetic system
$\mathcal{R}(Y, A_k^d)$	The set of reactions in reaction graph with maximal number of reactions
$\mathcal{R}(Y, A_k^s)$	The set of reactions in reaction graph with minimal number of reactions
$\mathcal{R}_C$	set of core reactions in a kinetic system
R_c	number of core reactions in a kinetic system
R_d	number of reactions in a dense realization of a kinetic system
R_s	number of reactions in a sparse realization of a kinetic system
N_c	number of core complexes in a kinetic system
$[M]_{ij}$	the entry in the ith row and jth column of the matrix M
$Y_{i,\cdot}$	ith row of matrix Y
$M_{\cdot,j}$	jth column of matrix M
$\mathrm{col}(Q)$	column expansion of matrix $Q \in \mathbb{R}^{n\times m}$ into a vector $\mathrm{col}(Q) \in \mathbb{R}^{n\cdot m\times 1}$
$\mathrm{row}(Q)$	row expansion of matrix $Q \in \mathbb{R}^{n\times m}$ into a vector $\mathrm{row}(Q) \in \mathbb{R}^{1\times n\cdot m}$
1_m	$m \times m$ matrix with all entries equal to one
I_m	$m \times m$ unit matrix

DOI:10.15774/PPKE.ITK.2015.008

With four parameters I can fit an elephant and with five I can make him wiggle his trunk. — attributed to John von Neumann

DOI:10.15774/PPKE.ITK.2015.008

Chapter 1

Introduction

1.1 The notion and significance of dynamical and nonnegative models

Dynamical models play an important role in many fields of science and engineering such as in physics, mathematics or control theory. Basically, these models are fundamental in every field where the main interest lays in the description of the time and/or space evolution of the modeled quantities. The time evolution of the modeled quantities is often described by (non)linear ordinary differential equations (ODEs). If the space evolution is also an important aspect of the model, then partial differential equations (PDEs) are used [10].

Over the years different powerful mathematical tools emerged to comprehensively analyze the properties of dynamical models, e.g. whether certain states can be achieved, observed or the state itself is stable, robust to external disturbances [35]. More importantly, input functions can be designed to alter the dynamics of the model and drive these dynamical models to certain states and maintain them at that state over time [106].

Dynamical models can be classified either by the possible values of the state variables or by the structure of the model or both. For example, the defining property of the class of nonnegative systems is that all state variables remain nonnegative if the trajectories start in the nonnegative orthant [53]. Some notable examples of nonnegative systems

DOI:10.15774/PPKE.ITK.2015.008

are transportation models, population dynamics, ecological, certain type of economical models and (bio)chemical models. Clearly, the state variables in these examples are nonnegative by definition.

There are distinguished system classes of the nonnegative models which are general enough to include a wide range of smooth nonnegative systems. For example, in the case of quasi-polynomial (QP) systems, the model is composed of quasi-monomial functions where the exponents of the quasi-monomials are real numbers [19, 20]. Another formalism, called S-systems can be reformulated as a QP system. The S-system is a popular platform for modeling biochemical metabolism, because the model explicitly accounts for the influxes and outfluxes of each state variable, although this limits the types of nonnegative systems which can be modeled with S-systems [95, 123].

Also, rational functions can be the building blocks of a nonnegative model, e.g. in case of Michaelis-Menten or Hill-function kinetics [3, 120]. Rational function modeling is frequently used for model reduction of monomial systems, (e.g. with Michaelis-Menten kinetics [85]), since in this case the model with rational functions is often able to describe more realistic dynamical features with less state variables than a polynomial model [92].

1.2 Kinetic systems

A widely used class within nonnegative systems is the class of kinetic systems which is a subclass of the quasi-polynomial systems, where the exponents of the monomials are nonnegative integers. Kinetic dynamical models originate from chemistry, but their range of applicability reaches far beyond (bio)chemical models as they are suitable to describe all important dynamical phenomena such as stability/instability and multiplicity of equilibria [28], bifurcation [81], oscillatory and even chaotic behavior [37, 38]. Many of these phenomena have actually been observed in real chemical experiments where the practical constraints are much more severe than in the case of pure mathematical models [73, 80].

Furthermore, kinetic models can effectively be used in the description of numerous natural processes such as disease dynamics, population dynamics, compartmental models, or certain transportation phenomena. On the top of that, kinetic systems can be used to describe pure chemical reactions or the complex dynamics of intracellular processes,

DOI:10.15774/PPKE.ITK.2015.008

metabolic or cell signaling pathways [51]. Kinetic models have also been useful in performing complex non-conventional computation tasks [1, 2, 34]. Moreover, their simple algebraic structure make these models attractive both for rigorous mathematical analysis and for efficient computational techniques [40, 59], as well as certain strong statements of the structural and dynamical properties of the model can be made about kinetic systems using Chemical Reaction Network Theory, even without knowing the parameters of the kinetic model [30, 103].

Necessary and sufficient conditions for a regular polynomial system to be kinetic were first reported in [56], where a constructive proof was given to build the so-called "canonic mechanism" of kinetic polynomial models. Furthermore, transformation of non-kinetic systems into a polynomial kinetic form is also possible using the so called X-factorable systems and a state-dependent time-rescaling, as shown in Chapter 3.

The directed graph structure assigned to the kinetic system gives us important information about the qualitative dynamical properties of the system. Although it was known since at least the 1970's that multiple different directed graph structures/parametrizations can describe exactly the same dynamics of the concentrations [38, 59]. This phenomenon is called *macro-equivalence* or *dynamical equivalence*. However, the exact geometric conditions of macro-equivalence were not studied until recently in [29]. The first optimization-based numerical procedures for generating macro-equivalent structures with prescribed (dynamically relevant) properties were reported in, e.g. [113, 114, 115, 116].

Dynamical equivalence enables us to compute directed graph structures with maximum or minimum number of edges. It has been shown that the dynamically equivalent directed graph with maximum number of edges defines a unique structure [113]. Furthermore, Chapter 3 illustrates the result from [113] on the non-uniqueness of a dynamically equivalent directed graph with minimal number of edges.

Besides the maximum or minimum number of reactions, other properties can also characterize these directed graphs; e.g. structural invariance of certain directed edges (called core reactions). If such edges do exist, then these are present in each dynamically equivalent directed graphs. An optimization method for computing such structurally invariant edges was reported in [114]. Core reactions and their properties will be utilized extensively in Chapters 3 and 4.

DOI:10.15774/PPKE.ITK.2015.008

Kinetic systems can be extended to accommodate parametric uncertainty, where the uncertainty is represented as a multi-dimensional interval in the space of monomial coefficients. Therefore, one can immediately represent the measurement uncertainty in the model or incorporate the effect of e.g. temperature change or machine wear off. In the case of metabolic networks a similar model has been suggested, e.g. in [71].

Core reactions can be easily defined within the class of uncertain kinetic systems. Thus, they can be utilized as 'certain' structural elements of uncertain systems, as shown in Chapter 4. One of the possible applications of uncertain kinetic systems with core reaction is (bio)chemical network reconstruction, which gained significant interest recently [12, 107].

All these advantageous properties of kinetic models and the associated graph structure explain the recent raising interest of mathematicians and engineers towards (bio)chemical reaction networks [5, 25, 105].

Dynamical modeling in the field of systems biology and synthetic biology are good examples where nonnegative, especially, kinetic systems are typically applied [122]. In systems biology the aim is to understand and eventually control biomolecular processes such as signal transduction or metabolism. Thus, dynamical models can support this process by accurately describing the observed phenomena and the inherent properties of the biological system [3]. On the other hand, in synthetic biology rational designing and creating novel interaction networks, e.g. gene regulatory networks is the main aim [69]. If these interaction networks or so called biocircuits are successfully designed and tested they may be capable of sensing external or internal signals, compute the necessary response and actuate the molecular system accordingly. Meanwhile, all of these steps are based on molecular computation [64].

Therefore, in both fields, but especially in synthetic biology dynamical models are becoming essential tools to carefully investigate and understand the biological processes, to predict the possible dynamical properties and to support the rational design process with appropriate feedback. Consequently, in these fields kinetic models with monomial or rational reaction rates are frequently applied [84]. Also, the underlying processes can be modeled using stochastic approaches [45]. Besides deterministic kinetic or stochastic models, other model types do exist as well such as Petri nets [77], probabilistic graphical models [43], and some other model structures are reviewed in [62].

DOI:10.15774/PPKE.ITK.2015.008

In many fields of engineering there exists some type of test bed for rapid assembly of components and test out the proof-of-concepts; such as the breadboard of electrical engineers or the wind tunnel for aerospace engineers. These test beds are usually well characterized, e.g. all the physical dimensions or electrical properties are known in advance. Hence, it is easy to carry out controlled experiments and determine the properties of the proof-of-concepts system.

Recently, a similar breadboarding concept was developed for synthetic biology, where the DNA segments or whole biocircuits can be rapidly assembled and tested [101, 110, 111]. The components of such a system are DNA segments, which are assembled together with standard molecular biology techniques, and a host environment, typically some type of crude cell-extract. Due to this molecular breadboarding environment there has been a significant decrease in the amount of time required for biocircuit assembly and subsequent testing. Despite of these developments, exploring the dynamical properties of biocircuits is still resource and time consuming. Therefore a software toolbox was developed to simulate the dynamics of breadboarding environment, to explore the possible operation regimes and to make predictions about the performance of biocircuits before implementing them in *in vitro* experiments [119, C4].

This software toolbox gives insight into the dynamics of unmeasured states of the molecular breadboard, especially accounts for resource usage. It is important, because at the current state of this particular molecular breadboard, it only allows for a single dose of resources, i.e. limits the time of the operation. The toolbox provides a general modeling framework for planning circuit layouts and gives predictive models for synthetic biomolecular circuits. A major benefit of the toolbox is that it gives feedback to the rational design step, and thus it increases our capacity to rationally design biomolecular circuits. The kinetic model developed for this software toolbox is introduced in Chapter 5.

Recent advances in measurement technology provide us a rich source of data for revealing the structure and behavior of biochemical processes which became the integral part of the software toolbox. Using real time measurements of both transcriptional and translational stages of gene expression give us the necessary insight to determine the parameters of the kinetic model.

DOI:10.15774/PPKE.ITK.2015.008

1.3 Challenges of structure and parameter estimation of kinetic systems

For simulations of kinetic systems or biochemical processes accurate model parameters are needed, e.g. initial concentrations or reaction rate coefficients. Usually, unit values or previously published figures from the literature are used, these numbers are aggregated in the BioNumber database [17], but there can be multiple parameter values for the same process and these values can show great variability. This may due to the fact they were measured or calculated under different conditions, e.g. with different type of equipment or protocol, not to mention the diversity of bacterial strains and available chemical compounds.

In another approach the parameters of the system can be determined from measurement data, but the process of parameter estimation is often challenging. Generally, these challenges can be classified into two main categories. First, the selected process model may have structural identifiability issues, namely the model structure is capable to produce exactly the same output for different sets of parameters [6, 126]. The second challenge stems from the poor excitation of the dynamics or by the poor quality of the available measurements, which is often labeled as practical identifiability problem [74]. Although these challenges are linked to either the model or the measured data, the concepts of structural and practical identifiability are sometimes mixed together [39, 50, 87].

Structural identifiability depends only on the structure of the model including the output functions. Therefore, structural identifiability analysis can be carried out before collecting the data, if the model structure is known. Unfortunately, this analysis is often neglected and still not a standard practice of modelers. On the other hand, several approaches and software tools exist for structural identifiability analysis and they are suitable for many different model structures [7, 26, 33].

If one of these approaches show that the model has structural identifiability problems, then generally, there are two ways to solve it. We can either try to change the output function, e.g. by making more states observable or change the structure of the model, e.g. by reducing the model.

DOI:10.15774/PPKE.ITK.2015.008

If the first approach is unsuccessful, i.e. all states are already observed or the output function can not be changed anymore, then the model structure has to be modified to achieve a structurally identifiable model. For example the first principle models—which tend to yield many parameters—can be restructured and simplified by normalization of the parameters. Besides that, model reduction is commonly executed to restructure or reduce the number of states and parameters of the model [55, 85, 125].

For kinetic systems dynamical equivalence may also hamper the structural identification process [114]. In essence, many different parameter sets can describe equally well the observed dynamics. By assuming the minimum number of edges in the reaction graph one could potentially alleviate this problem [12], however as we will see in Chapter 3 the minimal (sparse) structure may be non-unique.

Even if a model is structurally identifiable, the quality and the information content of the measured data has a huge impact on the parameter estimation process, i.e. there are practical identifiability issues with the dynamical model. Poor excitation of the dynamics often manifests in highly correlated parameters, large confidence intervals, which severely limit the usability of the model. Local parameter sensitivity analysis [134] or sample based global parameter sensitivity analysis [93] is commonly used to determine the parameters' influence on the dynamics of the model [82]. Parameter sensitivity is a good metric for experiment design or input design. Various optimization techniques exist to maximize the information content of the measured data by efficiently excite the dynamical model, i.e. this maximizes the parameter sensitivity and improves the quality of the parameter estimation [23, 42, 90, 124].

The input or experiment design was developed by the system and control theory community, where the technological processes or the electrical systems can be relatively easily manipulated by input functions, e.g. by changing the temperature or electrical current [70]. On the other hand, in the case of a biomolecular breadboarding system, most of the changes in the experiment conditions require re-optimization of the experiment protocol (e.g. restore pH or optimize salt content), which can be tedious work and it might unintentionally change other parameters of the system [110]. Therefore, it is still an open question how to efficiently do an experiment design for molecular breadboarding and ultimately do complete system identification on biocircuits.

DOI:10.15774/PPKE.ITK.2015.008

Finally, the last step of the system identification process is the evaluation of the estimated parameters [70]. This analysis can be carried out either in the space of parameters or in the space of cost functions. In the space of the parameters, usually statistical aspects of the parameters are investigated such as calculation of confidence intervals, coefficients of variations or p-value [9]. In the space of cost functions, generally, the shape of the cost function, curvature of the cost function or the value of residuals around the optimal values of the parameter set are the target of analysis [97]. Validation can also be used for testing the performance of the estimated parameter set with a different set of data which was used for the parameter estimation.

In conclusion, the above mentioned tools, techniques and the highlighted issues show that a successful parameter estimation for identifying the kinetic systems involves a complex set of procedures starting from structural identification, through experiment design, until the evaluation of the result of the parameter estimations [13, 89]. Given the complex nature of the identification process, software tools have emerged that can assist the users from the beginning of the modeling process till the end of the evaluation. These softwares can effectively explore the different aspects of the process model, the measured data as well as the estimated parameters [14, 27, C4].

1.4 Structure and objectives of the thesis

Most of the results in this thesis are built upon linear programing, modeling and parameter estimation of kinetic systems. Therefore, the necessary tools and techniques for mathematical optimization and kinetic systems along with selected references for further details are introduced in Chapter 2.

New scientific contributions about the class of kinetic systems are presented in three chapters. In Chapter 3 and 4, I present how to use Linear Programming and Mixed Integer Linear Programming to calculate different structural elements of chemical reaction networks. While in Chapter 5, I describe the construction and parameter estimation of the kinetic model.

Chapter 3 outlines how to develop an algorithm in order to find all dynamically equivalent reaction graphs with minimal number of reactions (sparse reaction graphs) using optimization techniques.

DOI:10.15774/PPKE.ITK.2015.008

In Chapter 4, I extend the previous results on structurally invariant reactions to uncertain kinetic systems. Here, I also define uncertain polynomial kinetic systems and show the development of a polynomial time algorithm which can be used to find a structurally invariant reactions of this system.

Chapter 5 presents an application oriented modeling method and the parameter estimation of an *in vitro* gene expression system. In this chapter, I introduce the measurement framework and the methods were performed in a molecular biology laboratory. Thanks to the high resolution measurements this model was populated with parameters estimated from time series measurement data.

Finally, the conclusion is drawn in Chapter 6 which also highlights some future directions of researches based on the presented results of this thesis. Moreover, this chapter contains a list of new scientific results with the corresponding peer-reviewed publications as well.

DOI:10.15774/PPKE.ITK.2015.008

Chapter 2

Background and Basic Notations

The objectives of this chapter are to introduce the mathematical tools and techniques which will serve as the foundation of the results presented in this thesis. These results in the later chapters highly rely on linear programming and least squares optimization, which belong to convex optimization. Therefore, the chapter starts with a brief summary of convex optimization along with an introduction to linear programming and least squares optimization. The latter one is used for parameter estimation of dynamical systems in this thesis. As the introduction highlighted the challenges associated with system identification of kinetic systems, one part of this chapter introduce the concept and the latest developments of structural identifiability. Finally, since majority of the results in this thesis are related to the class of kinetic systems, a comprehensive introduction to kinetic system modeling and its properties is given.

2.1 Convex Optimization

Convex optimization is an important area of mathematical optimization and used widely in engineering and in other fields. Just to give some examples: optimization of circuits design, portfolio optimization in finance, route or production planning or optimal resource allocation [31, 32].

For a proper introduction, we need to define the convex set and convex function.

Definition 2.1. A set C is convex if for any $x_1, x_2 \in C$ and any θ defined as $0 \leq \theta \leq 1$, we have $\theta x_1 + (1-\theta)x_2 \in C$ [18].

DOI:10.15774/PPKE.ITK.2015.008

Definition 2.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom} f$ is a convex set and if for all x, $y \in \mathrm{dom} f$ and θ with $0 \leq \theta \leq 1$, we have $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$ [18].

Based on the above definitions we can write a convex optimization problem as

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq b_i, \quad i = 1, \ldots, m, \end{aligned} \tag{2.1}$$

where the functions $f_0, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex. The variable $x \in \mathbb{R}^n$ is the optimization or decision variable of the problem. The convex function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective or cost function. The convex functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ are the constraints associated with the problem and vector $b \in \mathbb{R}^m$ is the bound for the constraints. The goal of the optimization problem is to find vector $x^* \in \mathbb{R}^n$ which satisfies the following: for any $z \in \mathbb{R}^n$ with $f_i(z) \leq b_i$, $i = 1, \ldots, m$ and $f_0(z) \geq f_0(x^*)$, i.e. there is no other vector in $\mathbb{R}^n$ which satisfies all the constraints and the cost function at that point has lower value than x^* has [18].

The linear programming and least squares covered in the next subsections are both special cases of convex optimization.

2.1.1 Linear Programming

Linear programing (LP) is a constrained convex optimization technique, where a linear objective function of the real-valued optimization variables is minimized (or maximized) with respect to linear equality and inequality constraints.

The linear programming problems that will be used in this thesis can be written in the following form

$$\begin{aligned} & \min_{x} c^T x \\ & \text{subject to:} \\ & Ax \leq b \\ & Gx = d \end{aligned} \tag{2.2}$$

DOI:10.15774/PPKE.ITK.2015.008

where $x \in \mathbb{R}^n$ is the vector of decision variables, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$ are known vectors and matrices and they encode the inequality constraints on x, the symbol '$\leq$' means element-wise non-strict inequality. The matrix $G \in \mathbb{R}^{q \times n}$ and $d \in \mathbb{R}^q$ are also known and they accommodate the equality constraints on x.

It should be noted that there exist many definitions of the linear programming problem and they are all equivalent to each other [31]. We choose the above definition because the type of problems we want solve can be easily encoded into this formulation.

Each equality and inequality constraint defines a hyperplane or a halfspace, respectively. The intersection of these hyperplanes and halfspace defines a polyhedron. The goal of a LP optimization is to minimize the cost function $c^T x$ over the polyhedron defined by the solution set of the constraints. The points where the cost function takes its minimum are on the boundary of the polyhedron, but depending on the type of the cost function a face of the polyhedron can contain the point where the minimum of the cost function is reached.

The existence of a feasible solution of the optimization problem defined in Equation (2.2) can be checked by applying the following optimization model

$$\begin{aligned} &\min_z \textstyle\sum_{i=0}^{p} z_i \\ &\text{subject to:} \\ &Ax + z = b \\ &x \geq 0, z \geq 0 \end{aligned} \tag{2.3}$$

where vector $z \in \mathbb{R}^p$ represents the so called auxiliary variables. The LP problem defined in Equation (2.2) has a feasible solution if and only if the LP in Equation (2.3) has minimal value 0 with $z_i = 0$ for $i = 1, \ldots, p$ [31].

The linear programming framework is very versatile, thus the practical applications are ranging from engineering to social sciences, e.g. production optimization, transportation and assignment problems, etc [18]. Moreover, many efficient solvers are available to solve linear programming problems even with millions of decision variables and hundreds of thousands constraints enable us to solve large problems efficiently. These solvers are based on the simplex method or lately on the interior point method, reviewed in e.g. [31].

DOI:10.15774/PPKE.ITK.2015.008

2.1.2 Mixed Integer Linear Programming

Some problems require decision variables with integer values. This constraint makes the optimization problem NP-hard, although thanks to efficient solvers many practical problems can be solved. A mixed integer linear programming (MILP) problem can be written as

$$\begin{aligned} &\min_x c^T x \\ &Ax \leq b \end{aligned} \tag{2.4}$$

where $x_i \in \mathbb{R}$ for $i = 1, \ldots, l$, $c \in \mathbb{R}^l$ and $x_j \in \mathbb{Z}$ for $j = l+1, \ldots, n$. The vector x represents the decision variables and for some elements accepted the optimal value is integer, these elements called integrality constraints. The matrix $A \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$ are the inequality constraints on x [31].

Linear programs with integrality constraints arise in many fields, e.g. in transportation, scheduling, etc. For example, the number of rail cars assigned to a train to transport certain amounts of goods or optimal allocation of people or machinery to perform certain tasks. Clearly, some of the resources in these problems have to have integer values.

Propositional Calculus and Linear Integer Programming A connection between linear integer programming and propositional calculus can be made [127]. Logical literals which are either true (T) or false (F) denoted by X_i represent certain facts, e.g. $x \geq 0$ or the sky is clear. Boolean algebra makes possible to connect these literals into compound statements with the so called connectives. The connectives and their truth table is given in Table 2.1. In addition, it is possible to transform compound statements into different connectives or give a simple form for a complex statement.

A propositional logic problem, where a statement X_1 must be proved to be true given a set of (compound) statements involving literals $X_1, X_2 \ldots, X_n$ can be solved by means of a linear program with integrality constraints, by translating the original compound statements into linear inequalities involving logical variables [16, 127]. These logical variables, denoted by $\delta_i \in \{0, 1\}$ are associated with the corresponding literals $X_i, i = 1, \ldots, n$. Based on that, a list of equivalent compound statements and linear equalities or inequalities is given in Table 2.2.

DOI:10.15774/PPKE.ITK.2015.008

TABLE 2.1: Truth table of the possible connectives

Literals		Connectives					
		NOT	OR	AND	Implies	If and only if	Exclusive or
X_1	X_2	$\sim X_1$	$X_1 \vee X_2$	$X_1 \wedge X_2$	$X_1 \rightarrow X_2$	$X_1 \iff X_2$	$X_1 \oplus X_2$
F	F	T	F	F	T	T	F
F	T	T	T	F	T	F	T
T	F	F	T	F	F	F	T
T	T	F	T	T	T	T	F

TABLE 2.2: List of equivalent compound statements and linear equalities or inequalities, taken from [127]

Compound statements	linear equalities or inequalities
$X_1 \vee X_2$	$\delta_1 + \delta_2 \geq 1$
$X_1 \wedge X_2$	$\delta_1 = 1,\ \delta_2 = 1$
$\sim X_1$	$\delta_1 = 0$
$X_1 \rightarrow X_2$	$\delta_1 - \delta_2 \leq 0$
$X_1 \iff X_2$	$\delta_1 - \delta_2 = 0$
$X_1 \oplus X_2$	$\delta_1 + \delta_2$

2.1.3 Least Squares Optimization

Least squares optimization is a special case of convex optimization, where the cost function is

$$\min_x \quad f_0(x) = ||Ax - b||_2^2 = \sum_{i=1}^{k} (\langle a_i^T x \rangle - b_i)^2 \tag{2.5}$$

where $A \in \mathbb{R}^{k \times n}$, a_i^T is the ith row of A and vector $x \in \mathbb{R}^n$ is the optimization variable.

The least squares problem originates from solving the system of equations denoted as $Ax = b$, where $A \in \mathbb{R}^{k \times n}$ and $b \in \mathbb{R}^k$, but it has no exact solution, typically because $k > n$. Roughly speaking, there are more equations then variables. But we can still search for an approximate solution, where the difference $r = b - Ax$ is minimal. This difference is called the residual vector or the residual, the goal of the optimizations is to find a parameter vector x which minimizes the residual. A commonly used metric to measure the size of the residual is the 2-norm, and a parameter vector minimize the 2-norm of the residual is called the least squares solution [9, 70].

DOI:10.15774/PPKE.ITK.2015.008

The least squares optimization problem has many interpretations in engineering, statistics, astrophysics and in many other fields [18]. One reason for that, if there is no constraints on the optimization variable x, then this optimization problem has a closed form solution as

$$x = (A^T A)^{-1}(A^T)b. \tag{2.6}$$

Equation (2.6) is called the normal equation and this 2-norm solution is a special interest because this is statistically the most likely solution if the data error are normally distributed [9].

Even though the optimization problem in Equation (2.5) has a closed form solution, it can be ill-posed, i.e. the column rank of A is less than n. In this case we are facing a rank deficient or ill-posed problem [9]. A solution of this problem is called regularization, where additional constraints are introduced on the optimization variable to improve the solution. Among many regularization techniques the notable examples are Tikhonov regularization and L1 regularization [44].

Further aspects of the least squares optimization and its application to parameter estimation of kinetic models will be explored in Section 2.4.

2.2 Introduction to modeling of Kinetic Systems

(Bio)chemical systems obeying the mass action law can be described by nonlinear polynomial ODEs where there are strict relations between the monomial exponents and coefficients guaranteeing nonnegativity of the solutions (in case of nonnegative initial conditions), and giving rise to a weighted directed graph structure called reaction graph [38, 41]. In this graph, the participating chemical complexes are the nodes in the network and the reactions which transform complexes into each other are represented by weighted directed edges. The reaction rates are directly proportional to the edge weights.

Linear programming based optimization techniques exist to calculate certain graph structures. Some of these structural properties are directly connected to the dynamical behavior of the kinetic system. Therefore, based on the structure of the graph some dynamical properties, e.g. stability can be determined.

DOI:10.15774/PPKE.ITK.2015.008

Since most of the contribution reported in this thesis are related to the class of kinetic systems, a thorough introduction of notations, definitions and key features of kinetic systems are presented in this section.

2.2.1 ODE description

The kinetic models studied in this thesis are given in the following polynomial form:

$$\dot{x} = M \cdot \psi(x), \tag{2.7}$$

where $x \in \mathbb{R}^n_+$, $M \in \mathbb{R}^{n \times p}$ and $\psi(x)$ is a monomial-type vector mapping which is defined as

$$\psi_j(x) = \prod_{i=1}^{n} x_i^{\alpha_{ij}}, \quad j = 1, \ldots, p \tag{2.8}$$

with $\alpha_{ij} \in \mathbb{N}$. In order to define a kinetic system, the following relation has to be fulfilled between M and α [38]:

$$\alpha_{ij} \geq 1 \text{ for any } i, j \text{ for which } M_{ij} < 0. \tag{2.9}$$

For computation purposes, we will use an appropriate factorization of Equation (2.7) as follows. Let us define $Y \in \mathbb{N}^{n \times m}$ as the complex composition matrix of the system. Additionally, $A_k \in \mathbb{R}^{m \times m}$ is a special compartmental matrix, the so-called Kirchhoff matrix belonging to the system. A_k is defined as:

$$[A_k]_{ij} = \begin{cases} k_{ji} & \text{if } i \neq j \\ -\sum\limits_{l=1, l \neq i}^{m} k_{il} & \text{if } i = j, \end{cases} \tag{2.10}$$

where $k_{ij} \geq 0 \ \forall i, j$. With the help of these two matrices, we can write Equation (2.7) as

$$\dot{x} = Y \cdot A_k \cdot \varphi(x), \tag{2.11}$$

where

$$\varphi_j(x) = \prod_{i=1}^{n} x_i^{Y_{ij}} \quad j = 1, \ldots, m. \tag{2.12}$$

DOI:10.15774/PPKE.ITK.2015.008

Note that the monomials in functions ψ and φ are not necessarily identical, although the right hand sides of Equations (2.7) and (2.11) determine the same dynamics.

2.2.2 Directed graph structure

We can associate a graph representation to kinetic models. A kinetic system equipped with this graph structure will be called a Chemical Reaction Network (CRN) as it is described in e.g. [41]. In this representation, a CRN is characterized by three sets:

1. $\mathcal{S} = \{X_1, \ldots, X_n\}$ is the set of *species* or chemical substances.

2. $\mathcal{C} = \{C_1, \ldots, C_m\}$ is the set of *complexes*. Formally, the complexes are represented as linear combinations of the species, i.e.

$$C_i = \sum_{j=1}^{n} \beta_{ij} X_j, \qquad i = 1, \ldots, m, \tag{2.13}$$

   where β_{ij} are nonnegative integers and are called the *stoichiometric coefficients*.

3. $\mathcal{R} = \{(C_i, C_j) \mid C_i, C_j \in \mathcal{C},\ i \neq j,$ and C_i is transformed to C_j in the CRN$\}$ is the set of *reactions*.

   The reaction $(C_i, C_j) \in \mathcal{R}$ will be denoted as $C_i \to C_j$. Moreover, a positive weight, the *reaction rate coefficient* denoted by k_{ij} is assigned to each reaction $C_i \to C_j$. According to our convention, $k_{ij} = 0$ indicates that the reaction $C_i \to C_j$ is not present in the CRN.

Given the sets $\mathcal{S}$, $\mathcal{C}$ and $\mathcal{R}$, a weighted directed graph (called the reaction graph) $\mathcal{G} = (\mathcal{R}, \mathcal{C})$ can be constructed, where the set $\mathcal{C}$ contains the vertices that represent the complexes of the reaction network, i.e. $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$. The set $\mathcal{R}$ contains the directed edges representing the reactions between the complexes and the corresponding *reaction rate coefficient* is assigned as weight to each edges. It is important to remark that loops and multiple edges with the same direction are not allowed in the reaction graph.

The relationship between the ODE model of a kinetic system in Equation (2.11) and its reaction graph is the following. The state vector x contains the species concentrations. The entries of matrix $Y \in \mathbb{N}^{n \times m}$ are $Y_{ij} = \beta_{ji}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$,

DOI:10.15774/PPKE.ITK.2015.008

and $[A_k]_{ij} = k_{ji}$ is the reaction rate coefficient corresponding to the reaction $C_j \to C_i$. $[A_k]_{ij} = 0$ means that the reaction $C_j \to C_i$ does not occur in the CRN. It is clear that matrices Y and A_k encode the stoichiometric composition and the weighted directed graph of a CRN, respectively, and these matrices are sufficient to completely characterize the kinetic dynamics described in Equation (2.7).

Now we briefly define the notions and properties of CRNs that will be used in the later chapters. More details can be found in [40], while basic notions of directed graphs are discussed in e.g. [15]. First of all, *linkage classes* are the maximal connected subgraphs (i.e. components) of $\mathcal{G}$. That is, complexes C_i and C_j belong to the same linkage class if and only if there exists a path from C_i to C_j in $\mathcal{G}$. Throughout this thesis, we do not treat isolated vertexes (complexes without reactions) as separate linkage classes, and we simply omit them from the CRN model (although we depict them in the figures for the sake of completeness). We call a reaction graph *weakly reversible*, if there exists a directed path from C_i to C_j whenever there is a directed path from C_j to C_i in the reaction graph.

We have to briefly discuss the usage of the so-called *zero complex* in our models. The zero complex is formally represented by a zero column vector in Y, i.e. it is a special complex containing no species. Similarly to [40], we can use it to uniformly represent the environment, i.e. a CRN containing a (non-removable) zero complex is actually an open system. Therefore, the reactions of the type that are commonly written in the literature as $S \to X$ and $X \to P$, where S is a species of constant concentration and P is an unreactive product (the concentration of which is not included into the dynamic model), will be written as $0 \to X$ and $X \to 0$, respectively, where '0' denotes the zero complex. Similarly, reactions like $X + S \to 2X$ will be simply written as $X \to 2X$. It is emphasized that this is only a notational convention simplifying the description of CRNs. The resulting kinetic dynamics describing the concentrations of the species in $\mathcal{S}$ are the same in both cases, and the two ways of representation can be easily transformed to each other, if necessary.

To each reaction $C_i \to C_j$, we can associate a *reaction vector* denoted by e_{ij} as

$$e_{ij} = [Y]_{\cdot,j} - [Y]_{\cdot,i}, \quad (C_i, C_j) \in \mathcal{R}, \tag{2.14}$$

DOI:10.15774/PPKE.ITK.2015.008

where $[Y]_{.,i}$ denotes the ith column of Y. The *rank of a network* is the dimension of span$\{e_{ij}\}, i = 1, \ldots, m, j = 1, \ldots, m$.

The structure of the reaction graph is directly connected to certain important dynamical properties of a kinetic system. The *deficiency* (a nonnegative integer number depending on the structure of the reaction graph and on complex composition but not on the particular values of reaction rate coefficients) is a good example for this [40]. The deficiency d of a CRN is given by the simple formula

$$d = m - l - s, \tag{2.15}$$

where m is the number of (non-isolated) complexes in the network, l is the number of linkage classes and s is the rank of the network. Roughly speaking, lower deficiencies (particularly 0 and 1) with certain structural properties like weak reversibility can guarantee an 'ordered' behavior of kinetic dynamics without 'exotic' phenomena such as periodic solutions or chaos. Some examples of the direct consequences related to deficiency number such as the Deficiency One and Deficiency Zero Theorems can be found in [40].

The concept of complex balance is offering a way to check stability properties of kinetic systems. A kinetic system realization, defined by (Y, A_k), is *complex balanced* at $x^* \in \mathbb{R}^n_+$ if

$$A_k \varphi(x^*) = 0. \tag{2.16}$$

The implication of the complex balanced property of kinetic systems can be found in [58].

Section 2.3.4 will illustrate the above defined CRN properties through a biomolecular example.

It is important to remark that similarly to [40], the class of deterministic kinetic systems is considered here as a general nonlinear system class, and it is much wider than the family of chemically actually meaningful kinetic systems. Therefore, we do not study the practical realizability of the obtained CRNs in this thesis. We note that the existence of thermodynamically feasible CRN structures can be examined in the same optimization framework that we use in this thesis by adding extra linear constraints (see, e.g. [115]).

DOI:10.15774/PPKE.ITK.2015.008

Moreover, the realization computation techniques that are summarized in subsection 2.3, were successfully applied to biochemical models known from the literature in [114].

2.2.3 Assigning the canonical CRN to a kinetic system

It is important to summarize when it is possible to assign a mass-action type CRN to a general polynomial dynamical system. If it is possible, we call it a dynamical system *kinetic*. There exists a necessary and sufficient condition to check this and it was first published in [56]. Consider an autonomous nonlinear system

$$\dot{x} = F(x), \quad x \in \mathbb{R}^n \tag{2.17}$$

with polynomial right hand side. The system in Equation (2.17) is kinetic if and only if the coordinate functions f_i of F fulfill

$$f_i(x) = -x_i g(x) + h(x), \quad i = 1, \ldots, n, \tag{2.18}$$

where g and h are polynomial functions with nonnegative coefficients. This means that all the negative monomial terms in the ith coordinate function of f must contain x_i, i.e. negative cross-effects are not allowed in kinetic models.

A constructive proof for the above condition can be found in [56], along a simple procedure is to build the so-called "canonical" CRN realization of a kinetic ODE. We briefly summarize this algorithm for convenience, because it is needed for the algorithm will be presented in Section 3.1.

Let us assume that the polynomial coordinates functions of the right hand side of Equation (2.17) is given in the following form

$$f_i(x) = \sum_{j=1}^{r_i} m_{ij} \prod_{k=1}^{n} x_k^{b_{jk}}, \quad i = 1, \ldots, n, \tag{2.19}$$

where r_i is the number of monomial terms in function f_i. Let us denote the transpose of the ith standard basis vector in $\mathbb{R}^n$ as e_i and let $B_{j,\cdot} = [b_{j1} \ \ldots \ b_{jn}]$. Then, the steps necessary to construct the canonical CRN realization are the following.

DOI:10.15774/PPKE.ITK.2015.008

ALGORITHM 1: Algorithm for building a canonical realization a kinetic system from [56].

Input : A non-negative polynomial system encoded by matrices M, B
Output: Y, A_k

$Y := 0_{n\times 1}$ // $n \times 1$ zero matrix;
$A_k := 0$;
for each $i = 1, \ldots, n$ **do**
  for each $j = 1, \ldots, r_i$ **do**
    $Q_{j,\cdot} := B_{j,\cdot} + \text{sign}(m_{ij}) \cdot e_i$;
    $m := \text{FindVector}(B^T_{j,\cdot}, Y)$;
    $z := \text{FindVector}(Q^T_{j,\cdot}, Y)$;
    $A_k := \text{AddEntry}(A_k, z, m, |m_{ij}|$ // Adds $|m_{ij}|$ to zth row mth column by adjusting the size of A_k, if necessary ;
  endfor
endfor
$A_k := \text{Kirchhoff}(A_k)$ // restores the Kirchhoff property of A_k by adjusting the diagonals ;
return Y, A_k ;

FindVector(V, Y)
  // $columns(Y)$ gives the number of columns in Y ;
  for each $p = 1, \ldots, columns(Y)$ **do**
    if $V = Y_{p,\cdot}$ **then**
      $k := p$;
    else
      $Y := Y \cup V$ // add vector V as the last column of Y;
      $k = columns(Y)$;
    end
  endfor
  return k;

2.2.4 Dynamical equivalence of kinetic systems

It has been known since at least the 1970's that multiple different structures (parametrizations) of a CRN can generate exactly the same dynamics of the concentrations [38, 59]. This phenomenon is called *macro-equivalence* or *dynamical equivalence*. However, the exact geometric conditions of macro-equivalence were not studied until relatively recently in [29]. Naturally, the phenomenon of dynamical equivalence may hamper the parameter identification process, since multiple structures can explain the modeled dynamics equally well [114].

Mathematically, dynamical equivalence means that the factorization in Equation (2.11) is non-unique. Therefore, the matrix pair (Y, A_k), where Y is a complex composition matrix and A_k is a Kirchhoff matrix, is called a *dynamically equivalent realization* of the kinetic system in Equation (2.7), if $M \cdot \psi(x) = Y \cdot A_k \cdot \varphi(x)\ \forall x \in \mathbb{R}^n_+$. We note that a

DOI:10.15774/PPKE.ITK.2015.008

given kinetic dynamics can generally be represented using different complex sets. However, there exists a simple procedure described in Section 3.2 that generates a possible dynamically equivalent realization called the *canonical structure* for any kinetic polynomial model. From now on, we assume that the set of complexes is known and fixed, therefore, all dynamical equivalent realizations can be characterized by the equation

$$Y \cdot A_k = M. \tag{2.20}$$

Clearly, if $A_k^{(1)}$ and $A_k^{(2)}$ give dynamically equivalent realizations with fixed Y and $[A_k^{(1)}] \neq [A_k^{(2)}]$, then $A_k^{(3)} = \frac{A_k^{(1)}+A_k^{(2)}}{2}$ also gives a valid dynamically equivalent realization with Y which is different from $A_k^{(1)}$ and $A_k^{(2)}$. In general, we can define a series $A_k^{(n+1)} = \frac{A_k^{(1)}+A_k^{(n)}}{2}$, where each element of the series is a dynamical equivalent realization. Therefore, a kinetic system with different dynamically equivalent realizations has infinitely many dynamically equivalent realizations.

From an optimization point of view the dynamical equivalence defines a polyhedron which contains all dynamically equivalent realizations. Therefore, we can define a linear programming problem where the constraint set contains the definition of dynamical equivalence, then we can search for realizations with special properties such as maximal or minimal number of edges in the reaction graph, weakly reversible realization, etc.

2.3 Known optimization methods for computing certain CRN realizations

We briefly recall the computation framework first described in [113] and some related results to lay the foundation of Chapter 3 and 4.

2.3.1 Computing Sparse and Dense Realizations

For a fixed complex composition matrix Y, *sparse realizations* contain the minimum number of nonzero off-diagonal elements (i.e. reactions) in the matrix A_k. Throughout the thesis, the set of reactions in a particular sparse realization is denoted as $\mathcal{R}(Y, A_k^s)$. Conversely, *dense realizations* contain the maximal number of nonzero off-diagonal elements in A_k. Similarly, the set of reactions in a dense realization is denoted as $\mathcal{R}(Y, A_k^d)$.

DOI:10.15774/PPKE.ITK.2015.008

The search for these matrices (graph) structures can be formulated as mixed integer linear programing (MILP) problems, where we assume that we have a canonical CRN and its parameters are known. Solving the MILP optimization, we want to find valid A_k Kirchhoff matrices that fulfill the given requirements with minimal (maximal) number of reactions.

The mass-action dynamics can be expressed as equality and inequality constraints as

$$\begin{aligned} Y \cdot A_k &= M && (2.21)\\ \sum_{i=1}^{m} [A_k]_{ij} &= 0, \quad j = 1, \ldots, m && (2.22)\\ [A_k]_{ij} &\geq 0 \quad i, j = 1, \ldots, m \quad i \neq j && (2.23) \end{aligned}$$

where the elements of A_k are the decision variables. We also put lower and upper bound constraints on the decision variables to make the optimization problem computationally tractable and to avoid unbounded feasible solutions

$$\begin{aligned} 0 \leq \quad [A_k]_{ij} &\leq l_{ij}, \quad i, j = 1, \ldots, m, i \neq j && (2.24)\\ l_{ii} \leq \quad [A_k]_{ii}, &\quad i = 1, \ldots, m. && (2.25) \end{aligned}$$

Using these constraints we can find such A_k matrices where the number of nonzero off-diagonal elements are minimal or maximal. To achieve this, we utilized the connection between proposition logic and linear integer programming (see Section 2.1.2 for details), and introduce logical variables denoted by δ and construct the following compound statements.

$$\delta_{ij} = 1 \Leftrightarrow [A_k]_{ij} > \epsilon, \quad i, j = 1, \ldots, m, i \neq j \qquad (2.26)$$

where '$\Leftrightarrow$' encodes the *if and only if* logical statement and ϵ is a sufficiently small positive value (i.e. solutions below ϵ are treated as zero). The inequalities in (2.24) and (2.26) can be combined into the following form [86]

$$\begin{aligned} 0 \leq \quad & [A_k]_{ij} - \epsilon\delta_{ij} && i, j = 1, \ldots, m, i \neq j && (2.27)\\ 0 \leq \quad & -[A_k]_{ij} + l_{ij}\delta_{ij}, && i, j = 1, \ldots, m, i \neq j. && (2.28) \end{aligned}$$

DOI:10.15774/PPKE.ITK.2015.008

The function summing the nonzero reaction rate coefficients is:

$$h(\delta) = \sum_{i,j=1 i \neq j}^{m} \delta_{ij}. \tag{2.29}$$

By maximizing (minimizing) $h(\delta)$, we are able to compute realizations with maximal (minimal) number of reactions, i.e. the dense (sparse) realizations of a canonical CRN.

As Section 3.3 will illustrate that sparse realizations are structurally non-unique, meanwhile with a fixed complex and constraint set, the dense realization is unique and determines a superstructure of all possible realizations [114]. In one hand, it is important to note that the computation of constrained dense realizations can be traced back to a series of pure LP steps [114] therefore it can be performed in polynomial time. On the other hand, the computation of a sparse realization without any prior knowledge about maximal and minimal number of reaction in the reaction graph still requires integer variables and MILP computations.

From now on, we will denote the number of reactions in the dense and sparse realizations with R_d and R_s, respectively.

2.3.2 Computing Constrained Realizations

We remark that it is straightforward to extend the notions of dense and sparse realizations to the constrained case, when some of the mathematically possible reactions are a priori excluded from the reaction network by setting the appropriate elements of A_k to zero. The simple constraint set denoted by $\mathcal{K}$ used for the exclusion of selected reactions from the CRN is given by:

$$\mathcal{K} = \{[A_k]_{i_1,j_1} = 0,\ \ldots,\ [A_k]_{i_s,j_s} = 0\}, \tag{2.30}$$

where s is the number of individual constraints, and $i_k \neq j_k$ for $k = 1, \ldots, s$. Then, a dynamically equivalent *$\mathcal{K}$-constrained realization* of a CRN (Y, A_k) is a reaction network $(Y, A_k^{\mathcal{K}})$ such that $Y \cdot A_k = Y \cdot A_k^{\mathcal{K}}$ and the prescribed constraints $\mathcal{K}$ in the form of Equation (2.30) are fulfilled for $A_k^{\mathcal{K}}$. A dynamically equivalent *$\mathcal{K}$-constrained dense realization* of a chemical reaction network (Y, A_k) is a $\mathcal{K}$-constrained realization that contains the maximal number of nonzero elements in $A_k^{\mathcal{K}}$.

DOI:10.15774/PPKE.ITK.2015.008

Similarly, a $\mathcal{K}$-*constrained sparse realization* is a $\mathcal{K}$-constrained realization with the minimal number of nonzeros in $A_k^{\mathcal{K}}$. Naturally, a dynamically equivalent constrained realization may not exist for certain constraint sets, therefore the existence of such realizations must be checked through the feasibility of the corresponding optimization problem.

2.3.3 Computing Core Reactions and Core Complexes

Besides maximal (minimal) number of positive off-diagonal elements in matrix A_k, the dynamically equivalent realizations share some other common structural elements that are structurally invariant. The so-called *core complexes* and *core reactions* are examples of such elements [114].

The *core complexes* are those vertices of the reaction graph that are present as reactants or products in any dynamically equivalent realization of a given CRN. It is easy to see that a complex is non-reacting (or isolated) in a CRN realization, if both the corresponding row and column of A_k contains only zeros (i.e. there are no incoming or outgoing directed edges to/from that complex in the reaction graph). Based on this, we can formulate a simple test to find core complexes.

The complex C_z is a core complex if and only if the constraint

$$\sum_{\substack{i=1\\ i\neq z}}^{m} [A_k]_{zi} + \sum_{\substack{j=1\\ j\neq z}}^{m} [A_k]_{jz} = 0 \tag{2.31}$$

together with Equations (2.21)-(2.23) is infeasible. Since no integer variable is involved in this constraint, core complexes can be found by using linear programming (LP).

A reaction $C_i \to C_j$ of a CRN is called a *core reaction*, if it is present—possibly with different rate coefficients—in any dynamically equivalent realization of a kinetic system. Whether a reaction belongs to the set of core reactions or not, can be tested with simple linear programming, too. Therefore, the reaction $C_i \to C_j$ is a core reaction if and only if Equations (2.21)-(2.23) with

$$[A_k]_{ji} = 0 \tag{2.32}$$

DOI:10.15774/PPKE.ITK.2015.008

yields an infeasible LP problem. The Algorithm 2 establishes core reaction set ($\mathcal{R}_C$) by testing the feasibility of a series of $\mathcal{K}$-constrained realizations defined by the off-diagonal elements of A_k.

ALGORITHM 2: The Algorithm computes the set of core reactions, $\mathcal{R}_c$, in a kinetic system defined by Y, A_k.

Input : Y, A_k
Output: $\mathcal{R}_C$

$\mathcal{R}_C := \emptyset$;
for each $(i, j) \in$ *OffDiagonal*(A_k) **do**
  $\mathcal{K} := \{[A_k]_{ij} = 0\}$;
  $\mathcal{R}(Y, A_k^{\mathcal{K}})$:= ComputeKconstrainedReal($\mathcal{R}(Y, A_k)$, $\mathcal{K}$) `// Procedure from Section 2.3.2` ;
  if $|\mathcal{R}(Y, A_k^{\mathcal{K}})| > 0$ **then**
    $\mathcal{R}_C := \mathcal{R}_C \cup (C_j \to C_i)$;
  end
endfor
return $\mathcal{R}_C$

In Algorithm 2, the procedure *OffDiagonal* finds the nonzero off-diagonal elements of the argument and gives back their indexes as a set.

Finally, we introduce some notation that we will use in the later chapters. Let us denote the number of core reactions and core complexes with R_c and N_c, respectively.

2.3.4 Introductory example

To give an example of a polynomial kinetic system we use a biomolecular model reported in [76]. In this example a dynamical model of positive feedback motif is used, the biomolecular mechanism behind the model and the detailed explanation for the dynamical behavior can be found in the original paper [76].

This model tracks the concentrations of a monomer protein (x_1), its dimer form (x_2), also its mRNA (x_5). The promoter which regulates the production of the protein is also included in the model (x_3 is the unoccupied promoter and x_4 is the occupied promoter). The protein (x_1) is made at a basal rate given by the concentration of x_4 then, this protein forms a dimer (x_2) and binds back to the promoter, forming an occupied promoter (x_4). This loop acts as a positive feedback and accelerates the production of the protein.

DOI:10.15774/PPKE.ITK.2015.008

The kinetic polynomial system representing this dynamics is given as

$$\begin{aligned}
\dot{x}_1 &= -2k_1x_1^2 + 2k_2x_2 + k_9x_5 - k_8x_1 \\
\dot{x}_2 &= k_1x_1^2 - k_2x_2 - k_3x_2x_3 + k_4x_4 \\
\dot{x}_3 &= -k_3x_2x_3 + k_4x_4 \\
\dot{x}_4 &= k_3x_2x_3 - k_4x_4 \\
\dot{x}_5 &= k_5x_4 + k_6x_3 - k_7x_5.
\end{aligned} \tag{2.33}$$

The following parameter values were used for the numerical computations

$$\begin{aligned}
&k_1 = 1,\ k_2 = 1,\ k_3 = 1,\ k_4 = 1,\ k_5 = 1, \\
&k_6 = 0.025,\ k_7 = 0.1,\ k_8 = 0.05,\ k_9 = 0.5.
\end{aligned} \tag{2.34}$$

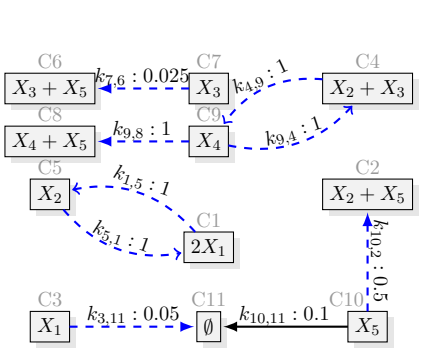
Using Algorithm 1, we can build the so-called canonical realization of the kinetic system in Equation (2.33). Then, using the procedures presented in Section 2.3, we can calculate realizations with prescribed properties.

Figure 2.1 shows two dynamically equivalent realizations of Equation (2.33), the reaction graph on the right panel shows the dense realization of the kinetic system. The core reactions which are structurally invariant under dynamical equivalence are shown in blue while non-core reactions are shown in black.

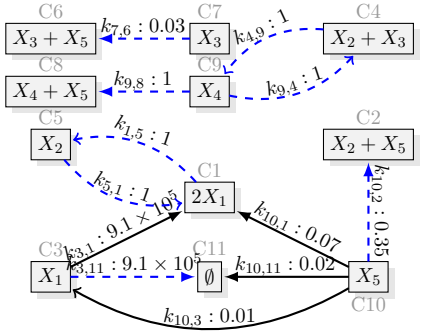
Moreover, we can calculate the linkage classes and deficiency of the two CRNs and demonstrate that they are realization properties of a kinetic system. The graph in Figure 2.1a has 4 linkage classes and its deficiency is 3, whereas the dense realization depicted in Figure 2.1b has 3 linkage classes and the deficiency is 4.

As we can see from this example, we can apply computational methods to calculate those realizations that fulfill predefined properties such as density or sparsity, weak reversibility, etc. Further details of these methods can be found in [113, 114, 116].

DOI:10.15774/PPKE.ITK.2015.008



(A) A graph representation of the positive feedback motif based on Equation (2.33)



(B) This graph shows the dense realization of the positive feedback motif. The graph was calculated via MILP optimization introduced in Section 2.3.1

FIGURE 2.1: Two dynamically equivalent realizations of the positive feedback motif. The blue edges represent the core reactions which are structurally invariant under dynamical equivalence. The black edges are the non-core reactions. The reader can notice that the graph on panel (a) is a subgraph of the graph on panel (b), this property was proven in [61].

2.4 Estimating Parameters of Kinetic Systems

One of the objectives of this thesis is to construct and validate a kinetic model for an *in vitro* transcription/translation system based on laboratory measurements which is introduced in Chapter 5. Parameter estimation of this kinetic system is a fundamental part of this task, therefore the essentials of this technique are summarized in this Section.

Let us assume we have a general nonlinear state space model where we can measure a projection (usually a subset) of the state variables denoted by $y^m(t,\theta) \in \mathbb{R}^r$

$$\begin{aligned} \dot{x} &= f(x(t),\theta) \\ x_0 &= x(0,\theta) \end{aligned} \tag{2.35}$$

where $f : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$, $\theta \in \mathbb{R}^p$. The parameter vector θ may not only include the dynamical parameters but also the unknown initial conditions, $x(0,\theta)$.

Often, not all state variables can be measured directly; therefore an observation function $h : \mathbb{R}^n \to \mathbb{R}^r$ defines which properties of the system are possible to measure. Moreover, the output of the systems can be measured at a certain frequency, thus we have the

DOI:10.15774/PPKE.ITK.2015.008

output of the system at discrete time instance.

$$y^m(k,\theta) \quad = \quad h(x(k),\theta),\ k=1,\ldots,N \tag{2.36}$$

where k is time instance of the output observation and N denotes the number of measurements.

Moreover, measurements have additive noise

$$y(k) \quad = \quad y^m(k,\theta^*)+\varepsilon_k,\ k=1,\ldots,N \tag{2.37}$$

where $\theta^* \in \mathbb{R}^p$ is the true parameter vector.

It is commonly assumed that the measurement error has Gaussian distribution with zero mean and σ_k is the variance

$$\epsilon_k \sim \mathcal{N}(0,\sigma_k),\ k=1,\ldots,N. \tag{2.38}$$

The goal of the parameter estimation is to find a suitable vector of parameters that generates the minimal distance between the model output and the measurements

$$e(k,\theta) := y^m(k,\theta)-y(k),\ k=1,\ldots,N. \tag{2.39}$$

Multiple metrics exist to measure the distance between the model output and the measurements, for example quadratic or absolute distance can be used [70]. The most commonly used one is the quadratic distance.

Once we have a way to measure the distance between the model output and the measurements at each time instance, we can define a function that gives a real number to each parameter vector

$$C(\theta) = e(\theta)^T Q e(\theta) \tag{2.40}$$

where $e(\theta) \in \mathbb{R}^N$ and $Q \in \mathbb{R}^{N\times N}$ is the weight matrix. Through the weight matrix, it is possible to assign different weight to each time instance. Often, the standard deviation

DOI:10.15774/PPKE.ITK.2015.008

of the measurements is used as weights

$$Q = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma_N^2} \end{bmatrix}. \tag{2.41}$$

If we use the weight matrix defined in Equation (2.41) and quadratic distance for Equation (2.39), then the cost function can be defined as

$$C(\theta) = \sum_{k=1}^{N} \sum_{j=1}^{r} \frac{(y_j(k) - y_j^m(k,\theta))^2}{\sigma_{kj}^2} \tag{2.42}$$

where $y_j(k)$ is the jth element of the measurement vector at time instance k. With all these assumptions, this is called the least squares optimization problem for dynamical systems.

In case of a kinetics system, the general state space model gets the following form

$$\begin{aligned} \dot{x} &= M \cdot \psi(x) \\ y^m(k,\theta) &= h(x(k),\theta),\ k = 1,\ldots,N. \end{aligned} \tag{2.43}$$

The vector $\theta \in \mathbb{R}^p$ denotes the parameters of the model. Usually, among other case specific parameters, the parameter vector includes the kinetic rate coefficients and the unknown initial concentrations of the species in the model.

Let assume that $h(x,\theta) = x$, i.e. all the state variables are directly measured. Then, the last step in this process is to classify the model structure regarding the linearity of the parameters.

Definition 2.3. A model structure is linear in its parameters if its outputs satisfy the superposition principle with respect to its parameters [126]:

$$\forall(\lambda,\mu) \in \mathbb{R}^2, \forall t \in \mathbb{R}^+, \quad y^m(t, \lambda\theta_1 + \mu\theta_2) = \lambda y^m(t,\theta_1) + \mu y^m(t,\theta_2). \tag{2.44}$$

According to this definition the model structure in Equation (2.43) with observation function $h(x,\theta) = x$ is linear in the coefficients contained in M, hence many standard

DOI:10.15774/PPKE.ITK.2015.008

parameter estimation techniques can be used [126]. However, the inference often remains a challenging task because of poor excitation of the dynamics, bad measurement quality or the lack of structural identifiability as it was detailed in Section 1.3.

In addition to these challenges, numerical issues may arise when minimizing Equation (2.42), this is due to the fact that the evaluation of the cost function requires simulation of the dynamical system in each step. Namely, the ODE defined in Equation (2.43) is solved numerically with relative and absolute tolerances on the integration error. At the same time the cost function defined in Equation (2.42) is usually part of a gradient based algorithm. When the gradient is calculated with finite differences, the gradient calculation may interfere with the integration error from the ODE solver and cause convergence problems.

A number of solutions exist to this problem, the first one is setting the ODE error tolerance and the finite difference step size apart by several order of magnitude, although this might lead to significant slow down on the solution of the ODE. The second one is the analytical calculation of the gradients by an automatic differentiation package, e.g. with CASAdi [4]. The third possibility is using gradient-free optimization methods, such as pattern search [11].

2.5 Structural Identifiability Analysis of Mathematical Models

Once we have selected a model structure like the one in Equation (2.35), the question of parameter identifiability has to be considered, too, i.e. whether it is theoretically possible to determine the model parameters based on the model structure and the observables ($h(x, \theta)$).

Definition 2.4. A parameterized model in Equation (2.35) is called structurally globally identifiable if

$$h(x, \theta') = h(x, \theta'') \quad \forall t \implies \theta' = \theta'', \tag{2.45}$$

where $h(x, \theta)$ is the observed output with parameter vector θ.

DOI:10.15774/PPKE.ITK.2015.008

Roughly speaking, a structurally non-identifiable system may produce exactly the same output for different parameterizations. Definition 2.4 has a local version where structural identifiability is only fulfilled for a neighborhood of θ in the parameter space, hence the name structurally locally identifiable.

There are many approaches to check structural identifiability of a nonlinear system and [26] contains a comprehensive survey of them. Recently, in [7] a different approach to structural identifiability was reported which is suitable for moderate sized problems as well, but this approach only gives local identifiability of the parameters. Also, procedures have been developed to check identifiability of certain type of kinetic systems [29, 33, 78].

In this thesis the generation series approach combined with identifiability tableau is used, which are fully implemented in the GenSSI software toolbox [27].

The generating series can be written where the observables are expanded in a series w.r.t inputs and time. The coefficients of the series are $h(x_0, \theta)$ and Lie derivatives of the observables along f evaluated at the initial time instance.

The Lie derivative of h along the vector field f is given by

$$L_f h(x, \theta) = \sum_{j=1}^{n} \frac{\partial h(x, \theta)}{\partial x_j} f_j(x(t), \theta) \tag{2.46}$$

where n is the number of states. If we have input to our system, then the Lie derivatives of the input function must be calculated as well [33]. Generally, if the number of parameters are larger than the number of first order Lie derivatives, then we need to calculate higher orders Lie derivatives as well

$$L_f^k h(x, \theta) = \sum_{j=1}^{n} \frac{\partial (L_f^{k-1} h(x, \theta))}{\partial x_j} f_j(x(t), \theta). \tag{2.47}$$

Let us denote by d the maximum order of the Lie derivatives. A disadvantage of the generating series approach is that there is no upper bound on d to determine structural identifiability. The lack of such a bound offers only sufficient, but not necessary condition to identifiability [26].

DOI:10.15774/PPKE.ITK.2015.008

A vector, denoted as $s(\theta)$, consists the Lie derivatives calculated up to dth order evaluated at the initial time instance and $h(x_0, \theta)$ can be formulated. The length of this vector is given by $n \times d$ where n is the number of state variables.

Based on Definition 2.4, one can try to solve the $s(\theta') = s(\theta'')$ symbolic equation for the parameters of the system. However, it is sufficient to solve only for $s(\theta') = c$ where c is an arbitrary constant vector (check [33] for the proof). Thus, we end up with a system of nonlinear algebraic equations. If this system of nonlinear equations has a unique solution then the model is structurally globally identifiable. If many distinct solutions exist the model is only structurally locally identifiable. Finally, if there is no solution, the model is structurally unidentifiable.

The identifiability tableau is a graphical representation of the non-zero elements of the Jacobian of the series coefficients with respect to the parameters. This representation gives us a graphical aid to determine structural identifiability and also the underlying system of nonlinear equations ($s(\theta) = c$) can be further reduced before solving it. The couple of these features are listed below.

If a column for the corresponding parameter is empty, then this tells us a parameter may not be identifiable. Since there is no upper bound for the order of the required Lie derivatives for identifiability, the number of non-zero coefficients are, in theory, infinite. Hence, in this setup a parameter is only non-identifiable if the higher order series coefficients are demonstrated to be zero [13].

If the rank of the Jacobian is equal to the number of parameters, then at least the model is structurally locally identifiable. Further reduction techniques exist for the identifiability tableau, thus for the underlying system of equations, which help the solvability of the system of nonlinear equations defined by the coefficients of the generating series. These techniques reported in [26] and fully implemented in the GenSSI Matlab toolbox [27].

It should be noted that calculating the generating series and solving the system of nonlinear equations formulated from the coefficients is computationally intensive, and the computation time rapidly grows with the increasing number of parameters checked by the algorithm. Hence, this method is limited to models with low number of parameters (generally less then about 20).

DOI:10.15774/PPKE.ITK.2015.008

Chapter 3

Computing All Dynamically Equivalent Sparse Chemical Reaction Network Structures

It is known that the state-variables of kinetic systems are always nonnegative valued, since they commonly describe the evolution of species concentrations in time. It is straightforward to show therefore that kinetic models mathematically belong to the family of nonnegative systems for which the nonnegative orthant is dynamically invariant [53]. However, this important property is not necessarily an obstacle to the kinetic description of e.g. electric or mechanical systems: Firstly, in [94], the so-called *X-factorable transformation* is introduced that (together with an appropriate coordinates-translation) allows the kinetic representation of a wide class of dynamical systems. The second possibility is a coordinates-translation followed by a state-dependent time-rescaling, that is also suitable to transform originally non-kinetic models into kinetic form [54, 117]. Both of these approaches will be applied in this chapter for the kinetic representation of general nonlinear systems.

The main purpose of this chapter is to study the structural non-uniqueness of the kinetic realizations of dynamical systems using the example of the well-known Lorenz system. For this, an optimization-based approach is proposed for effectively computing all dynamically equivalent sparse structures. These structures contain minimum number of reactions and thus provide a minimal parametrization of the system in terms of the

DOI:10.15774/PPKE.ITK.2015.008

reaction rate coefficients. Generally, minimal parametrization is often goal of the model building, because it can describe the dynamics with minimum number of parameters. Specifically in molecular biology it is often assumed the reaction network behind the observed phenomenon is 'sparse' [12, 21].

The kinetic realizations of certain chaotic systems have been studied previously in the literature. In [83], a kinetic model of the Lorenz system was given by applying variable-translation and the assumption of slow reaction steps and constant concentrations of certain species. The nuclear spin generator (NSG) system is examined and transformed into chemical forms in [128] using the methods published in [94] and [83], respectively. The first deterministic kinetic models of Chua's circuit were presented in [67] and [129], while a stochastic simulator was described in [68].

The structure of this chapter is the following. In Section 3.2, the notations and computational background for kinetic realizations of nonlinear systems are given. The new contributions related to the applied methodology and the computational results can be found in Sections 3.1 and 3.3, respectively, while Section 3.4 contains the conclusions of this chapter.

3.1 Computation of all possible sparse structures

In this Section, we present a new computational approach to calculate all possible dynamically equivalent sparse CRN structures. As a reminder, a sparse realization of a CRN means that the reaction graph has the minimum number of edges (reactions) while dynamically equivalent to the canonical CRN. Conversely, the dense realization has the maximum number of edges in the graph and also dynamically equivalent to the canonical CRN.

Firstly, we describe the main ideas behind the applied computation method. We take advantage of the fact that the reaction graph of the dense realization of a CRN is unique, and it contains the structures of all possible dynamically equivalent realizations as subgraphs (see Section 2.3.1). The first step is the computation of a sparse and the dense realization of the studied kinetic system to determine R_s and R_d, namely the number of reactions in the sparse and dense realizations, respectively. Then, we extract all possible sparse structures by 'pulling down' the dense one. For this, it is clear that we have

DOI:10.15774/PPKE.ITK.2015.008

to remove as many reactions from the dense realization (if possible) as the difference between the number of reactions in the dense and sparse realizations, i.e. $R_d - R_s$. By 'removing' a reaction from a CRN, we mean that the corresponding reaction rate coefficient in A_k is made zero while maintaining dynamical equivalence with the initial kinetic system, i.e. we calculate a $\mathcal{K}$-constrained realization of a CRN.

We recall that, the core reactions are structurally invariant reactions in the dynamically equivalent Kinetic Systems (see Section 2.3.1 for details). Hence, the core reactions are not removable from any realization, then the maximal theoretical number of possibilities to be checked is

$$N_{max} = \begin{pmatrix} R_d - R_c \\ R_d - R_s \end{pmatrix} = \frac{(R_d - R_c)!}{(R_d - R_s)!(R_s - R_c)!}. \tag{3.1}$$

This means that in the worst case, we have to check N_{max} possibilities for dynamical equivalence, where each checking requires the solution of an LP problem with constraints (2.21)-(2.25) and (2.30). As we will see later even in the case of the Kinetic Lorenz system, which is a small reaction network, it might be computationally intractable to check all possibilities individually. To ease the computational burden, we identify the set of reaction pairs (denoted by $\mathcal{R}_e^2 \subset (\mathcal{R}(Y, A_k^d) \setminus \mathcal{R}_c) \times (\mathcal{R}(Y, A_k^d) \setminus \mathcal{R}_c)$) which are not part of the core reaction set, but if any reaction pair from $\mathcal{R}_e^2$ is not present in a given reaction graph, then the corresponding dynamically equivalent realization cannot be sparse. In other words, the exclusion of any reaction pair from $\mathcal{R}_e^2$ always implies the inclusion of more than 2 other reactions in the reaction graph, and this clearly violates the sparsity constraint. Based on that, we can omit those realizations from the systematic search that do not contain any pair from $\mathcal{R}_e^2$, and this can drastically reduce the overall computation time. The computational background for determining $\mathcal{R}_e^2$ is simple: any pair of distinct reactions $\mathcal{R}_p = (C_i \to C_j, C_k \to C_l)$ belongs to $\mathcal{R}_e^2$ if and only if the constrained sparse realization not containing the elements of $\mathcal{R}_p$ contains more reactions than the original unconstrained sparse realization. This means that determining $\mathcal{R}_e^2$ requires $\begin{pmatrix} R_d - R_c \\ 2 \end{pmatrix}$ MILP optimizations steps.

Generally, we can investigate what would be the largest set of reactions which would help to reduce the search space for a given network size. For that we need to solve the following equation for K:

DOI:10.15774/PPKE.ITK.2015.008

$$\begin{pmatrix} R_d - R_c \\ R_d - R_s \end{pmatrix} = \begin{pmatrix} R_d - R_c \\ K \end{pmatrix}. \tag{3.2}$$

After simplification and rearrangements this leads to

$$K! = \frac{(R_d - R_s)!(R_s - R_c)!}{(R_d - R_c - K)!}, \quad K < R_d - R_c. \tag{3.3}$$

Using Stirling's approximation, we can rewrite the factorials as $\log(n!) \approx n\log(n) - n + O(\log(n))$ where $\log(n)$ is the natural logarithm of n and $O()$ is the Big O notation. In above equations we did not consider the computation time difference between one LP and one MILP problem for the given network size. Later on, we will check the crossing point for specific examples and compare the required computation time to establish $\mathcal{R}^j,\ j = 2, \ldots, K$.

Let us introduce the following additional notations. Let $\mathcal{P}(S)$ denote the power set of an arbitrary set S, and let $|S|$ be the cardinality of set S. Let $\mathcal{R}(Y, A_k)$ denote the set of reactions of a CRN with complex composition matrix Y and Kirchhoff matrix A_k. Furthermore, let $\mathcal{P}^k(S)$ denote the elements (sets) of $\mathcal{P}(S)$ containing k elements, i.e. $\mathcal{P}^k(S) = \{A \in \mathcal{P}(S) \mid |A| = k\}$. Now we can summarize the computation steps for determining all sparse realization structures with Algorithm 3. The sparse realizations with different structures are collected into the set called `Sparse_structs`.

For the generation of $\mathcal{R}^z$, we applied the fast algorithm described in Section 7.2.1.3 of [66]. We emphasize again that integer variables are required only in lines 3 and 11 of the above algorithm, and all other realizations can be computed by standard linear programming. We remark that after the initialization steps between line 1 and line 9, the necessary time for the subsequent computations can be pre-estimated quite precisely, knowing the average time required for a MILP or LP realization computation step.

3.2 Transformation of polynomial models into kinetic form

We address the transformation of general polynomial ODEs that are not in the form described in Equation (2.18) into kinetic form using two different methods. Namely, the state-dependent time-rescaling and the so called X-factorable transformation.

DOI:10.15774/PPKE.ITK.2015.008

ALGORITHM 3: Algorithm for computing all sparse dynamically equivalent realizations of a kinetic system.

Input : A non-negative polynomial system encoded by matrices M, B
Output: `Sparse_structs`

$\mathcal{R}(Y, A_k^c) :=$ BuildCanonicalReal(M, B) `// calling Algorithm 1` ;
$\mathcal{R}(Y, A_k^d) :=$ ComputeDenseReal$(\mathcal{R}(Y, A_k^c))$ `// Procedure from Section 2.3.1` ;
$\mathcal{R}(Y, A_k^s) :=$ ComputeSparseReal$(\mathcal{R}(Y, A_k^c))$ `// Procedure from Section 2.3.1` ;
$R_d := |\mathcal{R}(Y, A_k^d)|$;
$R_s := |\mathcal{R}(Y, A_k^s)|$;
$\mathcal{R}_c :=$ ComputeCoreReactions$(\mathcal{R}(Y, A_k^c))$ `// Procedure from Section 2.3.3`;
$R_c := |\mathcal{R}_c|$;
$\mathcal{R}_e^2 := \emptyset$;
`Sparse_structs`$:= \emptyset$;
for each $\{C_i \to C_j, C_k \to C_l\} \subseteq \mathcal{R}(Y, A_k^d) \setminus \mathcal{R}_c$ **do**
  $\mathcal{K}_1 := \{[A_k]_{ji} = 0,\ [A_k]_{lk} = 0\}$;
  $\mathcal{R}(Y, A_k^{\mathcal{K}_1}) :=$ ComputeKconstrainedSparseReal$(\mathcal{R}(Y, A_k^d), \mathcal{K}_1)$ `// Procedure from Section 2.3.2` ;
  if $|\mathcal{R}(Y, A_k^{\mathcal{K}_1})| > R_s$ **then**
    $\mathcal{R}_e^2 := \mathcal{R}_e^2 \cup (C_i \to C_j, C_k \to C_l)$;
  end
endfor
$z := R_d - R_s$;
$\mathcal{R}^z := \mathcal{P}^z(\mathcal{R}(Y, A_k^d) \setminus \mathcal{R}_c)$;
for each $\mathcal{R}_i = \{C_{i_1} \to C_{j_1}, \ldots, C_{i_z} \to C_{jz}\} \in \mathcal{R}^z$ **do**
  if $\mathcal{R}_i \cap \mathcal{R}_e^2 = \emptyset$ **then**
    $\mathcal{K}_2 := \{[A_k]_{j_1 i_1} = 0, \ldots, [A_k]_{j_z i_z} = 0\}$;
    $\mathcal{R}(Y, A_k^{\mathcal{K}_2}) :=$ ComputeKconstrainedReal$(\mathcal{R}(Y, A_k^d), \mathcal{K}_2)$;
    if $|\mathcal{R}(Y, A_k^{\mathcal{K}_2})| > 0$ **then**
      `Sparse_structs`$:=$`Sparse_structs` $\cup\ \mathcal{R}(Y, A_k^{\mathcal{K}_2})$;
    end
  end
endfor
return `Sparse_structs`

First of all, we have to ensure that the operating domain of the system's trajectories remain in the positive orthant. For this, the following simple translation of the state variables is used, if the positive orthant is not invariant for the original system's dynamics:

$$\bar{x} \quad = \quad x + w, \tag{3.4}$$

where the coordinates of $w = \begin{bmatrix} w_1 & w_2 & \ldots & w_n \end{bmatrix}^T \in \mathbb{R}^n$ are sufficiently large so that all trajectories of the translated system remain in the positive orthant if started from the studied initial conditions.

DOI:10.15774/PPKE.ITK.2015.008

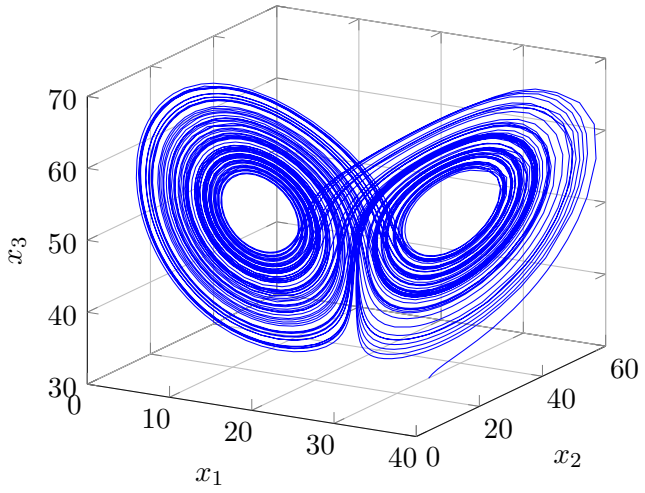


FIGURE 3.1: Phase-plane plot of the kinetic Lorenz system obtained via state-dependent time-rescaling.

3.2.1 State-dependent time-rescaling

Time-rescaling is a common operation in physical sciences. Its particular form depending on the positive state variables can be used e.g. for stability analysis of nonlinear systems (see [46, 117]) or for motion control [112]. In our case, the relationship between the original and transformed time-scales is defined by

$$dt = \prod_{i=1}^{n} {x_i}^{\chi_i} d\tau, \tag{3.5}$$

where t and τ are the time variables of the original and rescaled systems, respectively, and $\chi_i \in \{0, 1\}$ for $i = 1, \ldots, n$. It is important to stress that the positivity of the state variables implies that τ is a strictly monotonously increasing (and therefore invertible) function of t, and the phase-portraits of the original and rescaled models are identical. We will use the following notations for the derivative of x with respect to t and τ, respectively: $\dot{x} = \frac{dx}{dt}$, $x' = \frac{dx}{d\tau}$.

Clearly, the ODEs in Equation (2.17) corresponding to the new time-scale have the following form:

$$x' = F(x) \prod_{i=1}^{n} {x_i}^{\chi_i}, \tag{3.6}$$

from which it is easy to see that χ_i can always be chosen in such a way that the negative cross-effects are eliminated in the time-rescaled system.

DOI:10.15774/PPKE.ITK.2015.008

3.2.2 X-factorable transformation

Another method for transforming a polynomial system into kinetic form was proposed in [94]. A polynomial system given in Equation (2.17) is called *X-factorable* if its the right hand side can be factorized as

$$F(x) \quad = \quad D(x)G(x) \tag{3.7}$$

where $G \in \mathbb{R}^n \to \mathbb{R}^n$ is a polynomial vector field and $D(x) = \text{diag}\{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{n\times n}$ is a diagonal matrix. It can be shown that the positive orthant for any X-factorable system is invariant for the dynamics [53], and any X-factorable system is kinetic. Well-known examples of X-factorable systems are classical Lotka-Volterra models [118].

Let us suppose that the solutions of Equation (2.17) from a given set of initial conditions are strictly positive (possibly after a variable translation like in Equation (3.4)) but the model itself is not X-factorable. Then, we assign the following transformed X-factorable model to the original one given by Equation (2.17):

$$\dot{\tilde{x}} \quad = \quad D(\tilde{x})F(\tilde{x}), \tag{3.8}$$

where again, $D(\tilde{x}) = \text{diag}\{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$. It is obvious that the phase-portraits of the two systems in Equations (2.17) and (3.8) are not identical in this case. However, under mild conditions, the system trajectories in the interior of the positive orthant are 'sufficiently similar' in the sense that the "distortion is weak or negligible for trajectories far from the boundary" of the positive orthant, while "a substantial compression of trajectories occurs close to the boundary" [94]. Moreover, the behavior of the dynamical system defined in Equation (3.8) around the strictly positive equilibrium points is qualitatively dynamically equivalent to that of Equation (2.17) as it is described in [94]. The dynamics of the Lorenz system after the translation and X-factorable transformation can be seen in Figure 3.2.

DOI:10.15774/PPKE.ITK.2015.008

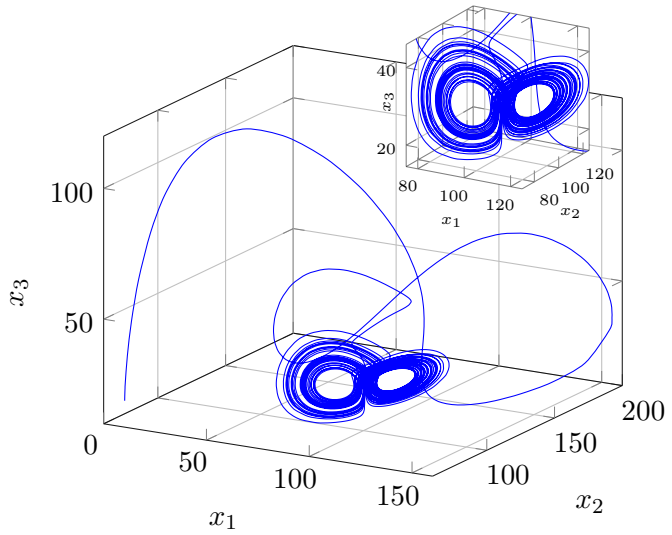


FIGURE 3.2: Phase-plane plot of the Kinetic Lorenz system obtained using the X-factorable transformation. The inset shows the local behavior of the Lorenz attractor.

3.3 Results: sparse kinetic realizations of the Lorenz system

The starting point is the classical set of equations corresponding to the Lorenz system:

$$\begin{aligned} \dot{x}_1 &= \sigma(x_2 - x_1) \\ \dot{x}_2 &= \rho x_1 - x_2 - x_1 x_3 \\ \dot{x}_3 &= x_1 x_2 - \beta x_3 \end{aligned} \tag{3.9}$$

with parameter values $\sigma = 10, \rho = 28, \beta = 8/3$ that are known to lead to chaotic behavior. It is also known that the nonnegative orthant is not invariant for the original dynamics described in Equation (3.9), therefore as a first step a coordinates-translation was applied with $w = \begin{bmatrix} 24 & 25 & 26 \end{bmatrix}^T$. The principles for selecting the elements of w were the following. First, the numbers should be as small as possible while allowing the shift of the studied operating domain to the strictly positive orthant. Second, these numbers should be different in order to avoid any monomial cancellations later in the

DOI:10.15774/PPKE.ITK.2015.008

kinetic models. The translated model reads:

$$\begin{aligned}
\dot{\bar{x}}_1 &= \sigma\bar{x}_2 - \sigma\bar{x}_1 + \sigma(w_1 - w_2) \\
\dot{\bar{x}}_2 &= (\rho + w_3)\bar{x}_1 - \bar{x}_2 + w_1\bar{x}_3 - \bar{x}_1\bar{x}_3 - \rho w_1 + w_2 - w_1 w_3 \\
\dot{\bar{x}}_3 &= \bar{x}_1\bar{x}_2 - w_2\bar{x}_1 - w_1\bar{x}_2 + w_1 w_2 - \beta\bar{x}_3 + \beta w_3
\end{aligned} \tag{3.10}$$

3.3.1 State-dependent time-rescaling

It is easy to see that the set of ODEs in Equation (3.10) is generally not kinetic. Therefore, we apply the most general time-scaling which contains all 3 state variables, namely $dt = \bar{x}_1\bar{x}_2\bar{x}_3 d\tau$. The rescaled kinetic equations are written as

$$\begin{aligned}
\bar{x}_1' &= \sigma\bar{x}_1\bar{x}_2^2\bar{x}_3 - \sigma\bar{x}_1^2\bar{x}_2\bar{x}_3 + \sigma(w_1 - w_2)\bar{x}_1\bar{x}_2\bar{x}_3 \\
\bar{x}_2' &= (\rho + c_3)\bar{x}_1^2\bar{x}_2\bar{x}_3 + (w_2 - w_1\rho - w_1 w_3)\bar{x}_1\bar{x}_2\bar{x}_3 - \bar{x}_1\bar{x}_2^2\bar{x}_3 - \bar{x}_1^2\bar{x}_2\bar{x}_3^2 + w_1\bar{x}_1\bar{x}_2\bar{x}_3^2 \\
\bar{x}_3' &= \bar{x}_1^2\bar{x}_2^2\bar{x}_3 - w_2\bar{x}_1^2\bar{x}_2\bar{x}_3 - w_1\bar{x}_1\bar{x}_2^2\bar{x}_3 + (w_1 w_2 + \beta w_3)\bar{x}_1\bar{x}_2\bar{x}_3 - \beta\bar{x}_1\bar{x}_2\bar{x}_3^2.
\end{aligned} \tag{3.11}$$

The simulated system trajectories of Equtation (3.11) are shown in Figure 3.1. Let us denote the species corresponding to the concentrations $\bar{x}_1, \bar{x}_2$ and $\bar{x}_3$ by X_1, X_2 and X_3, respectively. Then, Algorithm 1 for building the canonical structure (described in Subsection 2.2.3) generates the following set complexes for the canonical realization of Equation (3.11)

$$\begin{aligned}
&C_1 = X_1 + X_2 + X_3, C_2 = X_2 + X_3, C_3 = 2X_1 + X_2 + X_3, C_4 = X_1 + 2X_2 + X_3, \\
&C_5 = 2X_1 + 2X_2 + X_3, C_6 = X_1 + X_3, C_7 = X_1 + X_2 + 2X_3, C_8 = X_1 + 2X_2 + 2X_3, \\
&C_9 = 2X_1 + X_2 + 2X_3, C_{10} = 2X_1 + 2X_3, C_{11} = 2X_1 + X_2, C_{12} = X_1 + 2X_2, \\
&C_{13} = 2X_1 + 2X_2 + 2X_3.
\end{aligned} \tag{3.12}$$

The canonical reaction network corresponding to Equation (3.11) can be seen in Figure 3.3, which shows the reaction rate coefficients on the directed edges. In this chapter all figures showing CRN structures, the core reactions and core complexes are indicated by blue dashed arrows and gray rectangles, respectively. The six core reactions of the

DOI:10.15774/PPKE.ITK.2015.008

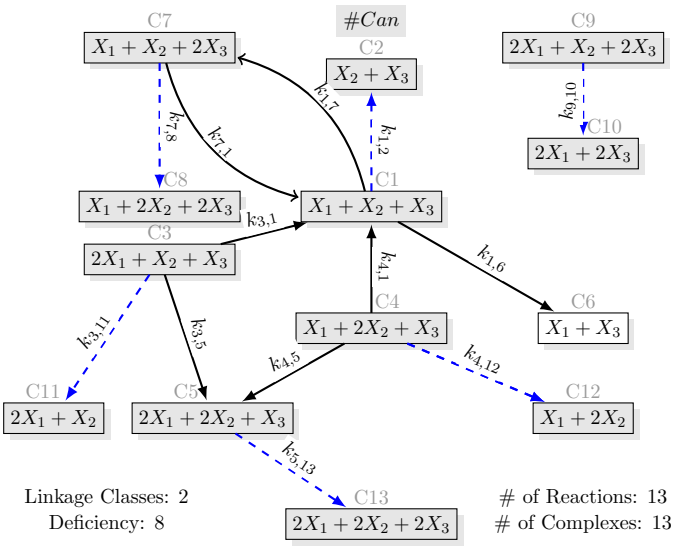


FIGURE 3.3: The canonical realization of the Lorenz system with state-dependent time-scaling. In this case, the canonical realization is also a sparse one. The parameters are: $k_{3,1} = \sigma$, $k_{4,1} = 1$, $k_{7,1} = \beta$, $k_{1,2} = \sigma$, $k_{3,5} = \rho + w_3$, $k_{4,5} = \sigma$, $k_{1,6} = |w_2 - w_1\rho - w_1w_3|$, $k_{1,7} = w_1w_2 + \beta w_3$, $k_{7,8} = w_1$, $k_{9,10} = \sigma$, $k_{3,11} = w_2$, $k_{4,12} = w_1$, $k_{5,13} = \sigma$.

Kinetic Lorenz system are the following:

$$C_1 \to C_2,\ C_3 \to C_{11},\ C_4 \to C_{12},\ C_5 \to C_{13},\ C_7 \to C_8,\ C_9 \to C_{10}.$$

Moreover, all complexes in Equation (3.12) are core complexes, except $C_6 = X_1 + X_3$. We remark that the canonical realization can be computed symbolically, but the optimization techniques for computing the forthcoming dense and sparse realizations are numerical, and they use the previously described model parameter values.

To compute all possible sparse structures of this CRN, we need the possible minimal (R_s) and maximal number (R_d) of reactions, the number of core reactions (R_c) and the number of core complexes (N_c), these are calculated with optimization methods described in Section 2.3.

In the case of the time-rescaled model, the Kinetic Lorenz System is characterized by $R_d = 51$, $R_s = 13$, $R_c = 6$, and $N_c = 12$. The value of N_{max} (i.e. the maximal number of required LP steps to check all possibilities) is $45,379,620$. This would require about 630 hours of computation time on a high-end desktop PC that we used in 2011, assuming approximately 0.05 sec for one LP step.

DOI:10.15774/PPKE.ITK.2015.008

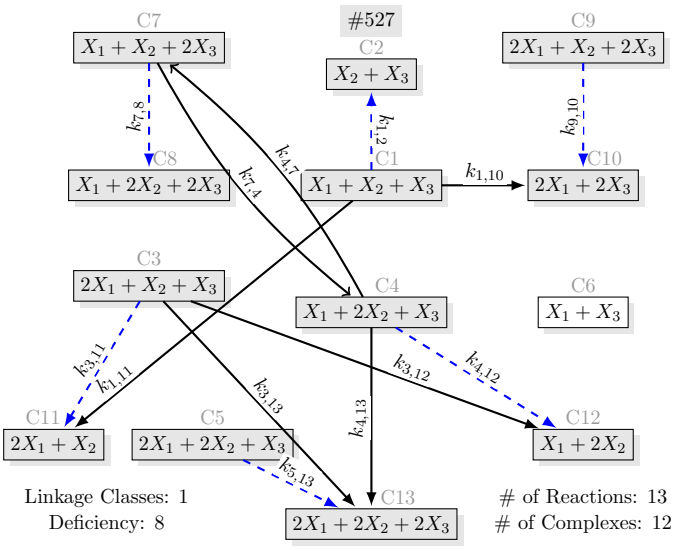


FIGURE 3.4: Sparse realization of the model in Equation (3.11) that contains only the core complexes. The reaction rate coefficients are: $k_{1,2} = 1882.7$, $k_{7,4} = 2.6667$, $k_{4,7} = 1$, $k_{7,8} = 21.333$, $k_{1,10} = 1271$, $k_{9,10} = 1$, $k_{1,11} = 601.67$, $k_{3,11} = 59$, $k_{3,12} = 10$, $k_{4,12} = 35$, $k_{3,13} = 44$, $k_{4,13} = 10$, $k_{5,13} = 1$.

In the case of the time-rescaled model in Equation (3.11), after the solution of 706 MILP optimization problems, which took approximately 1.2 hours, the set $\mathcal{R}_e^2$ is the following:

$$\begin{aligned}\mathcal{R}_e^2 \;=\; & \{(C_3 \to C_5, C_3 \to C_{13}), (C_4 \to C_5, C_4 \to C_{13}),\\ & (C_1 \to C_6, C_1 \to C_{10}), (C_1 \to C_6, C_1 \to C_{11})\}. \end{aligned} \tag{3.13}$$

Using $\mathcal{R}_e^2$, the search space was reduced to $442,454$ possibilities, that is a huge reduction compared to the initial which is more than 45 million.

We can investigate, if $\mathcal{R}_e^K$ with $K > 2$ would further reduce the computational burden. From Equation (3.3) we conclude that the upper bound for K in this example is around 8, but even $K = 3$ would require the solution of 8436 MILP optimization problems, which would take approximately 14 hours. Therefore, $K = 2$ is the optimal value for this network.

After checking the realization candidates with Algorithm 3, we found that 5376 different structures are valid dynamically equivalent sparse realizations, this procedure took about 6 hours on a desktop PC. Out of these, 504 are such that they contain only the 11 core complexes.

DOI:10.15774/PPKE.ITK.2015.008

3.3.2 X-factorable transformation

Besides the state-dependent time-scaling, we also investigated the 3-dimensional Lorenz system with the X-factorable transformation. Firstly, we performed a coordinates shift on the model in Equation (3.9) and then applied the X-factorable transformation. The resulting equations are

$$\begin{aligned}
\dot{\tilde{x}}_1 &= \sigma\tilde{x}_1\tilde{x}_2 - \sigma\tilde{x}_1^2 - \sigma(w_1 - w_2)\tilde{x}_1 \\
\dot{\tilde{x}}_2 &= (w_3 + \rho)\tilde{x}_1\tilde{x}_2 - \tilde{x}_2^2 + w_1\tilde{x}_2\tilde{x}_3 - \tilde{x}_1\tilde{x}_2\tilde{x}_3 + (w_2 - w_1\rho - c_1c_3)\tilde{x}_2 \\
\dot{\tilde{x}}_3 &= \tilde{x}_1\tilde{x}_2\tilde{x}_3 - w_1\tilde{x}_2\tilde{x}_3 - w_2\tilde{x}_1\tilde{x}_3 + (w_1w_2 + \beta w_3)\tilde{x}_3 - \beta\tilde{x}_3^2,
\end{aligned} \tag{3.14}$$

where $w = \begin{bmatrix} 100 & 101 & 1 \end{bmatrix}^T$, which is basically taken from [94], but w_2 is modified from 100 to 101 to avoid cancellation of the last monomial in the first equation.

Again, using Algorithm 1, the complexes of the canonical realization are given as

$$\begin{gathered}
C_1 = X_1, C_2 = 0, C_3 = 2X_1, C_4 = X_1 + X_2, C_5 = 2X_1 + X_2, C_6 = X_2, \\
C_7 = 2X_2, C_8 = X_1 + 2X_2, C_9 = X_2 + X_3, C_{10} = 2X_2 + X_3, C_{11} = X_1 + X_2 + X_3, \\
C_{12} = X_1 + X_3, C_{13} = X_3, C_{14} = 2X_3, C_{15} = X_1 + X_2 + 2X_3,
\end{gathered} \tag{3.15}$$

where the species X_1, X_2 and X_3 correspond to the state variables $\tilde{x}_1$, $\tilde{x}_2$ and $\tilde{x}_3$ of Equation (3.14), respectively. The reaction graph of the canonical realization is shown in Figure 3.5.

The list of core reactions is

$$C_1 \to C_2, C_6 \to C_2, C_{11} \to C_{15}, C_{13} \to C_{14}, \tag{3.16}$$

while the core complexes are C_1, C_2, C_4, C_6, C_{13}, C_{14} and C_{15}. The characteristic parameter values for this network—obtained via X-factorable transformation—are $R_s = 12$, $R_d = 44$, $R_c = 4$. and $N_c = 7$.

The value of N_{max} is $76,904,685$ that would require about 1068 hours of computation time to check each possible combination. To establish $\mathcal{R}_e^2$, we evaluated all the 780

DOI:10.15774/PPKE.ITK.2015.008

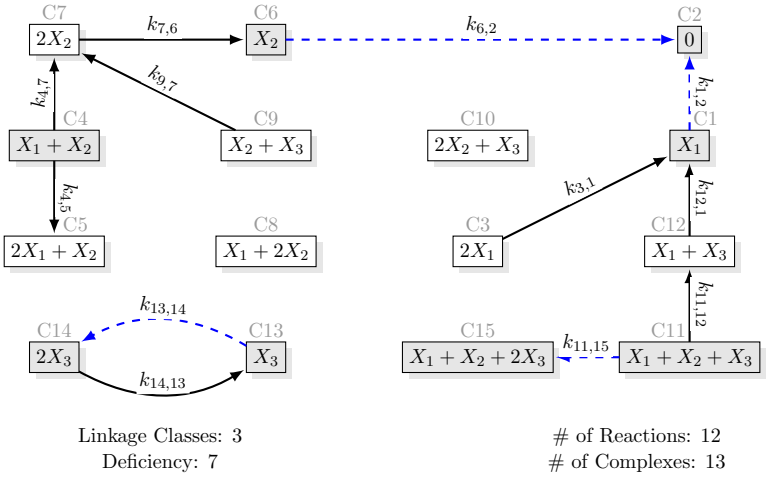


FIGURE 3.5: Canonical realization of the Lorenz system transformed with X-factorable transformation. This realization is sparse as well. The values of the rate coefficients are: $k_{1,2} = \sigma$, $k_{3,1} = \sigma$, $k_{4,7} = \rho + w_3$, $k_{4,5} = \sigma$, $k_{6,2} = w_1\rho + (w_1w_3 - w_2)$, $k_{7,6} = 1$, $k_{9,7} = 1$, $k_{11,15} = 1$, $k_{11,12} = 1$, $k_{12,1} = w_2$, $k_{13,14} = w_1w_2 + \beta w_3$, $k_{14,13} = \beta$.

possibilities, which took 1.3 hours, and obtained the following set of reaction pairs:

$$\begin{aligned}\mathcal{R}_e^2 \;=\; & \{(C_3 \to C_1, C_3 \to C_2), (C_{12} \to C_1, C_{12} \to C_3), (C_7 \to C_2, C_7 \to C_6),\\ & (C_{14} \to C_2, C_{14} \to C_{13}), (C_4 \to C_3, C_4 \to C_5), (C_4 \to C_5, C_4 \to C_8),\\ & (C_4 \to C_7, C_4 \to C_8), (C_9 \to C_7, C_9 \to C_{10})\}. \end{aligned} \tag{3.17}$$

After checking the possible sparse graph structures, we found that only 2864 do not contain any pair from $\mathcal{R}_e^2$, which is only 0.0037% of the original nearly 77 million possibilities. It was computationally tractable to check the remaining 2864 candidates, among which we found only 48 valid dynamically equivalent sparse realization structures in about 3 min on a desktop PC. In this case, there was no such realization that only contained the core complexes.

We also checked that, if we increase K in $\mathcal{R}_e^K$, then we can find all sparse solutions with fewer LP iterations. The upper bound for K is around 7, according to Equation (3.3). But similarly to the previous case, even $K = 3$ would require the solution of 9880 MILP optimization problems, which would take 16 hours. Therefore, we conclude that $K = 2$ is the optimal value for this network.

Finally, we can compare the computational time needed to compute all sparse structures for the two version of the kinetic Lorenz systems in Table 3.1.

DOI:10.15774/PPKE.ITK.2015.008

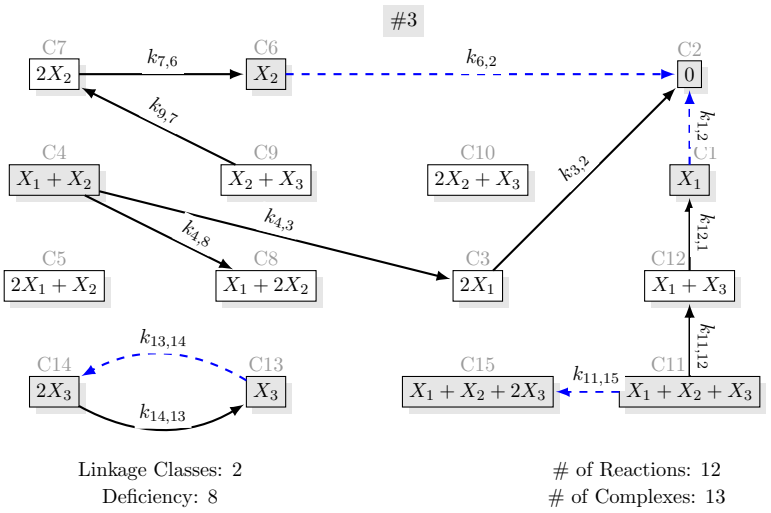


FIGURE 3.6: One of the 48 possible sparse realizations of Equation (3.14) that contains 13 complexes. The parameters are: $k_{12,1} = 101$, $k_{1,2} = 10$, $k_{3,2} = 5$, $k_{6,2} = 2799$, $k_{4,3} = 10$, $k_{7,6} = 1$, $k_{9,7} = 100$, $k_{4,8} = 39$, $k_{11,12} = 1$, $k_{14,13} = 2.6667$, $k_{13,14} = 10103$, $k_{11,15} = 1$.

Property	SD-TS	X-factorable
N_{max}	45,379,620	76,904,685
Time to check N_{max}	630 hours	1068 hours
$\lvert\mathcal{R}^2\rvert$	706	780
Time to compute $\mathcal{R}^2$	1.2 hours	1.3 hours
Reduced search space	442,454	2864
Time to check the remaining	6 hours	5 min
Total Computation Time	7.2 hours	1.3 hours

TABLE 3.1: Comparing the computational time required to compute all different sparse realizations for the kinetic Lorenz systems given by the two transformations. The time required to solve one LP and one MILP problem with the gives network size on our desktop PC is 0.05 sec and 6 sec, respectively. The 'SD-TS' and 'X-factorable' denote state-dependent time-scaling and X-factorable transformation, respectively.

Due to the large number of dynamically equivalent structures, the reaction graphs for all sparse realizations from both methods are provided in an electronic supplement that can be downloaded[1]. Therefore, only a few characteristic examples are included in this Chapter. The core reactions and core complexes are indicated by dashed arrows and gray boxes, respectively in the figures. The numbers of the complexes are written above the boxes containing the complexes. The unique identification numbers (serial numbers) of the sparse structures are indicated at the top of the figures. The isolated (unconnected) complexes are omitted from the models but they are drawn in the figures for easier comparison. The superstructure (i.e. the dense realization) for the time-scaling case is

[1] http://daedalus.scl.sztaki.hu/PCRG/works/Suppl2012_001.pdf

DOI:10.15774/PPKE.ITK.2015.008

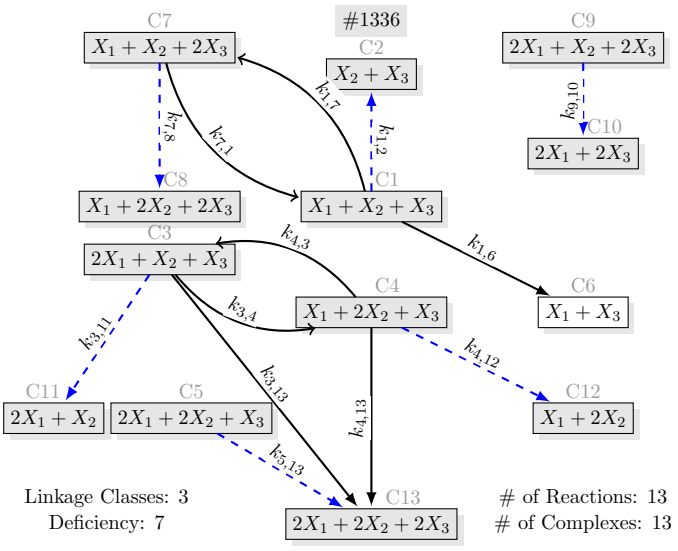


FIGURE 3.7: Sparse realization of the model in Equation (3.11) with 3 linkage classes. The reaction rate coefficients are: $k_{7,1} = 2.6667$, $k_{1,2} = 10$, $k_{4,3} = 1$, $k_{3,4} = 10$, $k_{1,6} = 1271$, $k_{1,7} = 699.33$, $k_{7,8} = 24$, $k_{9,10} = 1$, $k_{3,11} = 69$, $k_{4,12} = 33$, $k_{3,13} = 44$, $k_{4,13} = 9$, $k_{5,13} = 1$.

shown in Figure 3.8 and for the X-factorable case is shown in Figure 3.9. Figure 3.7 shows a sparse realization of (3.11) with the maximal number of linkage classes (three), while a sparse realization of (3.11) containing only the 12 core complexes can be seen in Figure 3.4. From the 48 sparse realizations of (3.14), Figure 3.6 shows one with the minimal number of complexes (13). The brief comparison of the time-rescaling and X-factorable transformation cases can be found in Table 3.2. As the table shows, we haven't found any weakly reversible realizations among the sparse ones. Moreover, the deficiencies of all obtained sparse CRNs are high (between 7 and 9) that is related to the dynamical complexity of the network.

3.4 Summary

This chapter have presented an algorithmic approach to calculate all distinct dynamically equivalent sparse realization structures of a kinetic system, the process has been illustrated with the well-known Lorenz System. The optimization-based computational framework proposed originally in [113] was successfully used to develop the proposed algorithm. The set of complexes for the kinetic Lorenz system realizations was generated using the procedure published in [56]. The original Lorenz system was transformed into a kinetic form using two different approaches known from literature: the state-dependent

DOI:10.15774/PPKE.ITK.2015.008

Feature	SD-TS	X-factorable
R_d	51	44
R_s	13	12
R_c	6	4
C_c	12	8
no. of complexes in the canonical realization	13	15
no. of valid sparse realizations	5376	48
no. of realizations containing only core complexes	504	0
minimal no. of linkage classes	1	1
maximal no. of linkage classes	3	3
no. of weakly reversible realizations	0	0
minimal deficiency	7	8
maximal deficiency	9	9

TABLE 3.2: Comparison table of the two approaches for transforming the Lorenz system into kinetic form. In the header line, 'SD-TS' and 'X-factorable' denote state-dependent time-scaling and X-factorable transformation, respectively.

time-rescaling that completely preserves the structure of the phase-space, and the X-factorable transformation. As a first step in the algorithm the structurally invariant components (i.e. the core reactions and core complexes) of the kinetic system have been determined. The studied realization computation problem is clearly of combinatorial nature, therefore an effective reduction of the search space was proposed based on the core reactions and the uniqueness of the dense realization. All different sparse structures are listed in an electronic supplement. The large number of valid solutions clearly illustrate the possible high degree of structural non-uniqueness of sparse chemical reaction networks. To the best of my knowledge, it has been the first attempt to enumerate all dynamically equivalent structures of a kinetic dynamical system with a given property.

DOI:10.15774/PPKE.ITK.2015.008

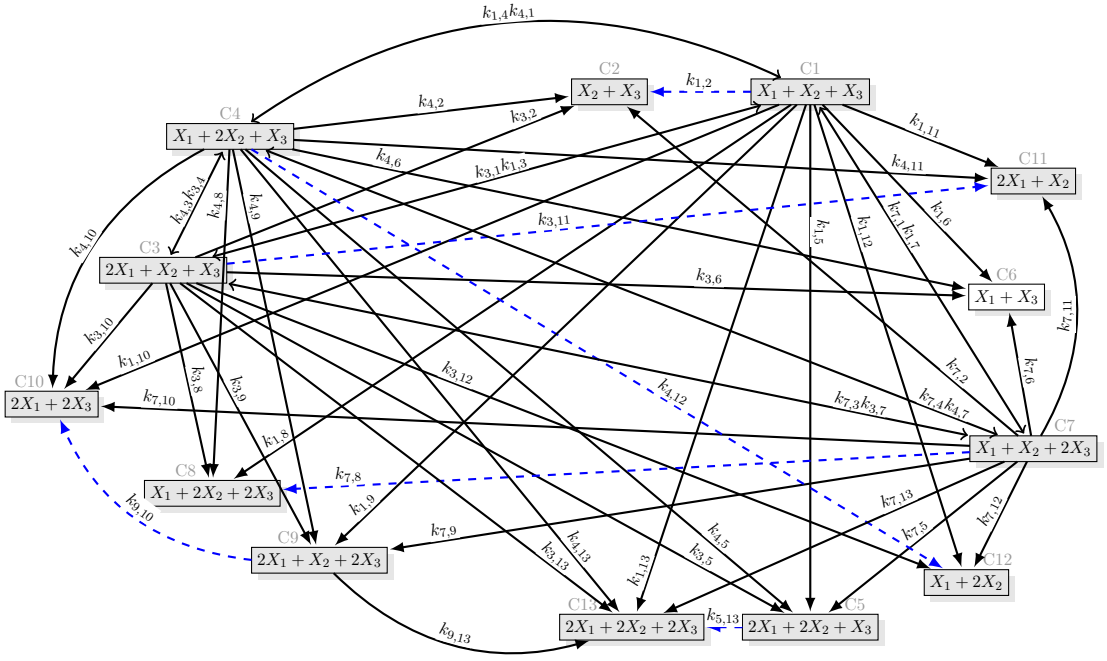
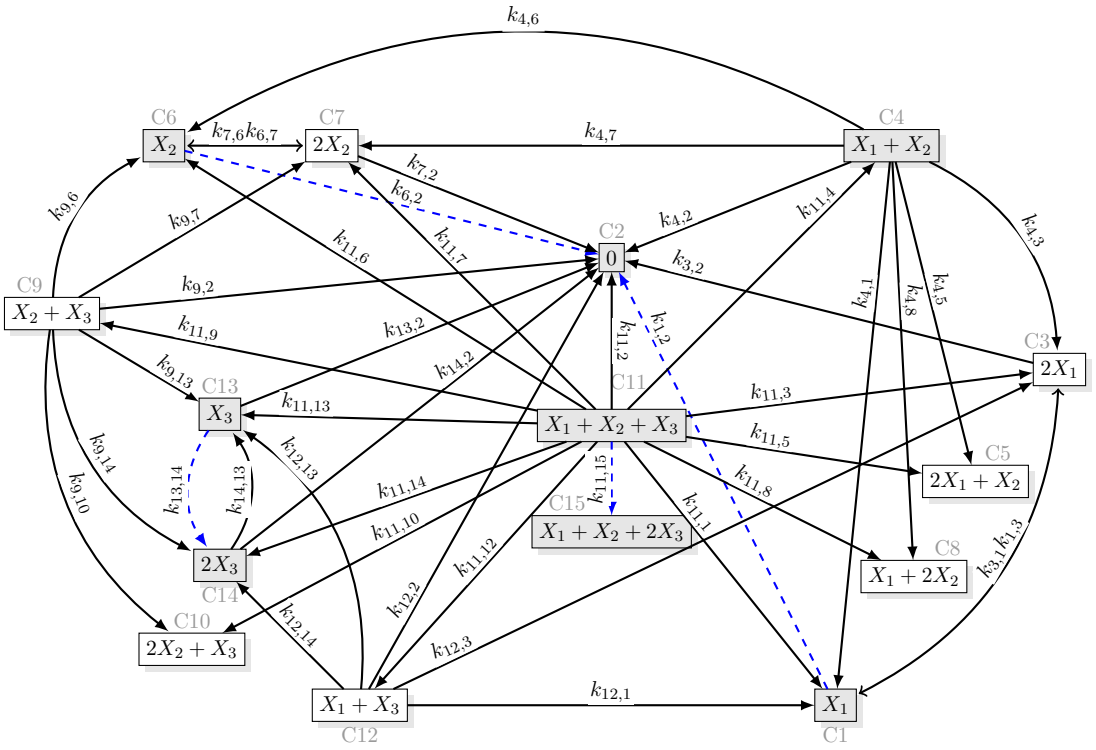


FIGURE 3.8: Dense realization of Equation (3.11) containing 51 reactions. The reaction rate coefficients are $k_{1,2} = 679.63, k_{1,3} = 0.1, k_{1,4} = 0.1, k_{1,5} = 0.1, k_{1,6} = 602.37, k_{1,7} = 0.1, k_{1,8} = 0.1, k_{1,9} = 0.1, k_{1,10} = 669.13, k_{1,11} = 0.1, k_{1,12} = 0.1, k_{1,13} = 0.1, k_{3,1} = 0.1, k_{3,2} = 0.1, k_{3,4} = 0.1, k_{3,5} = 44.6, k_{3,6} = 0.1, k_{3,7} = 0.1, k_{3,8} = 0.1, k_{3,9} = 0.1, k_{3,10} = 0.1, k_{3,11} = 16.2, k_{3,12} = 9.3, k_{3,13} = 0.1, k_{4,1} = 0.1, k_{4,2} = 0.1, k_{4,3} = 0.1, k_{4,5} = 9.6, k_{4,6} = 0.1, k_{4,7} = 0.1, k_{4,8} = 0.1, k_{4,9} = 0.1, k_{4,10} = 0.1, k_{4,11} = 0.1, k_{4,12} = 24.4, k_{4,13} = 0.1, k_{5,13} = 1, k_{7,1} = 0.1, k_{7,2} = 1.1833, k_{7,3} = 0.1, k_{7,4} = 0.1, k_{7,5} = 0.68335, k_{7,6} = 0.1, k_{7,8} = 23.217, k_{7,9} = 0.1, k_{7,10} = 0.1, k_{7,11} = 0.1, k_{7,12} = 0.1, k_{7,13} = 0.1, k_{9,10} = 1.1, k_{9,13} = 0.1.$

DOI:10.15774/PPKE.ITK.2015.008



Linkage Classes: 1
Deficiency: 11

# of Reactions: 44
# of Complexes: 15

FIGURE 3.9: Dense realization of Equation (3.14) with 44 reactions. The parameter values are the following: $k_{1,2} = 10.1, k_{1,3} = 0.1, k_{3,1} = 0.1, k_{3,2} = 4.95, k_{4,1} = 0.1, k_{4,2} = 0.1, k_{4,3} = 0.1, k_{4,5} = 10.2, k_{4,6} = 0.1, k_{4,7} = 0.1, k_{4,8} = 29.2, k_{6,2} = 2799.1, k_{6,7} = 0.1, k_{7,2} = 0.45, k_{7,6} = 0.1, k_{9,2} = 0.1, k_{9,6} = 0.1, k_{9,7} = 99.9, k_{9,10} = 0.4, k_{9,13} = 0.1, k_{9,14} = 0.1, k_{11,1} = 0.1, k_{11,2} = 0.1, k_{11,3} = 0.7, k_{11,4} = 0.1, k_{11,5} = 0.1, k_{11,6} = 0.1, k_{11,7} = 0.1, k_{11,8} = 0.1, k_{11,9} = 0.1, k_{11,10} = 0.1, k_{11,12} = 0.1, k_{11,13} = 0.1, k_{11,14} = 0.2, k_{11,15} = 2.2, k_{12,1} = 100.7, k_{12,2} = 0.1, k_{12,3} = 0.3, k_{12,13} = 0.1, k_{12,14} = 0.1, k_{13,2} = 0.1, k_{13,14} = 10103, k_{14,2} = 1.2833, k_{14,13} = 0.1$.

DOI:10.15774/PPKE.ITK.2015.008

Chapter 4

Computing Structural Properties of Uncertain Kinetic Polynomial Systems

The reconstruction of reaction network structure from measurement data and prior information is an important and intensively studied area [126, 133]. In certain chemical and biological problem statements it is generally assumed that the participating chemical species and the possible chemical complexes, i.e. the node set of the reaction graph is fixed [12]. Hence, an important remaining task is to determine the existence and the rate coefficients of the reactions between the participating complexes. This task unfortunately turns out to be computationally intractable as the number of complexes increase. The problem is well studied in the field of model and parameter estimation, often called as network reconstruction or inference.

As it was shown in the previous chapter, different reaction network structures (weighted directed graphs) may belong to exactly the same dynamics, therefore, the structural and parametric identification problem is generally not uniquely solvable without additional prior knowledge on the network structure, even if we have full and perfect measurements [29, 114]. In certain applications, the assumption of network sparsity may improve the solvability of the inference problem [12, 133]. However, in general, sparse structures corresponding to a given kinetic dynamics are not unique as it was illustrated in the previous chapter and in [113]. Therefore, we would like to further analyze the most

DOI:10.15774/PPKE.ITK.2015.008

"certain" structural elements of the network. These are called core reactions [114], and have the property that they are present in any reaction graph structure (realization) that is compatible with a given kinetic dynamics. The goal of this chapter is to extend the previous results in [114] on computing structural properties of kinetic systems to the case when there are uncertainties in the model. These uncertainties will be modeled as intervals for the coefficients of the monomials in the ODEs, similarly to the approach that was used for the estimation of fluxes of metabolic networks in [71]. We show that the computation of core reactions can be put into the framework of linear programming. Similarly, we show that sparse or dense realization of an uncertain kinetics system can be computed using mixed integer linear programming. Using illustrative examples we demonstrate the properties of the computed structures and the potential application of the method in the support of structural identification of biochemical networks.

4.1 Uncertain polynomial kinetic systems

We can extend the kinetic system model defined in Equation (2.7) to the uncertain case where the uncertainty will be modeled by intervals of the monomial coefficients. In this way, a family of kinetic systems is obtained that can be represented as

$$\dot{x} = M \cdot \psi(x) \tag{4.1}$$
$$[M_l]_{ij} \leq [M]_{ij} \leq [M_u]_{ij}\, i = 1, \ldots, n, \quad j = 1, \ldots, m$$

where matrices $M_l \in \mathbb{R}^{n \times m}$, $M_u \in \mathbb{R}^{n \times m}$ and $M \in \mathbb{R}^{n \times m}$ are kinetic matrices as defined in Equation (2.9).

It should be noted that just like in the regular case where a kinetic system is completely characterized by the matrix pair (Y, A_k), an uncertain polynomial kinetic system is completely described by the matrix triplet (Y, M_l, M_u).

With these upper and lower bounds we can represent the uncertainty of parameters individually. These bounds can represent the uncertainty of the parameters due to temperature change, measurement error, poor excitation of the dynamics. Also, it can

DOI:10.15774/PPKE.ITK.2015.008

represent operation regimes. In case of parameter estimation it can represent the parameter uncertainty, hence the quality of parameter estimation can be calculated via confidence intervals which are directly applicable to Algorithm 4.

Thanks to this representation we can calculate structural properties of an uncertain kinetic system that are valid for the whole family of kinetics systems. For example, a sparse realization will be a reaction graph that contains the least amount of reactions that are needed to generate all possible dynamics within the boundaries of M_l and M_u in Equation (4.1).

4.2 Computing Core reactions of Uncertain Polynomial Kinetic systems

We can take advantage of the fact that if core reactions exist, then they are present in any dynamically equivalent realization. Therefore, they are an essential part of the system's structure to produce the dynamics described by matrix pair (Y, A_k).

Formally, a reaction $(C_i, C_j) \in \mathcal{R}, i, j \in \{1, \ldots, m\}$ is a core reaction if and only if for any dynamically equivalent realization (Y, A_k), $[A_k]_{ji} > 0$ holds. This condition can be translated as a constraint set in a linear program as it was introduced in [114].

4.2.1 Algorithm for computing core reactions

Let us assume that we have an uncertain kinetic system characterized by matrix triplet (Y, M_l, M_u), then the following linear program (LP) can be formulated

$$\min \textstyle\sum_{ij} E_{ij}[A_k]_{ij} \tag{4.2}$$

s.t.

$$\textstyle\sum_{i=1}^{m} [A_k]_{ij} = 0, \quad j = 1, \ldots, m \tag{4.3}$$

$$[A_k]_{ij} \geq 0, i, j = 1, \ldots, m, \ i \neq j \tag{4.4}$$

$$Y \cdot A_k - M = 0 \tag{4.5}$$

$$[M_l]_{ij} \leq [M]_{ij} \leq [M_u]_{ij} \tag{4.6}$$

DOI:10.15774/PPKE.ITK.2015.008

where entries of matrices A_k and M are both variables of the optimization problem. The binary matrix $E \in \{0,1\}^{m\times m}$ selects the elements of A_k that are considered in the cost function.

The first two constraints (Equations 4.3 and 4.4) force A_k to be a Kirchhoff matrix (as defined in Equation (2.10)). The constraint in Equation 4.5 enforces dynamical equivalence (see Equation (2.20) for the definition). Finally, the last type of constraints (Equation 4.6) set up element-wise the lower and upper bounds for all entries of the kinetic matrix M.

The LP defined in Equation (4.2) is called as `DynEqLP` in the pseudo code below and this LP needs 4 inputs for operation, such as Y, M_l, M_u and E and calculates a realization characterized by matrices A_k and M.

ALGORITHM 4: The goal of the algorithm is to find core reactions of an Uncertain Kinetic System characterized by matrix triplet (Y,M_l,M_u). The Algorithm returns with the set of core reactions, $\mathcal{R}_c$.

Input : Y, M_l, M_u
Output: $\mathcal{R}_C$

$D := 0_m$ `// empty matrix`;
$E := 1_m - I_m$ `// only off-diagonals are non-zero` ;
$A_k := 0_m$ `// empty matrix` ;
while *true* **do**
  $A_k :=$ `DynEqLP`(Y,M_l,M_u,E);
  $D := E$;
  $E := (PositiveElements(A_k)$ & $D)$ `// & is an element-wise logical AND` ;
  if $D = E$ **then**
    break;
  end
end
$\mathcal{R}_C := \emptyset$;
for each $(i,j) \in PositiveElements(E)$ **do**
  $Z := 0_{m\times m}$;
  $[Z]_{i,j} := 1$;
  $A_k :=$ `DynEqLP`(Y,M_l,M_u,Z) ;
  if $[A_k]_{i,j} > 0$ **then**
    $\mathcal{R}_C := \mathcal{R}_C \cup (C_j \to C_i)$;
  end
endfor
return $\mathcal{R}_C$

In Algorithm 4, the procedure *PositiveElements* finds the nonzero elements of the argument and gives back a binary matrix containing ones where the original matrix elements are larger then zero.

DOI:10.15774/PPKE.ITK.2015.008

Algorithm 4 starts with binary matrix E where the off-diagonal elements are ones and the rest are zero. In the first iteration the `DynEqLP` tries to minimize the sum of off-diagonal elements of A_k. After that, in each iteration the zero off-diagonal elements are excluded from the cost function until no more off-diagonal elements can be excluded, i.e. matrix E remains the same between two iterations. In the final stage of the Algorithm, for each of the remaining non-zero off-diagonal elements (encoded in matrix E) the `DynEqLP` is executed where only one element from matrix E is minimized. If this element remains non-zero after the optimization, then this reaction (edge) is part of the core reaction set.

Property	Algorithm 2	Algorithm 4
maximum of LP iterations	(m^2-m)	$2+(m^2-m)$
minimum of LP iterations	(m^2-m)	2
LP variables	$\underbrace{m^2-m}_{offdiag(A_k)}$	$\underbrace{nm}_{M}+\underbrace{m^2-m}_{offdiag(A_k)}$
Inequality Constraints	$\underbrace{m^2-m}_{offdiag(A_k)}+\underbrace{m}_{diag(A_k)}$	$\underbrace{2nm}_{M_l,M_u}+\underbrace{m^2-m}_{offdiag(A_k)}+\underbrace{m}_{diag(A_k)}$
Equality Constraints	$\underbrace{m}_{Kirchhoff}+\underbrace{m}_{DynEq}+\underbrace{1}_{\mathcal{K}}$	$\underbrace{m}_{Kirchhoff}+\underbrace{m}_{DynEq}$

TABLE 4.1: Comparison of the main properties of two algorithms for core reaction set computation. Keywords: $offdiag(A_k)$—number of off-diagonal entries of A_k, $diag(A_k)$—number of diagonal entries of A_k, $Kirchhoff$—Equation (4.3), $DynEq$—Equation (4.5).

Finally, we can compare the main properties of Algorithm 2 and Algorithm 4 using Table 4.1. By comparing the minimum and maximum number of iterations required by the two algorithms we can clearly see that the newly proposed algorithm for calculation of core reactions usually outperforms the Algorithm 2, except for the theoretical worst case. Furthermore, the previous algorithm is based on systematic exclusion of reactions and testing for the feasibility of the resulting linear program, which might take place for different reasons than exclusion of a reaction. It should be noted that the size of the LP in Algorithm 4 is larger than the one in Algorithm 2, but this comes from the fact that Algorithm 4 is designed for uncertain kinetic systems and it can also operate with zero uncertainty as well ($M = M_l = M_u$).

DOI:10.15774/PPKE.ITK.2015.008

4.3 Examples

4.3.1 Example of an Uncertain Kinetic System

In this example, we revisit the positive feedback motif from Section 2.3.4 and investigate that how increasing intervals around matrix M affect the core reaction set, $\mathcal{R}_C$.

The positive feedback motif model is encoded with matrix pair (Y,A_k) as

$$Y = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{4.7}$$

and the only non-zero off-diagonal elements of A_k are

$$\begin{aligned} [A_k]_{5,1} &= k_1,\ [A_k]_{1,5} = k_2,\ [A_k]_{9,4} = k_3, \\ [A_k]_{4,9} &= k_4,\ [A_k]_{8,9} = k_5,\ [A_k]_{6,7} = k_6, \\ [A_k]_{11,10} &= k_7,\ [A_k]_{11,3} = k_8,\ [A_k]_{2,10} = k_9. \end{aligned} \tag{4.8}$$

First, the lower and upper bounds are set as $M_l = M_u = M$ (i.e. there is no parametric uncertainty in the system) and with Algorithm 4 we calculated that this system has $|R_C| = 8$ core reactions and these are

$$\begin{aligned} &C_1 \to C_5, C_3 \to C_{11},\ C_4 \to C_9,\ C_5 \to C_1 \\ &C_7 \to C_6, C_9 \to C_4,\ C_9 \to C_8,\ C_{10} \to C_2 \end{aligned} \tag{4.9}$$

and they are shown as blue dashed edges in Figure 2.1.

The matrix M defined by Equations (4.7) and (4.8) has 5×11 elements. In each step we symmetrically increase the interval around one element of M, then the invariant reaction set is calculated with the help of Algorithm 4. Meanwhile the kinetic property of matrix M is ensured by checking the sign constraints of M, defined by Equation (2.9). The result is summarized in Figure 4.1, in the top left corner there are 8 core reactions in the

DOI:10.15774/PPKE.ITK.2015.008

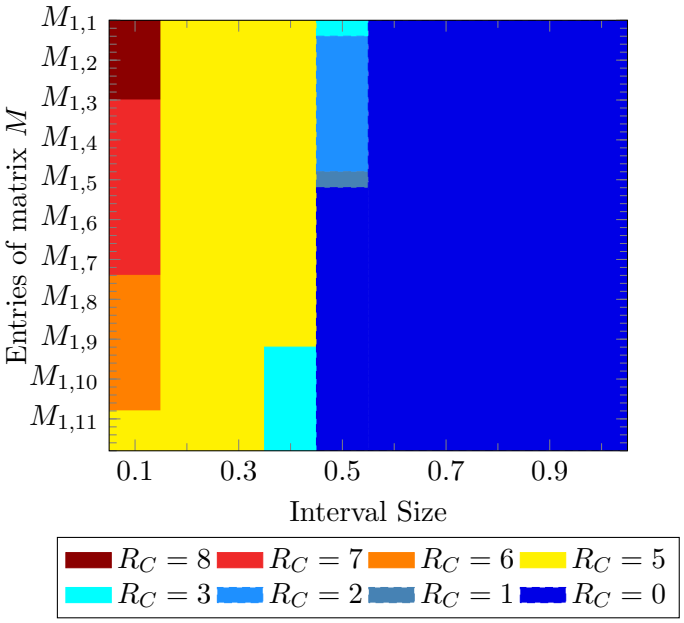


FIGURE 4.1: Figure shows how the increasing intervals around M is affecting the number of core reactions inside the interval. The vertical axis lists the elements of M and the horizontal axis shows the accumulation of the interval size around the values of M. The calculation start at the top left corner and goes down along the horizontal axis, then current interval gets increased and the calculation start at the top in the next column along the vertical axis. Each color represents the size of the core reaction set within the interval. In each iteration, the applied step size was 0.1.

system (dark red area) as we increase the bounding box around M the number of core reactions gradually decrease. Finally, the core reaction set becomes empty (dark blue area). It should be noted that in each step the current interval contains the previous interval as a subset, hence the core reaction set can either remains the same or become a subset of the previous one.

4.3.2 Network reconstruction example

Using a chemical reaction network example, we highlight a possible field of application of the algorithm presented in Section 4.2. In (bio)chemical modeling and many other fields, it is often the goal of the network reconstruction to find the sparsest network describing the measured dynamics which might be not a unique structure. Therefore, our goal is to find the structurally invariant elements that are characteristic for the dynamics while the parameters of the network are uncertain.

DOI:10.15774/PPKE.ITK.2015.008

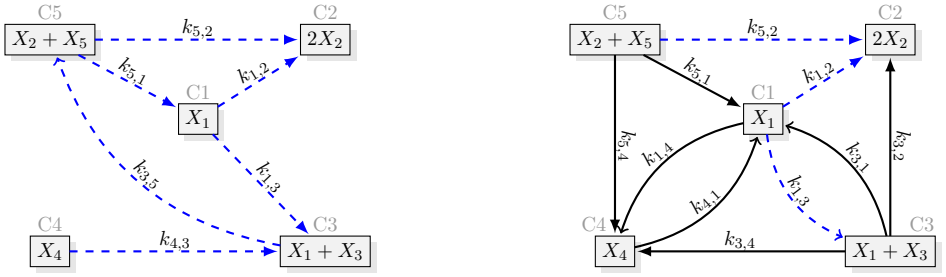


FIGURE 4.2: Comparison of the original network from [12] (left) and the network given by parameter estimation (right). The core reactions in each case are shown with blue dashed edges.

First, let us investigate a benchmark example from the literature which is based on Figure 2 in [12] and encoded as

$$Y = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.10}$$

and

$$A_k = \begin{bmatrix} -1.163 & 0 & 0 & 0 & 0.8492 \\ 0.3386 & 0 & 0 & 0 & 0.4290 \\ 0.8244 & 0 & -0.7364 & 0.5631 & 0 \\ 0 & 0 & 0 & -0.5631 & 0 \\ 0 & 0 & 0.7364 & 0 & -1.2782 \end{bmatrix}. \tag{4.11}$$

As a first step we establish the core reaction set for the dynamics represented by (Y, A_k). Algorithm 4 with boundary matrices $M_u = M_l = Y \cdot A_k$ tells us that all six reactions in the original network are core reactions (blue dashed edges on the left panel of Figure 4.2), hence this is the sparsest and also the only sparse realization of the given dynamics.

The goal of the parameter estimation is to obtain an interval model in the form of Equation (4.1) from time series measurements of the species concentration. During the parameter estimation procedure, we will exploit the fact that the type of kinetic systems we use in this thesis is linear in parameters. Finally, the parametric uncertainty in matrix M will be characterized by the covariance matrix of the estimator. Then, the

DOI:10.15774/PPKE.ITK.2015.008

core reaction set of the interval model will be calculated using the confidence intervals of the estimated parameters.

4.3.2.1 Parameter Estimation Procedure

We utilize the discrete Least Squares framework and for that we discretize Equation (2.7). Taking sufficiently small intervals between samples we apply the forward Euler method

$$x_i(k) = x_i(k-1) + hM_{i,\cdot}\psi_i(x(k-1)) \quad k = 2...T_{end}, i = 1,\dots,n \tag{4.12}$$

where x_i is the ith state variable from Equation (2.7), $\psi_i(x)$ is the ith element of vector mapping ψ and vector $M_{i,\cdot}$ is the ith row of matrix M. Finally, h is the sampling time, which is $h = 0.1$ *sec* in our case.

We assume that we can measure all state variables, then let us define the artificial outputs for the regressor model used for parameter estimation as

$$y_i(k) := x_i(k) - x_i(k-1),\ i = 1,\dots,n$$

and the model, which is used for the parameter estimation reads

$$y_i(k,\theta) = \theta_{i,\cdot}^T\varphi(k)$$

where $\theta_{i,\cdot}$ is the ith row of M and

$$\varphi(k) = \begin{bmatrix}\psi_1(x(k-1)) & \psi_2(x(k-1)) & \psi_m(x(k-1))\end{bmatrix}^T.$$

Further, let us define matrix $R \in \mathbb{R}^{m\times m}$ and vector $d_i \in \mathbb{R}^m$

$$R = \frac{1}{N}\sum_{k=2}^{N}\varphi(k)\varphi(k)^T \tag{4.13}$$

$$d_i = \frac{1}{N}\sum_{k=2}^{N}\varphi(k)y_i(k). \tag{4.14}$$

DOI:10.15774/PPKE.ITK.2015.008

Then, the estimate for the parameter vector θ_i can be calculated using the well-known formula as

$$\hat{\theta}_i = R^{-1} d_i. \tag{4.15}$$

which is the same as Equation (2.6).

The regression matrix R depends on all of the state variables, hence each row of M can be calculated in one step. By defining

$$C = \operatorname{diag}(R, \ldots, R) \tag{4.16}$$

where $C \in \mathbb{R}^{n \cdot m \times n \cdot m}$ is block diagonal matrix and

$$d = \begin{bmatrix} d_1 & \cdots & d_i & \cdots & d_n \end{bmatrix}^T \tag{4.17}$$

we can calculate all the elements of M in one optimization step. Let us define parameter vector $\theta = \operatorname{row}(M)^T$ which is the transpose of the row expansion of M. The row expansion of M is defined as $\operatorname{row}(M) = \begin{bmatrix} M_{1,\cdot}, M_{2,\cdot}, \ldots, M_{n,\cdot} \end{bmatrix}$ where $M_{i,\cdot}$ denotes the ith row of matrix M.

Then, the estimate for the parameter vector θ can be calculated as

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{2} ||C\theta - d||_2^2. \tag{4.18}$$

To ensure that $\hat{\theta}$ represents a proper kinetic matrix, the set Θ denotes the possible parameter vectors where the sign condition of M is fulfilled (see Equation (2.9) for details). Finally, the estimated matrix M is given as $\hat{\theta}/h$.

4.3.2.2 Estimation of matrix A_k

It is possible to improve the parameter estimation by exploiting the fact that with fixed matrix Y, the matrix M can be written as $Y \cdot A_k$. Hence, we can estimate A_k directly.

DOI:10.15774/PPKE.ITK.2015.008

For that, we define

$$\bar{Y} = \begin{bmatrix} Y_{1,\cdot} & 0 & \cdots & 0 \\ 0 & Y_{1,\cdot} & \cdots & 0 \\ 0 & 0 & \cdots & Y_{1,\cdot} \\ Y_{2,\cdot} & 0 & \cdots & 0 \\ 0 & Y_{2,\cdot} & \cdots & 0 \\ 0 & 0 & \cdots & Y_{2,\cdot} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{n,\cdot} & 0 & \cdots & 0 \\ 0 & Y_{n,\cdot} & \cdots & 0 \\ 0 & 0 & \cdots & Y_{n,\cdot} \end{bmatrix} \quad \bar{A}_k = \begin{bmatrix} [A_k]^T_{1,\cdot} \\ \vdots \\ [A_k]^T_{i,\cdot} \\ \vdots \\ [A_k]^T_{m,\cdot} \end{bmatrix}. \tag{4.19}$$

where $Y_{i,\cdot}$ denotes the ith row of matrix Y and vector $[A_k]_{j,\cdot}$ is the jth row of matrix A_k.

Then, we formulate the following constrained Least Squares optimization problem [126]

$$\begin{gathered} \min_{\bar{A}_k} \tfrac{1}{2}||C\bar{Y}\bar{A}_k - d||_2^2 \\ \text{s.t.} \\ \textstyle\sum_{i=1}^{m} [A_k]_{ij} = 0, \quad j = 1, \ldots, m \\ [A_k]_{ij} \geq 0, i, j = 1, \ldots, m, \ i \neq j. \end{gathered} \tag{4.20}$$

Finally, the estimated value of $\bar{A}_k$ is denoted as $\hat{\bar{A}}_k$. In this way we can have an estimate matrix M as $\bar{Y}\hat{\bar{A}}_k/h$, but imposing a set of constraints on the elements of A_k.

4.3.2.3 Results

Now, we attempt to (partially) restore the entries of matrix A_k which was defined in Equation (4.11) by applying the discrete Least Squares framework from Subsection 4.3.2.1. For this, we assume that we can measure all state variables of the kinetic system and to each state variable 5% additive Gaussian noise is added. A dataset with 25 different initial vectors is generated with uniform Latin hypercube sampling. Using the optimization defined in Equation (4.20), we calculate the graph structure from the

DOI:10.15774/PPKE.ITK.2015.008

noisy dataset as follows

$$A_k^{est} = \begin{bmatrix} -2.3 & 0 & 0.46 & 1.51 & 0.07 \\ 1.1 & 0 & 0.06 & 0 & 0.09 \\ 0.92 & 0 & -0.58 & 0 & 0 \\ 0.26 & 0 & 0.065 & -1.51 & 0.047 \\ 0 & 0 & 0 & 0 & -0.21 \end{bmatrix}. \tag{4.21}$$

The graph encoded by Equation (4.21) can be seen on the right panel of Figure 4.2.

Thanks to the factorization of M in Equation (4.20), we have an estimate of matrix M as

$$\hat{M} = \bar{Y}\hat{\bar{A}}_k/h. \tag{4.22}$$

The parametric uncertainty around matrix M is extracted from the estimator's covariance matrix. It is calculated as $Cov(\hat{\theta}) = C^{-1} \cdot \sigma^2$ where σ^2 is the standard deviation of the measurement noise. Then, for each element of the estimated matrix M, we establish the 95% confidence interval and cut off intervals that contradict the sign constraints of matrix M as follows

$$M_l = \hat{M} - 1.96\sqrt{\text{diag}(Cov(\hat{\theta}))} \tag{4.23}$$

$$[M_l]_i = \begin{cases} \max([M_l]_i, 0) & \text{if } [\text{row}(Y)]_i > 0 \\ [M_l]_i & \text{otherwsie} \end{cases} \quad i = 1, \ldots, n \cdot m \tag{4.24}$$

$$M_u = \hat{M} + 1.96\sqrt{\text{diag}(Cov(\hat{\theta}))} \tag{4.25}$$

where $\text{diag}(Cov(\hat{\theta}))$ selects the diagonal elements of the covariance matrix $Cov(\hat{\theta})$.

These modified confidence intervals define the interval model as it was introduced in Equation (4.1), then we execute Algorithm 4 with input matrices (Y,M_l,M_u). The result shows that 3 reactions are in the core reaction set, namely $C_1 \to C_2, C_1 \to C_3$ and $C_5 \to C_2$ (blue dashed edges on the right panel of Figure 4.2).

DOI:10.15774/PPKE.ITK.2015.008

This information can be utilized during the further refinement of the network reconstruction, e.g. in case of a new dataset, these reactions can be constrained as fixed parts of the graph. Even though the Least Squares approach performed poorly on the reconstruction of the original network structure, this performance was enough to recover a subset of the original core reaction set. This shows an important feature of this approach, namely it is capable of recovering the core reaction set of an uncertain model, even if the original network is only partially restored, which is often the case with other network reconstruction approaches [12, 132]

We can briefly comment on the performance on Algorithm 4. It took only 6 iterations for the algorithm to establish the core reaction set, compared to the previous algorithm, this network would require 20 iteration, so it means that in this case, Algorithm 4 is more than three times faster than Algorithm 2.

In this example, we used the simplest possible method to show how our methods can support the network reconstruction procedure. Any other type of network reconstitution can be used as long as some form of parameter uncertainty is available from the applied method or from a priori information.

4.4 Computing Sparse and Dense Realizations of Uncertain Kinetic Systems

In this section we want to investigate the question of non-uniqueness of the sparse realization in case of uncertain models. For that reason we revisit the positive feedback motif introduced in Section 2.3.4. Examples of multiple dynamically equivalent sparse realizations of the positive feedback motif was reported in [114] (see Figure 2 in [114]). In this model, there are 8 core reactions and a sparse structure has 9 reactions, this difference is a necessary condition to have more than one sparse structure.

Since we do not have measurement, we add roughly 10% uncertainty to each parameter. By applying Algorithm 4 on the positive feedback motif with $[M_l]_{ij} = [M]_{ij} - 0.1$ and $[M_u]_{ij} = [M]_{ij} + 0.1$ where $M = Y \cdot A_k$ yields a core reaction set with 5 elements (yellow region in Figure 4.1) and shown with blue dashed edges in Figure 4.3.

DOI:10.15774/PPKE.ITK.2015.008

Again, we can compare the performance of the two algorithms for computation of the core reactions on this example. Algorithm 4 calculated the core reaction set in 10 iterations, which is only 9% of the number of iterations would be required by the Algorithm 2.

To calculate the sparse structure of an uncertain kinetic polynomial system we need to formulate a MILP optimization based on a method described in Section 2.3.1 and Equation (4.2)

$$\begin{aligned}
&\text{minimize} \sum_{i,j=1,i\neq j}^{m} \delta_{ij} \\
&Y \cdot A_k - M = 0 \\
&\textstyle\sum_{i=1}^{m} [A_k]_{ij} = 0,\ j = 1,\dots,m \\
&[A_k]_{ij} \geq \epsilon\ i,j = 1,\dots,m\ i \neq j \\
&0 \leq [A_k]_{ij} - \epsilon\delta_{ij}\ i,j = 1,\dots,m, i \neq j \\
&0 \leq -[A_k]_{ij} + l_{ij}\delta_{ij},\ i,j = 1,\dots,m, i \neq j \\
&[M_l]_{ij} \leq [M]_{ij} \leq [M_u]_{ij}\ \ i,j = 1,\dots,m.
\end{aligned} \tag{4.26}$$

We calculate a sparse realization of the positive feedback motif by applying optimization problem defined in Equation (4.26). The resulting realization can be seen on the left panel of Figure 4.3.

The sparse realization on the left panel in Figure 4.3 has only one non-core reaction ($C_{10} \to C_2$), thus we can search $\mathcal{K}$-*constrained sparse realization* of an uncertain sparse realization by modifying the cost function in Equation (4.26) as

$$\sum_{i,j=1;i\neq j;i,j\notin\mathcal{K}}^{m} \delta_{ij}. \tag{4.27}$$

Let us add reaction $C_{10} \to C_2$ to set $\mathcal{K}$ and calculate another sparse realization of the uncertain positive feedback motif (depicted on the left panel of Figure 4.3). In this realization only reaction $C_{10} \to C_1$ is not part of the core reaction set. Again, we can redefine set $\mathcal{K} = [A_k]_{1,10}$ and calculate another uncertain sparse realization. The realization is the same as the realization shown on the left panel of Figure 4.3, therefore at this level of uncertainty only two sparse realizations exist for the positive feedback motif model.

DOI:10.15774/PPKE.ITK.2015.008

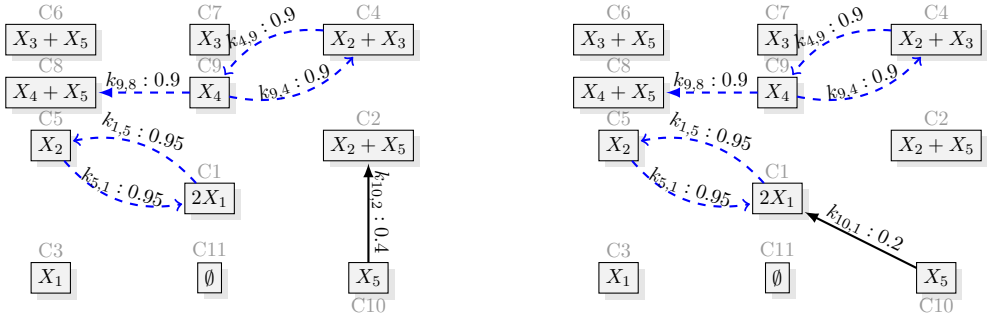


FIGURE 4.3: Two sparse realizations of the positive feedback motif with $[M_l]_{ij} = [M]_{ij} - 0.1$ and $[M_u]_{ij} = [M]_{ij} + 0.1$.

The reader can notice that by comparing Figures 2.1 and 4.3, in the later one some complexes became isolated, i.e. a complex has no incoming or outgoing edges. Its explanation lays in the bounds around M. By comparing the corresponding M matrix of the sparse realization on the left panel of Figure 4.3, denoted by M^{bound}, and the current lower bound M_l, it can be seen that, the current lower bounds on matrix M allows only the columns 2,3,6,7,8 and 11 to be zero (see matrix M_l in Equation (4.29)). Since we try to minimize the number of reactions in the realization, the optimization tries to push the values of matrix M toward the minimum bound (M_l). As a result of that isolated complexes may emerge.

It should be noted that there exists another case when a column in M^{bound} is zero, but the corresponding complex is not isolated. In both sparse cases in Figure 4.3, the complex C_8 is not an isolated complex. It is explained by the fact, complex C_8 is a product complex, i.e. it has only ingoing edges, which causes the associated column in the matrix M^{bound} in Equation (4.28) to become zero.

$$M^{bound} = \begin{bmatrix} -1.9 & 0 & 0 & 0 & 1.9 & 0 & 0 & 0 & 0 & 0.4 & 0 \\ 0.95 & 0 & 0 & -0.9 & -0.95 & 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & -0.9 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & -0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0 \end{bmatrix} \tag{4.28}$$

DOI:10.15774/PPKE.ITK.2015.008

$$M_l = \begin{bmatrix} -2.1 & -0.1 & -0.15 & 0 & 1.9 & 0 & 0 & 0 & 0 & 0.4 & 0 \\ 0.9 & 0 & 0 & -1.1 & -1.1 & 0 & 0 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & -1.1 & 0 & -0.1 & -0.1 & 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & -0.1 & -1.1 & 0 & 0 \\ 0 & -0.1 & 0 & 0 & 0 & -0.1 & 0 & -0.1 & 0.9 & -0.2 & 0 \end{bmatrix} \tag{4.29}$$

This example shows that under parametric uncertainty the sparse realization is generally non-unique, too, hence the core reactions are the only certain elements of the (uncertain) kinetic model.

4.5 Summary

We have given an effective algorithm to compute the so-called core reactions of uncertain kinetic polynomial models assuming a given complex set. The proposed method iteratively uses linear programming steps, and therefore it runs in polynomial time. Several numerical examples have been given to illustrate the capability of the algorithm: First, the effect of parametric uncertainty on the core reaction set was analyzed, then through another example, we have highlighted how to utilize the core reaction calculation in a simple network reconstruction problem. Second, we have shown that under parametric uncertainty, the sparse network structure is generally non-unique for the studied kinetic interval models. Finally, in these numerical examples we have shown that the proposed algorithm, in terms of computation speed, outperforms the previously applied one.

DOI:10.15774/PPKE.ITK.2015.008

Chapter 5

Modeling and Parameter Estimation of a Cell-free *in vitro* System

While the use of quantitative models in molecular biology has become commonplace in recent decades, the amount and types of experimental data suitable for model parameter estimation are often severely limited. In many cases, temporal resolution and/or the sensitivity of the measurement technique present significant obstacles for effective parameter estimation. Additionally, structural non-identifiability of the model is also a real possibility [26, 29, 114]. However, recent developments in real-time mRNA reporter technology have provided new, powerful tools, which together with fluorescent proteins made the concurrent tracking of the concentrations of mRNA and protein species of interest possible. As a result, we can now measure transcription and translation simultaneously with sufficiently high frequency and specificity to directly use the obtained time series data for parameter estimation [88].

This is particularly useful in the rapidly expanding field of synthetic biology, wherein biological 'parts' (e.g., promoters, terminators, genes) can be rapidly combined into 'biocircuits' [111] that may not otherwise exist. One of the major goals of synthetic biology is to apply rational engineering design to create functioning novel biocircuits, operating *in vitro* and *in vivo* [69]. This bottom-up approach represents a shift towards carrying

DOI:10.15774/PPKE.ITK.2015.008

out design in a systematic and hierarchical way using well characterized and reusable biomolecular parts (DNA/plasmids, RNAs, enzymes, proteins, membranes) [36].

Even though many engineering methodologies and philosophies are being adopted by synthetic biologists, modeling remains challenging due to the existence of a large number of coupled (both known and unknown) mechanisms [22]. Furthermore, synthetic circuits introduced into a host organism may interact with the existing reactions of the host in complex and emergent ways.

In synthetic biology an *in vitro* environment is usually the place where biomolecular breadboarding happens. One type of *in vitro* systems is called cell-free expression system which consists of crude cell extracts supplemented with certain buffers and resources [60, 102]. The crude cell extract contains functional transcription-translation machinery, with the genetic material (and hence the regulatory circuits they encode) from the host carefully removed. Therefore, cell-free extracts provide a platform for the characterization of biomolecular parts and circuits in relative isolation, free from natural regulation and "cross-talk" which is present in living cells [57]. A further advantage of cell-free extracts is the significantly shorter characterization and design cycles compared to *in vivo* approaches.

The cell-free system that we are modeling in this chapter was developed in [101]. This versatile *in vitro* system supports multiple stage cascades and bistable biocircuits, as well as the expression of the complete phage genome [100]. While elementary transcription and translation steps as well as degradation steps have been successfully modeled in [63], the previous work focused on the linear regime and the initial stage of protein expression where the finite resources within the system are not limiting factors.

In this chapter, by explicitly considering transcription, translation, and degradation machinery with energy sources, we show that it is possible to capture experimental observations more quantitatively and beyond the initial stages of experiment. This approach also provides insight into possible underlying reasons, such as resource limitations and enzyme loading [91, 96, 130], for previously unexplained behavior of the synthetic circuits being tested.

Several examples for modeling of the transcription and translation processes can be found in the literature, ranging from coarse grain [63],[109] to very detailed [49],[8],

DOI:10.15774/PPKE.ITK.2015.008

where they are focusing on different aspects of gene expression. Our aim here is to develop and analyze a kinetic model describing transcription and translation in a cell-free experimental environment using real measurement data.

The goal of this chapter to describe the construction and parameter estimation of a kinetic model which became the core part of a Matlab software toolbox, called TXTL toolbox and the toolbox was developed by the authors of paper [C4]. The toolbox is freely available at [119] and designed as a modeling tool for synthetic biologists who are interested in seeing the different dynamical behaviors of their biocircuit designs. Furthermore, the data collection methodology and experiment design presented in this chapter gives a procedure for the users to calibrate their copy of the TXTL toolbox. The reported parameters provide a default set of parameters for the TXTL toolbox as well.

This chapter is organized as follows: in Section 5.1 we briefly outline the biomolecular breadboarding system and introduce the experimental setup for data collection. In Section 5.2 we discuss the proposed nonlinear mathematical model that captures the dynamics of the cell-free system. Section 5.3 carries out an analysis of the applied model with time-scale separation and structural identifiability. Finally, the model which is the result of model reduction and analysis is calibrated via parameter estimation in Section 5.4.

5.1 Experimental background

5.1.1 Cell-free system

Cell-free gene expression systems are popular platforms for biocircuit design. A cell-free biomolecular "breadboard" system is a collection of *in vitro* protocols that can be applied for testing transcription and translation circuits in a set of systematically-constructed environments that explore different elements of the external conditions in which the biocircuits must operate [101].

The gene regulatory circuits or biocircuits expressed using this platform can be engineered in a molecular biology laboratory using standard molecular cloning techniques

DOI:10.15774/PPKE.ITK.2015.008

to create 'plasmid' DNA, or may be produced using Polymerase Chain Reaction (PCR) to create 'linear' DNA. The latter substantially decreasing the design cycle time.

These externally supplied genes are expressed in the host environment (*in vitro* or *in vivo*) where they interact with each other and/or the host environment via protein-protein, protein-DNA, etc interactions. With this interaction network, various tasks can be performed such as computation, sensing, actuating, etc.

The *in vitro* cell-free system used here has certain advantages that make it desirable as a tool for simplifying the study of biological circuit function. Due to the lack of back regulation of the host environment, it is possible to design and operate a synthetic circuit in different concentration regimes that may not be achievable in a living cell. This in turn may allow us to explore a much larger parameter space to characterize the range of dynamical behaviors that a circuit topology may be capable of producing.

On the other hand, the *in vitro* cell-free system also has some potential disadvantages over living cells, such as lack of growth, lack of sophisticated organization, and limited resources; though these do not necessarily pose a significant impediment to the prototyping function of this type of breadboard systems as shown in a recent work [24]. The biggest caveat is that the *in vitro* cell-free system has finite resources (e.g., RNA polymerase, Ribosome, NTPs, AAs), and the loss of the energy source adenosine triphosphate (ATP) in particular has a strong impact on the system performance [65]. Moreover, since resources are limited in this environment, competition for the finite resources can arise [131]. The second disadvantage of this type of cell-free system is that there is neither active growth nor waste removal process in the system, leading to the accumulation of reaction by-products such as non-functioning mRNA fragments and adenosine diphosphate (ADP), which slow down reactions and eventually cause them to stop. The third difficulty is lack of compartmentalization, because of that unintentional reactions (interactions) can emerge.

The finite resource and waste management problem can be partially addressed by an exchange dialysis system using membranes that allow exchange of fuel molecules and small wastes [101]. Furthermore, it is possible to implement compartmentalization and membranes in the form of tiny lipid vesicles in the cell free system [101]; however, the control of localization within vesicle surface remains primitive.

DOI:10.15774/PPKE.ITK.2015.008

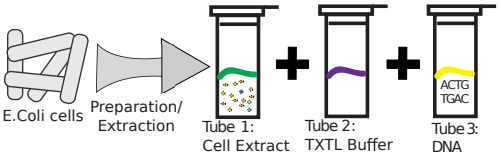


FIGURE 5.1: The cell-free *in vitro* system consists of three main components. Tube 1 has the extracted content of *E. coli* cells. Tube 2 contains all the amino acids and nucleotides, among other chemicals. Our biocircuit is placed in Tube 3, hence different DNA fragments. Finally, when each components are mixed together the gene expression is initiated.

All the experiments reported in this chapter were performed in a cell-free environment derived from *Escherichia coli* crude extract. This extract contains all the endogenous system components necessary for transcription and translation (e.g., ribosomes, RNA polymerase, translation initiation and elongation factors, etc.) but is free of structural components (e.g, cell wall) and genomic DNA, this is represented as 'Tube 1' in Figure 5.1.

The processed extract is supplemented with molecular energy sources: nucleotides, amino acids, and tRNAs, shown as 'Tube 2' in Figure 5.1. In this simple form the crude extract plus energy source mixture is a fully functional cell-free expression system. The detailed description of the system and the preparation steps of the crude extract can be found in [110]. The last necessary component for the operation of the cell-free system is the DNA itself, which contains the genetic code of the biocircuit subject to the testing ('Tube 3' in Figure 5.1).

5.1.2 Measurements

Plasmids and Bacterial strains The plasmid pBEST-Luc (Promega) was used as a template for the construct and was created using standard cloning methods. *E. coli* strains KL740 (which contains lambda repressor to control for Pr promoter) or JM109 were used and Qiagen Plasmid Midi prep kits were used for the DNA extraction. LB media with 100 μg/mL carbenicillin was used to culture cells. The genotype of the plasmid is given in Table 5.1.

Reporters Radiolabeling of mRNA and subsequent gel analysis can provide limited temporal resolution for mRNA dynamics [63]. Alternative methods using molecular

DOI:10.15774/PPKE.ITK.2015.008

TABLE 5.1: Genotype of the plasmid used in this study.

Plasmid name	Transcription unit (TU)	TU size (bp)	Backbone/resistance
pBEST-Pr-GFP	P_R:deGFP:T500	810	ColE1/Amp^R

beacon and probes showed improved temporal resolution but the dynamic range was limited [79, 109].

To determine mRNA dynamics at high sensitivity with high temporal resolution we used real-time fluorescence monitoring of mRNA dynamics by utilizing an RNA aptamer with a fluorescent dye, called malachite green (triphenylmethane) [48]. This aptamer contains a binding pocket for the malachite green dye, but the pocket is very short (35 bases), and thus, the presence of aptamer sequence does not affect the transcripiton of *GFP* gene. The binding of the dye into the aptamer's binding pocket enhances the dye's fluorescence, thus reports the presence of binding pockets, i.e. the concentration of mRNA.

The mRNA aptamer (MGApt) was placed in the $3'$ untranslated region (UTR) of a gene encoding Green Fluorescent Protein (GFP) 15 bases downstream of the stop codon (see panel A in Figure 5.2). The fluorescent protein GFP with T500 transcriptional terminator were previously designed to be more translatable in the cell-free system [102] and used as a reporter for the translational activities.

Production of the combined GFP-MGApt construct is driven by a strong constitutive promoter.

Fluorescence was measured in every 3 minutes for both MGApt (excitation: 610 nm; emission: 650 nm) and GFP (excitation: 485 nm; emission: 525 nm) in a Biotek plate reader. All measurements were background corrected to account for the autofluorescence of the malachite green dye. Further details of the measurements and sample preparation can be found in [J1].

Cell-free system The cell extract is a crude cytoplasmic extract from *E. coli* which contains soluble proteins, including the entire endogenous transcription-translation machinery, as well as mRNA and linear DNA degradation enzymes [101, 102]. Detailed

DOI:10.15774/PPKE.ITK.2015.008

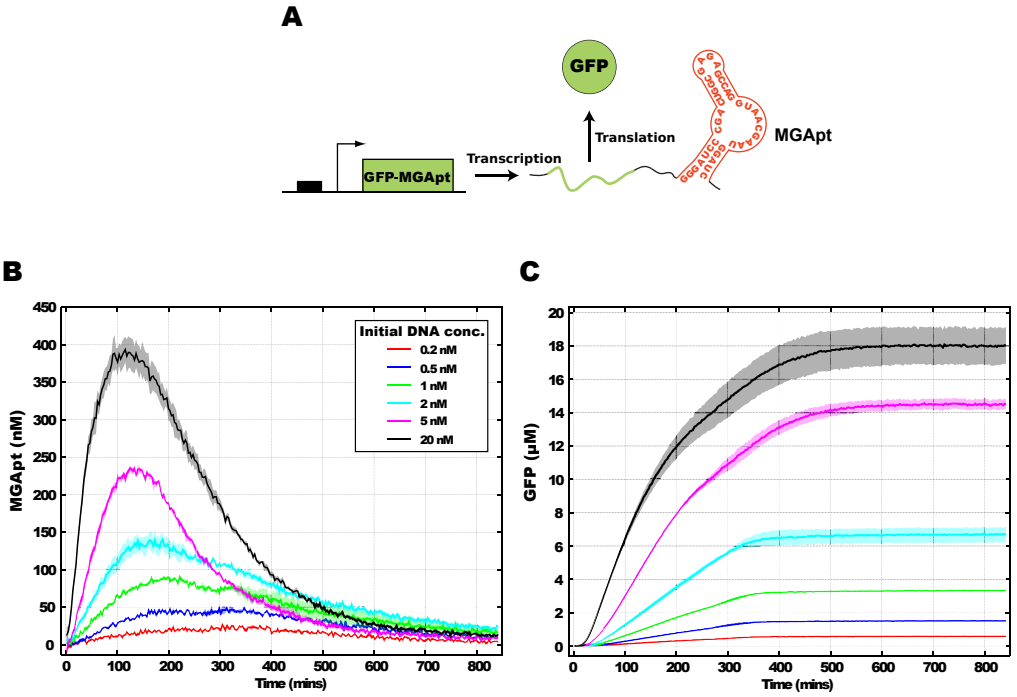


FIGURE 5.2: The plasmid DNA contains a constitutive promoter, then an untranslated region with aptamer and finally the GFP gene. After transcription, mRNA dynamics were measured by utilizing an RNA aptamer for the fluorescent dye malachite green [48]. Finally, translation creates a fluorescent GFP protein. Readers should note that the MGApt emission is in the far red region whereas the GFP emission is in the green spectra.

instructions on the cell extract preparation can be found in [110].

Experiments took place in a 10 μl reaction volume at 29°C over 14 hours, and each experiment was repeated three times. The initial plasmid DNA concentration was varied between 0.1 nM and 20 nM (see panel B and C in Figure 5.2 for MGApt and GFP measurements, respectively).

5.2 Process Model

The process model tries to capture the dynamics of the gene expression system with a set of mass action kinetic reactions. The process of gene expression consists of two major steps. First, the RNA is transcribed from the DNA, then this RNA is translated as an unfolded protein. Figure 5.3 shows that how this process model is built around the "central dogma" of molecular biology.

DOI:10.15774/PPKE.ITK.2015.008

5.2.1 Initial modeling steps

We start the model building from the simplest model for gene expression which contains only 4 reactions [122]

$$\text{DNA} \xrightarrow{\alpha} \text{DNA} + \text{mRNA}$$

$$\text{mRNA} \xrightarrow{\beta} \text{mRNA} + \text{Protein}$$

$$\text{mRNA} \xrightarrow{\gamma} \emptyset$$

$$\text{Protein} \xrightarrow{\rho} \emptyset \cdot$$

These four reactions facilitate the mRNA and protein production and also accounts for the degradation of mRNA and protein.

A simple state space model can be formulated for this reaction network

$$\begin{aligned} \dot{m} &= \alpha - \gamma m \\ \dot{P} &= \beta m - \rho P \end{aligned}$$

where m and P track the mRNA and protein concentrations in the system, respectively.

In this simple model the steady-state level for each state variables is dictated by the balance between the production and degradation.

We will extend this model in two directions in order to accurately describe our *in vitro* gene expression system. First, these reactions are facilitated by enzymes such as RNA polymerase and Ribosome. In addition to that enzymes are finite and shared resources in this system. Thus, modeling of the enzyme loading is an important as aspect of the final model. Second, the building blocks of mRNAs and proteins (nucleotides and amino acids) are also finite in the system. Both directions are equally important to accurately characterize the main operation regimes of the *in vitro* system.

Finally, there are other specialties of the gene expression system that we are explain in details at the corresponding step in the model building process.

DOI:10.15774/PPKE.ITK.2015.008

5.2.2 Transcription

Transcription, the process by which enzymes use nucleotide bases to create an RNA transcript from DNA, is modeled using the following chemical equations:

$$\mathrm{RNAP} + \sigma \underset{p_3}{\overset{p_2}{\rightleftharpoons}} \mathrm{RNAP} : \sigma \tag{5.1}$$

$$\mathrm{DNA} + \mathrm{RNAP} : \sigma \underset{p_8}{\overset{p_7}{\rightleftharpoons}} \mathrm{DNA} : \mathrm{RNAP} : \sigma$$

$$\mathrm{NTP} + \mathrm{DNA} : \mathrm{RNAP} : \sigma \underset{p_{10}}{\overset{p_9}{\rightleftharpoons}} \mathrm{NTP} : \mathrm{DNA} : \mathrm{RNAP} : \sigma$$

$$\mathrm{NTP} : \mathrm{DNA} : \mathrm{RNAP} : \sigma \xrightarrow{p_{12}} \mathrm{DNA} + \mathrm{RNAP} : \sigma + \mathrm{mRNA}\,.$$

In the above equations the $p_i \in \mathbb{R}_+$ are the kinetic rate coefficients of the reactions and these equations describe transcription as a three step process: the activation of the core RNA polymerase (RNAP) by a sigma factor protein (σ) to form the holoenzyme (RNAP:σ); the binding of this activated RNAP to DNA to rapidly form a larger 'enzyme-template' complex DNA:RNAP:σ; and the use of this complex with NTPs to produce mRNA. When other σ factors are utilized for transcriptional control (e.g., [101]), our approach can effectively capture the competition for core RNA polymerase (e.g., [104]). We do not model initiation, elongation and termination separately, instead we simply model the production of the whole mRNA transcript in a single enzymatic step, as shown in the last reaction in the list above. (A more detailed model of mRNA production, including competitive inhibition can be found in [8]). It should be noted that, as a modeling assumption we do not distinguish between the functional segments of the DNA (e.g, promoter, terminator, etc.), instead we model it just as a single species.

For a typical transcription reaction utilizing sigma factor σ^{70}, the complex RNAP:σ will effectively be at an equilibrium after an initial transient. Furthermore, the four RNA nucleotides (ATP, GTP, CTP and UTP) are summed and lumped into a single species (NTP,x_5). The NTP concentrations are determined from the buffer preparation protocol ([ATP]=[GTP]=1.5 mM, [CTP]=[UTP]=0.9 mM) [110]. The stoichiometry of the NTP usage in these reactions have numerical implementation issues, which are addressed in Section A.2.1.

The cell extract also contains the machinery for mRNA degradation (presumably several endonucleases and exonucleases remain active in cell-free extract), which has been

DOI:10.15774/PPKE.ITK.2015.008

studied in a previous work [63].

The degradation enzymes attack all the complexes containing mRNA, but we can only measure the aggregated effect of these degradation pathways. Thus, a model with multiple degradation pathways for mRNA can lead to structural identifiability issues, since many parameter combinations of the individually not measurable pathways can lead to the same overall rate of mRNA degradation. Because of that, we model mRNA degradation as a first order reaction:

$$\text{mRNA} \xrightarrow{p_{17}} \emptyset \cdot \tag{5.2}$$

Using purified mRNA, we have independently measured that the mRNA degradation in the cell-free extract, which follows first-order kinetics with a half life in range of 12-16 minutes, as it can be seen in Figure A.1, the details can be found in [J1].

5.2.3 Translation

The transcribed messenger RNA (mRNA) is then used as a template to create protein via translation. In detailed models, a ribosome (Ribo) sits on the mRNA transcript and charged transfer RNA (tRNA) molecules are used to transport specific amino acids to the Ribo:mRNA complex to initiate and elongate a polypeptide chain. The specific tRNA that binds to the elongation site, and hence the amino acid incorporated, depends on the nucleotide triplet (codon) on the mRNA being read. The endogenous enzyme, aminoacyl-tRNA-synthetase, charges the tRNA molecules with amino acids; we add 8 μM of additional tRNA into the buffer to ensure that there are sufficient tRNAs in the mixture. Furthermore, we ensure that there is an excess of AA in the mixture (1.5 mM each), so that the various tRNAs can be assumed to be constantly charged by their respective AAs. This assumption is valid till the concentrations of the AAs fall close to that of the tRNAs (on the order of 10 μM); typically NTPs run out before AAs. Thus, we approximate this saturation regime of tRNA-AA through a reaction where a single species of AA binds to the ribosome-mRNA complex with the value binding constant set close to total tRNA levels. The rate of this reaction lumps the time required for tRNA charging and the transport of tRNA to the ribosome by diffusion. In all, we

DOI:10.15774/PPKE.ITK.2015.008

model translation as follows:

$$\text{mRNA} + \text{Ribo} \underset{p_6}{\overset{p_5}{\rightleftharpoons}} \text{mRNA} : \text{Ribo} \tag{5.3}$$

$$\text{AA} + \text{mRNA} : \text{Ribo} \underset{p_{14}}{\overset{p_{13}}{\rightleftharpoons}} \text{AA} : \text{mRNA} : \text{Ribo}$$

$$\text{AA} : \text{mRNA} : \text{Ribo} \xrightarrow{p_{16}} \text{mRNA} + \text{Ribo} + \text{Protein}\,.$$

Here, we have lumped initiation, elongation, and termination into a single reaction of the ribosome-mRNA-AA complex leading to protein production and the dissociation of the complex. To avoid numerical issues and to keep track of material balance (stoichiometry), we also implemented an 'AA consumption' reaction (see A.2.1), whose rate depends on the translation rate and the length of protein. The energy requirement of translation is not modeled, and this is partially justified because ATP and GTP are provided in excess of CTP and UTP in anticipation of the additional usage for translation (about one ATP and one GTP are used for each amino acid incorporated). An explicit resource usage model for keeping track of each NTP separately is planned to be explored in a future work. The protein concentration does not achieve a steady state via balance of production and degradation, which is contrast with the simple model in Equation (5.1). Therefore, the final protein concentration—due to lack of protein degradation—remains at a constant level when the system runs out of resources.

The variant of GFP (x_7) that we used in the experiments requires 5-7 mins (p_4) to develop the fluorophore and become visible (x_8) [102]. This maturation process is modeled as a first order reaction: $\text{GFP} \xrightarrow{p_4} \text{GFP}^*$.

5.2.4 Resource degradation

The dynamics of *in vitro* systems are largely influenced by the finite amount of resources and the change of conditions (e.g., waste accumulation, pH change, etc.) during biocircuit operation. This has been known for some time; an early paper on cell-free expression highlighted how ATP degradation leads to a decrease in protein production [75]. Recent experimental and computational studies on resource effects have shown that the operational lifetime of a system can be extended by maintaining optimal pH and replenishing resources in such a way that enzymes in the system remain functional over long periods of time [101, 102, 109]. These earlier findings led us to incorporate

DOI:10.15774/PPKE.ITK.2015.008

degradation of transcriptional and translational resources as necessary components of our process model.

Based on that transcriptional resource (NTP) degradation modeled as a first order reaction which is sufficient to capture the decay of transcriptional activity.

Similar modeling assumption can be made for the translational resources. In our previous work we showed that with additional nucleotides, transcription produces significantly more mRNA, but translation output is roughly the same (Figure 3 in [J1]). To accommodate this observation in our model, we include a same type of degradation for translational resources (AA).

Finally, the transcription reaction slows down and eventually stops because the NTP levels become too low or the build up of ADP relative to ATP makes enzyme reactions energetically unfavorable [65]. Instead of modeling the various NTPs or their phosphorylation states separately, we model the decrease in ATP usability in a lumped manner through the first order degradation of NTP. This decrease in NTP concentration eventually leads to a point where the mRNA production rate falls below the mRNA degradation rate, and eventually causing the mRNA concentration to drop to zero.

5.2.5 State-space model

We chose a mass action kinetics (MAK) framework for modeling transcription and translation in the *in vitro* system. Unlike many of the common models used in synthetic biology, this model explicitly accounts for resource consumption in order to cover resource limits and resource sharing effects. Mass action–based modeling has a number of advantages, including the ease with which stochastic solvers can be applied (e.g., to investigate the dynamics in the non-deterministic regime [72]).

The model of transcription-translation is written in the following general non-linear form:

$$\begin{aligned} \dot{x} &= f(x, P), \\ x(0) &= x_0, \end{aligned} \tag{5.4}$$

DOI:10.15774/PPKE.ITK.2015.008

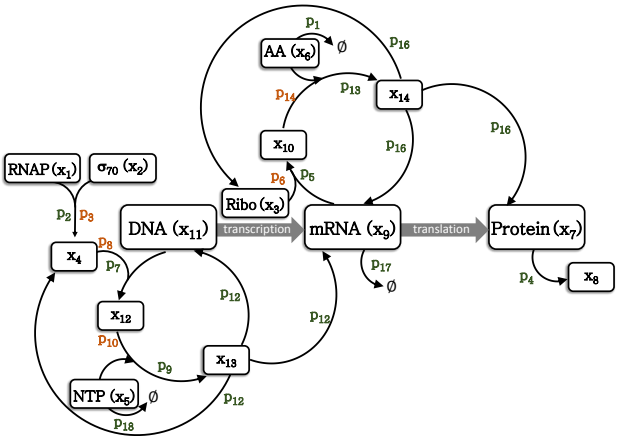


FIGURE 5.3: Overview of the process model. The model is built around the central dogma of molecular biology with additional step accounting for resource consumptions and degradations. Forward and reverse reaction rate coefficients are denoted with green and orange colors, respectively. The $x_1, \ldots, x_{14}$ are the species concentrations. The parameters p_{11} and p_{15} are not shown, see Appendix A.2.1 for details.

where $x : \mathbb{R} \to \mathbb{R}^n_+$ is the state vector, $P \in \mathbb{R}^m_+$ denotes the vector of model parameters, in our case $n = 14$ and $m = 18$. Table 5.2 lists the non-zero initial values used in this chapter.

The dynamics of the transcription and translation described in Subsections 5.2.2 and 5.2.3 is given by the following system of ODEs:

$$\begin{aligned}
\dot{x}_1 &= -F_1 & \dot{x}_2 &= -F_1 \\
\dot{x}_3 &= -F_3 + F_{10} & \dot{x}_4 &= F_1 - F_4 + F_7 \\
\dot{x}_5 &= -F_5 - F_{12} & \dot{x}_6 &= -F_8 - F_{13} \\
\dot{x}_7 &= -F_2 + F_{10} & \dot{x}_8 &= F_2 \\
\dot{x}_9 &= -F_3 + F_7 + F_{10} - F_{11} & \dot{x}_{10} &= F_3 - F_8 + F_9 \\
\dot{x}_{11} &= -F_4 + F_7 & \dot{x}_{12} &= F_4 - F_5 + F_6 \\
\dot{x}_{13} &= F_5 - F_6 - F_7 & \dot{x}_{14} &= F_8 - F_9 - F_{10}.
\end{aligned} \tag{5.5}$$

The $F_i, i = 1, \ldots, 13$ appearing in the above ODEs are the following:

$$\begin{aligned}
F_1 &= p_2 x_1 x_2 - p_3 x_4, \;\; F_2 = p_4 x_7, \\
F_3 &= p_5 x_3 x_9 - p_6 x_{10}, \;\; F_4 = p_7 x_4 x_{11} - p_8 x_{12}, \\
F_5 &= p_9 x_5 x_{12} - p_{10} x_{13}, \;\; F_6 = p_{11} x_{13}, \\
F_7 &= p_{12} x_{13}, \;\; F_8 = p_{13} x_6 x_{10} - p_{14} x_{14},
\end{aligned}$$

DOI:10.15774/PPKE.ITK.2015.008

$$
\begin{aligned}
F_9 &= p_{15}x_{14}, \;\; F_{10} = p_{16}x_{14}, \\
F_{11} &= p_{17}x_9, \;\; F_{12} = p_{18}x_5, \\
F_{13} &= p_1 x_6,
\end{aligned}
\tag{5.6}
$$

where $p_i \in \mathbb{R}_+, i = 1, \ldots, 18$ are the positive model parameters (reaction rate coefficients), i.e. $P = [p_1, \ldots, p_{18}]^T$. The elements of the state vector x are the species concentrations and Table A.1 lists which species is corresponding to which state variable.

TABLE 5.2: The table lists the species with non-zero initial concentrations in the model. Resource (R) type species are established by the crude-cell extract protocol [110]. The values in case of enzyme (E) type species are taken from the literature. We took the average concentration of the nucleotides (ATP, CTP, GTP, UTP) and denoted the average value as NTP. (The same goes for amino acids.)

Species	State	Initial value	Type	Source
[NTP]	x_5	1.2 mM	R	Protocol in [110]
[AA]	x_6	1.5 mM	R	Protocol in [110]
[RNAP]	x_1	100 nM	E	Table S5 in [101]
[Ribo]	x_3	1000 nM	E	Table S5 in [101]
$[\sigma_{70}]$	x_2	35 nM	E	Table S5 in [101]

5.2.6 Measured outputs

The measured MGApt signal is the experimental measure of mRNA concentration and thus the total concentration of mRNA within the system can be calculated via

$$
[mRNA]^{tot} = x_9 + x_{10} + x_{14}. \tag{5.7}
$$

The matured GFP protein is a final product in the system and its concentration is represented with state variable x_8. Thus, the observed outputs can be written as

$$
h_1(x) = S_1[mRNA]^{tot}, \tag{5.8}
$$

$$
h_2(x) = S_2 x_8. \tag{5.9}
$$

Each output was measured in relative fluorescence (a.u.) and then converted to nM via scaling factors $S_1 = 7.75$ a.u./nM and $S_2 = 1.723$ a.u./nM. These scaling factors were determined in [J1]. We note that these values are highly machine specific and valid

DOI:10.15774/PPKE.ITK.2015.008

only in the concentration range considered in this Chapter. This means that, the same calibration process has to be carried out as in [J1] to determine the scaling factors for the actual machine. On the top of that these scaling factors should be reevaluated overtime as the light intensity of the built in light source decreases.

5.3 Analysis of the Process Model

The goal of this section is twofold. First, we briefly examine the dynamics of the model and draw some conclusions from it. Then, we check the model structure itself whether it allows us (at least theoretically) to uniquely determine the model parameters.

5.3.1 mRNA dynamics

During the experiments we observed that a peak in mRNA production occurs around 150 min (see Figure 5.2, left panel). By doing simple calculations, we can find a relation between reaction rates that is valid in that time instant. The total mRNA concentration is given by Equation (5.7). From this, we can calculate that an extremum in mRNA concentration may occur when $\dot{x}_9 + \dot{x}_{10} + \dot{x}_{14} = 0$. Summing the corresponding right hand sides in Equation (5.5) we obtain

$$p_{12}x_{13} - p_{17}x_9 = 0. \tag{5.10}$$

In Equation (5.10) the first term is the transcription rate and the second part is the degradation rate of mRNA. On the other hand, the value of x_{13} depends on the concentration of NTP (x_5), which decreases over time. Therefore, the maximum of mRNA level occurs when the two terms in Equation (5.10) are equal as shown in Figure 5.4.

5.3.2 Steady state assumption

We can somewhat simplify the model by considering the fact that the sigma factor (x_2) binds with the RNAP (x_1) on a time-scale that is much faster than those of other reactions. Thus, we assume that $\dot{x}_1 = 0$ and $\dot{x}_2 = 0$, and then the differential equation

DOI:10.15774/PPKE.ITK.2015.008

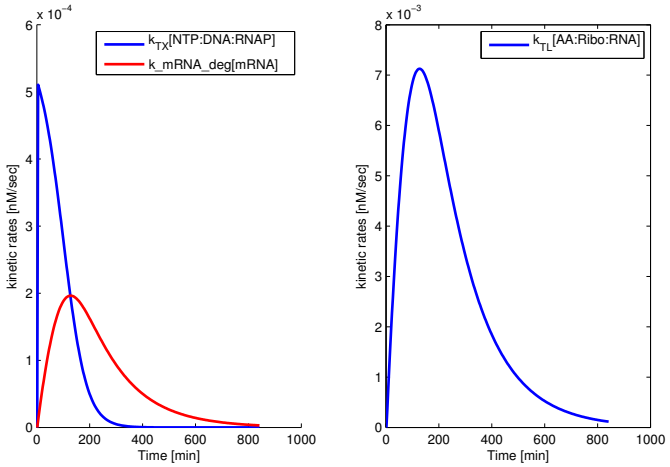


FIGURE 5.4: Simulation of reaction rates for the mRNA dynamics and transcription dynamics. The reaction rate for transcription ($p_{12}x_{13}$) decays over time and eventually crosses the rate of mRNA degradation ($p_{17}x_9$) causing a peak in mRNA production.

for state x_4 becomes

$$\dot{x}_4 = -F_4 + F_7. \tag{5.11}$$

This way, we do not have to consider the dynamics of x_1 and x_2 and estimate or find values from literature for their parameters (p_2, p_3). However, we need to estimate the initial value of x_4 denoted by x_{4init}.

5.3.3 Structural Identifiability

At this point we can check whether it is theoretically possible to determine the model parameters based on the model structure and the observables $(h_1(t), h_2(t))$.

For the structural identifiability analysis we use the generating series approach (see Section 2.5) and checked the model structure and its output with the GenSSI toolbox [26].

Calculating the generating series is computationally intensive, and the computation time rapidly grows with the increasing number of parameters checked by the algorithm. Thus, we had to limit the number of parameters in the identifiability analysis by assuming that the mRNA degradation (p_{17}) is known from our independent measurement of mRNA degradation (see Figure A.1 and [J1] for details). After that, we checked a model with 12 reaction rate coefficients and two initial value denoted by vector $P_{id} \in \mathbb{R}_+^{14}$. According

DOI:10.15774/PPKE.ITK.2015.008

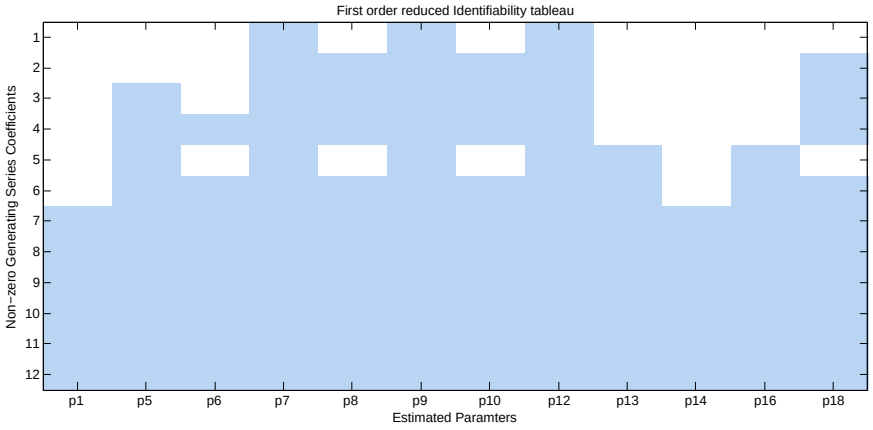


FIGURE 5.5: Result of the structural identifiability which is called the identifiability tableau. The Jacobian of generating series coefficient were calculated and the non-zero elements of the Jacobian are shown. This helps to reduce and eventually solve the underlying algebraic equations; see [26] for details. The GenSSI software does not show the parameters for initial values of the system on the identifiability tableau.

to the report generated by the GenSSI toolbox the set of nonlinear equations had more than one solution, we can thus conclude that our model is at least locally identifiable. The graphical output of the structural identifiability analysis can be seen in Figure 5.5. Each blue field in the matrix represents a non-zero element in the Jacobian of the non-zero generating series coefficients, which is used to solve system of nonlinear equations, see Section 2.5 for details.

5.4 Parameter Estimation

Our parameter estimation procedure is based on the commonly applied minimization of the distance between the measured and model computed output. The statistical evaluation and validation of the parameters is also carried out to investigate the quality of the parameter estimation and to identify parameters where further experiments may need to decrease uncertainty.

Figure 5.6 shows the mRNA and GFP measurements with four different initial DNA concentrations between 1 nM and 10 nM. These measurements were taken according to the procedure detailed in Section 5.1.2. For each measurement with different initial DNA concentration, the mean and standard deviation were calculated and shown in Figure 5.6 as colored solid lines and correspondingly colored shaded areas, respectively.

DOI:10.15774/PPKE.ITK.2015.008

5.4.1 Prediction error minimization

After model reduction and taking the parameter values from the literature into account we can formulate the $P_{est} \in \mathbb{R}_{+}^{15}$ parameter vector, which consists of 13 reaction rate coefficients $[p_1, p_5, \ldots, p_{10}, p_{12}, \ldots, p_{14}, p_{16}, \ldots, p_{18}]$ and two initial values ($x_{3_{init}}$ and $x_{4_{init}}$). Since a kinetic system requires positive parameter values, we restrict the range of possible parameters onto the positive orthant with appropriate lower limits. Furthermore, we utilize the measurement error from the independent measurement of the mRNA degradation (see [J1] for details) to restrict the possible values for mRNA degradation rate coefficient (p_{17}). Different starting points for the parameter estimation were generated with hypercube sampling from a uniform distribution [108].

The model has two measured outputs $h_1(t)$ and $h_2(t)$ for the mRNA and GFP concentrations, respectively. To incorporate the multiple outputs into the cost function we normalize each term with the maximum of the corresponding time series data $\bar{y}_k(t) = y_i^{(k)}(t)/max(y_i^{(k)}(t))$, where k is the index of consecutive experiments with different initial DNA concentrations and i denotes the measured outputs. We did the same normalization with the model output $\bar{h}_i^{(k)}(t) = h_i^{(k)}(t)/max(y_i^{(k)}(t))$. This leads to the following cost function

$$C(P_{est}) = \sum_{k=1}^{M} \sum_{t=1}^{T} \left[\bar{y}_1^{(k)}(t) - \bar{h}_1^{(k)}(x(t), P_{est})\right]^2 + \left[\bar{y}_2^{(k)}(t) - \bar{h}_2^{(k)}(x(t), P_{est})\right]^2, \tag{5.12}$$

where $M = 4$ is the number of different experiments we consider and T=280 is the number of samples and samples are taken in every 3 min. Then, we need to solve the following constrained least squares optimization problem

$$\begin{aligned} &\underset{P_{est} \in \mathcal{P}}{\operatorname{argmin}}\, C(P_{est}) \\ &0 \leq P_{est} \leq UB, \end{aligned} \tag{5.13}$$

where $\mathcal{P}$ is the set of feasible parameters and UB is the vector of upper bounds. The optimization problem stated in Equation (5.13) was performed with a gradient-free global optimizer implemented as a pattern-search [11]. The gradient-free optimization approach was selected to avoid interference between the accuracy of the ODE solver and

DOI:10.15774/PPKE.ITK.2015.008

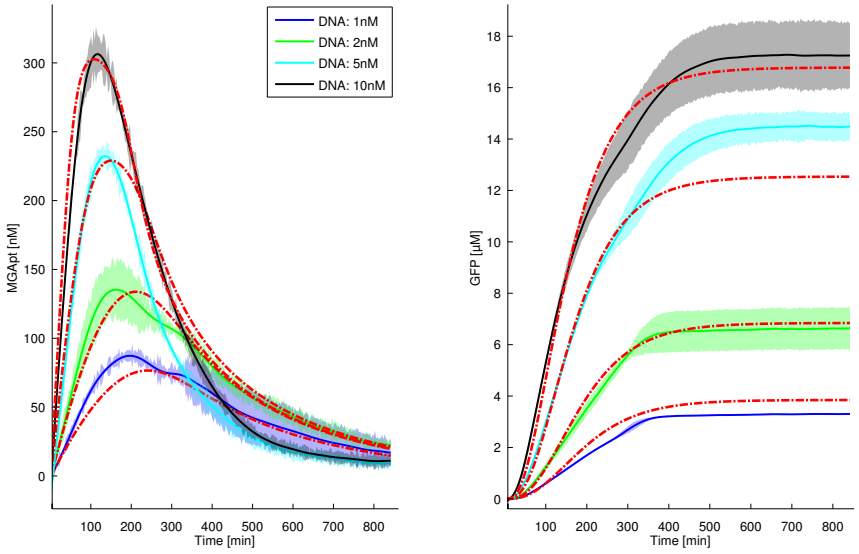


FIGURE 5.6: Simulations with the estimated parameter set is shown in red. The Figure contains time series measurement for both channels with different initial concentration DNA (1nM, 2nM, 5 nM and 10 nM plasmid DNA concentration, shown in green, red, cyan and black respectively). The left panel shows the dynamics of MGApt, which is proportional to the mRNA concentration. The GFP dynamics is shown on the right panel. The fluorescent counts for each channel have been converted by applying Equations (5.8) and (5.9) to nM and μM, respectively.

the finite differentiation for gradient calculation which commonly occur when gradient based optimization is applied in this setup (see Subsection 2.4 for details on this issue).

During the parameter estimation we found numerous local minima, where the mRNA degradation varied over several orders of magnitude. In our previous study [J1], we conducted independent measurement of mRNA degradation in the *in vitro* system (see Section 5.1 for the details). From that study, the measurement puts the mRNA half-life in a range of 12–16 min [101]. Thus, we used that information to restrict the mRNA degradation rate (p_{17}) in the parameter estimation process.

The result of the parameter estimation is summarized in Table 5.3 and simulations of the estimated parameter set (solid red lines) overlapped with the measurements is shown in Figure 5.6.

Finally, we validated the estimated parameters over a different set of data. Originally, we had measurements with different initial DNA concentrations between 0.1 nM and 20 nM. We divided this data set and used measurements from 1 nM to 10 nM for the parameter estimation. For validation, we used the 0.1 nM to 0.5 nM concentration range. We found that, on average we have 20% error in the final GFP production. This

DOI:10.15774/PPKE.ITK.2015.008

TABLE 5.3: Numerical result of the parameter estimation. The table has 13 reaction rate coefficients and two initial concentrations.

	Name	Value	Coefficient of variation	Unit
1	p_1	1.90×10^{-4}	9.88%	1/s
2	p_5	6.94×10^{-2}	29.9%	1/(s·nM)
3	p_6	8.43×10^{-1}	30.32%	1/s
4	p_7	7.00×10^{-3}	29.93%	1/(s·nM)
5	p_8	1.38×10^{2}	79.52%	1/s
6	p_9	6.20×10^{-2}	40.23%	1/(s·nM)
7	p_{10}	3.55×10^{-1}	108.84%	1/s
8	p_{12}	8.90×10^{-3}	6.34%	1/s
9	p_{13}	5.95×10^{-2}	28.43	1/(s·nM)
10	p_{14}	2.33×10^{5}	75.21%	1/s
11	p_{16}	3.02×10^{-1}	63.26%	1/s
12	p_{17}	1.07×10^{-3}	6.56%	1/s
13	p_{18}	2.98×10^{-4}	8.47%	1/s
14	x_{3_0}	375.58	16.62%	nM
15	x_{4_0}	276.13	6.18%	nM

is acceptable for three reasons. First, the qualitative features of the simulations are still acceptable in comparison with the measurements. Second, this mount of steady-state error is acceptable since most of the biocircuits are sensitive to fold-change [47]. Third, the presence of malachite green aptamer affects the protein production rate, as it can be seen in Figure A.2, so we can only have approximate figures for the parameters related to the final protein concentration.

Numerical Implementation The simulated ODE model is stiff [99] — most likely as a result of the NTP (x_5) and AA (x_6) consumption — thus we used the efficient CVODES solver [98] to solve the model ODEs, also ODEmex software was applied for further speed gain [121].

5.4.2 Statistical Analysis of the Parameter Estimation

To evaluate the quality of the parameter estimation we used a Markov-Chain Monte Carlo (MCMC) implementation to generate the joint posterior distribution of the parameters. The MCMC procedure was initiated at the result of the point estimation with uniform prior and with the same lower and upper bounds that was used in Equation (5.13). The detailed version of the algorithm can be found in [52]. From the results

DOI:10.15774/PPKE.ITK.2015.008

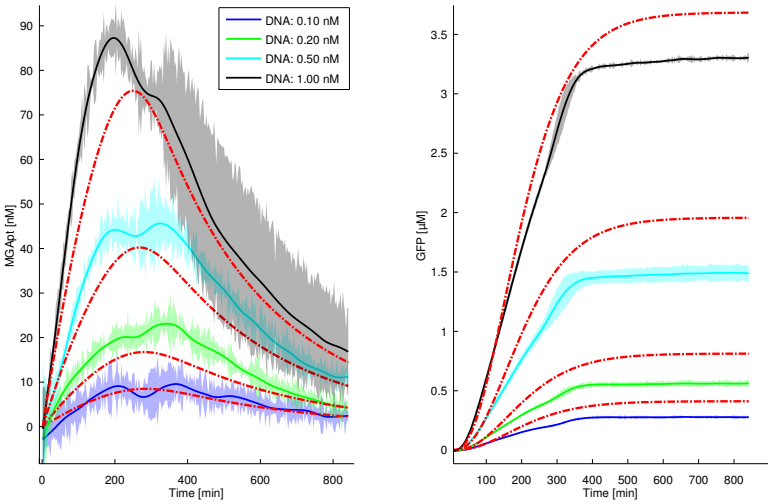


FIGURE 5.7: Validation of the estimated parameter set over another set of data (0.1 nM, 0.2 nM, 0.5 nM of plasmid DNA). The red curves show the corresponding simulations with the parameters from Table 5.3. On average there is 20% error in the final value of GFP, the simulation qualitatively matches the time series data. The 1 nM data (black curve) was used for estimation, shown here only for comparison.

of the MCMC algorithm, either through counting statistics or through the covariance matrix we can establish the coefficient of variation for the parameters [9].

We ranked and listed the cross-correlations larger than 0.5 in Table 5.4. It shows three groups of cross-correlations. In the first one, the translation rate (p_{16}), the translational resource binding (p_{14}) and the Ribosome bindings (p_5, p_6) are highly correlated. This may suggest that there is not enough information in the measurement data to determine the correct parameters for all stages of translation. In the second group, there is a cross-correlation between the promoter strength (p_8) and translational reactions (p_{14}, p_{16}). In the third group, the same translation reaction coefficients (p_{14}, p_{16}) are grouped with the initial concentration of the sigma factor activated RNA polymerase ($x_{4_{init}}$).

The result of the cross-correlation analysis will certainly help in designing future experiments to improve parameter estimations. Most of the parameters in Table 5.4 are related to translation. In order to get a better estimate of these parameters, we have to manipulate Ribosome binding strength and/or Ribosome concentration in the *in vitro* system. These are possible future directions to design targeted experiments in order to enhance the quality of the parameter estimation for some of the parameters. Considering the coefficient of variation of the parameters we can see that some of the parameters are accurately estimated, e.g. p_1, p_{18} (resource degradation), p_{12} (transcription rate). This suggests that we have a good estimate for the resource degradation and some of

DOI:10.15774/PPKE.ITK.2015.008

TABLE 5.4: The table contains parameter pairs with the strongest cross-correlation.

Parameter	Parameter	Correlation
p_{14}	p_{16}	0.942
p_5	p_6	0.929
p_8	p_{14}	0.702
p_8	p_{16}	0.645
p_{14}	$x_{4_{init}}$	0.574
p_{16}	$x_{4_{init}}$	0.551

the transcription related parameters, e.g. p_{12}, p_7 and p_9. On the other hand, coefficient of variation for p_8 (promoter strength), p_{10} (NTP binding), p_{14} (AA binding) and p_{16} (translation rate) are very large. Therefore, we need another way to decrease our coefficient of variation in these parameters. One possibility was highlighted at the cross-correlation analysis, but these resource binding related parameters can be measured with tedious experiments (if measurement is possible at all). Therefore, it may be necessary to do some level of model reduction to improve the estimation result.

Besides analyzing the coefficient of variation and evaluating the cross-correlation, we can take advantage of the fact that we calculated the full joint posterior distribution of the parameters. Hence, we can select regions from the parameter distribution (via confidence intervals) and take all the parameters from a selected region and simulate them. Roughly speaking, we can visualize how the dynamics is 'spreading' by the variations of the parameters. The results computed by using the parameters in Table 5.3 is shown in Figure 5.8. where the general shape of the dynamics is the same over different confidence intervals (99%, 95%, 90%, 50%). Although higher uncertainty arises around the mRNA peak time and peak value and around the steady-state level of GFP.

5.5 Summary

We have demonstrated a reduced order model for our cell-free system, where the modeled states captured the observable dynamics with low model complexity. The proposed model is capable to simulate synthetic biocircuits. The model also has predictive capability due to the detailed modeling of the resources (NTP, AA). Finally, the experimental setup enabled us to monitor the mRNA level in real time, this leads to a better estimation of the model parameters.

DOI:10.15774/PPKE.ITK.2015.008

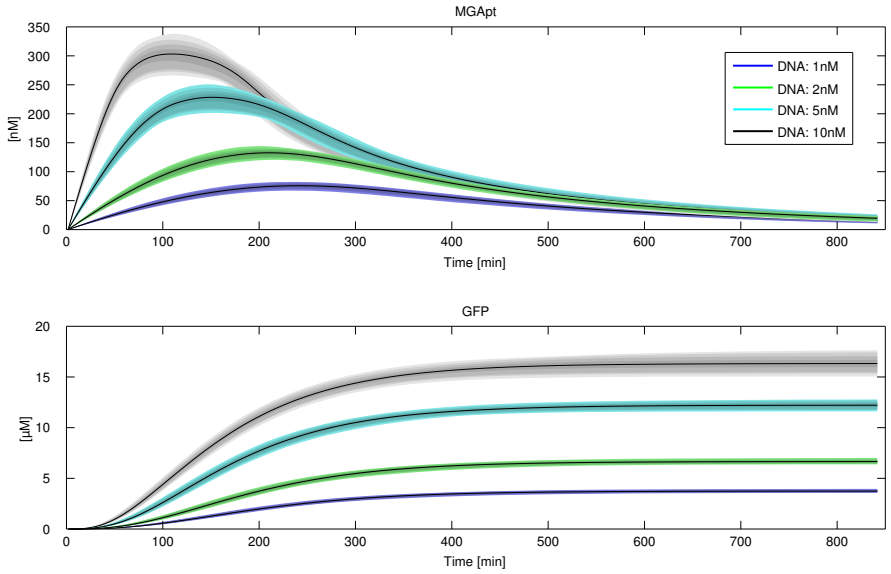


FIGURE 5.8: Samples for the joint posterior distribution of the parameters were drawn with different confidence intervals (99%, 95%, 90%, and 50%), we simulated the dynamics with these parameter sets.

The demonstrated model analysis has explained the observed mRNA dynamics and has allowed the steady state assumption-based model reduction. To ensure proper model structure, we checked—the often neglected—structural identifiability of the improved model. This model turned out to be at least locally structurally identifiable, that provides a good foundation for parameter estimation. For the parameter estimation, we applied a derivative-free pattern search method accommodating multi-channel multi-experiments data. The resulting parameter set was statistically evaluated and validated on a different data set. Statistical analysis revealed some uncertain parameters that we attempted to explain from biochemical process and the experimental setup points of view. Based on these results, we can focus our future work on improving the experiment design and possibly further reducing the dynamical model itself.

DOI:10.15774/PPKE.ITK.2015.008

Chapter 6

Conclusions

Building upon a set of mathematical optimization based approaches to investigate properties of kinetics systems and the associated directed graph structures, this thesis has presented two new optimization based algorithms.

It has been known that the sparse directed graph structure is not necessarily unique which is in contrast with the unique dense structure. This non-uniqueness may hamper the successful identification of a kinetic system because a unique sparse structure is often implicitly assumed. The first algorithm shows an efficient way to calculate all sparse directed graph structures of a kinetic system.

The second algorithm has been developed within the newly introduced class of uncertain kinetic systems. This algorithm with polynomial time complexity computes the core reactions of the uncertain kinetic system by applying linear programming. In these kinetic systems the uncertainty is represented as a multi-dimensional interval in the space of monomial coefficients. The versatility of this type of model makes it possible to accommodate various types of parametric uncertainties such as temperature change or different operation regimes.

Application examples have been demonstrated to show the operation of the algorithms. The capability of the first procedure has been illustrated by a well-known nonlinear system, the so-called Lorenz system. This system is suitable to exhibit chaotic behavior. To show the non-unique sparse structures of the Lorenz system, it has been transformed into a kinetic system with two different approaches in order to make it compatible with the algorithm. In the case of the second algorithm, a simple network reconstruction

DOI:10.15774/PPKE.ITK.2015.008

benchmark example was used to illustrate the computation of the core reaction set of an uncertain kinetic system. Moreover, it has been illustrated that the sparse structure of a kinetic system with predetermined uncertainty may not be unique.

The last part of this thesis is focusing on the modeling process of an *in vitro* system. The process has started with a list of molecular laboratory protocols to prepare the different molecular probes. Then, series of experiments has been performed to collect data about the cell-free gene expression system. From these experiments, a first principle model has been developed to capture the transient behavior of the gene expression in the *in vitro* the cell-free system. As a next step, the dynamical and structural properties of the model has been investigated which includes a structural identifiability analysis. Finally, the quality of the time series data enabled us to estimate and validate the parameters of the developed kinetic model.

6.1 New scientific results

Thesis I. ***I have developed a mathematical optimization based efficient algorithm to compute all dynamically equivalent sparse realizations of a kinetic system.***

Using combinatorial and mathematical optimization techniques, I have developed the first algorithm in the literature to compute all sparse realizations of dynamically equivalent kinetic systems. This algorithm uses mixed integer linear programming (MILP) and linear programming (LP) steps to compute all the sparse realizations.

Corresponding publications: [J2], [C2].

Thesis I.a

I have proposed an effective reduction of the combinatorially possible search space by using appropriate constraint-pairs and the properties of constrained sparse realizations.

The special properties of dense and sparse realizations made it possible to reduce the original search space which consists of all directed graphs with fixed set of nodes to a computationally tractable number of candidate structures.

DOI:10.15774/PPKE.ITK.2015.008

Thesis I.b

By applying state-dependent time-rescaling and X-factorable transformation, I have computed all sparse CRN structures for two different kinetic models of the well-known Lorenz system showing chaotic behavior. I have compared the obtained realizations from a structural point of view and determined the minimal and maximal number of linkage classes and deficiencies corresponding to the sparse realizations. I have shown that the complete search space was reduced to 0.01 and 0.0037 percents in the case of state-dependent time-rescaling and X-factorable transformation, respectively.

The Lorenz system was a good candidate to show the application of the developed algorithm. Two distinct approaches yielded different number of sparse structures in each case, but with similar structural properties. It was computationally checked that, the chaotic behavior of the system was preserved in all representations.

Thesis II. ***I have developed structural analysis tools for kinetic systems with parametric uncertainty.***

I have developed optimization based tools for the structural analysis of uncertain kinetic systems. The uncertainty in these systems are represented as a multi-dimensional interval in the space of monomial coefficients.

Corresponding publications: [C5], [C1].

Thesis II.a

I have proposed a new algorithm for the computation of dense and sparse reaction network structures for kinetic polynomial models, where the uncertainties are represented as parameter intervals. The problem is traced back to mixed integer linear programming where the parameter uncertainties are given by linear inequalities.

The current computational framework has been extended to accommodate parametric uncertainties while certain structural properties of the kinetic system—as well as the associated directed graph structure—can be effectively calculated. This approach opens

DOI:10.15774/PPKE.ITK.2015.008

up the possibility to extend several previous optimization based results to uncertain kinetic systems such as weak reversibility or complex balance.

Thesis II.b

I have developed an algorithm with polynomial time complexity to calculate the structurally invariant elements, called core reactions that are present in any reaction network belonging to the model set defined by the uncertainties. The proposed algorithm is based on linear programming and incorporates the parametric uncertainty of the system as element-wise boundary constraints.

The core reactions are one of the most important elements of the reaction graph, since if the set of core reactions is non-empty, its elements are present in each dynamically realization. Thus, the computation of distinct dynamically equivalent reaction network structures satisfying a given property can effectively be supported by utilizing the core reaction set.

Thesis III. ***I have built a kinetic model for an in vitro cell-free gene expression system.***

I have built a first principle kinetic model for an in vitro cell-free gene expression system using a specific experimental setup. I have shown that with the estimated and validated parameters the kinetic model effectively captures the dynamical features of the cell-free system.

Corresponding publications: [C4], [J1], [C3].

Thesis III.a

I have built combinations of molecular probes containing RNA aptamer and fluorescent proteins for the investigation of the cell-free system. I have designed and carried out a comprehensive study of the cell-free system utilizing concurrent measurement of transcription and translation.

DOI:10.15774/PPKE.ITK.2015.008

Concurrent measurement of transcription and translation was important for the development of a reliable kinetic model. The library of molecular probes made possible to study the dynamical features of the cell-free system in detail. As a result of that, this measurement set up serves as a benchmark for testing different versions of the cell-free system.

Thesis III.b

Based on the observations from the experiments, I have built a kinetic model for the studied cell-free system, which is capable of capturing the transient behavior of the system taking into consideration the finite resources. I have shown that the model is structurally identifiable using the applied measurement setup. Finally, I have determined and validated the parameters of the dynamical models using constrained least squares based parameter estimation.

Utilizing the library of molecular probes and domain knowledge about the process of gene expression, the resulting dynamical model is structurally identifiable. Also, the fact that the developed model structure is linear in parameters and the quality of the data enabled us to use the least squares based parameter estimation.

DOI:10.15774/PPKE.ITK.2015.008

6.2 Possible directions of future work

Given the broad topic of this thesis, several directions of future work are possible in different fields. In Chapter 3, we investigated the non-uniqueness of the dynamically equivalent sparse structure of a kinetic systems. The result illustrates that the sparse structure of kinetic systems may be non-unique. Moreover, the Lorenz example has a large number of sparse realizations. A possible direction of future work might be the graph theoretical investigation of these sparse structures. That could include the average incoming and outgoing connection of a complex in each realizations or the reaction distribution considering a complex in all the realizations.

As another direction, we could extend the proposed algorithm in Chapter 3 to accommodate uncertain kinetic systems which were introduced in Chapter 4. This would offer an algorithmic way to calculate all dynamically equivalent sparse realizations at a given level of uncertainty. This algorithm would be a powerful tool for network reconstruction where usually only noisy time series data and the set of vertices of the directed graph is known.

In Chapter 4, we briefly investigated the application of the core-reactions in network reconstruction. As a result of that, we have concluded that the core reaction set can be at least partially restored from the time series data. On the other hand, we did not extend the concept of core complexes to uncertain kinetic systems and did not include it to the computation. Thus, the application of the core reactions and the core complexes in an algorithm where they are calculated in an alternating fashion could be the backbone of an improved method for network reconstruction within the class of kinetic systems.

The introduction elaborated on the challenges of kinetic system identification. Among many important open questions, the relevance of optimization based experiment design for a molecular breadboard was highlighted. Chapter 5 laid the foundation of that work by introducing a dynamical model for the cell-free system on which the molecular breadboard is built on. Utilizing this kinetic model, an optimization based framework could be developed where the optimization task would consider all the physical limitations of the molecular breadboard.

DOI:10.15774/PPKE.ITK.2015.008

Appendix A

A.1 Special measurements

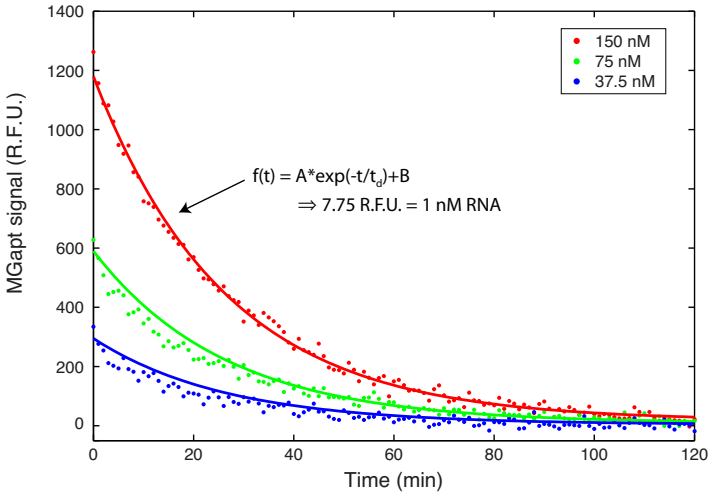


FIGURE A.1: Measurement of mRNA degradation in the gene expression system. The figure shows that, within the measured rage, mRNA degradation follows first order kinetics.

DOI:10.15774/PPKE.ITK.2015.008

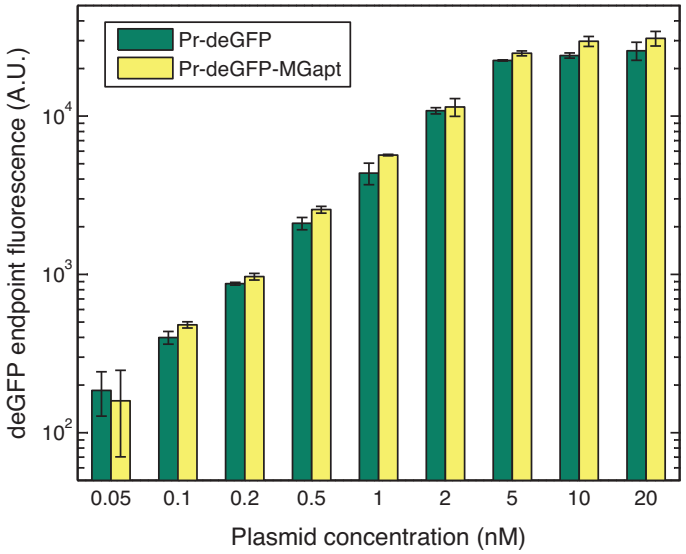


FIGURE A.2: The experiment compares the two constructs with MGApt and without it. The endpoint measurements (final concentrations level) show that the presence of MGApt increases the overall protein production.

TABLE A.1: Species - state variables

Species	State variables
[RNAP]	x_1
$[\sigma]$	x_2
[Ribo]	x_3
[RNAP : σ]	x_4
[NTP]	x_5
[AA]	x_6
[GFP]	x_7
[GFP*]	x_8
[mRNA]	x_9
[mRNA : Ribo]	x_{10}
[DNA]	x_{11}
[DNA : RNAP : σ]	x_{12}
[NTP : DNA : RNAP : σ]	x_{13}
[AA : mRNA : Ribo]	x_{14}

DOI:10.15774/PPKE.ITK.2015.008

A.2 TXTL Software toolbox

A.2.1 Implementation considerations

Both transcription and translation are implemented by ignoring the individual stages of initiation, elongation and termination. Under a naïve implementation scheme, transcription, for instance, would look like:

$$\mathrm{n}\cdot\mathrm{NTP}+\mathrm{DNA}:\mathrm{RNAP}:\sigma \rightleftharpoons \mathrm{n}\cdot\mathrm{NTP}:\mathrm{DNA}:\mathrm{RNAP}:\sigma$$

$$\mathrm{n}\cdot\mathrm{NTP}:\mathrm{DNA}:\mathrm{RNAP}:\sigma \longrightarrow \mathrm{DNA}+\mathrm{RNAP}:\sigma+\mathrm{mRNA}$$

with a large stoichiometric coefficient n in front of the NTP (AA) term in the chemical equation. This leads to numerical difficulties. First, the ODEs describing these reactions become very stiff due to an extremely large exponent in the NTP term. Second, the reaction order and hence rate depends on the length of the mRNA, because n is simply this length in nucleotide bases (this would also not be an accurate model, since the whole mRNA is not assembled instantaneously, like it would suggest).

We first alleviated these problems by lumping the NTP (AA) into units of 100, bringing reaction orders down to 10 (most of our genes are on the order of a thousand base pairs). We further implemented a heuristic approach whereby we used only one unit of NTP (AA) for each transcription (translation) reaction and coupled this to a 'consumption' reaction. This reaction effectively consumes NTP (AA) at a rate of $p_{11} = (n-1)p_{12}$ where p_{12} is the original transcription rate, and n is the length of the RNA (protein) sequence

$$\mathrm{NTP}:\mathrm{DNA}:\mathrm{RNAP}:\sigma \xrightarrow{p_{12}} \mathrm{DNA}+\mathrm{RNAP}:\sigma+\mathrm{mRNA}$$

$$\mathrm{NTP}:\mathrm{DNA}:\mathrm{RNAP}:\sigma \xrightarrow{p_{11}} \mathrm{DNA}+\mathrm{RNAP}:\sigma\cdot$$

A similar an auxiliary reaction (F_9) for AA consumption can be introduced with reaction rate coefficients $p_{15} = n/3p_{16}$

$$\mathrm{AA}:\mathrm{mRNA}:\mathrm{Ribo} \xrightarrow{p_{16}} \mathrm{mRNA}+\mathrm{Ribo}+\mathrm{Protein}$$

$$\mathrm{AA}:\mathrm{mRNA}:\mathrm{Ribo} \xrightarrow{p_{15}} \mathrm{mRNA}+\mathrm{Ribo}\cdot$$

DOI:10.15774/PPKE.ITK.2015.008

Publications Related to The Thesis

[J1] D. Siegal-Gaskins, Z. A. **Tuza**, J. Kim, V. Noireaux, and R. M. Murray. "Gene Circuit Performance Characterization and Resource Usage in a Cell-Free "Breadboard"". In: *ACS Synth Biol Impact Factor: 3.951* 3 (2014), pp. 416–25. DOI: 10.1021/sb400203p.

[C1] G. Szederkényi, Z. A. **Tuza**, and K. M. Hangos. "Determining biochemical reaction network structures for kinetic polynomial models with uncertain coefficients". In: *AIP Conference Proceedings-American Institute of Physics* 1479 (2012), pp. 2427–2430. DOI: 10.1063/1.4756685.

[C2] G. Szederkényi, Z. A. **Tuza**, and K. M. Hangos. "Dynamical Equivalence and Linear Conjugacy of Biochemical Reaction Network Models". In: *Biological and Medical Systems 2012*. Vol. 8. 2012, pp. 125–130. DOI: 10.3182/20120829-3-HU-2029.00042.

[C3] Z. A. **Tuza**, D. Siegal-Gaskins, J. Kim, and G. Szederkényi. "Analysis-Based Parameter Estimation of an in Vitro Transcription-Translation System". In: *European Control Conference 2015 (ECC2015)*. 2015.

[C4] Z. A. **Tuza**, V. Singhal, J. Kim, and R. M. Murray. "In Silico Modeling Toolbox for Rapid Prototyping of Circuits in a Biomolecular "Breadboard" System". In: *Decision and Control (CDC), 2013 51th IEEE Conference on*. 2013, pp. 1404–1410. DOI: 10.1109/CDC.2013.6760079.

[C5] Z. A. **Tuza** and G. Szederkényi. "Computing Core-Reactions of Uncertain Polynomial Kinetic Systems". In: *23rd Mediterranean Conference on Control and Automation*. IEEE, 2015, pp. 1140–1147.

[J2] Z. A. **Tuza**, G. Szederkényi, K. M. Hangos, A. A. Alonso, and J. R. Banga. "Computing All Sparse kinetic Structures for a Lorenz System using Optimization." In: *I. J. Bifurcation and Chaos Impact Factor: 1.017* 23.8 (2013). DOI: 10.1142/S0218127413501411.

DOI:10.15774/PPKE.ITK.2015.008

Other Publications of The Author

[J3] B. Ács, G. Szederkényi, Z. A. **Tuza**, and Z. Tuza. "Computing Linearly Conjugate Weakly Reversible Kinetic Structures Using Optimization and Graph Theory". In: *MATCH Communications in Mathematical and in Computer Chemistry Impact Factor: 1.829* (2015).

[C6] J. Rudan, Z. **Tuza**, and G. Szederkényi. "Using LMS-100 laser rangefinder for indoor metric map building". In: *IEEE International Symposium on Industrial Electronics - ISIE 2010, Bari, Italy, 4-7 July*. 978-1-4244-6391-6/10. IEEE, 2010, pp. 525–530. DOI: 10.1109/ISIE.2010.5637842.

[C7] Z. **Tuza**, J. Rudan, and G. Szederkényi. "Developing an integrated software environment for mobile robot navigation and control". In: *2010 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2010; Zurich; 15 September 2010 through 17 September 2010*. ISBN: 978-142445864-6. 2010, no. 5647506 (on CD). DOI: 10.1109/IPIN.2010.5647506.

DOI:10.15774/PPKE.ITK.2015.008

References

[1] A. Adamatzky, B. De Lacy Costello, and L. Bull. "On polymorphic logical gates in subexcitable chemical medium". In: *International Journal of Bifurcation and Chaos* 21 (2011), pp. 1977–1986.

[2] A. Adamatzky, B. De Lacy Costello, and T. Shirakawa. "Universal computation with limited resources: Belousov-Zhabotinsky and physarum computers". In: *International Journal of Bifurcation and Chaos* 18 (2008), pp. 2373–2389.

[3] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall, CRC, 2007, p. 320.

[4] J. Andersson. "A General-Purpose Software Framework for Dynamic Optimization". PhD thesis. Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium: Arenberg Doctoral School, KU Leuven, Oct. 2013.

[5] D. Angeli. "A tutorial on chemical network dynamics". In: *European Journal of Control* 15 (2009), pp. 398–406.

[6] M. Anguelova. "Nonlinear observability and identifiability: General theory and a case study of a kinetic model for S. Cerevisiae". PhD thesis. Chalmers University Of Technology University Of Gothenburg., 2004.

[7] M. Anguelova, J. Karlsson, and M. Jirstrand. "Minimal output sets for identifiability." eng. In: *Math Biosci* 239.1 (Sept. 2012), pp. 139–153. DOI: 10.1016/j.mbs.2012.04.005.

[8] S. Arnold, M. Siemann, K. Scharnweber, M. Werner, S. Baumann, and M. Reuss. "Kinetic Modeling and Simulation of *In Vitro* Transcription by Phage T7 RNA Polymerase". In: *Biotechnology and Bioengineering* 72 (2001), pp. 548–561.

[9] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter Estimation and Inverse Problems*. Elsivier, 2012.

[10] K. J. Åström and R. M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2011.

[11] C. Audet and J. J. E. Dennis. "Analysis of Generalized Pattern Searches". In: *SIAM J. Optim.* 13 (2006), pp. 889–903.

[12] E. August and A. Papachristodoulou. "Efficient, sparse biological network determination". In: *BMC Systems Biology* 3 (2009), p. 25.

DOI:10.15774/PPKE.ITK.2015.008

[13] E. Balsa-Canto, A. A. Alonso, and J. R. Banga. "An iterative identification procedure for dynamic modeling of biochemical networks." eng. In: *BMC Syst Biol* 4 (2010), p. 11. DOI: 10.1186/1752-0509-4-11.

[14] E. Balsa-Canto and J. R. Banga. "AMIGO, a toolbox for advanced model identification in systems biology using global optimization." eng. In: *Bioinformatics* 27.16 (Aug. 2011), pp. 2311–2313. DOI: 10.1093/bioinformatics/btr370.

[15] J. Bang-Jensen and G. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer, 2001.

[16] A. Bemporad and M. Morari. "Control of systems integrating logic, dynamics, and constraints". In: *Automatica* 35 (1999), pp. 407–427.

[17] *BioNumbers Database*.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[19] L. Brenig. "Complete factorisation and analytic solutions of generalized Lotka-Volterra equations". In: *Physics Letters A* 133 (1988), pp. 378–382.

[20] L. Brenig and A. Goriely. "Universal canonical forms for the time-continuous dynamical systems". In: *Phys. Rev. A* 40 (1989), pp. 4119–4122.

[21] A. P. Burgard, S. Vaidyaraman, and C. D. Maranas. "Minimal Reaction Sets for Escherichia coli Metabolism under Different Growth Requirements and Uptake Environments". In: *Biotechnology Progress* 17.5 (2001), pp. 791–797. ISSN: 1520-6033. DOI: 10.1021/bp0100880.

[22] S. Cardinale and A. P. Arkin. "Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems." In: *Biotechnol J* 7.7 (July 2012), pp. 856–866.

[23] A. Chakrabarty, G. T. Buzzard, and A. E. Rundell. "Model-based design of experiments for cellular processes." eng. In: *Wiley Interdiscip Rev Syst Biol Med* 5.2 (2013), pp. 181–203. DOI: 10.1002/wsbm.1204.

[24] J. Chappell, K. Jensen, and P. S. Freemont. "Validation of an entirely *in vitro* approach for rapid prototyping of DNA regulatory elements for synthetic biology". In: *Nucleic Acids Research* 41 (2013), pp. 3471–3481. DOI: 10.1093/nar/gkt052.

[25] V. Chellaboina, S. P. Bhat, W. M. Haddad, and D. S. Bernstein. "Modeling and Analysis of Mass-Action Kinetics – Nonnegativity, Realizability, Reducibility, and Semistability". In: *IEEE Control Systems Magazine* 29 (2009), pp. 60–78.

[26] O. Chis, J. R. Banga, and E. Balsa-Canto. "Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods". In: *PLoS ONE* 27 (2011), pp. 2610–2611.

[27] O. Chis, J. Banga, and E. Balsa-Canto. "GenSSI: a software toolbox for structural identifiability analysis of biological models". In: *Bioinformatics* (2011). DOI: 10.1093/bioinformatics/btr431.

[28] C. Conradi and D. Flockerzi. "Multistationarity in mass action networks with applications to ERK activation". English. In: *Journal of Mathematical Biology* 65.1 (2012), pp. 107–156. ISSN: 0303-6812. DOI: 10.1007/s00285-011-0453-1.

DOI:10.15774/PPKE.ITK.2015.008

[29] G. Craciun and C. Pantea. "Identifiability of chemical reaction networks". In: *Journal of Mathematical Chemistry* 44 (2008), pp. 244–259.

[30] G. Craciun, Y. Tang, and M. Feinberg. "Understanding bistability in complex enzyme-driven reaction networks". In: *Proc. of the National Academy of Sciences of the USA* 103 (23) (2006), pp. 8697–8702.

[31] G. B. Dantzig and M. N. Thapa. *Linear programming 1: Introduction*. Springer-Verlag, 1997.

[32] G. B. Dantzig and M. N. Thapa. *Linear Programming 2: Theory and Extensions*. Springer-Verlag, 2003.

[33] F. Davidescu and S. Jørgensen. "Structural parameter identifiability analysis for dynamic reaction networks". In: *Chemical Engineering Science* 63.19 (2008), pp. 4754–4762. ISSN: 0009-2509. DOI: 10.1016/j.ces.2008.06.009.

[34] B. De Lacy Costello and A. Adamatzky. "On multitasking in parallel chemical processors: Experimental findings". In: *International Journal of Bifurcation and Chaos* 13 (2003), pp. 521–533.

[35] G. E. Dullerud and F. Paganini. *A Course in Robust Control Theory*. Vol. 36. Texts in Applied Mathematics. Springer New York, 2000. DOI: 10.1007/978-1-4757-3290-0.

[36] D. Endy. "Foundations for engineering biology". In: *Nature* 438 (2005), pp. 449–453.

[37] I. R. Epstein and J. A. Pojman. *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns and Chaos (Topics in Physical Chemistry)*. Oxford University Press, 1998.

[38] P. Érdi and J. Tóth. *Mathematical Models of Chemical Reactions. Theory and Applications of Deterministic and Stochastic Models*. Manchester, Princeton: Manchester University Press, Princeton University Press, 1989.

[39] K. Erguler and M. P. H. Stumpf. "Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models." eng. In: *Mol Biosyst* 7.5 (May 2011), pp. 1593–1602. DOI: 10.1039/c0mb00107d.

[40] M. Feinberg. "Chemical reaction network structure and the stability of complex isothermal reactors - I. The deficiency zero and deficiency one theorems". In: *Chemical Engineering Science* 42 (10) (1987), pp. 2229–2268.

[41] M. Feinberg. *Lectures on chemical reaction networks*. Notes of lectures given at the Mathematics Research Center, University of Wisconsin, 1979.

[42] G. Franceschini and S. Macchietto. "Model-based design of experiments for parameter precision: State of the art". In: *Chemical Engineering Science* 63.19 (2008). Model-Based Experimental Analysis, pp. 4846–4872. ISSN: 0009-2509. DOI: http://dx.doi.org/10.1016/j.ces.2007.11.034.

[43] N. Friedman. "Inferring cellular networks using probabilistic graphical models." eng. In: *Science* 303.5659 (Feb. 2004), pp. 799–805. DOI: 10.1126/science.1094068.

DOI:10.15774/PPKE.ITK.2015.008

[44] A. Gábor and J. Banga. "Improved Parameter Estimation in Kinetic Models: Selection and Tuning of Regularization Methods". English. In: *Computational Methods in Systems Biology*. Ed. by P. Mendes, J. Dada, and K. Smallbone. Vol. 8859. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 45–60. ISBN: 978-3-319-12981-5. DOI: 10.1007/978-3-319-12982-2_4.

[45] D. T. Gillespie. "Stochastic simulation of chemical kinetics." eng. In: *Annu Rev Phys Chem* 58 (2007), pp. 35–55. DOI: 10.1146/annurev.physchem.58.032806.104637.

[46] I. M. Gléria, A. Figueiredo, and T. M. R. Filho. "Stability properties of a general class of nonlinear dynamical systems". In: *Journal of Physics A - Mathematical and General* 34 (2001), pp. 3561–3575.

[47] L. Goentoro, O. Shoval, M. W. Kirschner, and U. Alon. "The incoherent feed-forward loop can provide fold-change detection in gene regulation." eng. In: *Mol Cell* 36.5 (Dec. 2009), pp. 894–899. DOI: 10.1016/j.molcel.2009.11.018.

[48] D. Grate and C. Wilson. "Laser-mediated, site-specific inactivation of RNA transcripts". In: *Proc Natl Acad Sci USA* 96 (1999), pp. 6131–6136.

[49] S. J. Greive, J. P. Goodarzi, S. E. Weitzel, and P. H. von Hippel. "Development of a "modular" scheme to describe the kinetics of transcript elongation by RNA polymerase." eng. In: *Biophys J* 101.5 (Sept. 2011), pp. 1155–1165. DOI: 10.1016/j.bpj.2011.07.042.

[50] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna. "Universally Sloppy Parameter Sensitivities in Systems Biology Models". In: *PLoS Comput Biol* 3.10 (Oct. 2007), e189. DOI: 10.1371/journal.pcbi.0030189.

[51] J. Haag, A. Wouver, and P. Bogaerts. "Dynamic modeling of complex biological systems: a link between metabolic and macroscopic description". In: *Mathematical Biosciences* 193 (2005), pp. 25–49.

[52] H. Haario, E. Saksman, and J. Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7 (2001), pp. 191–379.

[53] W. M. Haddad, V. Chellaboina, and Q. Hui. *Nonnegative and Compartmental Dynamical Systems*. Princeton University Press, 2010.

[54] K. M. Hangos and G. Szederkényi. "Mass action realizations of reaction kinetic system models on various time scales". In: *Journal of Physics: Conference Series (5th International Workshop on Multi-Rate Processes and Hysteresis (MURPHYS 2010))* 268 (2011), p. 012009. DOI: 10.1088/1742-6596/268/1/012009.

[55] R. Hannemann-Tamás, A. Gábor, G. Szederkényi, and K. M. Hangos. "Model complexity reduction of chemical reaction networks using mixed-integer quadratic programming". In: *Computers & Mathematics with Applications* 65.10 (2013), pp. 1575–1595. DOI: 10.1016/j.camwa.2012.11.024.

[56] V. Hárs and J. Tóth. "On the inverse problem of reaction kinetics". In: *Qualitative Theory of Differential Equations*. Ed. by M. Farkas and L. Hatvani. Vol. 30. Coll. Math. Soc. J. Bolyai. North-Holland, Amsterdam, 1981, pp. 363–379.

DOI:10.15774/PPKE.ITK.2015.008

[57] C. E. Hodgman and M. C. Jewett. "Cell-free synthetic biology: thinking outside the cell". In: *Metab. Eng.* 14.3 (2012), pp. 261–269.

[58] F. Horn. "Necessary and sufficient conditions for complex balancing in chemical kinetics". In: *Archive for Rational Mechanics and Analysis* 49 (1972), pp. 172–186.

[59] F. Horn and R. Jackson. "General mass action kinetics". In: *Archive for Rational Mechanics and Analysis* 47 (1972), pp. 81–116.

[60] M. C. Jewett, K. A. Calhoun, A. Voloshin, J. J. Wuu, and J. R. Swartz. "An integrated cell-free metabolic platform for protein production and synthetic biology". In: *Molecular Systems Biology* 4 (2008), p. 220.

[61] M. D. Johnston, D. Siegel, and G. Szederkényi. "Dynamical equivalence and linear conjugacy of chemical reaction networks: new results and methods". In: *MATCH Commun. Math. Comput. Chem.* accepted (2012), to appear.

[62] G. Karlebach and R. Shamir. "Modelling and analysis of gene regulatory networks". In: *Nature Reviews Molecular Cell Biology* 9 (Oct. 2008), pp. 770–780. DOI: doi:10.1038/nrm2503.

[63] E. Karzbrun, J. Shin, R. H. Bar-Ziv, and V. Noireaux. "Coarse-Grained Dynamics of Protein Synthesis in a Cell-Free System". In: *Phys. Rev. Lett.* 106 (2011), p. 048104.

[64] A. S. Khalil and J. J. Collins. "Synthetic biology: applications come of age." In: *Nat Rev Genet* 11.5 (May 2010), pp. 367–379.

[65] D.-M. Kim and J. R. Swartz. "Regeneration of Adenosine Triphosphate from Glycolytic Intermediates for Cell-Free Protein Synthesis". In: *Biotechnology and Bioengineering* 74.4 (2001), pp. 309–316.

[66] D. E. Knuth. *The Art of Computer Programming, Volume 4A, Combinatorial Algorithms, Part 1*. Addison-Wesley Professional, 2008.

[67] Q. Li and W. Xu. "A chemical model for Chua's equation". In: *International Journal of Bifurcation and Chaos* 12 (2002), pp. 877–882.

[68] Q. Li and R. Zhu. "Stochastic simulation of chemical Chua system". In: *International Journal of Bifurcation and Chaos* 14 (2004), pp. 1053–1057.

[69] M. E. W. A. Lim. "Build life to understand it". In: *Nature* 468 (2010), pp. 889–890.

[70] L. Ljung. *System Identification (2Nd Ed.): Theory for the User*. Ed. by L. Ljung. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1999. ISBN: 0-13-656695-2.

[71] F. Llaneras, A. Sala, and J. Pico. "Dynamic estimations of metabolic fluxes with constraint-based models and possibility theory". In: *Journal of Process Control* 22 (2012), pp. 1946–1955. DOI: 10.1016/j.jprocont.2012.09.001.

[72] S. Macnamara, A. M. Bersani, K. Burrage, and R. B. Sidje. "Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation." eng. In: *J Chem Phys* 129.9 (Sept. 2008), p. 095105. DOI: 10.1063/1.2971036.

DOI:10.15774/PPKE.ITK.2015.008

[73] G. Marlovits, M. Wittmann, Z. Noszticzius, and V. Gáspár. "A new chemical oscillator in a novel open reactor - the CLO2-I-2-Acetone system in a membrane fed stirred tank reactors". In: *Journal of Physical Chemistry* 99 (1995), pp. 5359–5364.

[74] I. Maros. *Computational Techniques of the Simplex Method*. Kluwer, 2003.

[75] S. V. Matveev, L. M. Vinokurov, L. A. Shaloiko, C. Davies, E. A. Matveeva, and A. YuB. "Effect of the ATP level on the overall protein biosynthesis rate in a wheat germ cell-free system." eng. In: *Biochim. Biophys Acta* 1293.2 (Apr. 1996), pp. 207–212.

[76] Y. Mileyko, R. I. Joh, and J. S. Weitz. "Small-scale copy number variation and large-scale changes in gene expression". In: *Proceedings of the National Academy of Sciences of the United States of America* 105 (2008), pp. 16659–16664.

[77] M. Mincheva and M. R. Roussel. "Graph-theoretic methods for the analysis of chemical and biochemical networks. I. Multistability and oscillations in ordinary differential equation models". In: *Journal of Mathematical Biology* 55 (2007), pp. 61–86.

[78] J. Nemcová. "Structural identifiability of polynomial and rational systems." eng. In: *Math Biosci* 223.2 (Feb. 2010), pp. 83–96. DOI: 10.1016/j.mbs.2009.11.002.

[79] H. Niederholtmeyer, L. Xu, and S. J. Maerkl. "Real-Time mRNA Measurement during an in Vitro Transcription and Translation Reaction Using Binary Probes". In: *ACS Synth Biol* 2 (2013), pp. 411–417.

[80] Z. Noszticzius and J. Bódiss. "Contribution to the chemistry of the Belousov-Zhabotinskii (BZ) type reactions". In: *Berichte der Bunsen-Gesellschaft - Physical Chemistry Chemical Physics* 84 (1980), pp. 366–369.

[81] I. Otero-Muras, J. R. Banga, and A. A. Alonso. "Characterizing Multistationarity Regimes in Biochemical Reaction Networks". In: *PLoS ONE* 7.7 (July 2012), e39194. DOI: 10.1371/journal.pone.0039194.

[82] T. M. Perumal and R. Gunawan. "Understanding dynamics using sensitivity analysis: caveat and solution." eng. In: *BMC Syst Biol* 5 (2011), p. 41. DOI: 10.1186/1752-0509-5-41.

[83] D. Poland. "Cooperative catalysis and chemical chaos: a chemical model for the Lorenz equations". In: *Physica D: Nonlinear Phenomena* 65 (1993), pp. 86–99.

[84] A. Polynikis, S. J. Hogan, and M. di Bernardo. "Comparing different ODE modelling approaches for gene regulatory networks". In: *Journal of Theoretical Biology* 261 (2009), pp. 511–530. DOI: doi:10.1016/j.jtbi.2009.07.040.

[85] T. P. Prescott and A. Papachristodoulou. "Guaranteed error bounds for structured complexity reduction of biochemical networks". In: *Journal of Theoretical Biology* 304 (2012), pp. 172–182. DOI: 10.1016/j.jtbi.2012.04.002.

[86] R. Raman and I. Grossmann. "Modelling and computational techniques for logic based integer programming". In: *Computers and Chemical Engineering* 18 (1994), pp. 563–578.

DOI:10.15774/PPKE.ITK.2015.008

[87] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood." eng. In: *Bioinformatics* 25.15 (Aug. 2009), pp. 1923–1929. DOI: 10.1093/bioinformatics/btp358.

[88] A. Raue et al. "Lessons Learned from Quantitative Dynamical Modeling in Systems Biology". In: *PLOS ONE* (2013).

[89] M. Rodriguez-Fernandez, J. R. Banga, and F. J. Doyle. "Novel global sensitivity analysis methodology accounting for the crucial role of the distribution of input parameters: application to systems biology models". In: *International Journal of Robust and Nonlinear Control* 22.10 (2012), pp. 1082–1102. ISSN: 1099-1239. DOI: 10.1002/rnc.2797.

[90] M. Rodriguez-Fernandez, S. Kucherenko, C. Pantelides, and N. Shah. "Optimal experimental design based on global sensitivity analysis". In: *17th European Symposium on Computer Aided Process Engineering* (2007).

[91] Y. Rondelez. "Competition for catalytic resources alters biological network dynamics". In: *Phys Rev Lett* 108.1 (2012), p. 018102.

[92] M. Sabouri-Ghomi, A. Ciliberto, S. Kar, B. Novak, and J. J. Tyson. "Antagonism and bistability in protein interaction networks." eng. In: *J Theor Biol* 250.1 (Jan. 2008), pp. 209–218. DOI: 10.1016/j.jtbi.2007.09.001.

[93] A. Saltelli et al. *Global Sensitivity Analysis. The Primer*. John Wiley & Sons, Ltd, 2008.

[94] N. Samardzija, L. D. Greller, and E. Wassermann. "Nonlinear chemical kinetic schemes derived from mechanical and electrical dynamical systems". In: *Journal of Chemical Physics* 90 (4) (1989), pp. 2296–2304.

[95] M. A. Savageau. "Introduction to S-systems and the underlying power-law formalism". In: *Mathl. Comput. Modelling* 11 (1988), pp. 546–551.

[96] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa. "Interdependence of Cell Growth and Gene Expression: Origins and Consequences". In: *Science* 330.6007 (2010), pp. 1099–1102.

[97] G. A. F. Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, Inc., 1989. DOI: 10.1002/0471725315.

[98] R. Serban and A. C. Hindmarsh. "CVODES: the Sensitivity-Enabled ODE Solver in SUNDIALS". In: *Proceedings of IDETC/CIE 2005*. Proceedings of IDETC/-CIE 2005 (2005).

[99] L. Shampinee. "Measuring Stiffness". In: *Applied Numerical Mathematics* 1 (1985), pp. 107–119.

[100] J. Shin, P. Jardine, and V. Noireaux. "Genome replication, synthesis and assembly of the bacteriophage T7 in a single cell-free reaction". In: *ACS Synth. Biol.* 1 (2012), pp. 408–413.

[101] J. Shin and V. Noireaux. "An *E. coli* Cell-Free Expression Toolbox: Application to Synthetic Gene Circuits and Artificial Cells". In: *ACS Synth. Biol.* 1 (2012), pp. 29–41.

DOI:10.15774/PPKE.ITK.2015.008

[102] J. Shin and V. Noireaux. "Efficient cell-free expression with the endogenous *E. Coli* RNA polymerase and sigma factor 70". In: *Journal of Biological Engineering* 4 (2010), p. 8.

[103] G. Shinar and M. Feinberg. "Structural Sources of Robustness in Biochemical Reaction Networks". In: *Science* 327 (2010), pp. 1389–1391.

[104] D. Siegal-Gaskins, V. Noireaux, and R. M. Murray. "Biomolecular resource utilization in elementary cell-free gene circuits". In: *American Control Conference (ACC), 2013*. 2013, pp. 1531–1536.

[105] E. Sontag. "Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction". In: *IEEE Transactions on Automatic Control* 46 (2001), pp. 1028–1047.

[106] E. D. Sontag. *Mathematical Control Theory: Deterministic and Finite Dimensional Systems*. Springer, 1998.

[107] E. Sontag, A. Kiyatin, and B. N. Kholodenko. "Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data". In: *Bioinformatics* 20 (2004), pp. 1877–1889.

[108] M. Stein. *Large Sample Properties of Simulations Using Latin Hypercube Sampling*. Vol. 29. 2. Taylor and Francis, 1987, pp. 143–151. DOI: 10.1080/00401706.1987.10488205. eprint: http://www.tandfonline.com/doi/pdf/10.1080/00401706.1987.10488205.

[109] T. Stögbauer, L. Windhager, R. Zimmer, and J. O. Rädler. "Experiment and mathematical modeling of gene expression dynamics in a cell-free system". In: *Integr. Biol.* 4 (2012), pp. 494–501. DOI: 10.1039/C2IB00102K.

[110] Z. Z. Sun, C. A. Hayes, J. Shin, F. Caschera, R. M. Murray, and V. Noireaux. "Protocols for implementing an *Escherichia coli* based TX-TL cell-free expression system for synthetic biology". In: *Journal of Visualized Experiments* 79 (2013), e50762.

[111] Z. Z. Sun, E. Yeung, C. A. Hayes, V. Noireaux, and R. M. Murray. "Linear DNA for Rapid Prototyping of Synthetic Biological Circuits in an Escherichia coli Based TX-TL Cell-Free System." In: *ACS Synth Biol* 3 (2014), pp. 387–397.

[112] E. Szádeczky-Kardoss and B. Kiss. "On-line trajectory time-scaling to reduce tracking error". In: *Intelligent Engineering Systems and Computational Cybernetics*. Springer, 2009, pp. 3–14.

[113] G. Szederkényi. "Computing sparse and dense realizations of reaction kinetic systems". In: *Journal of Mathematical Chemistry* 47 (2010), pp. 551–568. DOI: 10.1007/s10910-009-9525-5.

[114] G. Szederkényi, J. R. Banga, and A. A. Alonso. "Inference of complex biological networks: distinguishability issues and optimization-based solutions". In: *BMC Systems Biology* 5 (2011), p. 177. DOI: 10.1186/1752-0509-5-177.

[115] G. Szederkényi and K. M. Hangos. "Finding complex balanced and detailed balanced realizations of chemical reaction networks". In: *Journal of Mathematical Chemistry* 49 (2011), pp. 1163–1179. DOI: 10.1007/s10910-011-9804-9.

DOI:10.15774/PPKE.ITK.2015.008

[116] G. Szederkényi, K. M. Hangos, and Z. Tuza. "Finding weakly reversible realizations of chemical reaction networks using optimization". In: *MATCH Commun. Math. Comput. Chem.* 67 (2012), pp. 193–212.

[117] G. Szederkényi, K. Hangos, and A. Magyar. "On the time-reparametrization of quasi-polynomial systems". In: *Physics Letters A* 334 (2005), pp. 288–294. DOI: 10.1016/j.physleta.2004.11.026.

[118] Y. Takeuchi. *Global Dynamical Properties of Lotka-Volterra Systems*. Singapore: World Scientific, 1996.

[119] Z. A. Tuza and V. Singhal. *TXTL Matlab toolbox http://sourceforge.net/projects/txtl/*.

[120] J. J. Tyson, K. C. Chen, and B. Novak. "Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell". In: *Current Opinion in Cell Biology* 15 (2003), pp. 221–231.

[121] J. Vanlier, C. A. Tiemann, P. A. J. Hilbers, and N. A. W. van Riel. "An integrated strategy for prediction uncertainty analysis." eng. In: *Bioinformatics* 28.8 (Apr. 2012), pp. 1130–1135. DOI: 10.1093/bioinformatics/bts088.

[122] D. del Vecchio and R. M. Murray. *Biomolecular Feedback Systems*. 2012.

[123] E. O. Voit. "Biochemical Systems Theory: A Review". In: *ISRN Biomathematics* 2013 (2012), p. 53. DOI: 10.1155/2013/897658.

[124] É. Walter and L. Pronzato. "Qualitative and Quantitative Experiment Design for Phenomenological Models A Survey". In: *Aut* 26 (1990), pp. 195–213.

[125] E. Walter and Y. Lecourtier. "Unidentifiable Compartmental Models: What to Do?" In: *Mat* 56 (1981), pp. 1–25.

[126] É. Walter, L. Pronzato, and J. Norton. *Identification of parametric models from experimental data*. Communications and control engineering. Translated from an updated French version : Identification de modèles paramétriques à partir de données expérimentales. Berlin: Springer Paris, 1997. ISBN: 978-1-84996-996-3.

[127] H. P. Williams. *Model Building in Mathematical Programming*. Wiley, 2013, p. 432.

[128] W. Xu and Q. Li. "Chemical chaotic schemes derived from NSG system". In: *Chaos, Solitons and Fractals* 15 (2002), pp. 663–671.

[129] W. Xu and Q. Li. "Another chemical chaotic attractor for Chua's equation". In: *International Journal of Bifurcation and Chaos* 13 (2003), pp. 2715–2718.

[130] E. Yeung, J. Kim, Y. Yuan, J. Gonçalves, and R. Murray. "Quantifying Crosstalk in Biochemical Systems". In: *Proceedings of the IEEE Conference on Decision and Control*. 2012.

[131] E. Yeung, J. Kim, Y. Yuan, J. Goncalves, and R. M. Murray. "Quantifying crosstalk in biochemical systems". In: *51st IEEE Conference on Decision and Control (CDC)*. Maui, HI, Dec. 2012, pp. 5528–5535.

[132] Y. Yuan, G.-B. Stan, S. Warnick, and J. Goncalves. "Robust dynamical network structure reconstruction". In: *Automatica* 47.6 (2011). Special Issue on Systems Biology, pp. 1230–1235. ISSN: 0005-1098. DOI: http://dx.doi.org/10.1016/j.automatica.2011.03.008.

DOI:10.15774/PPKE.ITK.2015.008

[133] M. M. Zavlanos, A. A. Julius, S. P. Boyd, and G. J. Pappas. "Inferring stable genetic networks from steady-state data". In: *Automatica* 47.6 (2011), pp. 1113–1122.

[134] Z. Zi. "Sensitivity analysis approaches applied to systems biology models." eng. In: *IET Syst Biol* 5.6 (Nov. 2011), pp. 336–336. DOI: 10.1049/iet-syb.2011.0015.