Analysis of Biochemical Reaction Networks using Tropical and Polyhedral Geometry Methods

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Samal, Satya Swarup

aus

Bhubaneswar, Indien

Bonn, 2016

**Angefertigt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn**

1. Gutachter: Prof. Dr. Andreas Weber

2. Gutachter: Prof. Dr. Ovidiu Radulescu

Tag der Promotion: 25. November 2016

Erscheinungsjahr: 2016

Abstract

The field of systems biology makes an attempt to realise various biological functions and processes as the emergent properties of the underlying biochemical network model. The area of computational systems biology deals with the computational methods to compute such properties. In this context, the thesis primarily discusses novel computational methods to compute the emergent properties as well as to recognize the essence in complex network models. The computational methods described in the thesis are based on the computer algebra techniques, namely tropical geometry and extreme currents. Tropical geometry is based on ideas of dominance of monomials appearing in a system of differential equations, which are often used to describe the dynamics of the network model. In such differential equation based models, tropical geometry deals with identification of the metastable regimes, defined as low dimensional regions of the phase space close to which the dynamics is much slower compared to the rest of the phase space. The application of such properties in model reduction and symbolic dynamics are demonstrated in the network models obtained from a public database namely Biomodels. Extreme currents are limiting edges of the convex polyhedrons describing the admissible fluxes in biochemical networks, which are helpful to decompose a biochemical network into a set of irreducible pathways. The pathways are shown to be associated with given clinical outcomes thereby providing some mechanistic insights associated with the clinical phenotypes. Similar to the tropical geometry, the method based on extreme currents is evaluated on the network models derived from a public database namely KEGG. Therefore, this thesis makes an attempt to explain the emergent properties of the network model by determining extreme currents or metastable regimes. Additionally, their applicability in the real world network models are discussed.

Acknowledgements

In the beginning of January 2013, I was introduced to the wonderful and exciting field of computer algebra and specifically Tropical Geometry (TG) and Extreme Currents (ECs) by Prof. Dr. Andreas Weber (AW) and Prof. Dr. Dima Grigoriev (DG). My sincere gratitude to AW and DG for giving me this opportunity. In that following summer, Prof. Dr. Ovidiu Radulescu (OR) visited Bonn and introduced me to the applications of TG to real world biological problems especially in model reduction and other related ideas. I was also fortunate to visit him in Montpellier, France and get his guidance and advice on various subjects. My sincere gratitude to OR for the supervision and guidance throughout. While working on the ideas of TG and ECs, I was fortunate to receive the guidance of Dr. Holger Fröhlich to associate the the ECs with clinical and experimental data. Under his supervision, I got the opportunity to work in the area -"omics" data analysis as well. I am thankful to the IDENTIREST project (in collaboration with Prof. Dr. Björn Scheffler) which facilitated this endeavour. My gratitude to HF for his guidance and advice. Therefore, my thesis under the guidance of a computer scientist (AW), mathematician (DG), system biologist (OR) and binformatician (HF) taught me to work in a truly interdisciplinary environment. I would also extend my gratitude to Prof. Dr. Sergey Vakulenko from Russia for his ideas and stimulating scientific discussions. I am grateful to Prof. Dr. Frank Bertoldi and Prof. Dr. Thomas Schultz for accepting to be the part of the examination committee.

I had a fun time working with my colleagues Gloria Bahamóndez, Khalid Abnaof, Ashar Ahmed, Paurush Praveen, Benjamin Engelhardt, Zahra Narimani, Joao Dinis, Hassan Errami and Nikhil Vinod Mallela in B-IT, Bonn. My gratitude to my friends (both in Bonn and India) for their support and advice. Their presence made my time enjoyable in Bonn (both in and out of the office). My gratitude to the office staff of B-IT for their support and co-operation throughout.

My parents, in-laws and the inspiration from my late grandmother always gave the emotional and mental support during difficult times. Lastly, I would like to thank my wife Ankita, for supporting me in the difficult times with lot of patience and helping me in every possible way to complete the thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

SBML	**S**ystem **B**iology **M**arkup **L**anguage
EC	**E**xtreme **C**urrents
TE	**T**ropical **E**equilibrations
ODE	**O**rdinary **D**ifferential **E**quations
MR	**M**etastable **R**egimes
TGFβ	**T**ransforming **G**rowth **F**actor **B**eta

Dedicated to my Gurus (Teachers)

Chapter 1

Introduction

The main purpose of this thesis is to develop computational methods to enable the analysis of biological functions and processes from a systems perspective. Here, systems perspective generally refers to the analysis of biochemical reaction networks where the parts (e.g. chemical species, genes, proteins, etc) along with their interactions (e.g. reactions) are taken into account to describe the biological functions or processes. Hence, such an approach makes an attempt to take a holistic view.

In the context of healthcare, the idea of system level thinking can be dated back to traditional healthcare practices. For example, the medications based on the theory of *Ayurveda*, developed in India around 1000 B.C.E. define the healthy state of an individual as the balance (*saamya* in Sanskrit) of three bodily humours namely *Vata, Pita, Kapha* and disease state as the loss of this balance. The diagnosis involves the determination of that bodily humour(s) which causes loss of this balance, and medications are aimed to restore it back (Sharma and Dash, 1981).

In modern science, Claude Bernard may be considered as the "first system biologist" because of his concept of the constancy of the internal environment (le milieu intérieur) proposed in 1865 (Noble, 2008). In addition, Claude Bernard in his work explained "a complex organism should be looked upon as an assemblage of simple organisms", pointing towards a system level description of complex organism. Unfortunately, his work remained largely unrecognized to the scientific community of that time (Gross, 1998).

In general the idea of systems theory was first developed by Alexander Bogdanov in Russia which is called as *Tektology*. The work was first published in 1912. In his theory of organization he classified three systems namely organized complex, disorganized complex and neutral complex. In organized complex, the whole is greater than sum of its parts, in disorganized complex, whole is smaller than sum of its parts and in neutral complex both organizing and disorganizing complex cancel each other out. The organization as per Bogdanov was the resultant of formation and regulation (Capra, 1997). It can be seen that his idea of organized complex appears to be relevant for the understanding of systems biology, where the whole is generally more than the sum of its parts because it involves additionally the interactions between the parts (Kitano, 2002a).

Thereafter, the first theoretical framework for systems theory was put forward by Ludwig von Bertalanffy. He considered the systems to be open and operate far from equilibrium, which is common to the living organisms. A general introduction and a brief history of systems thinking can be found in his book (Bertalanffy, 1968).

Ilya Prigogine gave the idea of self organization through dissipative structures. After that the field of cybernetics grew and the two prominent figures in this area were Norbert Wiener proposing the idea of self-regulation by feedback loops and John von Neumann who proposed the use of mathematical logic to understand brain functions. A good introductory review on the history of system theory can be found in (Capra, 1997).

On one hand, there were developments in the general systems theory and on the other hand, molecular biology was taking a reductionist approach, which was much popularized by the discovery of DNA as a genetic material in 1944. This ultimately led to the Human Genome

project and hence it gave rise to *"high throughput data rich biology"* (Westerhoff and Palsson, 2004).

After the Human Genome project, the time was ripe to integrate the "*data poor insilico biology*" (Westerhoff and Palsson, 2004) resulting from the general systems approach with the data generated from next generation sequencing technologies which got a boost from the human genome project and the subsequent developments. The reductionist approach inferred the parts of the system and the general systems theory developed the relevant mathematics and concepts to be used to model systems. The next steps included the elucidation of various interactions of the parts and creation of the network models (Barabasi and Oltvai, 2004) from the high throughput data. Therefore, currently, systems biology is in a position to integrate mathematical modelling approaches with high throughput data in order to understand the living systems at a molecular level which was not possible at the times of Claude Bernard. Also, Claude Bernard and Louis Pasteur contributed to the field of enzymology, though at those time this field was in its infancy. The next crucial thing is to create a sync between the theoretical research and the experimental research in order to translate the findings from the computational systems biology models to the various experimental or clinical settings and back.

The core idea of computational systems biology is to view biological systems as *complex non-linear dynamical networks* of molecular parts (Kitano, 2002a; Kitano, 2002b). These networks have emergent collective behaviour that is explainable only if the constituting parts of the system and their mutual interactions are considered (Bhalla and Iyengar, 1999). These interactions are biochemical in nature and the network models are made of pathways of biochemical reactions. In general, such emergent properties can be understood as collective behaviour of dynamical systems. The core idea of collective behaviour means that the behaviour of individual molecular parts are influenced (or dominated) by their neighbours resulting in the emergence of a common pattern. Such collective behaviour is analysed in statistical physics models (Vicsek, 2001a; Vicsek, 2001b). In dynamical systems, such ideas are also investigated. For example, René Thomas (Thomas and d'Ari, 1990) associated collective behaviours (e.g. multistationarity and oscialltions) arising from biochemical networks with biological function. Often, small changes in certain critical parameters cause a change in the collective behaviour which results in a qualitative change (or regime change) of the system. In biological, physical and chemical systems, such aspects are studied in the context of phase transitions, birfurcation theory and center manifold theory e.g. (Brehme et al., 2016; Jirsa et al., 2014; Chen et al., 2016) in which the critical regions are determined where such change in the regime occurs. In this thesis, the computation of mathematical entities from dynamical systems representing such emergent collective behavior will be focused.

As described above, the biological systems considered here are mainly mechanistic models of biochemical reaction networks, essentially modelling the biological processes. Biological systems can be modelled using different mathematical formalisms, an overview for the same can be found in (Machado et al., 2011). Such models facilitate quantitative predictions of emergent behaviour of networks as well as in the control and manipulation of the associated biological processes (Craver, 2006). In addition, mechanistic models have cause-effect relationships which help to address the effects of different interventions (Woodward, 2005) e.g. drugs. Having quantitative predictions of emergent behaviour of network models is important for biology and medicine.

In spite of recent developments in the mathematical theory of biochemical networks (Feinberg, 1987) predicting the dynamical properties of dynamical networks remains a formidable task, especially for large models. Furthermore, predictive models of the biological function need to be multiscale, involving several intracellular biochemical pathways as well as interactions between a large number of cells in a tissue. In order to make them tractable, a new mathematical tool needs to be developed allowing to recognize where the essence resides and tame an overwhelmingly complex picture of interactions (Radulescu et al., 2008).

For large networks with ODE based dynamics, the concept of *attractors* (Milnor, 1985) is of relevance as it represents a stable balanced state of the system towards which nearby trajectories are drawn. These can be fixed points, limit cycles (oscillations) or even chaotic (MacArthur, Ma'ayan, and Lemischka, 2009). A general introduction to ODE based dynamical systems can be found in (Hirsch, Smale, and Devaney, 2012; Strogatz, 2014). For ODE based systems with multiple timescales it is reasonable to consider that a typical trajectory consists of a succession of qualitatively different slow segments separated by faster transitions. The slow segments, are called Metastable Regimes (MRs) (Samal et al., 2016). Metastability is also relevant for the behaviour of stochastic systems and is studied in the mathematics of Markov processes (Bovier and Hollander, 2015). The notion of metastability generalizes the notion of attractor (Milnor, 1985). A system remains in the proximity of an attractor after entering its basin of attraction, but can leave a MR after a more or less long time. Such a phenomenon is illustrated in Fig. 1.1. An example of such behavior is the set of bifurcations of MRs guiding the orderly progression of the cell cycle (Noel et al., 2012). In cell cycle models, biological distinct stages such as interphase and mitosis correspond to relatively slow segments of the dynamics and are separated by fast transitions. Other examples are provided by the evolution of stem cells to differentiated cells (MacArthur, Ma'ayan, and Lemischka, 2009), or by successive transformations occurring during cancerogenesis (Huang, Ernberg, and Kauffman, 2009) where each cell type can be associate to some dynamical attractor and the development to transitions from one attractor to another. Historically, such an idea can be dated back to Waddington's epigenetic landscape (Waddington, 1940; Waddington, 1957) to explain development of cell types. Thus, the MRs are representative of the emergent collective behaviour of the biological system. The switch among MRs is often associated with qualitative change in various biological function. Furthermore, one can study the dynamics close to MRs by computing a simpler description of the original system (usually involving less variables than the original model), popularly referred to as model reduction (Radulescu et al., 2008). The simpler description is possible due to the timescale separation of variables into fast and slow variables and thereafter representing only the slow dynamics by making use of Differential Algebraic Equations (DAEs). Therefore, one of the main objective of this thesis is to provide a computational approach to determine such MRs in an efficient and scalable manner.

In many practical applications, the kinetic rate constants appearing in the mechanistic models of biochemical reaction networks are unknown or highly uncertain thereby making the computation of MRs difficult or only the underlying biochemical reaction network structure is represented as a stoichiometric matrix. Under the conditions of steady state of the reaction network and the non-negativity of reaction fluxes, such a network can be algebraically decomposed into a set of irreducible steady state pathways often referred to as extreme currents (ECs) or Elementary Flux Modes (EFMs). Such pathways (essentially, a set of reactions) individually or in combination are often found to be associated with biological functions (Papin et al., 2003) and also associated with emergent collective behaviour of the network such as for instance in the oscialltion due to the occurence of Hopf bifurcation (Gatermann, Eiswirth, and Sensse, 2005a). Furthermore, these pathways have applications in drug target identification and network robustness analysis (Papin et al., 2003). To this aim, an established method of computing ECs is described and a statistical framework is used to associate them with different clinical phenotypes.

The work presented in the thesis makes an attempt to describe the emergent properties of biochemical reaction networks in terms of two mathematical entities, namely ECs and TEs. Mathematically, these entities are related to the steady states of the system. The core difference between them is that, while ECs solve the system in reaction or flux space, TEs solve the system in the species concentration space. Due this difference, the algorithmic techniques to compute them are not the same. Furthermore, the biological motivation and objectives considered in the thesis to compute ECs and TEs are different as shown in the corresponding application

scenarios. It will be a topic of future research to explore the mathematical connection between them. Nevertheless, from a biological perspective they are both associated with the emergent behaviour of the system.

To reiterate, the main objective of the thesis is to provide a computational approach to compute the mathematical objects namely MRs or ECs for a given biochemical reaction network, which are associated with the emergent collective behaviour of the system. The benefit of this approach is threefold, namely

1. Model reduction i.e. determining the key variables and parameters of the system corresponding to MRs.

2. Coarse graining of smooth ODE models for qualitative understanding of the network dynamics.

3. Linking the clinical or biological phenotypes with MRs and ECs, thereby providing mechanistic insights into disease states.

The work described in the thesis is partly published in peer reviewed journals and conferences. Here, we make the references to such publications. The ideas of model reduction based on tropical geometry are explored in (Samal et al., 2015a). The work related to metastability and the analysis of TGFβ signalling model are explored in (Radulescu et al., 2015; Samal et al., 2016). The algorithmic ideas to compute the tropical equilibrations are described in (Samal et al., 2015c). The manuscript describing the work of associating ECs with clinical phenotypes is in the process of submission to a peer reviewed journal.

The thesis is organised in three main parts. Firstly, the mathematical formalisms to represent biological systems are described in Chapters 2 and the background on tropical geometry and ECs is provided in Chapters 3. Tropical geometry is the main mathematical tool to compute the MRs. Secondly, the algorithms to compute tropical solutions and ECs are discussed in Chapters 4, 5. In addition, to test the algorithms, a benchmark is performed with models obtained from various public databases. Lastly, the application of tropical geometry concepts is shown in two contexts. First, in model reduction where the dynamics of a smooth ODE based model is described on a lower dimensional manifold which is determined by timescale separation of variables resulting in slow-fast variables. This is described in Chapter 6 and exemplified with different biochemcial reaction networks. Second, an approach is described to coarse grain the dynamics of a smooth biochemical reaction network to discrete symbolic dynamics, which is described in Chapter 7. The symbolic dynamics is described through a finite state automaton whose states correspond to MRs. Furthermore, the association of biological phenotypes namely epithelium-like (non-aggressive) and mesenchymal-like (aggressive) cell lines with MRs is shown exempliying its potential application in disease research. Likewise, in the context of pathway analysis, the association of ECs with categorical clinical phenotypes namely beningn and tumour prostate cancer samples and continuous clinical phenotypes namely, survival times of the Glioblastoma Multiforme (GBM) samples are shown. This is described in Chapter 8.

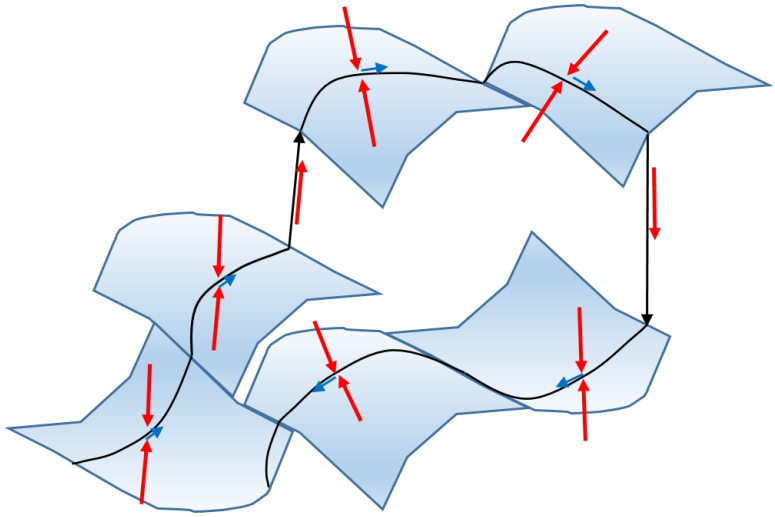


FIGURE 1.1: A representation of MRs as itinerant trajectory in a patchy phase space landscape. Dominant vector fields (red arrows) confine the trajectory to low dimensional patches on which act weak uncompensated vector fields (blue arrows). A typical trajectory contains succession of slow segments separated by fast transition. Figure reproduced from (Samal et al., 2015a).

Chapter 2

Formal Representation of Biochemical Reaction Networks

The main purpose of this chapter is to introduce the notations used to describe the biochemical reaction network in a mathematical manner. In this thesis, the rate of change of chemical species with respect to time will be described as a system of differential equations. The reaction fluxes will occur as terms or monomials in the equations. These reaction fluxes can be polynomials or rational functions of chemical species where reaction rate constants occur as parameters. The reaction rate constants will sometimes be called kinetic constants. For many networks, the reaction flux function is unknown or the corresponding kinetic constants are uncertain. In such scenarios, the dynamics is represented in terms of reaction fluxes which will be the variables of the equation system. Below, these two types of representations are described.

2.1 ODE based representation

The biochemical networks can be described by the following differential equations based on mass action laws

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j\in[1,r]} k_j S_{ij} \boldsymbol{x}^{\boldsymbol{\alpha_j}},\ i \in [1,n]. \tag{2.1}$$

where $k_j > 0, j \in [1,r]$ are the kinetic constants, r is the number of reactions, S_{ij} are the elements of the so-called stoichiometric matrix, $\boldsymbol{\alpha}_j = (\alpha_1^j, \ldots, \alpha_n^j) \in \mathbb{Z}_+^n$ are the multi-indices, $\boldsymbol{x}^{\boldsymbol{\alpha_j}} = x_1^{\alpha_1^j} \ldots x_n^{\alpha_n^j}$ and x_i, $i \in [1,n]$ are the species concentrations, n being the number of species. The purpose is to study the steady state behavior of such a system which means equating the above equation to zero, resulting in a system of non-linear polynomial equations with unknowns as x_i.

The polynomial equations (2.1) can result from the mass action law. For instance, a reaction $A + B \to C$ of kinetic constant k and satisfying the mass action law, has $S_{11} = -1, S_{21} = -1, S_{31} = 1$, $\boldsymbol{\alpha}_1 = (1,1,0)$, which correspond to the kinetic equations

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t} &= -kx_1x_2, \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} &= -kx_1x_2, \\ \frac{\mathrm{d}x_3}{\mathrm{d}t} &= kx_1x_2, \end{aligned} \tag{2.2}$$

where x_1, x_2, x_3 are the concentrations of A, B, C, respectively.

It is to be noticed that the mass action law implies tight relations between α_j and S_{ij}, namely $\alpha_i^j = -S_{ij}$ if $S_{ij} < 0$, otherwise $\alpha_i^j = 0$. These relations are not needed in our approach. Furthermore, our method can be extended to a more general case when the reaction rates are rational functions of the concentrations. Typically, we can use the least common denominator

of reaction rates to express the right hand sides of the kinetic equations as ratios of polynomials and apply the method to the numerators. This extension is briefly discussed in (Noel et al., 2012). At this point, it is to be mentioned that there exist approaches to describe such a polynomial system in a graph theoretic way, i.e., by a weighted directed graph and a weighted bipartite graph and to study the number of positive solutions depending on the graph structure (Gatermann and Huber, 2002). This approach uses decompositions of Newton polytopes to find that parts of the directed graph are related to the existence of positive steady state solutions. In the current thesis such graph theoretic considerations are not used.

A system of differential equations can have linear conservation laws that are linear combinations of the form $c(\boldsymbol{x}) = c_1x_1 + c_2x_2 + \ldots + c_nx_n$ that is identically constant on trajectories of the system, $c_1x_1 + c_2x_2 + \ldots + c_nx_n = K$. They provide constraints that have to be imposed if one wants to compute steady states, for instance. In Chapter 4 the conservation laws are either preserved or the system is transformed to one without conservation laws, by eliminating some variables (for an overview to eliminate variables using conservation laws, we refer to (Gunawardena, 2003)). Specifically, linear conservation law of a system of differential equations is a linear form $C(\boldsymbol{x}) =< \boldsymbol{c}, \boldsymbol{x} >= c_1x_1 + c_2x_2 + \ldots + c_nx_n$ that is identically constant on trajectories of the system. It can be easily checked that vectors in the left kernel $Ker^l(S)$ of the stoichiometric matrix $\boldsymbol{S}$ provide linear conservation laws of the system (2.1). Indeed, system (2.1) reads $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{SR}(\boldsymbol{x})$, where the components of the vector $\boldsymbol{R}$ are $R_j(\boldsymbol{x}) = k_j x^{\alpha_j}$. If $\boldsymbol{cS} = 0$, then $\frac{\mathrm{d}<\boldsymbol{c},\boldsymbol{x}>}{\mathrm{d}t} = \boldsymbol{cSR}(\boldsymbol{x}) = 0$, where $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$.

In addition, there exists several other ways to model the rate laws (including rational functions of the species). A comprehensive list can be found in (Dräger et al., 2015).

2.2 Reaction flux based representation

The biochemical networks can also be represented by

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j\in[1,r]} S_{ij}V_j,\ i \in [1,n]. \tag{2.3}$$

where $V_j \geq 0, j \in [1,r]$ are reaction fluxes, r is the number of reactions, S_{ij} are the elements of the stoichiometric matrix. It is to be noted here that such a description of the biochemical reaction system is useful when the kinetic constants in (2.1) are unknown. The goal is to analyse the system where reaction fluxes are variables as opposed to the chemical species being as variables in (2.1). To study the steady state behavior of such a system with non-negativity constraints on V_j, the framework of extreme currents is often used (cf. Section 3.2). In this particular formalism, the reversible reactions are split into separate forward and backward reactions respectively.

Chapter 3

Background

This chapter is divided into two sections. In the first section, a brief introduction to tropical geometry is provided along with important mathematical results mostly adopted from (Maclagan and Sturmfels, 2015; Sturmfels, 2002; Richter-Gebert, Sturmfels, and Theobald, 2005). The framework of tropical geometry is applied in the context of model reduction and determining the symbolic dynamics of smooth ODE based models derived from the biochemical reaction networks. In the second section, an introduction to the extreme currents is provided which uses methods from polyhedral geometry for the enumeration of the vertices of the polyhedron. This has many interesting applications in biological pathway analysis and stoichiometric network analysis.

3.1 A Brief Introduction to Tropical Geometry

Tropical geometry as a mathematical technique has a growing number of applications in nonlinear equation solving (Rojas, 2002), algebraic statistics (Pachter and Sturmfels, 2004), biochemical reaction networks (Noel et al., 2012; Noel et al., 2014) and in general optimization problems e.g. job scheduling, transportation networks, decision making, traffic optimization (Aubin, 2010; Krivulin, 2014). Furthermore, tropical geometry is closely related to max-plus algebra (Heidergott, Olsder, and Woude, 2006) which is shown in the context of mean payoff games (Grigoriev and Podolskii, 2013). In this section, the important concepts of tropical geometry are reviewed.

3.1.1 Tropical Arithmetic

In tropical arithmetic, tropical addition (denoted by $\oplus$) and tropical multiplication (denoted by $\odot$) of two numbers is their minimum and sum in classical arithmetic as shown below

$$\begin{aligned} x \oplus y &= \min(x, y), \\ x \odot y &= x + y. \end{aligned} \tag{3.1}$$

For example, $1 \oplus 2 = 1$, $1 \odot 2 = 3$.

The basic structure in tropical arithmetic is the *tropical semiring* which is a set defined by $(\mathbb{R} \cup \{\infty\}, \oplus, \odot)$ where ∞ denotes infinity and $\mathbb{R}$ is the set of real numbers. It is called semiring due to the absence of tropical subtraction i.e. absence of additive inverse. For example, consider the tropical subtraction of two numbers 2 and 1 and the result be x, this is equivalent to $1 \oplus x = 2$, which has no solution for x. The symbol ∞ plays the role of additive identity and 0 as multiplicative identity in such a semiring. For example, $x \oplus \infty = x$ and $x \odot 0 = x$. It should be pointed out that, $\min(x, y)$ can be replaced with its dual form $\max(x, y)$ resulting in a semiring defined by $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$. However, in this thesis the choice of $\min(x, y)$ is adopted.

Tropical geometry can also be understood as the limit of classical geometry in the following manner:

Let x and y be the powers of an auxiliary variable ε represented as ε^x and ε^y, where ε is a positive real number. Tropical addition can be described as $x \oplus y = \log_\varepsilon(\varepsilon^x + \varepsilon^y)$ which evaluates to $\min(\mathrm{x}, \mathrm{y})$ if $\varepsilon \to 0$ and if $\varepsilon \to \infty$ it evaluates to $\max(\mathrm{x}, \mathrm{y})$. Similarly, tropical multiplication can be described as $x \odot y = \log_\varepsilon(\varepsilon^x \varepsilon^y)$ which evaluates to $x + y$.

3.1.2 Tropical Polynomial

Let $x_1, \ldots, x_n$ represent variables in the tropical semiring. A tropical monomial is the tropical multiplication of these variables where repetitions are allowed. Hence, tropical monomial is a linear function from $\mathbb{R}^n$ to $\mathbb{R}$ with integer coefficients. For example, the tropical monomial $x_1^2 \odot x_2^3$ represents the linear function $2x_1 + 3x_2$ in classical arithmetic. In the similar spirit, a tropical polynomial is a finite linear combination of such tropical monomials. To this aim, tropical polynomial can be defined in the following manner:

Definition. A *Tropical polynomial* is a piecewise linear concave function which is given as the minimum of a finite set of linear functions with integer coefficients.

Let us exemplify, the idea of tropical polynomial in with a polynomial in two variables $x,\ y$ whose coefficients are rational functions of a small parameter ε as represented below

$$f(x, y, \varepsilon) = \sum_{(i,j)\in A} a_{ij}(\varepsilon) x^i y^j \tag{3.2}$$

The coefficients $a_{ij}(\varepsilon) = \overline{a_{ij}}\varepsilon^{\gamma_{ij}}$ where $\overline{a_{ij}}$ lie over the complex field $\mathbb{C}$. The variables $x,\ y$ also lie over $\mathbb{C}$. The field $\mathbb{C}$ is algebraically closed. The set A denotes a finite subset in $\mathbb{Z}^2$ (where $\mathbb{Z}$ denotes the set of integers).

The tropicalization of $f(x, y, \varepsilon)$, denoted by $T(f(x, y, \varepsilon))$ is an arithmetic operation that replaces the classical addition and multiplication with the tropical addition and multiplication, resulting in a tropical polynomial as shown below

$$\min_{(i,j)\in A} (\gamma_{ij} + ix + jy) \tag{3.3}$$

Remark. The tropicalization operation translates a non-linear polynomial into a piecewise linear function denoted by a tropical polynomial. This has a direct implication in the algorithmic development as it is much easier to handle linear functions based on the methods from polyhedral geometry.

3.1.3 Tropical Variety

The tropical zeros are determined by computing the points at which the minimum of the tropical polynomial is attained *at least* twice. For example, in the bivariate polynomial defined in (3.2), consider any two points (i', j') and (i'', j'') in A, the computation of tropical zeros translates to solving the following systems of linear inequalities

$$\gamma_{i'j'} + i'x + j'y = \gamma_{i''j''} + i''x + j''y \leq \gamma_{ij} + ix + jy \quad \text{for} \quad (i, j) \in A \tag{3.4}$$

where (i', j') and (i'', j'') range over the distinct points in A, leading to a disjunction of linear inequality systems. Such linear inequality systems are called as solution polytopes as because a linear inequality system is actually a convex polytope in H-representation. Here H-representation of a polytope means a bounded solution set of a finite systems of linear inequalities which can be represented by $P(A, b) = \{x \in \mathbb{R}^d \mid a_i^T x \leq b_i \text{ for } 1 \leq i \leq m\}$, where

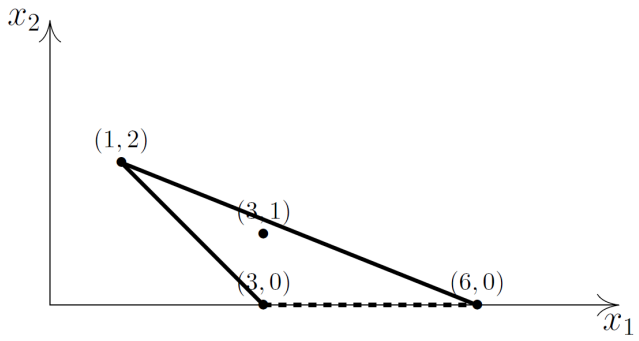


FIGURE 3.1: Newton polytope for the polynomial $-x_1^6 + x_1^3x_2 - x_1^3 + x_1x_2^2$. Note that the point $(3, 1)$ is not in the convex hull. The dotted edge corresponds to the monomials of same sign.

$A \in \mathbb{R}^{m\times d}$ is a real matrix with rows a_i^T and $b \in \mathbb{R}^m$ is a real vector with entries b_i. Boundedness means that there is a constant N such that $\|x\| \leq N$ holds for all $x \in P$ (defnition adopted from (Henk, Richter-Gebert, and Ziegler, 2004)).

In the case of bivariate polynomial, the *union* of such solution polytopes is called as a *tropical curve* and in multivariate case it generalizes to *tropical hypersurface*, as defined below

Definition. The set of tropical zeros (i.e. union of solution polytopes) of a tropical polynomial is called a tropical hypersurface.

The algorithmic way to compute the tropical zeros is through the technique of *Newton polytopes*. The Newton polytope associated with (3.2) is the convex hull of point (i, j, γ_{ij}) for each point (i, j) in A. The edges of the Newton polytope are orthogonal to the vectors satisfying the inequalities in (3.4).

Example. Let us consider a polynomial in two variables namely, $f(x_1, x_2) = -x_1^6 + x_1^3x_2 - x_1^3 + x_1x_2^2$. The corresponding Newton polytope is shown in Fig. 3.1 whose edges correspond to the disjunction of three linear inequality systems (cf. (3.4)) leading to three cases, as shown below

case 1: $6x_1 = 3x_1,\ 3x_1 \leq x_1 + 2x_2,$
case 2: $x_1 + 2x_2 = 3x_1,\ 3x_1 \leq 6x_1,$
case 3: $x_1 + 2x_2 = 6x_1,\ 6x_1 \leq 3x_1.$

It is to be noted that the Newton polytope construction simplifies the task of enumerating the cases by automatically eliminating some of them. For example, in the above example the case corresponding to the point $(3, 1)$ is not required as it is absent in the convex hull and hence, will never satisfy the condition in (3.4). The three edges of Newton polytope are the normal vectors satisfying the above cases and constitute the *tropical curve*. Below are the solutions for x_1, x_2 corresponding to above cases

$x_1 = 0, \quad x_2 \geq 0,$
$x_1 \geq 0, \quad x_2 = x_1,$
$x_1 \leq 0, \quad x_2 = \frac{5}{2}x_1.$

For applications considered in this thesis, the edges corresponding to opposite signs monomials are treated differently from the ones with same the sign ones (the justification for it will be given in Chapter 4). To illustrate it, in Fig. 3.1, the edge corresponding to the same sign monomials is represented as a dotted line. Further definitions of the properties of polytopes can be found in (Henk, Richter-Gebert, and Ziegler, 2004).

The computation of tropical zeros i.e. determining the points at which the minimum of a tropical polynomial is attained at least twice, is associated with important concepts namely, *tropical variety* and *tropical prevariety*. The distinction between them becomes clearer in case

of a system of multivariate polynomials. In context of a system of multivariate polynomials denoted by $f_1, f_2, \ldots, f_k$ where $f_i \in \mathbb{C}[x_1, x_2, \ldots, x_n]$, tropical prevariety and tropical variety are defined below

Definition. A *tropical prevariety* is defined as the intersection of a finite number of tropical hypersurfaces, denoted by

$$V(T(f_1, f_2, \ldots, f_k)) = \cap_{i\in[1,k]} T(f_i) \tag{3.5}$$

where $T(f_1, f_2, \ldots, f_k)$ and $V(T(f_1, f_2, \ldots, f_k))$ represent the set of tropicalization of the multivariate polynomials and the common tropical zeros respectively.

Definition. A *tropical variety* is the intersection of all tropical hypersurfaces that belong to the ideal I generated by the polynomials

$$f_1, f_2, \ldots, f_k, V(T(I)) = \cap_{f\in I} T(f) \tag{3.6}$$

where $T(I)$ represents the set of tropicalization of the elements of I and $V(T(I))$ denotes their common tropical zeros.

Remark. The tropical variety is within the tropical prevariety, but the reciprocal property is not always true. The tropical prevariety and tropical variety coincide in case of a single univariate or multivariate polynomial (cf. Section 3.1.5). Furthermore, the tropical variety and tropical prevariety are polyhedral complexes (Richter-Gebert, Sturmfels, and Theobald, 2005) meaning that the intersection between any two solution polytopes is either empty or is a face of each other.

The tropicalization of the ideal $T(I)$ can be represented as

$$T(I) = \cap_{i\in[1,r]} T(g_i) \tag{3.7}$$

where polynomials $g_1, g_2, \ldots, g_r$ denote a finite generating set of the ideal I (which is generated by the original set of polynomials namely, $f_1, f_2, \ldots, f_k$) and are called the *tropical basis* of I.

The relation between the tropical variety and algebraic variety can be represented in the following manner

$$\{\overline{T(V(I))}\} = V(T(I)) \tag{3.8}$$

where $V(I)$ is algebraic variety corresponding to the ideal I. $\{\overline{T(V(I))}\}$ denotes the closure of tropicalization of $V(I)$. In literature, tropical variety is also analysed in the context of Maslov dequantization and idempotent analysis (Maslov and Kolokoltsov, 1994; Litvinov, 2007; Litvinov, Maslov, and Shpiz, 2001) as the logarithmic limit of algebraic variety.

Computing the Tropical Variety

An algorithm to compute the tropical basis is described in (Bogart et al., 2007). However, the complexity of this algorithm can be double-exponential in the size of the system, both in time and in space. It should be pointed out that any universal Gröbner basis of I is not a tropical basis as pointed out in (Bogart et al., 2007). Furthermore, tropical basis could not be unique.

In this context, a fast heuristic to check whether an element in tropical prevariety belongs to the tropical variety is proposed here. The core idea is to define an augmented system consisting of finite number of polynomials from the ideals I which are randomly enumerated along with the polynomials $f_1, f_2, \ldots, f_k$. The tropical prevariety of this augmented system with a *high probability* represents the tropical variety $V(I)$. Although the ideal contains an infinite number of elements, a reasonable choice is to take the sums of products of $f_1, f_2, \ldots, f_k$ by arbitrary polynomials. The number of arbitrary polynomials is fixed at $2k$. An overview about choosing

the arbitrary polynomials can be found in (Hept and Theobald, 2009). The procedure can be formalised in the following steps

1. Compute the tropical prevariety for the polynomials $f_1, f_2, \ldots, f_k$.
2. Sample $2k$ elements from the I denoted by $g_1, g_2, \ldots, g_{2k}$ where $g_i = h_1 f_1 + \ldots + h_k f_k$, h_k are arbitrary polynomials, $h_i \in \mathbb{C}[x_1, x_2, \ldots, x_n]$.
3. Compute the tropical prevariety for the polynomials $g_1, g_2, \ldots, g_{2k}$ by augmenting one element at a time from the tropical prevariety obtained in Step 1. If the obtained tropical prevariety is empty it is called infeasible. Repeat this for all the elements in the tropical prevariety.
4. Elements of tropical prevariety from Step 1 resulting in infeasible tropical prevariety in Step 3, do not belong to the tropical variety. However, if an element in the tropical prevariety survives then it belongs to the tropical variety with a high probability but there is no guarantee. In practice, this would be sensitive to the selection of $g_1, g_2, \ldots, g_{2k}$.

An example demonstrating the above steps will be presented in Section 3.1.5 and the implementation details in Chapter 4.

3.1.4 Upper Bound on the Number of Tropical Zeros

The computation of the tropical prevariety or tropical variety involves the intersection of tropical hypersufaces. In case of tropical curves, the bound on the number of intersection points is given by the theorems of Bézout and Bernstein.

Theorem. *Tropical Bézout theorem: Two general tropical curves C and D of degrees c and d intersect in $c \cdot d$ points, counting multiplicities.*

The intersections are stable if the curves intersect transversely. Otherwise, the curves intersect in infinitely many points. Nevertheless, in such a situation the curves can be perturbed such that they intersect in finitely many points. This is explained with an illustration in (Maclagan and Sturmfels, 2015). In the terminology of Newton polytopes, the number of intersection points is given by Bernstein theorem as defined below

Theorem. *Bernstein Theorem: The number of intersection points of two tropical curves C and D with prescribed Newton polygons P_C and P_D equals to the mixed area of these polygons.*

The mixed area is denoted by $MA(P_C, P_D) = Area(P_C + P_D) - Area(P_C) - Area(P_D)$ where $Area(.)$ is the Euclidean area in $\mathbb{R}^2$ and $P_C + P_D$ is the Minkowsky sum. The Minkowsky sum of two polytopes P_C and P_D is defined as $P_C + P_D = \{c + d \mid c \in P_C,\ d \in P_D\}$ This means that the union of tropical curves is dual to the mixed subdivision of Minkowsky sum of their Newton polygons. The transversal intersection of tropical curves is represented by the mixed cells of the mixed subdivision of Newton polygons. In addition, approaches based on discrete mixed volume (Bihan, 2014) also determine such bounds.

In case of multivariate system of polynomials denoted as

$$f_1, f_2, \ldots, f_k \tag{3.9}$$

where $f_i \in \mathbb{C}[x_1, x_2, \ldots, x_n]$, Bézout theorem is generally valid when $k = n$. A general version of this result in the case of of over determined systems i.e. $k \geq n$ is analysed in (Davydow and Grigoriev, 2015).

3.1.5 Newton-Puiseux Series and Tropical Geometry

In this section, the relationship between the tropical solutions and Newton-Puiseux series solutions is discussed. The Newton-Puiseux series is used to represent the zeroes of polynomial whose coefficients are series with fractional exponents of a small parameter ε. In other words, the main objective is to solve polynomials over the field of the Newton-Puiseux series defined by $\mathbb{C}\{\{\varepsilon\}\}$, where ε plays the role of indeterminate in the formal power series. The elements in $\mathbb{C}\{\{\varepsilon\}\}$ are formal power series in ε with rational exponents and common denominator which are bounded below. These can be represented as $x(\varepsilon) = \tau_1\varepsilon^{a_1} + \tau_2\varepsilon^{a_2} + \cdots$, where $\tau_i \in \mathbb{C}$, and $a_1 < a_2 < \cdots$ are rational numbers with common denominator. The series is convergent in some neighborhood of the origin, the origin is excluded if the exponent in less than 0. Let us demonstrate the main idea with an univariate polynomial as shown below

$$f(x,\varepsilon) = A_d(\varepsilon)x^d + A_{d-1}(\varepsilon)x^{d-1} + \ldots + A_2(\varepsilon)x^2 + A_1(\varepsilon)x + A_0(\varepsilon) \tag{3.10}$$

The number of roots in such a situation is given by the Newton-Puiseux theorem which is defined below

Theorem. *Puiseux theorem: The field of Puiseux series denoted by $\mathbb{C}\{\{\varepsilon\}\}$ is algebraically closed and the polynomial $f(x,\varepsilon)$ has d roots counting multiplicities, in the field of $\mathbb{C}\{\{\varepsilon\}\}$.*

As the coefficients of (3.10) are power series with fractional exponents in ε. The i^{th} coefficient can be written as the Newton-Puiseux series expansion in the following manner:

$$A_i(\varepsilon) = \bar{A}_i\varepsilon^{\gamma_i} + \text{higher order terms in } \varepsilon \tag{3.11}$$

The roots of $f(x,\varepsilon)$ will be expressed in the similar manner

$$x(\varepsilon) = \bar{x}\varepsilon^{a_1} + \text{higher order terms in } \varepsilon \tag{3.12}$$

In order to compute the lowest order terms of Newton-Puiseux series for the roots, the values of $a_1, \bar{x}$ in $\mathbb{Q} \times \mathbb{C}$ for which $f(x,\varepsilon) = 0$ need to be determined. First, let us compute the possible values of a_1 by rewriting (3.10) with the lowest order terms as shown below

$$\bar{A}_d\bar{x}_1^d\varepsilon^{\gamma_d+da_1} + \bar{A}_{d-1}\bar{x}^{d-1}\varepsilon^{\gamma_d+(d-1)a_1} + \ldots + \bar{A}_2\bar{x}^2\varepsilon^{\gamma_2+2a_1} + \bar{A}_1\bar{x}^1\varepsilon^{\gamma_1+a_1} + \bar{A}_0\varepsilon^{\gamma_0} = 0 \tag{3.13}$$

The possible values of a_1 satisfying (3.13) can be computed by solving the following system of inequalities

$$\min(\gamma_d + da_1, \gamma_{d-1} + (d-1)a_1, \ldots, \gamma_2 + 2a_1, \gamma_1 + a_1, \gamma_0) \tag{3.14}$$

where the min is attained at least twice. This is required, else (3.13) will be nonzero due to the presence of a leading term. Furthermore, if one looks for *positive* and *real* solutions i.e. $x \in \mathbb{R}^+$, then it follows that at least two terms corresponding to the minimum should have the opposite signs. This condition is a necessary, but not a sufficient condition for positive real solutions (Radulescu, Vakulenko, and Grigoriev, 2015). Furthermore, (3.14) can be also seen as a tropical polynomial obtained by tropicalization of (3.10). Therefore, in this case, computing the possible values of a_1 means computing for the tropical zeros of this tropical polynomial (cf. Section 3.1.3). Representing the zeros of (3.10) as formal power series is equivalent to *lifting* of the tropical zeros to a Newton-Puiseux series. In the univariate case, this is always possible by the Puiseux theorem and in the multivariate case, this is ensured by the theorem of Kapranov (Einsiedler, Kapranov, and Lind, 2006).

After determining the possible values of a_1, the possible values of $\bar{x}$ can be determined by solving the following equation for non-zero solutions

$$\begin{gathered}\sum_{i,\gamma^i+ia=m} A_i\bar{x}^i = 0 \\ m = \min_i(\gamma_1^i + ia_1)\end{gathered} \tag{3.15}$$

Although this step involves solving of a polynomial equation, but in general this involves solving a system of binomial equations i.e. polynomial with exactly two terms, for which polynomial time algorithm exists (Grigoriev and Weber, 2012; Millán et al., 2012).

To this end, the core idea depicting the relation between the classical and tropical algebra can be shown in the following manner

$$\text{Classical Algebra} \overset{\text{Tropicalization}}{\underset{\text{Lifting}}{\longleftrightarrow}} \text{Tropical Algebra} \tag{3.16}$$

Some examples explaining the computation of Newton-Puiseux series using the framework of tropical geometry are discussed below.

Example. Consider the univariate polynomial $f(x, \varepsilon)$

$$x^2 + x - \varepsilon^3 = 0 \tag{3.17}$$

Step 1: Rewrite the system with lowest order terms (cf. (3.13)) as shown below

$$\bar{x}^2\varepsilon^{2a_1} - \bar{x}\varepsilon^{a_1} - \varepsilon^3 = 0 \tag{3.18}$$

Step 2: Solve the above equation for tropical zeros based on the technique of Newton polytope (cf. Section 3.1.3). The inequalities corresponding to the edges of Newton polytope are $2a_1 = a_1 \leq 3$ or $2a_1 = 3 \leq a_1$or $3 = a_1 \leq 2a_1$ resulting in two points namely $a_1 = 0$ or $a_1 = 3$. This denotes the tropical variety (recall for a single univariate polynomial the tropical prevariety and tropical variety coincide).

Step 3: The coefficients of Newton-Puiseux series i.e. $\bar{x}$ corresponding to the possible values of a_1 can be computed using (3.15) as exemplified below

For $a_1 = 0$, the following equation needs to be solved

$$\bar{x}^2 + \bar{x} = 0 \tag{3.19}$$

This results in a single non-zero value, $\bar{x} = -1$

Similarly, for $a_1 = 3$, the following equation needs to be solved

$$\bar{x} - 1 = 0 \tag{3.20}$$

This results in $\bar{x} = 1$

Therefore, two different Newton-Puiseux series solutions are obtained with -1 and ε^3 as the lowest order terms. For the application scenario considered in this thesis, the focus will be to compute the possible values of a_1 only.

In case of a system of multivariate polynomial equations denoted by $f_1, f_2, \ldots, f_k$ where $f_i \in \mathbb{C}[x_1, x_2, \ldots, x_n]$, there is no analogue of Kapranov theorem. In this case, the tropical zeros are the necessary condition for the existence of Newton-Puiseux solutions. Recall from Section 3.1.3 that the set of solutions obtained from solving the disjunction of min cases result into tropical prevariety. In case of a single polynomial, the elements of such a prevariety can be lifted to the series solution whereas for systems of equations this is always not the case. The computation of tropical variety involves the computation of the *tropical basis* whose algorithmic complexity can be double-exponential in the size of the system, both in time and in space.

Nevertheless, a heuristic method to check whether an element of the tropical prevariety belongs to the variety is presented in Section 3.1.3. Below is an example demonstrating this heuristic along with the steps to solve for the lowest order of the Newton-Puiseux series for a system of equations.

Example. Consider two polynomials in two variables

$$\begin{aligned} x - y - \varepsilon x^4 &= 0 \\ y - x + \varepsilon y^2 &= 0 \end{aligned} \tag{3.21}$$

Step 1: Rewrite the system as in (3.10) as shown below:

$$\begin{aligned} \bar{x}\varepsilon^{a_1} - \bar{y}\varepsilon^{a_2} - \bar{x}^4\varepsilon^{4a_1+1} &= 0 \\ \bar{y}\varepsilon^{a_2} - \bar{x}\varepsilon^{a_1} + \bar{y}^2\varepsilon^{2a_2+1} &= 0 \end{aligned} \tag{3.22}$$

Step 2: The inequalities corresponding to the edges of Newton polytope (considering edges with the opposite sign terms) for the 1st equation are $a_1 = a_2 \leq 4a_1 + 1$ or $a_1 = 4a_1 + 1 \leq a_2$. Similarly, for the 2nd equations , the inequalities are $a_2 = a_1 \leq 2a_2 + 1$ or $a_1 = 2a_2 + 1 \leq a_2$. A subset of tropical prevariety (as only terms of opposite signs are considered) is given by the intersection of tropical hypersurfaces resulting in a disjunction of min cases are mentioned below (in a combinatorial manner):

$$\begin{aligned} &a_1 = a_2 \leq 4a_1 + 1 \quad \text{and } a_2 = a_1 \leq 2a_2 + 1 \text{ or,} \\ &a_1 = 4a_1 + 1 \leq a_2 \quad \text{and } a_2 = a_1 \leq 2a_2 + 1 \text{ or,} \\ &a_1 = a_2 \leq 4a_1 + 1 \quad \text{and } a_1 = 2a_2 + 1 \leq a_2 \text{ or,} \\ &a_1 = 4a_1 + 1 \leq a_2 \quad \text{and } a_1 = 2a_2 + 1 \leq a_2. \end{aligned} \tag{3.23}$$

The possible values of a_1 and a_2 are $a_1 = a_2 \geq -\frac{1}{3}$ or $a_1 = -\frac{1}{2}, a_2 = -1$. The first solution is an infinite branch whereas the second solution is an isolated point. This constitutes the tropical prevariety.

In order to test which solution in the tropical prevariety belongs to tropical variety and hence lifts to Newton-Puiseux series, the procedure explained in Section 3.1.3 will be adopted. Let us generate an element from ideal by adding both the equations in (3.22) resulting in $\bar{y}^2\varepsilon^{2a_2+1} - \bar{x}^4\varepsilon^{4a_1+1}$. The corresponding inequality for this equation is $2a_2 + 1 = 4a_1 + 1$ (cf. Step 2 above) when added to the tropical prevariety results in two isolated points namely $(-\frac{1}{2}, -1)$ and $(0, 0)$.

It can be verified that these two points namely, $(-\frac{1}{2}, -1)$ and $(0, 0)$ are *liftable* to the series solution in the following manner

Adding both the equations in (3.21), we get

$$y = \pm x^2 \tag{3.24}$$

Substituting it back in the 1st equation of (3.21), we get

$$\pm x - 1 - \varepsilon x^3 = 0 \text{ or } x = 0 \tag{3.25}$$

Solving $\pm x - 1 - \varepsilon x^3$ using Newton potytope approach, it can be shown that the possible values of a_1 are 0 or $-\frac{1}{2}$. By Puiseux theorem, these points are liftable to the Newton-Puiseux series. Similarly, one can show that the values of a_2 are 0 and -1. Therefore, there are two liftable solutions namely $(a_1, a_2) = (0, 0)$ and $(-\frac{1}{2}, -1)$.

It is to be noted here that, for this specific example a single element from ideal I was enough to determine the liftable points. In general, it is advisable to use $2k$ elements from the ideal as explained in Section 3.1.3.

3.2 Extreme Currents

Metabolic networks are usually represented as a collection of enzyme catalysed biochemical reactions. There are numerous databases with such information (Kanehisa and Goto, 2000; Croft et al., 2011; Le Novére et al., 2006). One way to study such a network is by decomposing the network into sub-networks or pathways in an unbiased manner. Such pathways represent different metabolic routes for the production of given metabolites along with the essential enzymes. Extreme Currents (ECs) (Clarke, 1988), Elementary Flux Modes (EFMs) (Schuster, Fell, and Dandekar, 2000) and Extreme Pathways(EPs) (Schilling, Letscher, and Palsson, 2000) are three widely used techniques in this context. The common assumption is that the underlying biochemical reaction system is in a steady state, and with additional constraints on reaction fluxes, the solution space can be represented as a polyhedron. The vertices of such a polyhedron have the biochemical interpretation of being pathways in the network. A comparison between different metabolic pathway techniques can be found in (Llaneras and Picó, 2010). One major drawback of pathway enumeration is that the number of pathways can explode in a combinatorial fashion with the size of the network (Klamt and Stelling, 2002). Hence, optimization techniques are frequently used for larger networks . For an overview (Rezola et al., 2015) is recommended.

In short, the decomposed pathways are basically steady state reaction-flux distributions which capture a wide range of behaviour that the network is capable to exhibit. Essentially, these pathways are invariants of the network and hence do not require the kinetic rate parameters to be known or estimated, which is often very difficult in practice.

Consider the dynamics of biochemical reaction networks modeled using reaction fluxes (cf. Section 2.3) where reaction fluxes are represented as a vector V and the stoichiometric matrix as S. It is important to note here that while determining S, the reversible reactions in the network are split into separate forward and backward reactions.

Solving such a system of equations for steady state (i.e. at $\frac{\mathrm{d}x_i}{\mathrm{d}t} = 0$) along with non-negativity constraint on the fluxes (i.e. on the variables V) results in a convex polyhedron (also known as flux cone) and is defined as follows:

$$P = \{V \in \mathbb{R}^n \colon SV = 0, V_i \geq 0\} \tag{3.26}$$

The vertices of polyhedron P are called extreme currents (ECs) (Clarke, 1988) or convex bases which can be understood as pathways in the reaction network. The non-negativity constraint can be weakened by asking its satisfaction for a subset of reactions that are irreversible and do not have negative fluxes. Reversible reactions can have negative fluxes. Sometimes, a reversible reaction is decomposed artificially into irreversible reactions of opposite stoichiometry. Thus, resulting in all fluxes being positive. In steady state, the fluxes are linear combinations of ECs and can be written as

$$V = jE, \tag{3.27}$$

where E is a $k \times r$ extreme current matrix and j is a k dimensional row vector of non-zero real numbers called *convex parameters*. In literature, such pathways are also referred to as elementary flux modes (EFMs) or extreme pathways (EPs), depending on the way in which reversible reactions are split (affecting the non-negativity constraint on the fluxes). As we split the reversible reactions into separate forward and backward reactions, therefore, ECs, EFMs and EPs are equivalent (Llaneras and Picó, 2010).

Chapter 4

Algorithms for the Computation of Tropical Equilibrations

In this chapter algorithms for computing the set of TEs are discussed. TEs are a subset of tropical prevariety. In addition, a heuristic is provided to check for the tropical equilibrations which could be lifted to the tropical variety and hence, to the Newton-Puiseux series solution. To evaluate the performance of the proposed algorithms, a benchmark is shown on a set of models obtained Biomodels databases for the described algorithms.

4.1 Computation of Tropical Equilibrations

In this section, algorithms computing the tropical equilibrations will be discussed. The idea of tropical equilibration is based on the notion of tropical prevariety (cf. Chapter 3). We now present the heuristics to study the notion of tropical equilibrations.

For biochemical reaction networks represented as ODEs (cf. (2.1)) with multiple timescales, it is reasonable to consider that the kinetic parameters have different orders of magnitudes.

Therefore, it is assumed that the parameters of the kinetic models (2.1) can be written as

$$k_j = \bar{k}_j \varepsilon^{\gamma_j}. \tag{4.1}$$

The exponents γ_j are considered to be integer. For instance, the following approximation produces integer exponents:

$$\gamma_j = \text{round}(\log(k_j)/\log(\varepsilon)), \tag{4.2}$$

where round stands for the closest integer (with half-integers rounded to even numbers). Without rounding to the closest integer, changing the parameter ε should not introduce variations in the output of the method.

In contrast, species orders vary in time and have to be computed. To this aim, the species concentrations are first represented by orders of magnitude defined as

$$a_j = lim_{\varepsilon \to \infty} \frac{\log(x_j)}{\log(\varepsilon)}. \tag{4.3}$$

More precisely, one has $x_j = \bar{x}_j \varepsilon^{a_j}$, where $\bar{x}_j$ has zero order (unity). Because $\log(\varepsilon) < 0$, Equation (4.3) means that species orders and concentrations are anti-correlated (large orders mean small concentrations and vice versa).

We assume that the kinetic parameters are fixed. In contrast, the orders of the species vary in the concentration space and have to be calculated as solutions to the tropical equilibration

problem. To this aim, the network dynamics is first described by a rescaled ODE system

$$\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t} = \sum_j \varepsilon^{\mu_j - a_i} \bar{k}_j S_{ij} \bar{\boldsymbol{x}}^{\boldsymbol{\alpha_j}}, \tag{4.4}$$

where

$$\mu_j(\boldsymbol{a}) = \gamma_j + \langle \boldsymbol{a}, \boldsymbol{\alpha_j} \rangle, \tag{4.5}$$

and $\langle , \rangle$ stands for the dot product.

The r.h.s. of each equation in (4.4) is a sum of multivariate monomials in the concentrations. The orders μ_j indicate how large are these monomials, in absolute value. A monomial of order μ_j dominates another monomial of order $\mu_{j'}$ if $\mu_j < \mu_{j'}$.

The tropical equilibration problem consists in finding a vector $\boldsymbol{a}$, which will be called tropical equilibration solution such that

$$\min_{j, S_{ij}>0} (\gamma_j + \langle \boldsymbol{a}, \boldsymbol{\alpha_j} \rangle) = \min_{j, S_{ij}<0} (\gamma_j + \langle \boldsymbol{a}, \boldsymbol{\alpha_j} \rangle) \tag{4.6}$$

If γ_j are integers, the solutions $\boldsymbol{a}$ are rational. For the purpose considered here, the classical notion of tropical prevariety (cf. Chapter 3) is slightly modified. A *tropical equilibration* is defined as a vector $\boldsymbol{a} \in \mathbb{R}^n$ such that $\langle \boldsymbol{\alpha}, \boldsymbol{a} \rangle$ attains its minimum at least twice for monomials of different signs, for each polynomial in the (4.4). Thus, tropical equilibrations are subsets of the tropical prevariety. The rationale behind selecting monomials with different sign conditions is explained in Sub-section 3.1.5.

The system in (4.6) can be represented as a set of linear inequalities resulting into a convex polytope. The solutions of this system have a geometrical interpretation. Let us define the extended order vectors $\boldsymbol{a}^e = (1, \boldsymbol{a}) \in \mathbb{R}^{n+1}$ and extended exponent vectors $\boldsymbol{\alpha}_j^e = (\gamma_j, \boldsymbol{\alpha}_j) \in \mathbb{Z}^{n+1}$. Let us consider the equality $\mu_j = \mu_{j'}$. This represents the equation of a n dimensional hyperplane of $\mathbb{R}^{n+1}$, orthogonal to the vector $\boldsymbol{\alpha_j}^e - \boldsymbol{\alpha_{j'}}^e$:

$$\langle \boldsymbol{a}^e, \boldsymbol{\alpha_j}^e \rangle = \langle \boldsymbol{a}^e, \boldsymbol{\alpha_{j'}}^e \rangle, \tag{4.7}$$

where $\langle , \rangle$ is the dot product in $\mathbb{R}^{n+1}$. The minimality condition on the exponents μ_j implies that the normal vectors $\boldsymbol{\alpha_j}^e - \boldsymbol{\alpha_{j'}}^e$ are edges of the so-called Newton polytope (Henk, Richter-Gebert, and Ziegler, 2004; Sturmfels, 2002).

For each equation i, let us define

$$M_i(\boldsymbol{a}) = \operatorname*{argmin}_j(\mu_j(\boldsymbol{a}), S_{ij} > 0) = \operatorname*{argmin}_j(\mu_j(\boldsymbol{a}), S_{ij} < 0), \tag{4.8}$$

in other words M_i denote the set of monomials having the same minimal exponent μ_i.

Tropically truncated system can be defined as the system obtained by keeping only the dominating monomials in (4.4), as follows:

$$\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t} = \varepsilon^{\mu_i - a_i} \Big(\sum_{j \in M_i(\boldsymbol{a})} \bar{k}_j S_{ij} \bar{\boldsymbol{x}}^{\boldsymbol{\alpha_j}} \Big). \tag{4.9}$$

The tropical truncated system is uniquely determined by the index sets $M_i(\boldsymbol{a})$, therefore, by the tropical equilibration $\boldsymbol{a}$. Reciprocally, two tropical equilibrations can have the same index sets $M_i(\boldsymbol{a})$ and truncated systems. Two tropical equilibrations $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ are said to be equivalent iff $M_i(\boldsymbol{a}_1) = M_i(\boldsymbol{a}_2)$, for all i. Equivalence classes of tropical equilibrations are called as *branches*. For each branch there exists a unique convex polytope, cf. (4.6). The union of branches are subsets of tropical prevariety. It is a subset because we are interested in tropical equilibration of at least two monomials of different signs for the reasons discussed above. This is computed

using the method *equal_polyhedra* implemented in the software package polymake (Gawrilow and Joswig, 2000).

In case of rational expressions e.g. Michaelis-Menten enzymatic mechanism, the least common denominator of reaction rates can be used to express the right hand sides of the kinetic equations as ratios of polynomials and apply the method to the numerators. This extension was briefly discussed in (Noel et al., 2012).

Minimal Branches A branch B with an index set M_i is *minimal* if $M_i' \subset M_i$ for all i where M_i' is the index set of a branch B' implies $B' = B$ or $B' = \emptyset$. In the terminology of convex polytopes, this means a branch B with a convex polytope P_i is *minimal* if $P_i \subset P_i'$ for all i where P_i' is the convex polytope for branch B' implies $B' = B$ or B' is empty. For each index i, relation (4.7) defines a hyperplane, the tropical equilibration branches are on the intersections of k such hyperplanes where k is number of polynomial equations representing the right hand side of (2.1). Minimal branches are maximal (w.r.t. inclusion) polytopes in the tropical prevariety. This computation is done using *included_polyhedra* method implemented in polymake. Such inclusions are represented using directed graphs.

Connected Components of Minimal Branches Two minimal branches represented by index sets M_i and M_j are connected if there exists a branch with index set M_k such that $M_i \subset M_k$ and $M_j \subset M_k$. In the terminology of convex polytopes, this amounts to checking the intersection between two convex polytopes P_i and P_j (corresponding to minimal branches M_i and M_j) if whether $P_i \cap P_j$ is non void for all $i \neq j$. The connected component is the zeroth homology group of the polyhedral complex corresponding to the minimal branches. It indicates the possible transitions between the minimal branches. Geometrically, the minimal branches can intersect at a common face or the intersection be empty (recall, minimal branches are subsets of tropical prevariety and hence is a polyhedral complex). Therefore, for a given pair of convex polytopes corresponding to minimal branches, their intersection can be void or not. This check was performed in polymake environment and connected component(s) were computed. Specifically, this means that an undirected graph is constructed whose vertices are minimal branches and there exists an edge if the intersection between the two vertices is non-void.

4.2 Justification of the Tropical Equilibrations

The justification for the heuristics of tropicalization is described in Section 3.3 of (Noël, 2012). Here, we reproduce the main results. Consider the system in (4.4). We suppose that the cone $R_> = \{\bar{x} : \bar{x}_i \geq 0\}$ is invariant under the dynamics exhibited by (4.4) and the initial data is positive meaning, $\bar{x}_i(0) > \delta > 0$. The sign of terms in (4.1) determines if they are production (i.e. positive sign) or degradation (i.e. negative sign). Given that biochemical networks very often have multiple well separated timescales, so choice of (4.1) is justified. In this context, we are interested for determining the stable functioning of biochemical model and in order to study that we use the permanency condition as defined below

Definition. The system defined in (4.4) is permanent, if there are two constants $C_- > 0$ and $C_+ > 0$, and a function T_0, such that

$$C_- < \bar{x}_i(t) < C_+, \text{ for all } t > T_0(\bar{x}(0)) \text{ and for every } i \tag{4.10}$$

We assume that $C_\pm$ and T_0 are uniform and do not depend on ε as $\varepsilon \to 0$.

For such permanent systems, Proposition 3.3.2 in (Noël, 2012) states that if the dynamics of (4.4) is structurally stable in the domain $\Omega_{C_-,C_+} = \bar{x} : C_- < |\bar{x}| < C_+$, then the corresponding

tropical truncated system defined in (4.9) is also permanent and there is an orbital topological equivalence $\tilde{x}_i = h_\varepsilon(\bar{x})$ between the trajectories $\bar{x}(t)$ and $\tilde{x}_t$ of the corresponding Cauchy problems with same initial conditions. The solution of the tropically truncated system in (4.9) is denoted by $\tilde{x}$. The homemorphism h_ε is close to identity as $\varepsilon \to 0$.

For permanency condition, the necessary condition is that if a tropically truncated system is permanent, then for each $i \in \{1, \ldots, n\}$, the i-th equation of this system contains at least two terms. The terms should have different signs for the coefficients k_{ij} i.e. one term should be production and the other being the degradation term. This provides the justification to consider monomials of opposite signs for computing the tropical equilibration. Now, we state the sufficient condition for permanency in the following way

In a generic case, one can expect that all μ_i from (4.9) are mutually different, namely

$$\mu_1 < \mu_2 < \cdots < \mu_{n-1} \leq \mu_n. \tag{4.11}$$

In such a situation, let us consider the first equation of (4.9) with $i = 1$ and let try to denote $y = \tilde{x}_1$, $z = (\tilde{x}_2, \ldots, \tilde{x}_n)$. In this new notation, the first equation becomes

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y) = b_1(z)y^{\beta_1} - b_2(z)y^{\beta_2},\ b_1, b_2 > 0,\ \beta_i \in \mathbb{R} \tag{4.12}$$

It has been shown in Lemma 3.3.4 in (Noël, 2012) that (4.12) has permanency if and only if $\beta_1 < \beta_2$. For fixed z, in these cases we have $y(t, z) \to y_0$, as $t \to \infty$.

If the timescales are not all well separated, for instance if the orders of last two slowest timescales are allowed to be equal, the permanency condition is given by following theorem

Theorem. *Assume $\mu_1 < \mu_2 < \cdots < \mu_{n-1} \leq \mu_n$ holds. If the procedure, described above, leads to the permanency property at each step, where $i = 1, 2, \ldots, n-2$, and the last two equations have a globally attracting hyperbolic rest point or globally attracting hyperbolic limit cycle, then the tropically truncated system is permanent and has an attractor of the same type. Moreover, for sufficiently small ε the initial system also is permanent for initial data from some appropriate domain $W_{\varepsilon,a,A}$ and has an analogous attracting hyperbolic rest point (limit cycle) close to the attractor of the truncated system. If the rest point (cycle) is not globally attracting, then we can say nothing on permanency but, for sufficiently small ε, the initial system still has an analogous attracting hyperbolic rest point (limit cycle) close to the attractor of truncated system and the same topological structure.*

Tropical equilibrations satisfying the permanency condition imply invariant manifold. Given such invariant manifolds, model reduction techniques can be applied to reduce the number of variables and obtain the dynamics close to the original dynamics (Gorban and Karlin, 2005). This has been showed in the Lemma 3.3.6 in (Noël, 2012). The application of tropical equilibration for model reduction is covered in Chapter 6.

4.3 Implementation and Results

In this section, an algorithm is described to compute the tropical equilibrations, test the equilibrations for the equivalence classes (i.e., *branches*) and compute the *minimal branches* as described in the previous section. The algorithm to compute the tropical equilibrations is implemented in two different ways namely linear programming and convex polytope approach (cf. Section 4.3.4 and 4.3.5). For testing the implementations a data set comprising of models from a public database is created (cf. Section 4.3.1).

4.3.1 Data Source

For benchmarking, 36 models are selected from the r25 version of the Biomodels database (Le Novere et al., 2006). All models that have polynomial vector field (mass action kinetics) and satisfy some technical constraints imposed by the SBML parser (no function definitions, for instance) are selected. Models with zero valued parameters or which do not have at least one positive and one negative monomial per ODE are also filtered out (there are two of those). The model SBML files are parsed and the polynomial vector fields are extracted. Thereafter, the conservation laws (that are the sum of the variables whose total concentration is invariant) are computed. The vector field along with the conservation laws are the input to the tropical geometry based algorithm to compute the minimal branches. It should be noted here that due to the conservation laws the number of equations may exceed the number of chemical species. Additionally, as described in Chapter 3, conservation laws can be used to eliminate variables of the ODE system. The computation of minimal branches are also done for such *reduced* systems and the results are compared. The algebraic expressions in the equation system are processed using computer algebra system Maple.

4.3.2 Newton Polytope and Edge Filtering

Given the input polynomial in the form of (4.4), for each equation and species i, a Newton polytope $\mathcal{N}_i$ is defined, that is the convex hull of the set of points $\boldsymbol{\alpha}_j^e$ such that $S_{ij} \neq 0$ and also including together with all the points the half-line emanating from these points in the positive ϵ direction. This is the Newton polytope of the polynomial in right hand side of (4.4), with the scaling parameter ε considered as a new variable.

As explained above, the tropical equilibrations correspond to vectors $\boldsymbol{a}^e = (1, \boldsymbol{a}) \in \mathbb{R}^{n+1}$ satisfying the optimality condition as per (4.6). This condition is satisfied automatically on hyperplanes orthogonal to edges of Newton polytope connecting vertices $\boldsymbol{\alpha}_{j'}^e$, $\boldsymbol{\alpha}_{j''}^e$ satisfying the opposite sign condition. Therefore, a subset of edges from the Newton polytope is selected based on the filtering criteria which tells that the vertices belonging to an edge should be from opposite sign monomials as explained in (4.13).

$$E(P) = \{\{v_1, v_2\} \subseteq \tbinom{V}{2} \mid \mathrm{conv}(v_1, v_2) \in F_1(P) \\ \wedge \mathrm{sign}(v_1) \times \mathrm{sign}(v_2) = -1\}, \tag{4.13}$$

where v_i is the vertex and V is the vertex set of the Newton polytope, $\mathrm{conv}(v_1, v_2)$ is the convex hull of vertices v_1, v_2 and $F_1(P)$ is the set of 1-dimensional face (edges) of the Newton polytope, $\mathrm{sign}(v_i)$ represents the sign of the monomial which corresponds to vertex v_i. Figure 3.1 shows an example of Newton polytope construction for a single equation. Further definitions about properties of a polytope and Newton polytope can be found in (Henk, Richter-Gebert, and Ziegler, 2004; Sturmfels, 2002).

4.3.3 Computing Tropical Equilibrations

Using the Newton polytope formulation, one can then solve the tropical equilibration problem in (4.6) using the edges of Newton polytope (as in (4.8)). A feasible solution is a vector $(a_1, \ldots, a_n)$ satisfying all the equations of system (4.6) and lies in the intersection of hyperplanes (or convex subsets of these hyperplanes) orthogonal to edges of Newton polytopes obeying the sign conditions. Of course, not all sequences of edges lead to non-void intersections and, thus, feasible solutions. This can be tested by the following linear programming problem resulting

from (4.6):

$$\begin{aligned}\gamma_j(i) + \langle \boldsymbol{a}, \boldsymbol{\alpha_j(i)} \rangle = \gamma'_j(i) + \langle \boldsymbol{a}, \boldsymbol{\alpha'_j(i)} \rangle \leq \gamma''_j + \langle \boldsymbol{a}, \boldsymbol{\alpha''_j} \rangle), \\ \text{for all } j'' \neq j, j', \nu_{j''i} \neq 0, \quad i = 1, \ldots, n\end{aligned} \tag{4.14}$$

where $j(i), j'(i)$ define the chosen edge of the ith Newton polytope. The set of indices j'' can be restricted to vertices of the Newton polytope, because the inequalities are automatically fulfilled for monomials that are internal to the Newton polytope. From (4.14), the sequence of edges leading to a feasible solution is actually a set of linear inequalities and hence constitutes a feasible solution system (convex polytope), which is computed using Algorithm 1. Such feasible solution systems are actually convex polytopes (cf. solution polytopes in Chapter 3) as defined in (4.6). For instance, in the example of the preceding section, the choice of the edge connecting vertices $(1, 2)$ and $(6, 0)$ leads to the following linear programming problem:

$$a_1 + 2a_2 = 3a_1 \leq 6a_1,$$

whose solution is a half-line orthogonal to the edge of the Newton polygon. The pseudo-code is presented in Algorithm 1. It is clear from the above that the possible choices are exponential. In order to improve the running time of the algorithm, the pruning strategy evaluates (4.14) in several steps (cf. Algorithm 1 and Fig. 4.1). It starts with an arbitrary pair of edges and proceeds to add the next edge only when the inequalities (4.14) restricted to these two pair of edges are satisfied. The pruning method is a heuristic to filter out the infeasible set of edge combinations. A variant of algorithm computing only the minimal branches is also implemented (cf. Algorithm 2). A similar approach was undertaken in (Emiris and Canny, 1995; Sommars and Verschelde, 2016). As the tropical solutions are either isolated points or bounded or unbounded polyhedra, changing the parameter ε is just a way to approximate the position of these points and polyhedra by lattices or in other words by integer coefficients vectors. Finding the value of ε that provides the best approximation is a complicated problem in Diophantine approximation. For that reason, an experimental approach is preferred which consisted of choosing several values of ε and checking the robustness of the results.

4.3.4 Linear Programming Approach

In this approach, if the *LinearSolve* function in Algorithm 1 is true, a single feasible solution is computed and added to the solution set using the standard linear programming appraoch using Gurobi (Gurobi Optimization, 2012) software in Java programming environment. In other words, this means picking a single sample point from the feasible solution system. As discussed above such feasible solution systems are convex polytopes and can have dimension greater than zero, resulting in infinite tropical equilibrations. However, one solution per branch is enough for the identification of variable timescales and reduced models, which are the main applications of the tropical geometry approach considered in this thesis. Moreover, as this approach does not take into account the polytope structure hence in case of non-empty intersection among solution polytopes, the program might select sample points at the intersection of solution polytopes (i.e. within their common faces). Nevertheless, this provides a first insight into the working of the algorithm and the running times. Table 4.1 presents the benchmark results on the biomodels.

4.3.5 Convex Polytope Approach

In this approach, if the *LinearSolve* function in the Algorithm 2 is true, the convex polytope corresponding to the feasible system of inequalities is augmented to every element in the solution set at every iteration step (i.e. for each successive edge of k-th edge set). Upon completion

Algorithm 1: SolveOrders: Steps of tropical equilibration algorithm implementing the linear programming approach.

Input: List of edge sets $ne_1, ne_2, \ldots, ne_n$ (cf. Fig. 4.1), and the corresponding vertices of Newton polytope

Output: Set of tropical equilibrations corresponding to orders of the variables $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$

```
begin
    solutionset ={}; integer k=1; equation = {}; inequalities = {}
    SolveOrders(equation, inequalities, k, ne_1, .., ne_n, vertices)
    if k > n then
        return
    for l = 1 to number of entries in ne_k do
        equation(k)* = vertices in l^th row
        inequalities(k)* = all vertices other than l^th row
        if LinearSolve(equation,inequalities)** is feasible then
            if k = n then
                add the solution to solutionset
            SolveOrders(equation, inequalities, k + 1, ne_1, .., ne_n, vertices)***
```

*The equations and inequalities are initialised as per (4.14).

**Solves the system of equations and inequalities using the linear programming software.

***Recursive call to the function.

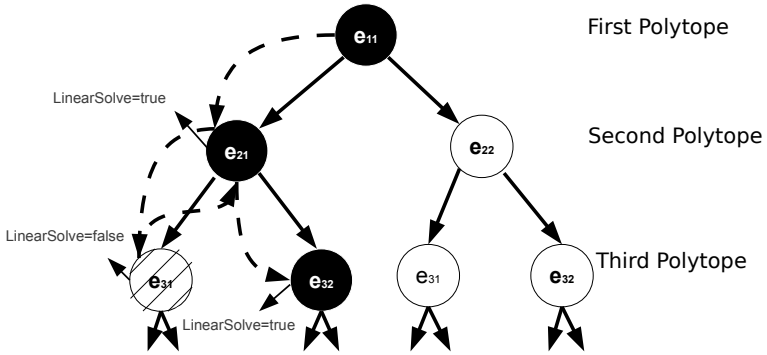


FIGURE 4.1: Pruning strategy. The possible combinations of edges are represented in a tree representation where e_{ji} represents ith edge from jth Newton polytope. An edge set ne_i is the set of edges for ith Newton polytope. The algorithm starts by testing for feasible solution for first pair of edge sets. If a feasible solution is found, the algorithm proceeds further to other edge sets or it backtracks. In the figure, e_{11} and e_{21} are selected from edge sets ne_1, ne_2 and are checked for a feasible solution satisfying (4.14). If such a solution exists, it moves to $e_{31}$from the next edge set and again checks for feasible solution, if not then it backtracks to e_{21} and then to e_{32} which results in a feasible solution. Therefore, the sub-tree with root node e_{31} is discarded from future searches and this improves the running time. Likewise the branches e_{11} and e_{22} are explored. This approach is similar to the branch and bound algorithm technique. The dashed arrows show the flow of the program

TABLE 4.1: Summary of analysis on Biomodels database based on the implementation of Algorithm 1. The benchmarked models have a number of dimensions (i.e. number of variables along with number of conservation laws) ranging from 2 to 41. Model BIOMD0000000289 has tropical branches at ε value 1/5, 1/7, 1/9, 1/11 but none at 1/17, 1/19, 1/23. Model BIOMD0000000080 has no solution only at ε value 1/19.

ε value	Total models considered	Models without tropical equilibrations	Models with tropical equilibrations	Timed-out models	Average running time (in secs)	Average number of tropical equilibrations	Min number of tropical equilibrations	Max number of tropical equilibrations
1/5	36	0	33	3	128.15	5.15	1	35
1/7	36	0	33	3	267.09	6.63	1	35
1/9	36	0	33	3	232.85	6.45	1	33
1/11	36	0	33	3	258.30	5.36	1	33
1/17	36	1	30	5	211.79	6.41	0	61
1/19	36	2	30	4	246.22	7.06	0	61
1/23	36	1	30	5	90.79	5.90	0	61

of the iterations, the equivalence classes of the convex polytope in the solution set gives the number of branches. Furthermore, from the branches, the minimal branches and the connected components can be subsequently computed.

Nevertheless, for the applications considered here, the most interesting ones are minimal branches, so instead of storing the equivalence classes of convex polytope at every iteration step, equivalence classes of only the maximal polytopes (with respect to inclusion) are stored. In the end, i.e. after all edge sets are evaluated, the solution set consists of only the minimal branches. In addition, the dimensions of the polytopes corresponding to the minimal branches are also computed and shown in Fig. 4.13. The dimension here refers to the dimension of the affine hull of the polytope. The affine hull of a finite set of points $X = \{x_1, \ldots, x_n\}$ in $\mathbb{R}^d$ can

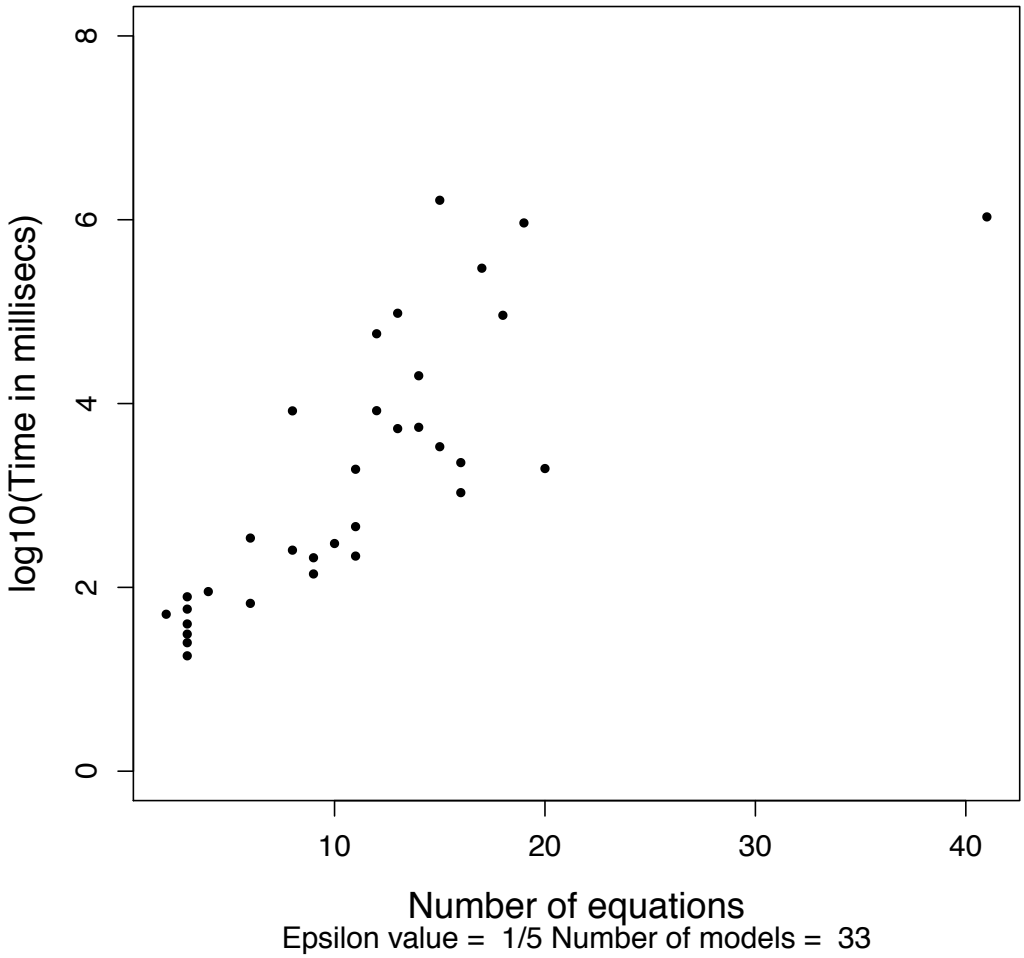


FIGURE 4.2: Plot of CPU running time against number of equations in the model for the Algorithm implementing linear programming approach (cf. Section 4.3.4).

be shown in the following manner

$$\text{aff}(S) = \{\sum_{i=1}^{i=n} \mid x_1, \ldots, x_n \in S, \ \sum_{i=1}^{i=n} \lambda_i = 1\} \tag{4.15}$$

This represents the smallest affine subspace of $\mathbb{R}^d$ containing S.

It is observed that the number of equivalence classes of maximal polytopes at each iteration step is sensitive to the ordering of input edge sets (cf. the Input in Algorithm 1) and thereby affects the CPU running times. To evaluate this further, three strategies for ordering of the list of edge sets are investigated. They are (i) Increasing, (ii) Decreasing, and (iii) Greedy orderings respectively.

In increasing and decreasing ordering the input list of edge sets consists of edge sets sorted in the increasing or decreasing order of their cardinalities. In greedy ordering, an edge set is dynamically selected at each iteration step, meaning that for a given solution set, that k-th edge set is selected which gives the minimal number of equivalence classes of maximal polytopes. It is seen that by following this approach, the intermediate growth in the number of equivalence classes is low compared to increasing and decreasing ordering criteria. This is shown in Fig. 4.7. This approach resembles greedy strategy in algorithmic design where the algorithm makes locally optimal choice but gives no guarantee that this leads to the global optimal choice. Therefore, at best this is a heuristic. Of course, the ordering of the edge sets do not affect the final solution i.e. the minimal branches. Table 4.2, 4.3, 4.4 presents the benchmark results on the biomodels for increasing, decreasing and greedy ordering strategies. The running times versus number of equations for same set of ordering is depicted in Fig. 4.3, 4.4, 4.5.

Although for applications in the thesis, computing of the minimal branches and the corresponding connected components are sufficient. Nevertheless, some additional investigations are done to get insights into the tropical equilibration solution structure as described below

Inclusion relations As described before, equivalence classes of tropical equilibrations are the branches and there can be inclusion relations among them (i.e. a polytope may be contained within another). This is investigated via an inclusion graph (directed graph) as shown in Fig. 4.8.

Connected components The connected components corresponds to the polyhedral complex of minimal branches. It is depicted as an undirected graph whose vertices are the minimal branches and there exists an edge if the intersection between the two vertices is non-void as shown in Fig. 4.9.

Dependency on the choice of ε In order to investigate the effect of different ε values on the number of minimal solutions, the value of ε is varied and the minimal branches are recomputed. A boxplot is presented in Fig. 4.10 describes this computation. A large number of models appear to be robust to the change in the ε values demonstrating the robustness of tropical approach in the context of biochemical models.

Visualizing tropical equilibrations One of the main applications of tropical equilibrations is in model reduction (cf. Chapter 6). The model reduction proceeds by timescale separation i.e. identifying certain species which are slow based on the rescaled orders which are computed from tropical equilibrations (cf. (6.9)). Here, heatmaps for two models are presented in Fig. 4.12 to show such rescaled orders.

Comparison with linear programming approach The linear programming approach finds at least one solution per minimal branch. However, the solutions may also belong to the intersections of the solution polytopes as explained above. Furthermore, the solutions may also belong to the non-minimal branches (i.e. solution polytopes which are included within the minimal branches). Therefore, the number of solutions obtained by linear programming approach may not be equal to the number of minimal branches. In order to test this aspect, the number of minimal branches is compared with the number of solutions obtained from the linear programming approach. Such a comparison is shown in Fig. 4.11.

Variable elimination using conservation laws As explained in Section 2.1, the variables in the ODE system can be eliminated using conservation laws. Theoretically, the minimal branches are a subset of tropical prevariety, which depends on the dominant terms of equation system and hence it is expected that these might be different after the elimination procedure whereas the tropical variety is not affected by such algebraic manipulations as the ideal of the equation system remains unchanged.

The number of variables that can be eliminated equals to the number of conservation laws of the ODE system. In practice, the conservation laws are computed by computing the left kernel of the stoichiometric matrix $\boldsymbol{S}$ with an additional constraint, enforcing non-negativity on each variable. Thus, giving rise to a systems of equalities and inequalities which can be represented as a polyhedron and whose vertices denote the conservation laws. Let us suppose that there are l conservation laws, for the elimination process, l distinct variables from the conservation laws are chosen randomly and the system of linear conservation laws is solved with respect to the chosen set of l variables. However, not every selection of l variables may give rise to a non-empty solution, therefore the process of randomly selecting l variables and solving the conservation laws with respect to them is repeated till a non-empty solution is found.

However, for the applications only the elements from tropical prevariety i.e. minimal branches or tropical equilibrations are used, therefore, it is important to compare the minimal branches before and after the elimination procedure in order to find a better strategy. The summary on the number of minimal branches and plot of the CPU running times are shown in Table 4.5 and Fig. 4.6 respectively. In practice, the variable elimination results in increased number of terms in the equation system which in turn leads to increased CPU running times and also the number of minimal branches sensitive to the selection of ε as compared to the same models without the elimination procedure.

4.3.6 Sample Point for Minimal Branches

For the applications considered here, sample points need to be computed from the minimal branches. In order to do so, the facets and affine hull of the polytopes corresponding to the minimal branches are computed resulting in a set of inequalities and equalities. From such a set of inequalities, a sample point $(a_1, \ldots, a_n)$ is computed using Satisfiability Modulo Theories (SMT) solver called Microsoft Z3 software (De Moura and Bjørner, 2008) in python programming environment. With Microsoft Z3, one can generate the sample point belonging exclusively to a minimal branch and not at the intersections of minimal branches (recall minimal branches may intersect at a common face, cf. Chapter 3). Therefore, the sample point should satisfy the inequalities and equalities corresponding to the facets and affine hull where the non-strict inequalities are replaced by their strict counterparts. For the purpose here, the benefit of using Z3 over any existing linear programming software is that it distinguishes strict and non-strict inequality conditions, which allows us to generate the sample point belonging exclusively to a minimal branch.

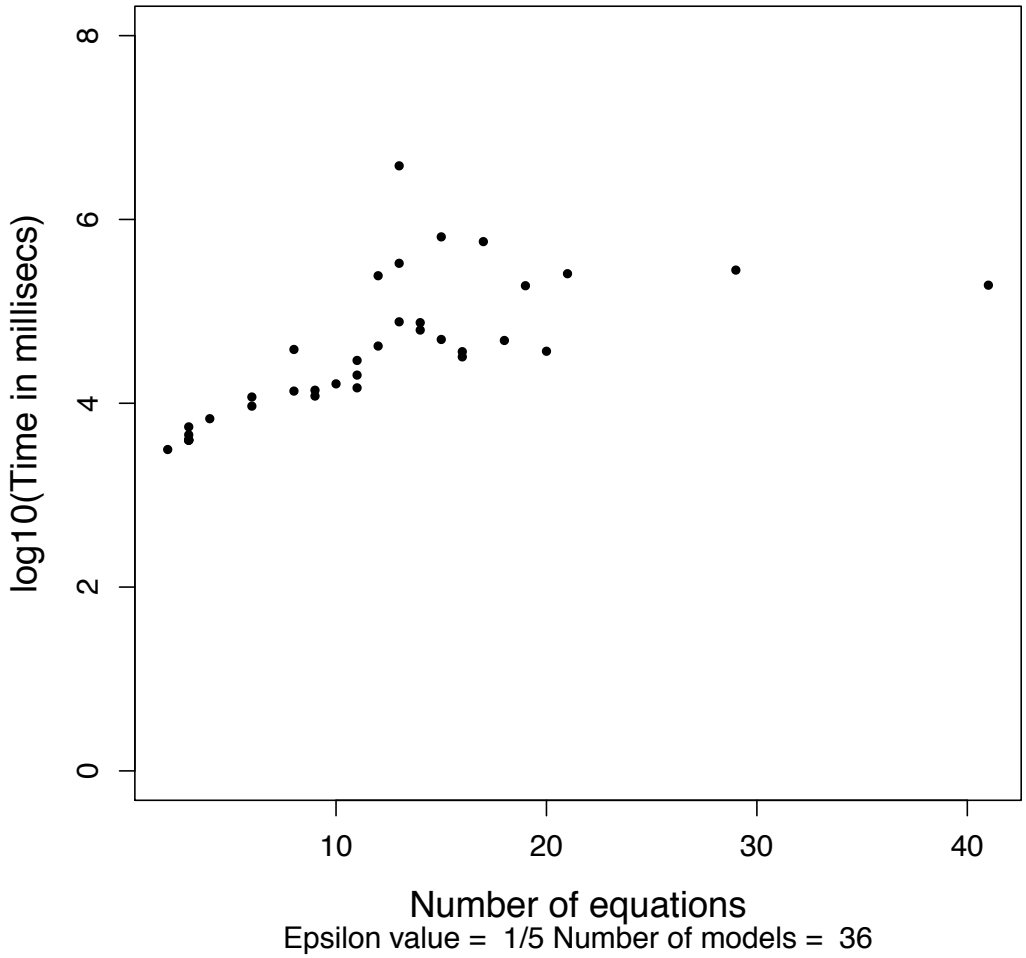


FIGURE 4.3: Plot of CPU running time against number of equations in the model for the Algorithm implementing convex polytope approach where the edge sets follow increasing ordering strategy (cf. Section 4.3.5).

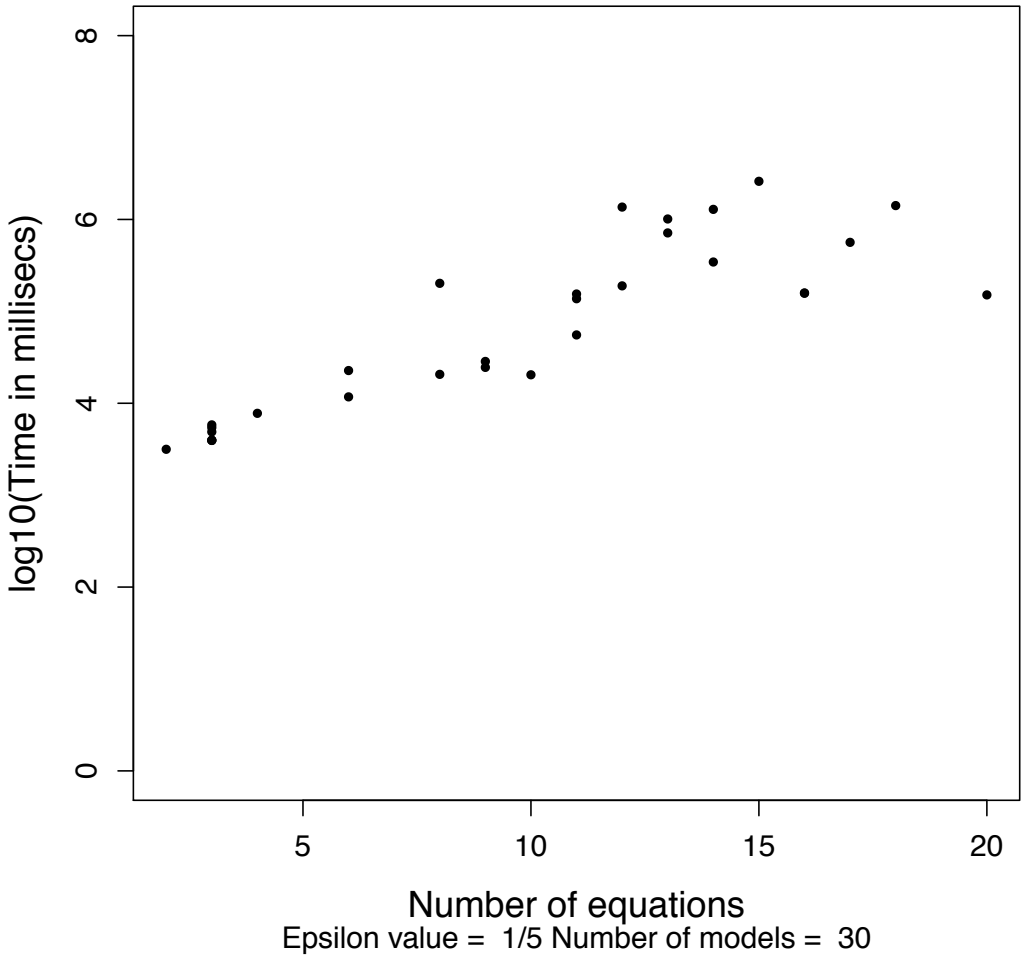


FIGURE 4.4: Plot of CPU running time against number of equations in the model for the Algorithm implementing convex polytope approach where the edge sets follow decreasing ordering strategy (cf. Section 4.3.5).

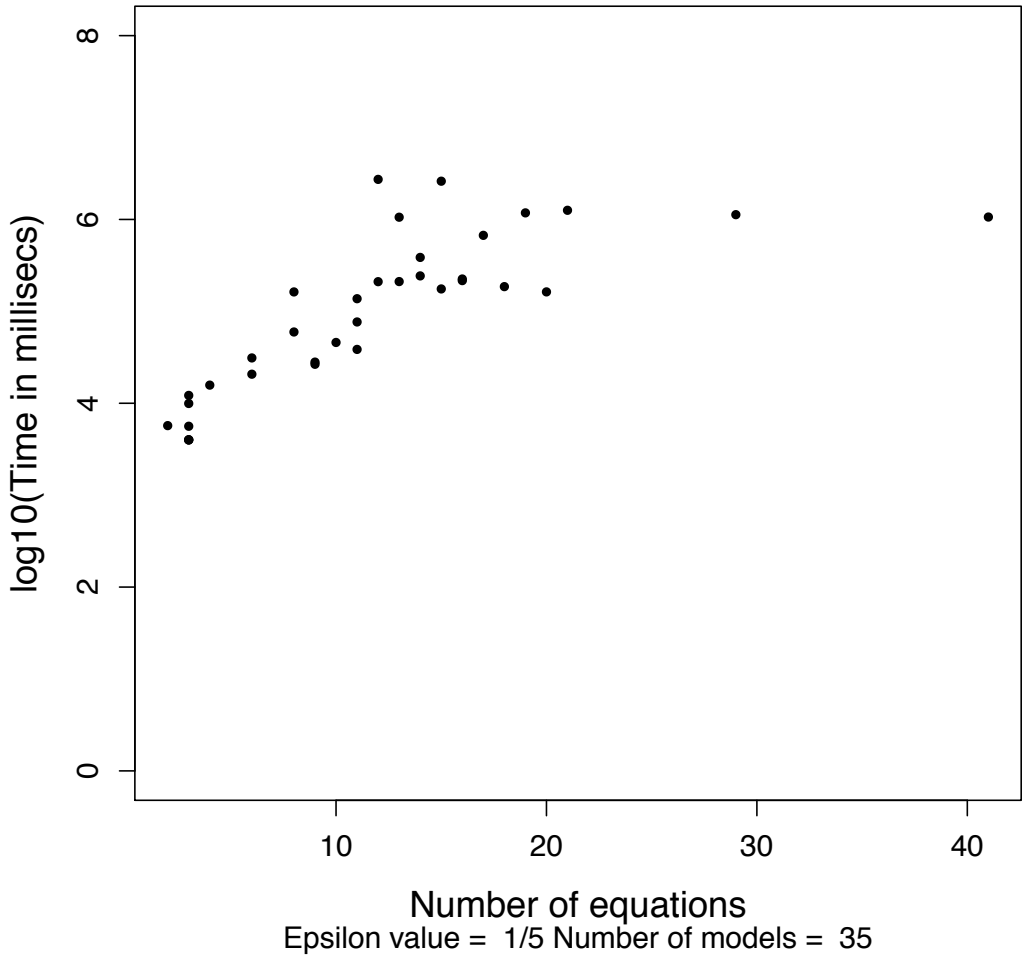


FIGURE 4.5: Plot of CPU running time against number of equations in the model for the Algorithm implementing convex polytope approach where the edge sets follow greedy ordering strategy (cf. Section 4.3.5).

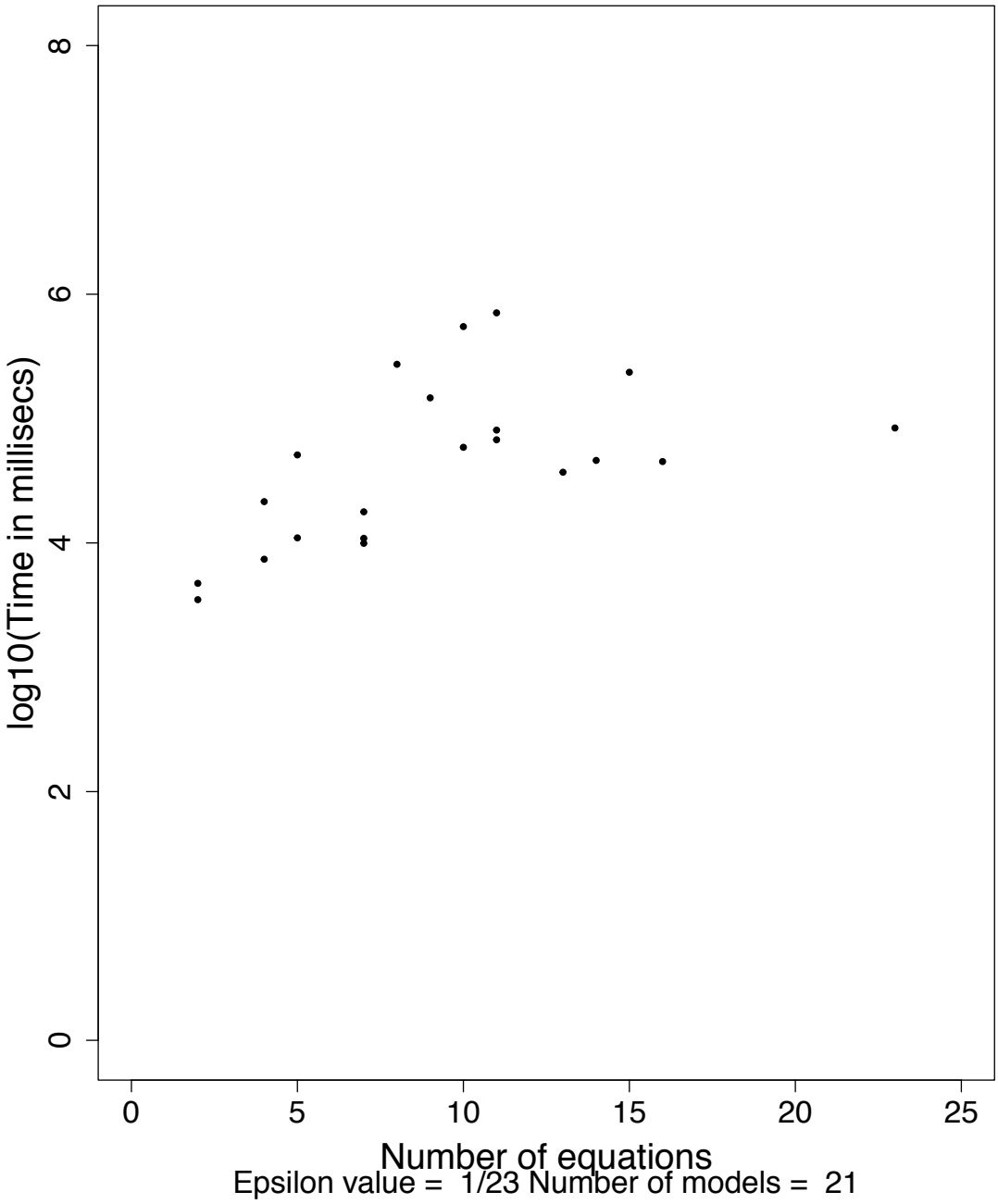


FIGURE 4.6: Plot of CPU running time against number of equations in the model for the Algorithm implementing convex polytope approach where the edge sets follow greedy ordering strategy (cf. Section 4.3.5). The models in the approach correspond to the ones whose variables are eliminated using the approach described in Section 4.3.5.

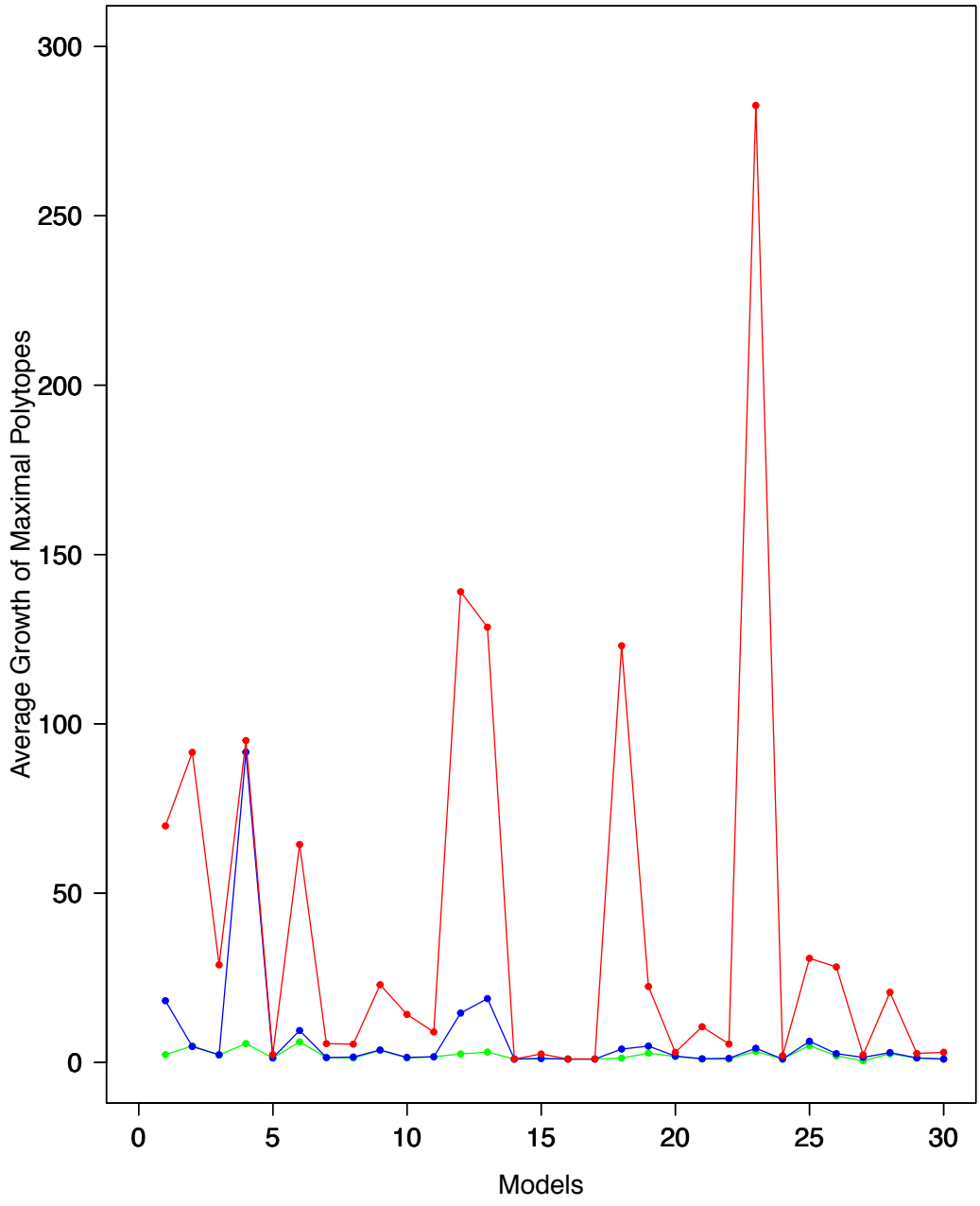


FIGURE 4.7: Plot showing the average of the number of maximal polytopes at each iteration for the Algorithm 2 implemented in three variants (cf. Section 4.3.5). Red denotes the decreasing, blue denotes increasing and green denotes greedy edge set ordering strategies respectively. The computations are done at $\varepsilon = 1/23$

Algorithm 2: SolveOrdersMB: Steps of tropical equilibration algorithm using the convex polytope approach.

Input: List of edge sets $ne_1, ne_2, \ldots, ne_n$ (cf. Fig. 4.1), and the corresponding vertices of Newton polytope

Output: Set of minimal branches

```
1  begin
2      solutionset ={};
3      for k = 1 to n do
4          for l = 1 to number of entries in ne_k do
5              equation* = vertices in l^th row
6              inequalities* = all vertices other than l^th row
7              if LinearSolve(equation,inequalities)** is feasible then
8                  Augment the convex polytope to the solutionset***
```

9 *The equations and inequalities are initialised as per (4.14) and are solved using the linear programming software.

10 **Solves the system of equations and inequalities using the linear programming software.

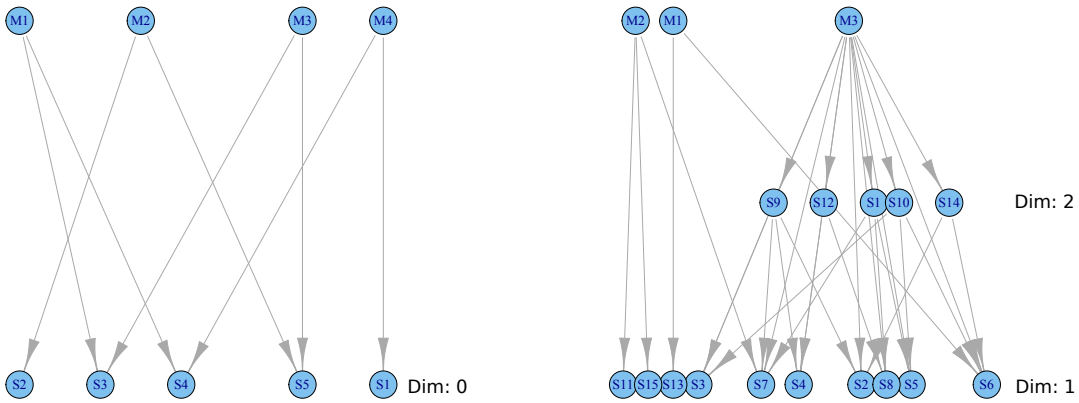
11 ***This means that the H-polytope is augmented to the existing ones (also H-Polytopes) in the solutionset at each iteration step. By augmenting it is meant that the system of inequalities and equations corresponding to the H-polytope are added to the existing ones.

TABLE 4.2: Summary of analysis on Biomodels database based on the implementation of Algorithm 2 with increasing ordering strategy. The benchmarked models have a number of dimensions (i.e. number of variables along with number of conservation laws) ranging from 2 to 41. Model BIOMD0000000289 has tropical branches at ε value 1/5, 1/7, 1/9, 1/11 but none at 1/17, 1/19, 1/23. Similarly, Model BIOMD0000000080 has no solution only at ε value 1/19.

ε value	Total models considered	Models without tropical equilibrations	Models with tropical equilibrations	Timed-out models	Average running time (in secs)	Average number of tropical equilibrations	Min number of tropical equilibrations	Max number of tropical equilibrations
1/5	36	0	36	0	200.76	15.08	1	423
1/7	36	0	36	0	177.01	14.41	1	406
1/9	36	0	36	0	195.01	13.02	1	340
1/11	36	0	36	0	169.59	12.36	1	322
1/17	36	1	35	0	175.56	11.02	0	287
1/19	36	2	34	0	187.42	11.08	0	287
1/23	36	1	35	0	184.80	11.08	0	287

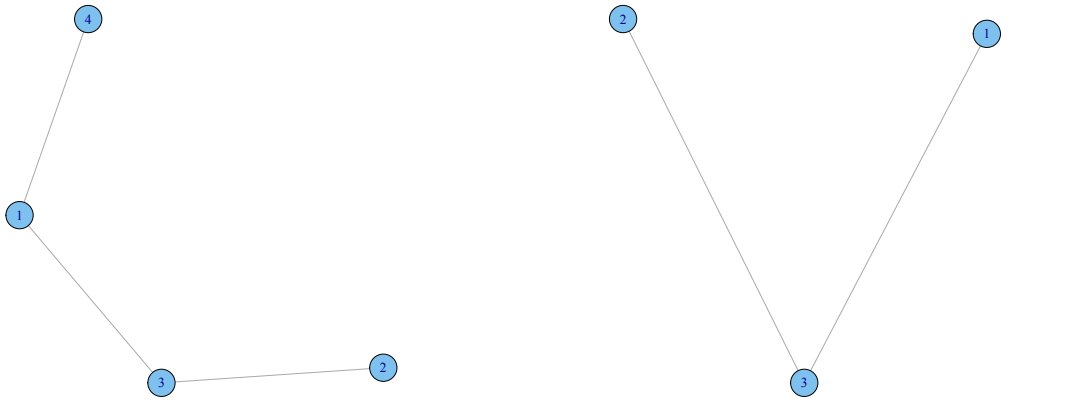
4.4 Estimating Liftable Tropical Equilibrations

The steps to check whether an element from tropical prevariety can be lifted to the tropical variety is explained in Section 3.1.3 in Chapter 3. These steps are implemented in the Algorithm 3 to test whether tropical minimal branches can be lifted to tropical variety. The algorithm is a



BIOMD0000000035 BIOMD0000000072

FIGURE 4.8: A directed graph in layered form showing the inclusion relations among the different solution branches (for $\varepsilon = 1/11$) for two models from Biomodels namely BIOMD00000000-35,72. Vertices in the graph comprise of polytopes corresponding to solution branches and an directed edge between i and j means j is included in i. The topmost layer contain the minimal solution branches, thereafter the bottom layers are "included" solution branches. The layers of the included solution branches are based on the dimension of the corresponding polytopes (arranged in descending order). Therefore, included solutions in one layer are of same dimensional polytope



BIOMD0000000035 BIOMD0000000072

FIGURE 4.9: Graph of connected components (at $\varepsilon = 1/11$) for four models namely BIOMD00000000-35,72. All of them have one connected component. The vertices are minimal solution branch and there exists an edge if the intersection between the two vertices is non-void

variant of Algorithm 2 computing the minimal branches with greedy ordering strategy. Similar to it, the solution set stores the equivalence classes of only the maximal polytopes at each k-th

TABLE 4.3: Summary of analysis on Biomodels database based on the implementation of Algorithm 2 with decreasing ordering strategy. The benchmarked models have a number of dimensions (i.e. number of variables along with number of conservation laws) ranging from 2 to 41. Model BIOMD0000000289 has tropical branches at ε value 1/5, 1/7, 1/9, 1/11 but none at 1/17, 1/19, 1/23. Similarly, Model BIOMD0000000080 has no solution only at ε value 1/19.

ε value	Total models considered	Models without tropical equilibrations	Models with tropical equilibrations	Timed-out models	Average running time (in secs)	Average number of tropical equilibrations	Min number of tropical equilibrations	Max number of tropical equilibrations
1/5	36	0	30	6	355.83	2.76	1	19
1/7	36	0	30	6	377.53	2.8	1	19
1/9	36	0	30	6	383.49	3.03	1	14
1/11	36	0	30	6	399.04	2.8	1	14
1/17	36	1	29	6	354.11	2.66	0	15
1/19	36	2	28	6	361.87	2.6	0	15
1/23	36	1	29	6	349.23	2.7	0	15

TABLE 4.4: Summary of analysis on Biomodels database based on the implementation of Algorithm 2 with greedy ordering strategy. The benchmarked models have a number of dimensions (i.e. number of variables along with number of conservation laws) ranging from 2 to 41. Model BIOMD0000000289 has tropical branches at ε value 1/5, 1/7, 1/9, 1/11 but none at 1/17, 1/19, 1/23. Similarly, Model BIOMD0000000080 has no solution only at ε value 1/19.

ε value	Total models considered	Models without tropical equilibrations	Models with tropical equilibrations	Timed-out models	Average running time (in secs)	Average number of tropical equilibrations	Min number of tropical equilibrations	Max number of tropical equilibrations
1/5	36	0	35	1	411.36	3.42	1	19
1/7	36	0	35	1	308.01	3.22	1	19
1/9	36	0	35	1	360.67	3.65	1	17
1/11	36	0	35	1	326.92	3.51	1	17
1/17	36	1	34	1	307.14	3.14	0	18
1/19	36	2	33	1	318.98	3.2	0	18
1/23	36	1	34	1	411.36	3.2	0	15

iteration step. The input is a list of edge sets $ne_1, ne_2, \ldots, ne_{2n}$ corresponding to the edges of Newton polytopes of the random polynomials from the ideal. For keeping the CPU running times practical, the number of terms in the random polynomials is fixed at 1. Although, this is an oversimplification but gives some insights of this heuristic and motivation of future work in this direction. The output is a list of sets where each set contains the maximal polytopes resulting from the addition of random elements from the ideal to a given tropical minimal branch of the model. The set is either empty meaning that after adding elements from ideal the corresponding minimal branch becomes infeasible suggesting that the minimal branch does not lift to the tropical variety. However, if the set is not empty, there is a certain probability that it may be liftable to the tropical variety but there is no guarantee. The test is implemented as explained in Algorithm 3 which returns a list of sets for a given model where each set contains the maximal polytopes resulting from the addition of random elements from the ideal to a given

TABLE 4.5: Summary of analysis on Biomodels database based on the implementation of Algorithm 2 with greedy ordering strategy on Biomodels after variable elimination using conservation laws (cf. Section 4.3.5). For ε values 1/11,1/17,1/19, model BIOMD0000000077 has no opposite sign monomials for at least one equation. This may happen due to Newton polytope construction at different values of ε.

ε value	Total models considered	Models without tropical equilibrations	Models with tropical equilibrations	Timed-out models	Average running time (in secs)	Average number of tropical equilibrations	Min number of tropical equilibrations	Max number of tropical equilibrations
1/5	23	1	20	2	225.95	15.28	0	117
1/7	23	1	20	2	209.02	17.80	0	117
1/9	23	1	20	2	216.17	15.33	0	117
1/11	23	2	19	2	204.28	16.14	0	117
1/17	23	2	19	2	109.24	9.04	0	40
1/19	23	2	19	2	111.05	8.9	0	40
1/23	23	0	21	2	117.68	9.71	1	50

tropical minimal branch of the model. A boxplot is presented in Fig. 4.14, which shows the cardinalities of the sets belonging to such a list for a given model. It is to be pointed out that the implementation of this test on the minimal branches obtained from the Biomodels did not resulted in empty sets for any given model. The CPU running times of the implementation is shown in Fig. 4.15. The results remain inconclusive for Biomodels and more in depth analysis of parameters of the program (e.g. number of random polynomial, terms in each random polynomial, etc) is a topic of future work. Some preliminary work in selecting elements from ideal was done in (Radulescu, Vakulenko, and Grigoriev, 2015) by computing the fast cycles (which are not random but have a biochemical interpretation).

It has been conjectured in (Radulescu, Vakulenko, and Grigoriev, 2015) that the minimal branches which are liftable to tropical variety are better candidates for performing model order reduction. More precisely, prevarieties can provide reductions for transient, metastable, but relatively faster modes (see Chapter 7), whereas tropical variety is more likely to provide reductions for the last, slowest modes.

4.5 Discussions

In this chapter, the algorithmic approaches are discussed to compute the tropical equilibrations and organise them into *branches* and *minimal branches*. Two different implementations are explained for computing the equilibrations namely the linear programming and the convex polytope approach. The linear programming approach, is comparatively fast to compute the first tropical solution but takes considerable amount of CPU time to compute all solutions as seen in the timing out of 3 models (i.e. models taking more than $10,000$ secs of computation time). Also as described, this approach does not takes into account the polytope structure of the solutions explicitly and hence may have sample points belonging to the common face of the minimal branches representing their intersection. On the other hand, the convex polytope approach explicitly computes the H-representation of the solution polytopes for determining the equivalence classes and minimal branches. Although, this seems to be computationally more expensive but in practice it is seen that for a majority of the models the computations were

Algorithm 3: EstimateTV: Algorithm for estimating tropical variety based on the heuristic approach outlined in Sub-section 3.1.3.

Input: Tropical minimal branches $m_1, m_2, \ldots, m_p$, List of edge sets $ne_1, ne_2, \ldots, ne_{2n}$, and the corresponding vertices of Newton polytope corresponding to random polynomials from the ideal

Output: List of sets of length p

begin

    varietylist =

(4.16)

    for $r = 1$ *to* p **do**

        solutionset ={};

        for $k = 1$ *to* $2n$ **do**

            for $l = 1$ *to number of entries in* ne_k **do**

                equation* = vertices in l^{th} row

                inequalities* = all vertices other than l^{th} row augmented along with H-polytope corresponding minimal branch m_r

                if *LinearSolve(equation,inequalities)** is feasible* **then**

                    Augment the convex polytope to the solutionset***

        varietylist(r) = solutionset

*The equations and inequalities are initialised as per (4.14)

**Solves the system of equations and inequalities using the linear programming software.

***This means that the H-polytope is augmented to the existing ones (also H-Polytopes) in the solutionset at each iteration step. By augmenting it is meant that the system of inequalities and equations corresponding to the H-polytope are added to the existing ones.

quite efficient. We believe that this is probably due to low number of minimal branches which in turn results into efficient prunning of the search space. For example, the number of minimal branches at $\varepsilon = 1/5$ ranged from 1 to 423 whereas the number of tropical equilibrations by Algorithm 1 ranged from 1 to 35 (the model with 423 minimal branches is timed-out here). The high number of minimal branches i.e. 423 is displayed only by a single model for others it is less 18. Importantly, the ordering of equations of the input ODE affected the running times. It is seen that the increasing order strategy computed all the models within the $10,000$ secs of computation time whereas the greedy strategy gave the least number of intermediate maximal polytopes. We believe that for larger models the greedy ordering will be a better strategy if it is parallelized over multiple cores, thereby speeding up the computation times. This requires re-implementation of certain parts of algorithm and will be explored as a topic of future research. It is also important to note that the biochemical reaction networks exhinit special network structure e.g. scale-free property (Albert, 2005), low tree width (Nabli et al., 2016) and developing heuristics to improve the running times using such special properties remains a topic of future research.

In order to investigate the dependency of minimal branches on the choice of ε, the computations are done with several values of ε. For several models the number of minimal branches is robust whereas for some this number varies. As the computation times for convex polytope are not very sensitive to the selection of ε compared to the linear programming approach, therefore going for lesser values is not a serious bottleneck of the approach.

We also described the structure of tropical equilibrations by computing the connected components of the connectivity graph derived from the minimal branches. We believe that such connectivity among minimal branches influences the dynamics of the system i.e. providing an estimate for the possible dynamical transitions between them. We investigate it in the context of symbolic dynamics as described in Chapter 6. Furthermore, the directed graphs are presented showing the inclusion relations between different branches of tropical equilibration solutions. These computations reveal the rich structure of the polytopes associated with tropical equilibrations. The tropical equilibrations determine the concentration orders of the chemical species of the ODE model. It is shown in Chapter 6 that this information will be used to identify timescales of the system leading to model reduction. The distribution of such timescales are depicted through heatmaps. The heatmaps show that different minimal branches may result into different timescale ordering of the chemical species. Therefore, our method provides several possible reductions corresponding to the different minimal branches. Thus, the overall solution structure provides insights into the dynamics of the system.

As the dominant terms in the polynomial system are the same for all the tropical solution on branches, it could be that the branches correspond to the invariant manifolds. This idea will be pursued in future work.

A heuristic was also implemented to check if a minimal branch be lifted to the tropical variety (in other words, liftable to Newton-Puiseux series solution of system). The results from this are not very convincing as for all models none of the minimal branch are discarded after the addition of random elements of the ideal. Fixing the random polynomials to be of a single term may be one of the reasons. Nevertheless, it was seen that the number of maximal polytopes in the tropical variety increases considerably by this procedure. A detail evaluation and efficient sampling of random polynomials or taking direct combinations of the polynomials will be an interesting future direction. There exists software e.g. Gfan (Jensen, 2006) to compute the full tropical variety. It will be our future endeavour to compare our heuristic approach with such software for smaller dimensional examples. However, for larger dimensional models, computing the tropical variety will be a challenge due to the high computational complexity involved.

Lastly, there exists results of the direct application of tropical geometry to differential equation systems (Grigor'ev and Singer, 1991; Bruno, 2000) which are not explored in this thesis.

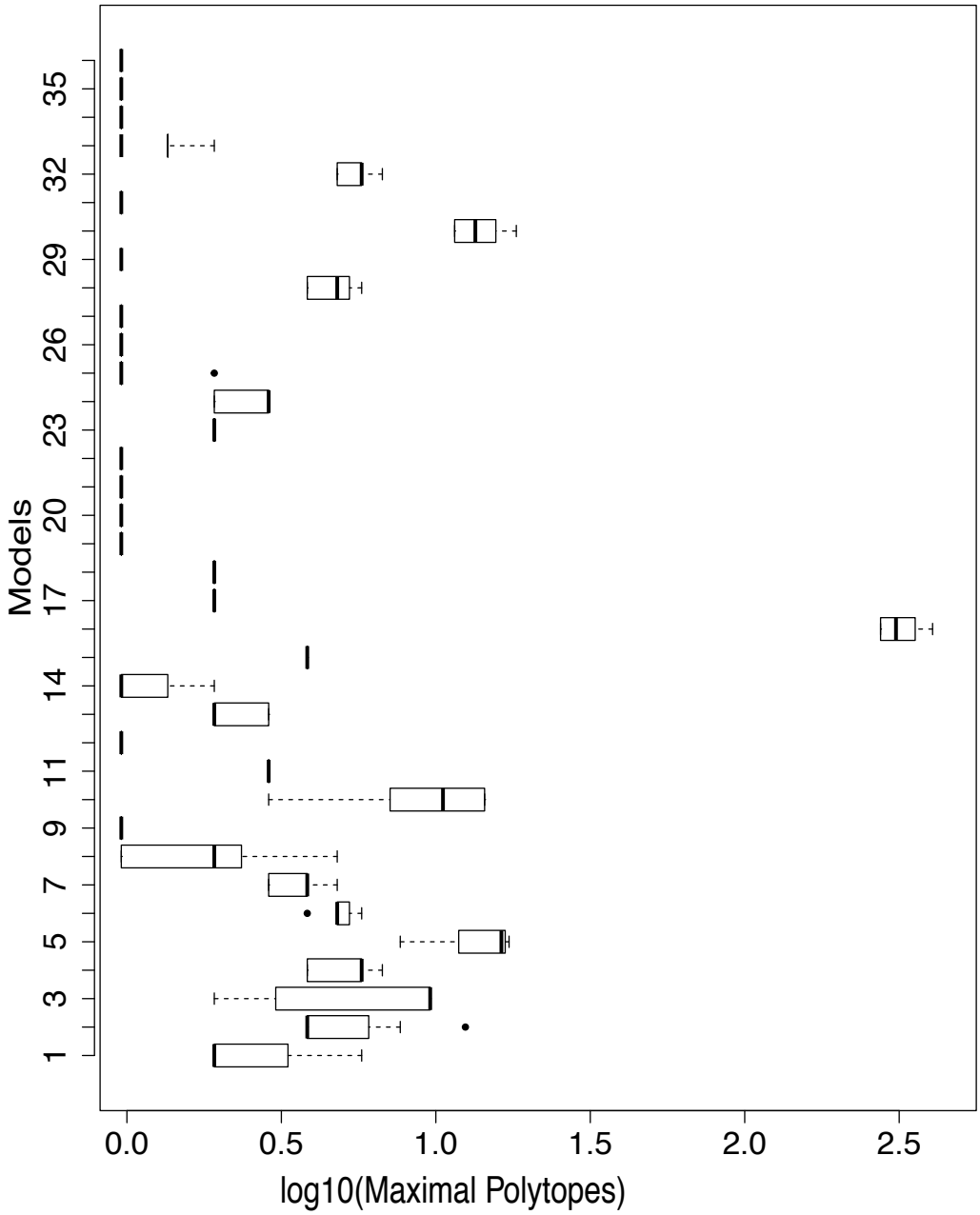


FIGURE 4.10: Tropical minimal branches are computed at different ε values (recall, the tropical minimal branches are maximal polytopes). The boxplot represents the distribution of number of minimal branches for a given biomodel at different values of ε values, namely 1/5,1/7,1/9,1/23.

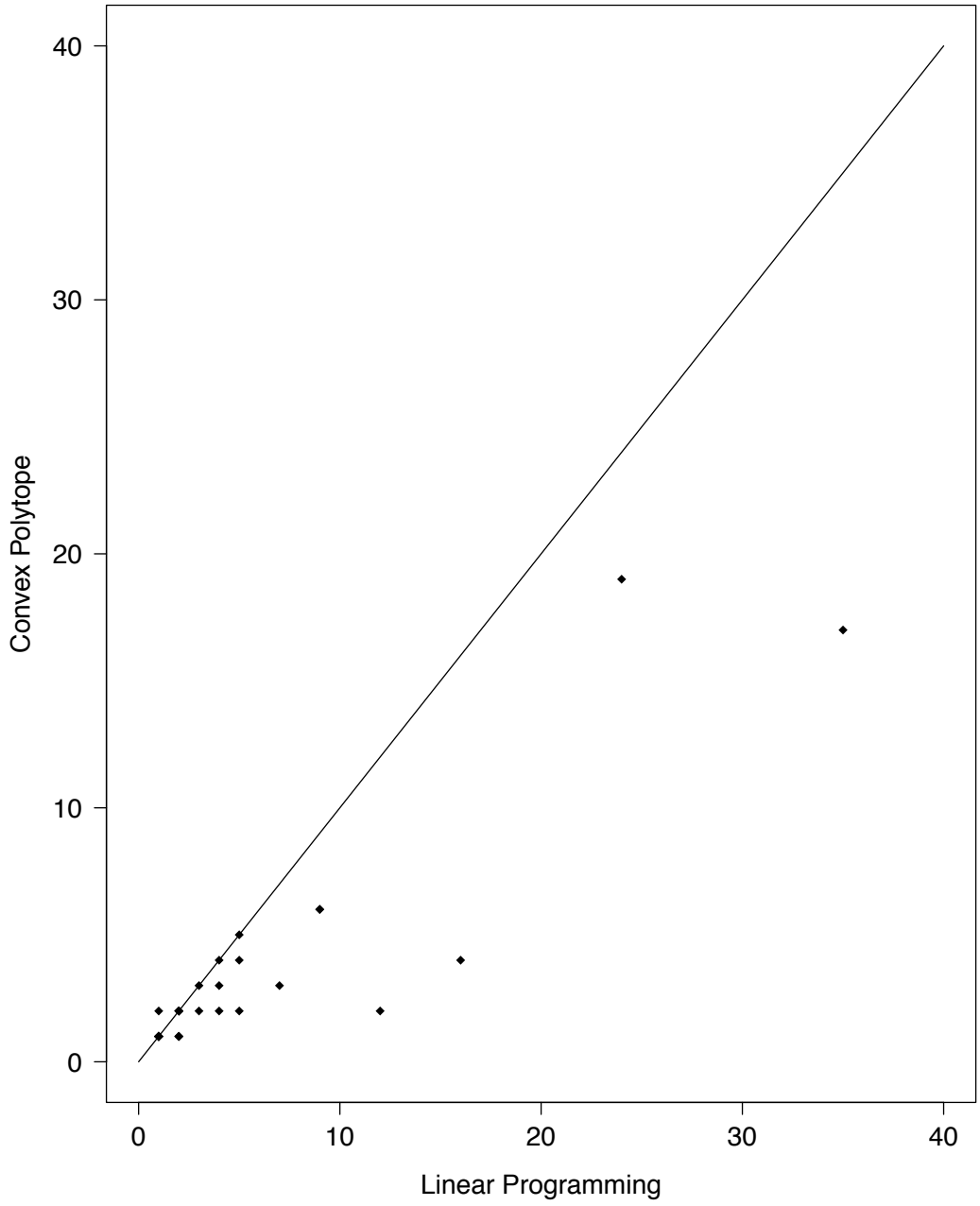
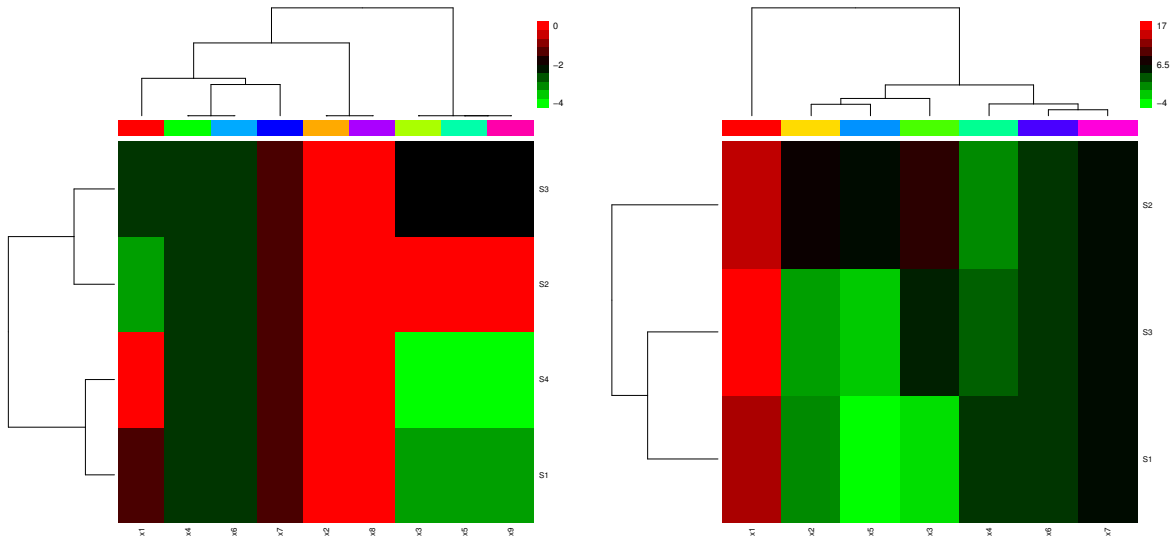


FIGURE 4.11: The tropical equilibrations obtained from linear programming approach (colored in red) is compared with the number of of minimal branch obtained using convex polytope approach (colored in blue). ε values: 1/5



BIOMD0000000035 BIOMD0000000072

FIGURE 4.12: A sample point (representing the concentration orders cf. (4.3)) is picked from the each tropical minimal branch corresponding to two models from Biomodels database namely BIOMD00000000-35,72. The heatmaps represent the rescaled orders computed from the the sample points at $\varepsilon = 1/11$ (cf. (6.9)). The rescaled order captures the timescale of the variable. The horizontal axis represents the variables and the vertical axis represents the tropical minimal branches. Additionally, hierarchical clustering is performed on the variables and also the tropical minimal branches.

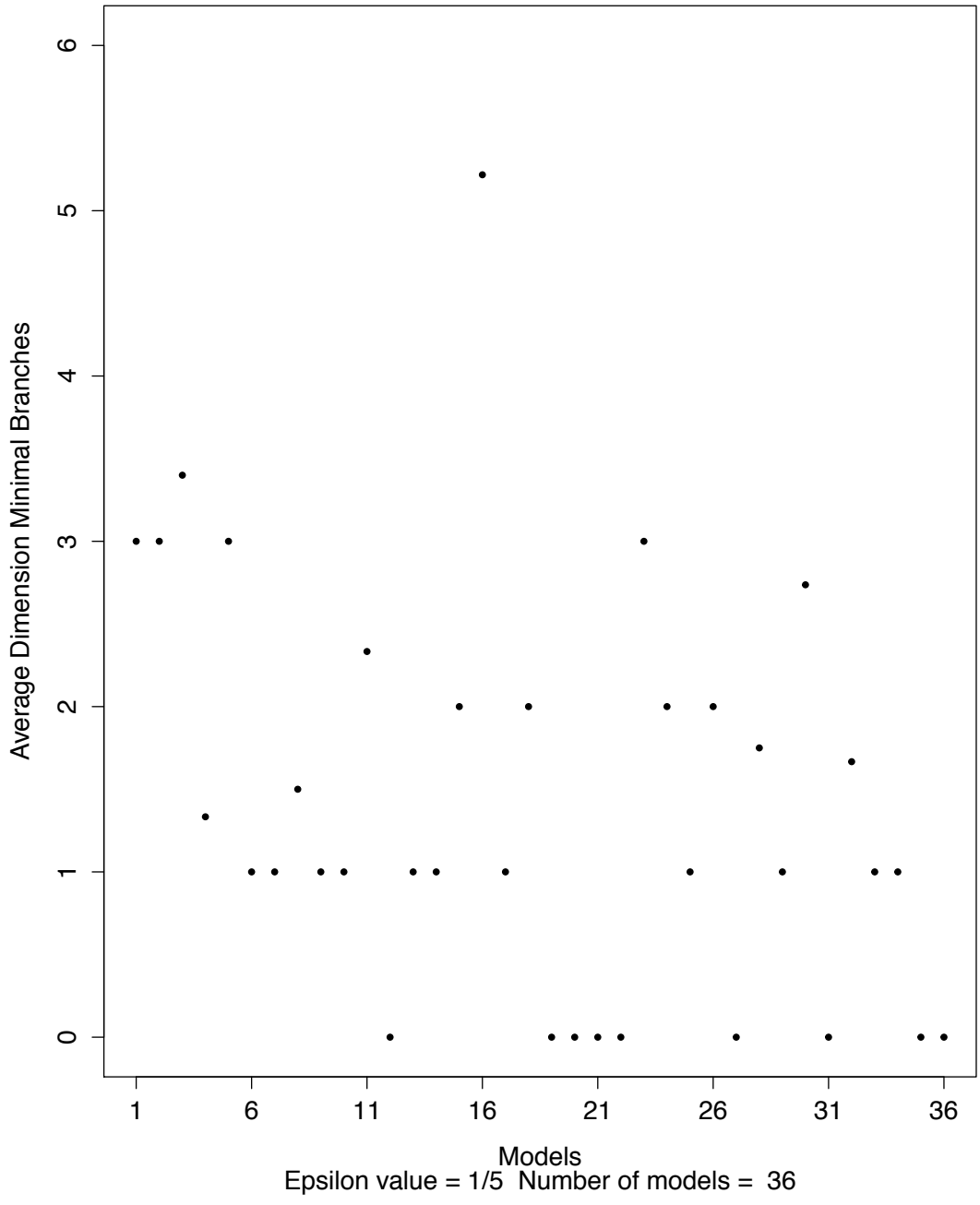


FIGURE 4.13: A plot showing the average dimension of tropical minimal branches for the 36 models for $\varepsilon = 1/5$ analysed in Table 4.2. 9 models have average dimension exactly zero suggesting that these polytopes are points. The dimension here refers to the dimension of the affine hull of the polytopes (cf. Section 4.3.5).

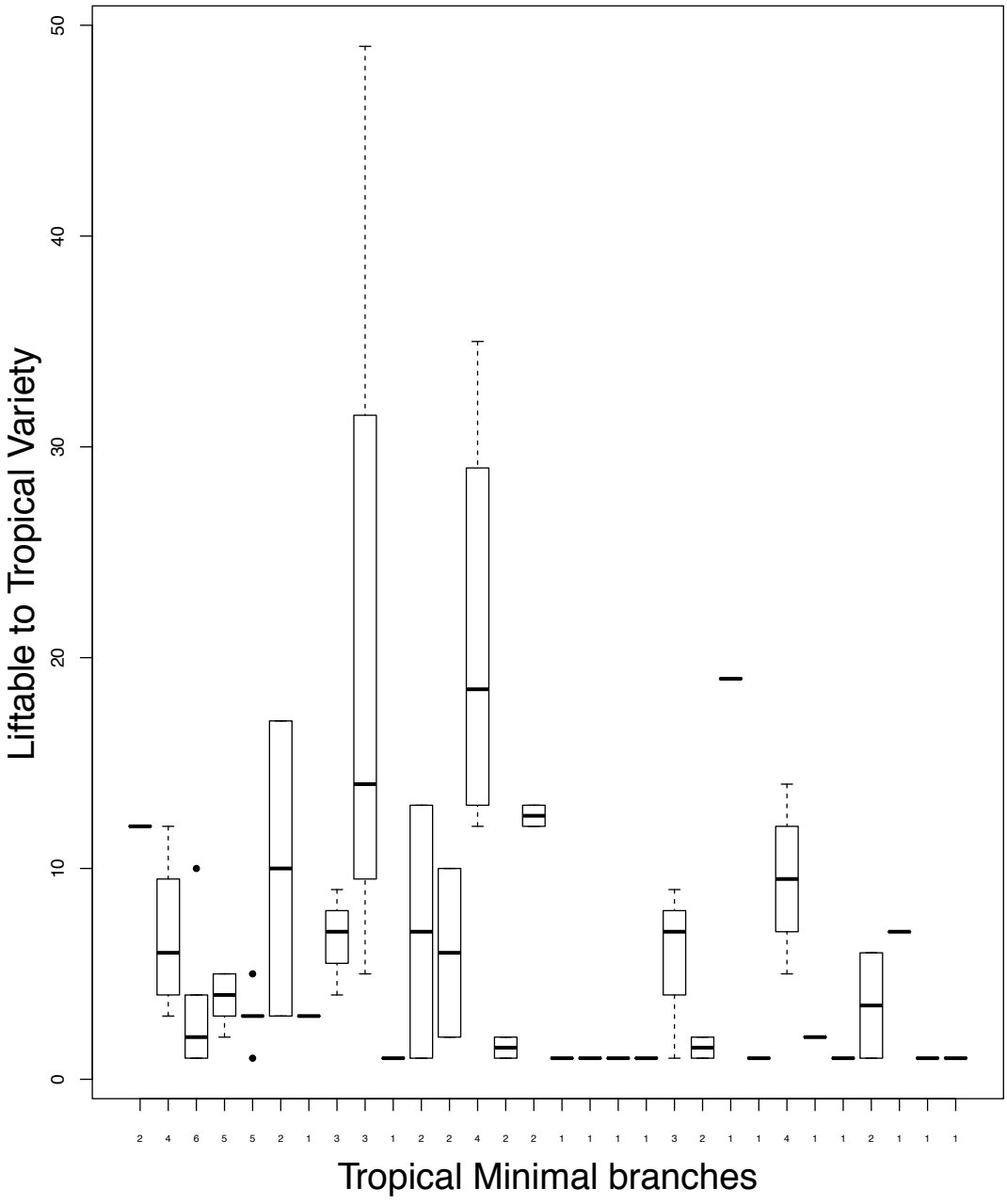


FIGURE 4.14: The boxplot shows the cardinalities of the sets belonging to such a list which is the output of Algorithm 3. Each set in such a list of sets for a given model contains the maximal polytopes resulting from the addition of random elements from the ideal to a given tropical minimal branch of the model. The computation are done at ε values: 1/5.

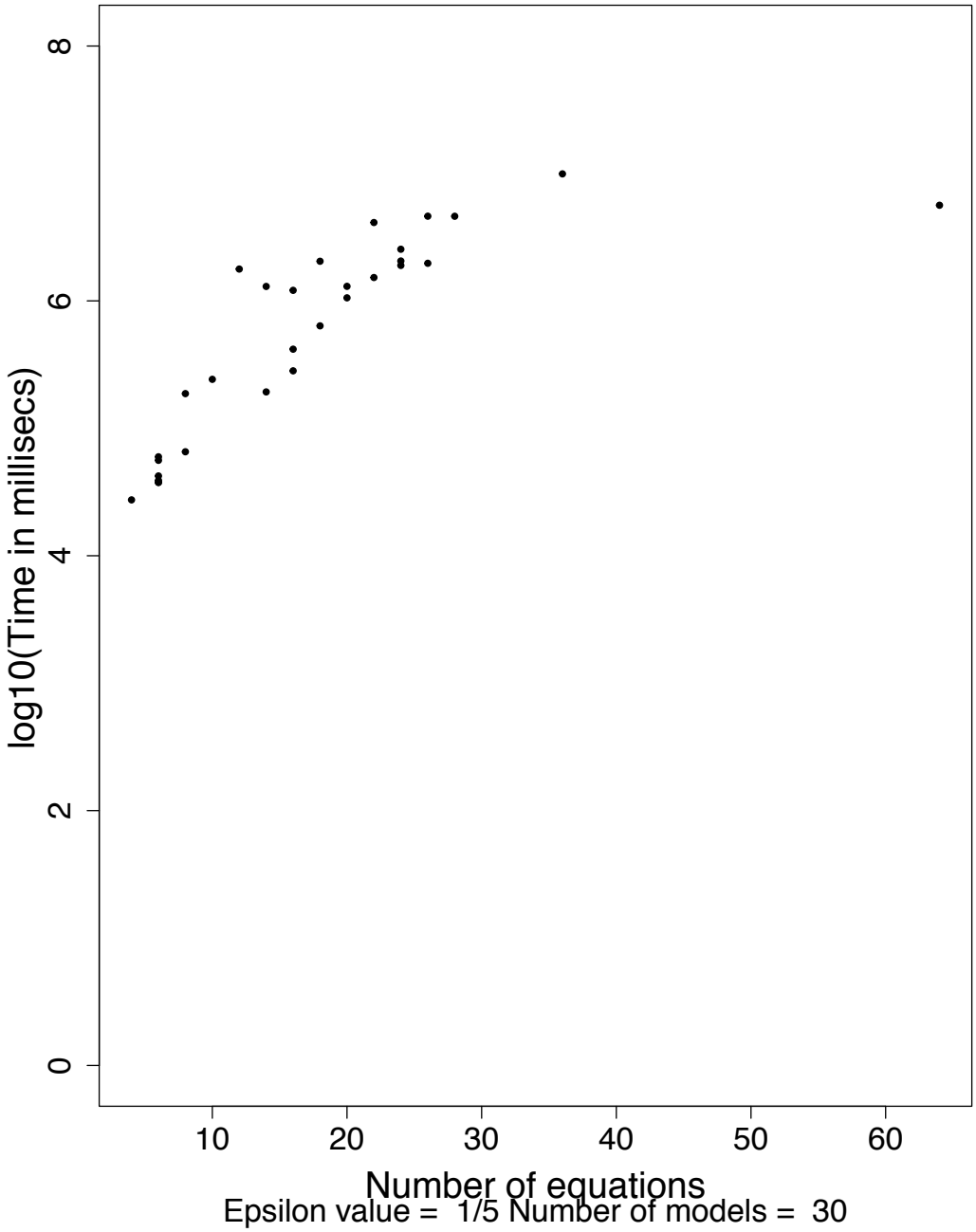


FIGURE 4.15: Plot of CPU running time against number of equations in the model for the Algorithm implementing tropical variety approach (cf. Algorithm 3) where the edge sets follow increasing ordering strategy (cf. Section 4.3.5).

Chapter 5

Computation of Extreme Currents

The chapter discusses an algorithm to compute the ECs. In addition, a benchmark is provided on the metabolic models obtained from KEGG database.

5.1 Algorithm

A brief background on extreme currents (ECs) is described in Section 3.2. Here, the computational method to compute the ECs is described. Recall, from Section 2.2 that biochemical networks can be also represented through the reaction fluxes (cf. (2.3)) with the constraint that the reaction fluxes are non-negative. Such a system of equations and inequalities results in a convex polytope in H-representation, also called as H-polytope (cf. Sub-section 3.1.3). Alternatively, a polytope can also be represented in V-representation or called as V-polytope, meaning that the convex hull of a finite set $X = \{x_1, ..., x_n\}$ of points in $\mathbb{R}^d$ (Henk, Richter-Gebert, and Ziegler, 2004) can be represented as

$$P = \text{conv}(X) = \{\sum_{i=1}^{i=n} \lambda_i x_i \mid \lambda_i \geq 0, \ \sum_{i=1}^{i=n} \lambda_i = 1\} \tag{5.1}$$

Now, let us state the main theorem of polytope theory from (Henk, Richter-Gebert, and Ziegler, 2004)

Theorem. *The definitions of* V*-polytopes and of* H*-polytopes are equivalent. That is, every* V*-polytope has a description by a finite system of inequalities, and every* H*-polytope can be obtained as the convex hull of a finite set of points (its vertices).*

Here, we are interested in convex hull computation which determines the vertices of the V-polytope from the given H-polytope. One of the ways to perform such a computation is the double description algorithm (Fukuda and Prodon, 1996; Terzer, 2009). In our case the vertices are computed from the H-polytope represented in (2.3) and are referred to as ECs. Unfortunately, the worst case time complexity of the algorithms computing the ECs can be exponential in the worst case (Terzer, 2009). The ECs are computed using the PoCaB software (Samal, Errami, and Weber, 2012).

5.2 Benchmarking

We chose 69 metabolic models from KEGG and compute the ECs to determine the CPU running times. A plot is provided in Fig. 5.1 describing the CPU running times versus the number of reactions of the models in the context of EC computation. Additionally, the distribution of number of ECs is shown in Fig. 5.2 as a histogram. The ECs are computed by considering only the main reaction pairs (a pre-processing step described in Section 8.4). For every split of a reversible reaction there appears an EC, denoting only the forward and backward reaction

components of the reversible reaction (Wagner and Urbanczik, 2005). Such cycles are often referred to as spurious cycles (Wagner and Urbanczik, 2005) and are filtered out in our approach. However, the spurious cycles may represent futile cycles which are involved in regulatory processes of the pathways (Qian and Beard, 2006) and to determine the role of such futile cycles in the context of our method remains a topic of future research.

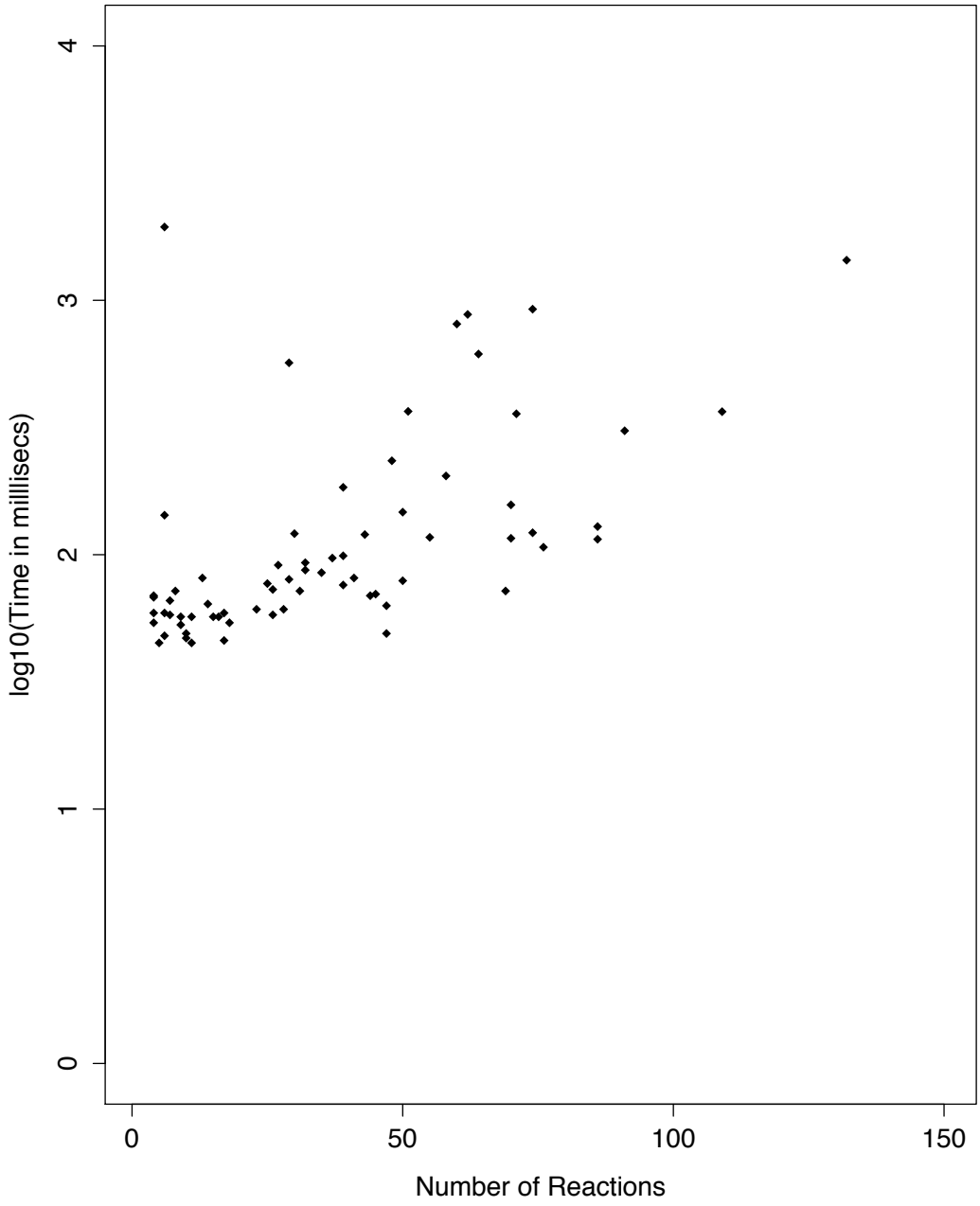


FIGURE 5.1: Plot of CPU running time against the number of reactions in the model for the algorithm implementing the extreme current (cf. Chapter 5). 69 metabolic models from KEGG are chosen for computing the CPU running times.

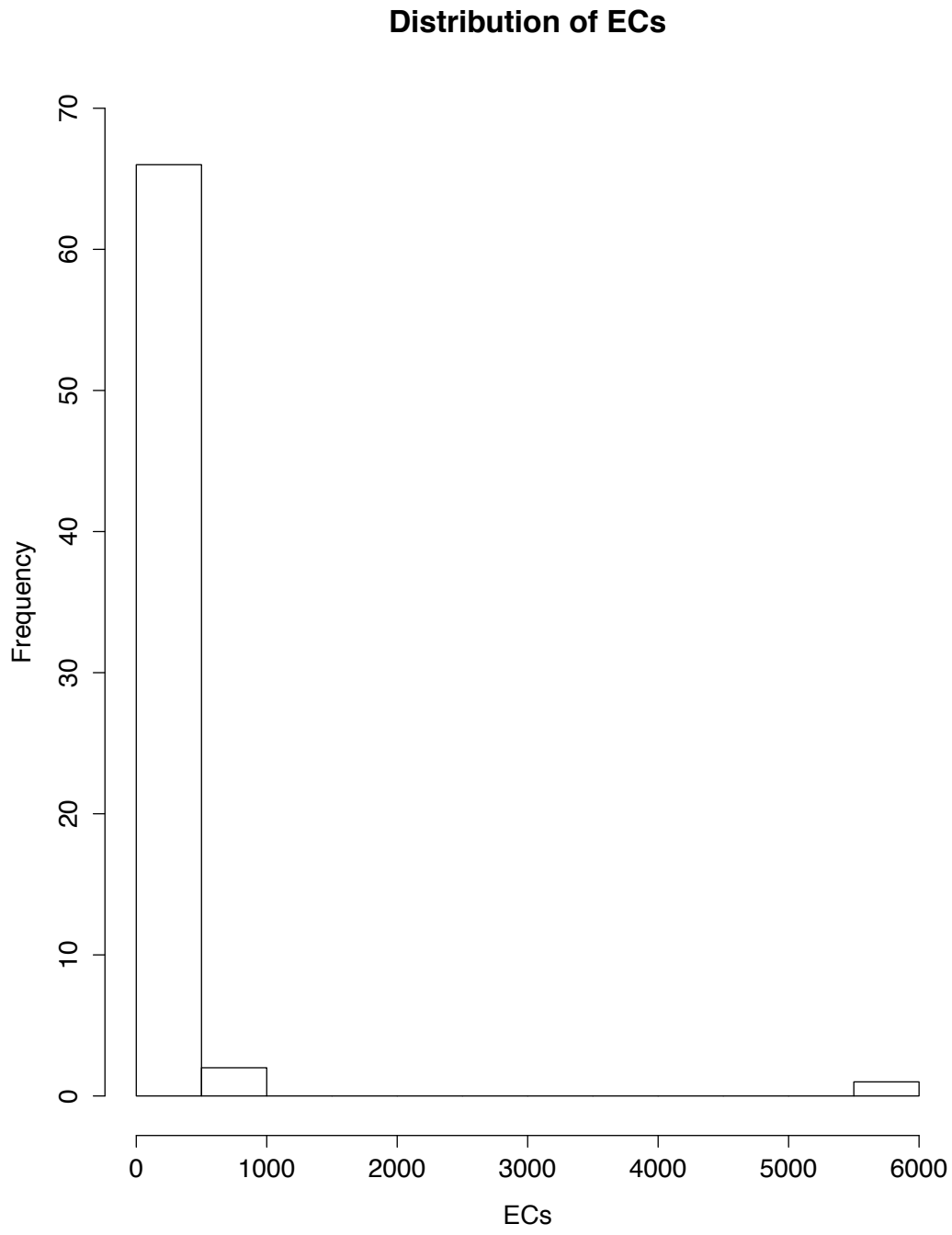


FIGURE 5.2: Plot of histogram showing the distribution of the number of ECs across 69 metabolic models chosen from KEGG.

Chapter 6

Model Reduction based on Tropical Equilibrations

The main motivation of computing tropical equilibrations or tropical minimal branches is to provide an efficient and scalable algorithmic technique to perform timescale decomposition of biochemical species. Its application in model reduction is discussed in this chapter. In model reduction the aim is to derive a simpler representation of a given biochemical reaction network.

6.1 Background

Model reduction is an important problem in computational biology. The main aim of model reduction is to represent the essential dynamics of the model with fewer species and reactions (Okino, and Mavrovouniotis, 1998). This is helpful in the following ways

1. Addressing the numerical issues associated with the simulation of stiff systems resulting from multiple timescales.
2. Reducing the number of parameters, which are often difficult to measure in a typical laboratory experimental setting.
3. Getting insights into the behaviour of the model by studying the key species and reactions of the model.

There are several methods for reducing networks of biochemical reactions. Broadly the model reduction techniques can be categorised into three categories, as explained below

1. The formal model reduction techniques can be based on conservation laws, exact lumping (Feret et al., 2009), and more generally, symmetry (Clarke, Grumberg, and Peled, 1999; Rowley and Marsden, 2000).
2. The sensitivity analysis techniques (Turányi, Bérces, and Vajda, 1989) which compute the effect of a species on the concentrations of other important species and based on this it is either included or neglected in the model.
3. The time scale analysis techniques which mainly compute slow invariant manifolds (Gorban and Karlin, 2003). Such methods exploit the separation of timescales of various processes and variables and compute a low dimensional invariant manifold. These include Intrinsic low dimensional manifold (ILDM, (Maas and Pope, 1992)), Singular perturbation techniques (Lin and Segel, 1988), Computational singular perturbation (CSP, (Lam and Goussis, 1994)), Geometric singular perturbation theory (Fenichel, 1979), Quasi-stationary and quasi-equilibrium analysis in chemical kinetics (Schnell and Maini, 2002).

In this chapter a method is proposed to identify the slow and fast variables in a biochemical kinetic model with polynomial rate functions, without the simulation of the trajectories. This method can be categorised as a time scale analysis technique, which identifies the slow-fast chemical species and reactions based on computation of tropical equilibrations (cf. Section 4.1). In this context, the application of tropical equilibrations (TEs) was shown recently in the analysis of systems of polynomial or rational differential equations with applications to cell cycle modelling (Noel et al., 2012). As a matter of fact, the most common kinetic laws in computational biology are monomial or rational functions of the concentrations of the species. Tropical methods can be extended to S-systems (Savageau and Voit, 1987), in which case the reaction rates are generalized, real degree multivariate monomial functions. Therefore, these methods can, at least in principle be applied to practically all the biochemical network models. Tropical equilibration was previously used in an interesting study by Savageau (Savageau et al., 2009) as a design tool for the network steady states. The focus of the current application is different because it is concerned with dynamics and model reduction. Tropical equilibrations are interesting because they provide the lowest order approximations to the invariant manifolds. Computing tropical equilibrations from the orders of magnitude of the model parameters is a NP-hard problem, cf. (Theobald, 2006). However, for problems of limited size it is possible to obtain full sets of solutions by branch-and-bound strategies combined with efficient pruning of the branching tree to reduce the search time. Worse complexity cases will of course remain, this being unavoidable by the NP-hardness of the problem. The algorithm using the Newton polytope is provided for the same in Section 4.1 to solve the tropical equilibration problem efficiently for large biochemical networks. An alternative algorithm for finding tropical equilibrations, based on constraint logic programming was proposed in (Soliman, Fages, and Radulescu, 2014). However, when there are infinite branches of equilibrations, logic programming has no other alternative but the exhaustive enumeration of solutions between arbitrary bounds, whereas the Newton polytope method detects one solution per branch which is enough for identifying variable timescales and reduced models.

In addition to introducing and testing a new slow-fast decomposition algorithm, a fundamental principle of living systems, namely the *compression of complexity* when going from interactome to physiome can be considered within this framework. Variants of this idea were proposed by several authors in theoretical and computational biology. One way to quantify network functional complexity is to count the number of different attractors generated by a network. This question is biologically important because each different attractor of a gene network can, at least theoretically, be associated to a cell type. Kaufmann proposed the idea that a (random) gene network of size N can have $O(\sqrt{N})$ attractors. This is nicely compatible with the observation that there are 30000 genes and 265 different cell types in human (Kauffman, 2004). Looking at functional complexity differently, many authors counted the number of critical parameters of large regulatory networks. These studies lead to ideas such as the *von Dassow robustness* (Dassow et al., 2000), *sloppy sensitivity* (Gutenkunst et al., 2007), summation laws in *metabolic control* (Reder, 1988) or in mechanisms of *clock compensation* (Rand, 2008), meaning that among the numerous regulators of biological networks only very few (or even none when taken individually) are sensitive and can be used to control the system. The high parametric robustness of biological networks was also related to the low dimension of the invariant manifold attracting trajectories generated from many initial conditions as computed by locally linear embedding and Laplacian eigenmaps (Barbano et al., 2007), elastic principal manifolds (Radulescu, Zinovyev, and Lilienbaum, 2007) model reduction (Gorban and Radulescu, 2007; Radulescu et al., 2008). Generating robustness by a high compression ratio between the number of regulators and interactions (interactome) and the number of dynamic variables, sensitive parameters and types of behaviour could be understood as a form a redundancy allowing to compensate for uncertainty in the elements (Von Neumann, 1956). This general phenomenon finds a natural mathematical interpretation in the *Gromov-Talagrand concentration of measure*

in metric species: objects from high dimension look *thin* in projection on small dimension; similarly, measurable and physiologically important properties of large biochemical networks such as the number and phase space positions of attractors, relaxation times and oscillation periods, have small variability in spite of variability of network parameters (Gorban and Radulescu, 2007). Similar ideas, relating entropy and robustness were introduced in the context of network evolution (Demetrius and Manke, 2005). The results in this section allow to automatically determine the compression ratio from the number of regulators to the dimension of the invariant manifold carrying the slow dynamics of the system and to check if this dimension is generically low for biochemical networks.

6.2 Dynamical equations and slow-fast decomposition

Consider the biochemical model represented in (2.1). The n-dimensional space with all possible solutions of system $(x_1(t), \ldots, x_n(t))$ denotes the *phase space* of the system. The evolution of the system with defined initial values corresponds to a *trajectory* in the phase space. The evolution here refers to a family of smooth evolution functions which map any point at time t in the phase space back into the same phase space. In other words, the trajectory is one solution denoted by $(x_1(t), \ldots, x_n(t))$ for a given set of initial conditions $(x_1(0), \ldots, x_n(0))$. Thus, the set of all possible trajectories denotes the phase space. A low dimensional *differentiable manifold* in the phase space means a manifold which can be continuously and smoothly parameterizable and its dimension is d, where $d < n$. This means that at every point of this d-dimensional manifold which is embedded in n-dimensional phase space, the following equation holds

$$x = h(y) \tag{6.1}$$

where x is a set of $d - n$ variables, y is a set of d variables, h is a differentiable function. As the manifold is continuously and smoothly parameterizable, so, such a h is possible (at least locally).

Definition. An invariant set is any set of points in a dynamical system which are mapped into other points in the same set by the evolution function.

Remark. Therefore, a trajectory (as described above) is an invariant set.

Definition. An invariant manifold is an invariant set that is also a differentiable manifold.

The above basic introduction to invariant manifold is adopted from (Roussel, 2005). The invariant manifold is called slow if a typical trajectory of the system in the phase space tend to approach the invariant manifold governed by fast timescales and along this manifold the motion is governed by relatively slow timescale. Such a manifold is also attractive as it attracts the fast motion towards it. (Barillot et al., 2012; Maas and Pope, 1992). In (6.1) if variables y are slow and x are fast variables then the low dimensional manifold will be a slow manifold. The determination of slow-fast variables in the context of tropical equilibrations.

Eigenvector analysis is one of ways to identify the existence of such a manifold. For example, the existence of a lower dimensional *center manifold* is demonstrated in a simplified model of AIDS disease as shown in (Roussel, 1997). The method involves computing of the *centre eigenspace* i.e. the space spanned by the eigenvectors whose eigenvalues have a zero real part. However, in case of non-linear systems, the Jacobian matrix decomposition might change along the trajectory which makes the computations of such manifolds local in nature. Generally, the center manifolds are constructed by series expansions, close to the attractor of the system as many interesting dynamical behaviours take place in such a region.

CSP and ILDM methods provide numerical approximations of the invariant manifold close to an attractor. These methods have been successfully applied to reduce networks of reactions

in chemical engineering. Other reduction methods utilize the quasi-steady state (QSS) or quasi-equilibrium (QE) approximations (Gorban, Radulescu, and Zinovyev, 2010; Radulescu et al., 2008; Radulescu et al., 2012). QE and QSS methods require the knowledge of which species and reactions are fast. This knowledge can result from the slow/fast decompositions performed numerically by CSP or ILDM methods, or from the calculation of a slowness index (Radulescu et al., 2008), in all cases relying on the trajectory simulation. The application of these methods to computational biology is possible when the model parameters are known. When the parameters are unknown, or if they are known only by their orders of magnitudes, formal model reduction is needed. In addition, it is convenient to find the reductions without having to simulate the trajectories.

The biochemical networks stemming from cell biology integrate processes evolving on very different time scales. For instance, the changes of messenger RNA concentrations are usually faster compared to the changes of protein concentrations and the post-transcriptional modifications of proteins (for instance phosphorylation) are faster than the protein synthesis. For this reason, here slow-fast systems that have variables evolving on very different timescales are considered. Formally, variables $\boldsymbol{x}$ are much faster than variables $\boldsymbol{y}$ if the logarithmic derivatives $\frac{\mathrm{d}\log(\boldsymbol{x})}{\mathrm{d}t}$ are much larger in absolute values than $\frac{\mathrm{d}\log(\boldsymbol{y})}{\mathrm{d}t}$. After time rescaling, the differential equations describing the dynamics of a system with fast variables $\boldsymbol{x}$ and slow variables $\boldsymbol{y}$ read as:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \tfrac{1}{\eta}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) \tag{6.2}$$

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}), \tag{6.3}$$

where η is a small positive parameter and $\boldsymbol{f}$, $\boldsymbol{g}$ are functions not depending of η. We consider the asymptotic behaviour of the system in the limit $\eta \to 0$.

In dissipative systems, fast variables relax rapidly to a low dimensional attractive manifold called invariant manifold (Gorban and Karlin, 2005) that carries the slow mode dynamics. A projection of dynamical equations onto this manifold provides the reduced dynamics (Maas and Pope, 1992; Gorban and Karlin, 2005). This simple picture can be complexified to cope up with the hierarchies of invariant manifolds and with phenomena such as transverse instability, excitability and itineracy. Firstly, the relaxation towards an attractor can have several stages, each with its own invariant manifold. During relaxation towards the attractor, invariant manifolds are usually embedded one into another (there is a decrease of dimensionality) (Chiavazzo and Karlin, 2011). Secondly, invariant manifolds can lose local stability, which allow the trajectories to perform large phase space excursions before returning in a different place on the same invariant manifold or on a different one (Haller and Sapsis, 2010). The set of slow variables can change from one place to another. For all these reasons, even for fixed parameters, nonlinear models can have several reductions.

In biochemical networks, the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are (positive) species concentrations. Therefore, the functions $\boldsymbol{f}$, $\boldsymbol{g}$ are defined on the positive orthant. Furthermore, for most of the kinetic laws, the functions $\boldsymbol{f}$, $\boldsymbol{g}$ are polynomial or rational in the species concentrations. Although the method proposed in this thesis for model reduction apply for both polynomial and rational functions, for the sake of simplicity we will consider that $\boldsymbol{f}$ and $\boldsymbol{g}$ are polynomial functions. The system (6.2),(6.3) is endowed with positive initial conditions for all variables:

$$\boldsymbol{x}(0) = \boldsymbol{x}_0,\ \boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{6.4}$$

Let us suppose that the fast dynamics (6.2) has a unique stable state $\boldsymbol{x}^*(\boldsymbol{y})$ for all fixed $\boldsymbol{y}$ values. Let $\boldsymbol{J}(\boldsymbol{y})$ be the linear operator (Jacobian) that gives the linearization of $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ at

fixed $\boldsymbol{y}$, namely

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{J}(\boldsymbol{y})(\boldsymbol{x}-\boldsymbol{x}^*(\boldsymbol{y})) + O(|\boldsymbol{x}-\boldsymbol{x}^*(\boldsymbol{y})|^2).$$

The stable state $\boldsymbol{x}^*(\boldsymbol{y})$ is said to be uniformly hyperbolic if all the eigenvalues in the spectrum $Spec_{\boldsymbol{J}(\boldsymbol{y})}$ of $\boldsymbol{J}(\boldsymbol{y})$ have strictly negative real parts and are at a distance from the imaginary axis larger than a value $d > 0$, namely

$$\text{there is } d > 0 \text{ such that } Re(\lambda) < -d \text{ for all } \lambda \in Spec_{\boldsymbol{J}(\boldsymbol{y})} \text{ for all } \boldsymbol{y}. \tag{6.5}$$

Tikhonov's theorem (Tikhonov, 1952) says that if the above conditions are satisfied, then after a quick transition the system evolves approximately according to the following differential-algebraic equation:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}), \tag{6.6}$$

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}) = 0. \tag{6.7}$$

More precisely, the difference between solutions of (6.2),(6.3) and solutions of (6.6),(6.7) starting from the same initial data satisfying (6.7) (i.e. $\boldsymbol{y}(0) = \boldsymbol{y}_0$, $\boldsymbol{x}(0) = \boldsymbol{x}_0^*$, where $\boldsymbol{x}_0^*$ is the unique solution of $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}_0) = 0$) vanishes asymptotically like a positive power of η when $\eta \to 0$. In the case when (6.2) has several stable steady states, then which one of these states should be chosen as solution of (6.7) will depend on the initial conditions (6.4) of the full model.

Equation (6.7) means that the fast variables are slaved by the slow ones. In this case, and given the condition (6.5) on the Jacobian of $\boldsymbol{f}$ one can implicitly solve (6.7) and transform (6.6) into an autonomous reduced model for the slow variables. This approach is known as the quasi-steady state approximation.

The first and most important step in the implementation of this reduction method is to find the slow-fast decomposition (6.2),(6.3), which means to identify $\boldsymbol{x}$, $\boldsymbol{y}$ and η. For small models this can be done by rescaling variables and kinetic constants and by identifying the small parameter η as a ratio of kinetic constants or initial values of the variables. A well known example is the quasi-steady state approximation of the Michaelis-Menten enzymatic mechanism, when $\boldsymbol{x}$ is the concentration of the enzyme-substrate complex, $\boldsymbol{y}$ is the substrate concentration and η represents the ratio of the enzyme to the substrate concentrations (Noel et al., 2014). More generally, η can be interpreted as the ratio of fast to slow timescales. Numerical methods such as ILDM (Maas and Pope, 1992) use the Jacobian of the full system to obtain the slow-fast decomposition. In such methods η can be interpreted as the gap separating in logarithmic scale, the timescales of slow and fast variables obtained from the spectrum of the Jacobian. In this thesis, tropical equilibrations will be used to perform the same decomposition resulting into a slow invariant manifold. In addition, the proposed method will be compared with an existing numerical method namely ILDM (Intrinsic Low Dimensional Manifolds).

6.3 Approach

The approach for the model reduction can be divided into following steps as described in (Radulescu, Vakulenko, and Grigoriev, 2015; Samal et al., 2015a)

1. Computation of the tropical equilibrations, slow-fast decomposition of variables and defining of the truncated system (i.e. the equation system only with the dominant terms).

2. Computation of the conservation laws of truncated system denoted as fast cycles and defining the augmented system i.e. new variables corresponding to the fast cycles along with truncated system. For a consistent reduction these new variables should be at a slower timescale.

3. Elimination of the fast variables and expressing the model with only the slow variables along with the conservation laws of truncated system resulting in a reduced model.

In the rest of this section, the above steps will be discussed in detail.

6.3.1 Step 1: Determination of slow-fast variables from tropical equilibrations

The tropically truncated system from Section 4.1 is briefly reviewed here. We call *tropically truncated system* the system obtained by pruning the system (4.4), i.e. by keeping only the dominating monomials as shown below

$$\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t} = \varepsilon^{\nu_i}\Big(\sum_{j\in D(i)} \bar{k}_j S_{ij} \bar{\boldsymbol{x}}^{\boldsymbol{\alpha}_j}\Big), \tag{6.8}$$

where $D(i) = \underset{j}{\operatorname{argmin}}(\mu_j, S_{ij} \neq 0)$ selects the dominating rates of reactions acting on species i and

$$\nu_i = \min\{\mu_j | S_{ij} \neq 0\} - a_i. \tag{6.9}$$

The tropically truncated equations contain generically two monomial terms of opposite signs (in special cases they can contain more than two terms among which two have opposite signs). Polynomial systems with two monomial terms are called binomial or toric. In systems biology, toric systems are known as S-systems and were used by Savageau (Savageau and Voit, 1987) for modeling metabolic networks.

The truncated system (6.8) indicates how fast is each variable, relatively to the others. The inverse timescale of a variable x_i is given by $\frac{1}{x_i}\frac{\mathrm{d}x_i}{\mathrm{d}t} = \frac{1}{\bar{x}_i}\frac{\mathrm{d}\bar{x}_i}{\mathrm{d}t}$ that scales like ε^{ν_i}. Thus, if $\nu_{i'} < \nu_i$ then $x_{i'}$ is faster than x_i.

Let us assume that $\nu_1 \leq \nu_2 \leq \ldots \leq \nu_n$ (this may require species re-indexing but is always possible) and the following gap condition is fulfilled:

$$\text{there is } m < n \text{ such that } \nu_{m+1} - \nu_m > 0, \tag{6.10}$$

meaning that two groups of variables have separated timescales. The variables $\boldsymbol{X}_r = (x_1, x_2, \ldots, x_m)$ are fast (change significantly on timescales of order of magnitude $\varepsilon^{-\nu_m}$ or shorter). The remaining variables $\boldsymbol{X}_s = (x_{m+1}, x_{m+2}, \ldots, x_n)$ are slow (have little variation on timescales of order of magnitude $\varepsilon^{-\nu_m}$). Then, the parameter $\eta = \varepsilon^{\nu_{m+1}-\nu_m}$ represents the fast/slow timescale ratio in the Tikhonov's theorem from the preceding section. The gap condition means that η should be small. With these conditions, it was shown in (Noel et al., 2014; Radulescu, Vakulenko, and Grigoriev, 2015) that the quasi-steady state approximation can be applied. A further complication arises when the system has fast cycles and this will be described in the next step.

For systems with hierarchical relaxation, the separation between fast and slow variables is mobile within the cascade of relaxing modes. In the extreme case this means that all the species timescales are distinct and separated by large enough gaps. Let us consider that we are interested in changes on timescales θ or slower. The timescale θ defines a threshold order value by the equation

$$\mu_{\text{threshold}} = -\log(\theta/\tau)/\log(\varepsilon), \tag{6.11}$$

where τ are the time units from the model. Then, from (6.8) it follows that all variables x_i with $\nu_i \geq \mu_{\text{threshold}}$ are slow. Perturbations in the concentrations of these species relax to an attractor slower or as slow as θ. The remaining species are fast and the perturbations in their concentrations relax to equilibrated values much faster than θ.

Partial tropical equilibrations. It is useful to extend the tropical equilibration problem to partial equilibrations, that means solving (4.6) only for a subset of species. This is justified

by the fact that slow species do not need to be equilibrated. In order to have a self-consistent calculation the species timescales are computed by (6.9). A partial equilibration is *consistent* if $\nu_i < \nu$ for all non-equilibrated species i. $\nu > 0$ is an arbitrarily chosen threshold indicating the timescale of interest. Computation and interpretation of partial tropical equilibrations is a topic of future work.

6.3.2 Step 2: Reduction with fast cycles

Tropical truncation is useful for identifying the slow and fast variables of a system of polynomial differential equations. However, the truncation alone is not always enough for accurate reduction. As discussed in (Noel et al., 2014; Radulescu, Vakulenko, and Grigoriev, 2015), there are situations when the truncated system is not a good approximation. Typically, truncation could eliminate all the reactions exiting a fast cyclic subnetwork. Thus new conserved quantities appear, that were not conserved by the full model. Truncation is in this case accurate at short times, but introduces errors at large times. In order to cope with fast cycles pruning, the recipe discussed in (Gorban, Radulescu, and Zinovyev, 2010) is adopted for the quasi-equilibrium approximation. This recipe allows one to recover the terms that were neglected by truncation, but which are important for large time dynamics.

First, let us remind some definitions. The linear conservation law of a system of differential equations is represented as a linear form $C(\boldsymbol{x}) =< \boldsymbol{c}, \boldsymbol{x} >= c_1x_1 + c_2x_2 + \ldots + c_nx_n$ that is identically constant on trajectories of the system. It can be easily checked that vectors in the left kernel $Ker^l(S)$ of the stoichiometric matrix $\boldsymbol{S}$ provide linear conservation laws of the system (2.1). Indeed, system (2.1) reads $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{SR}(\boldsymbol{x})$, where the components of the vector $\boldsymbol{R}$ are $R_j(\boldsymbol{x}) = k_j x^{\alpha_j}$. If $\boldsymbol{cS} = 0$, then $\frac{\mathrm{d}<\boldsymbol{c},\boldsymbol{x}>}{\mathrm{d}t} = \boldsymbol{cSR}(\boldsymbol{x}) = 0$, where $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$.

Let us assume that the truncated system (6.8), restricted to the fast variables has a number of independent, linear conservation laws, defined by the left kernel vectors $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_d$, where $\boldsymbol{c}_k = (c_{k1}, c_{k2}, \ldots, c_{kf})$. These conservation laws can be calculated by recasting the truncated system as the product of a new stoichiometric matrix and a vector of monomial rate functions and further computing left kernel vectors of the new stoichiometric matrix. It is assumed that the fast conservation laws are not conserved by the full system (2.1).

The new slow variables $\boldsymbol{Y} = (y_1, \ldots, y_d)$ are defined, where $y_k = \sum_{i=1}^{f} c_{ki}x_i$. and eliminate the fast variables $x_1, x_2, \ldots, x_f$ by using the system :

$$\sum_{j \in D(i)} k_j S_{ij} \boldsymbol{x}^{\boldsymbol{\alpha_j}} = 0,\ i \in [1, f], \tag{6.12}$$

$$\sum_{i=1}^{f} c_{ki}x_i = y_k,\ k \in [1, d]. \tag{6.13}$$

Reactions of the initial model that were pruned by truncation have to be restored if they act on the new slow variables $\boldsymbol{Y}$, i.e. if $\sum_{i=1}^{f} c_{li}S_{ik} \neq 0$, for some $l \in [1, d]$, where k is the index of the reaction to be tested. Finally, the kinetic laws of these reactions have to be redefined in terms of the slow variables $\boldsymbol{X}^s, \boldsymbol{Y}$.

The rigorous justification of the reduction procedure for models with fast cycles can be found in (Radulescu, Vakulenko, and Grigoriev, 2015).

6.3.3 Step 3: Elimination of the fast species

The reduced model should contain only the slow variables obtained in Step 2 namely $\boldsymbol{X}_s$, $\boldsymbol{Y}$. The fast variables $\boldsymbol{X}_r$ need to be algebraically eliminated. In other words, the fast variables

are expressed via slow variables in the following manner

$$\boldsymbol{X}_r = f(\boldsymbol{X}_s, \boldsymbol{Y}) \tag{6.14}$$

where f are rational functions. Intuitively, this relation suggests that the motion of the slow variables in the reduced model are constrained by fast variable. By elimination it means that using the above relation, the fast variables are substituted in the equation system as a function of slow variables.

6.4 Examples

As an illustration of the method simple models are chosen that (i) have polynomial dynamics and (ii) contain fast cycles that ask for the reduction steps described in Section 6.3.2. The Michaelis-Menten model of enzymatic reactions as well as a cell cycle model proposed by Tyson (Tyson, 1991) satisfy both these conditions.

6.4.1 The Michaelis-Menten model

The irreversible Michaelis-Menten kinetics consist of three reactions:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\to} P + E, \tag{6.15}$$

where S, ES, E, P represent the substrate, the enzyme-substrate complex, the enzyme and the product, respectively.

The corresponding system of polynomial differential equations reads:

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 x_3 + k_{-1} x_2, \\ \dot{x}_2 &= k_1 x_1 x_3 - (k_{-1} + k_2) x_2, \\ \dot{x}_3 &= -k_1 x_1 x_3 + (k_{-1} + k_2) x_2, \\ \dot{x}_4 &= k_2 x_2, \end{aligned} \tag{6.16}$$

where $x_1 = [S]$, $x_2 = [ES]$, $x_3 = [E]$, $x_4 = [P]$.

The system (6.16) has two conservation laws $x_2 + x_3 = e_0$ and $x_1 + x_2 + x_4 = s_0$. The values e_0 and s_0 of the conservation laws result from the the initial conditions, namely $e_0 = x_2(0) + x_3(0)$ and $s_0 = x_1(0) + x_2(0) + x_4(0)$.

The conservation laws can be used to eliminate the variables x_3 and x_4 and obtain the reduced system as follows

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 (e_0 - x_2) + k_{-1} x_2, \\ \dot{x}_2 &= k_1 x_1 (e_0 - x_2) - (k_{-1} + k_2) x_2. \end{aligned} \tag{6.17}$$

There are two types of approximations and reductions for the Michaelis-Menten model, the quasi-steady state and the quasi-equilibrium approximation (Meiske, 1978; Segel, 1988; Segel and Slemrod, 1989; Gorban, Radulescu, and Zinovyev, 2010; Gorban and Shahzad, 2011). It is discussed here how these approximations can be related to tropical equilibrations (see also (Noel et al., 2014; Soliman, Fages, and Radulescu, 2014) where the same model is analysed using tropical curves).

Let us introduce orders of variables and parameters as follows $x_i = \bar{x}_i \epsilon^{a_i}$, $1 \leq i \leq 2$, $k_1 = \bar{k}_1 \epsilon^{\gamma_1}$, $k_{-1} = \bar{k}_{-1} \epsilon^{\gamma - 1}$, $e_0 = \bar{e}_0 \epsilon^{\gamma_e}$.

Then, we get the tropical equilibration equations by equating minimal orders of positive monomials with minimal orders of negative monomials in (6.17):

$$\gamma_1 + \gamma_e + a_1 = \min(\gamma_1 + a_1, \gamma_{-1}) + a_2, \tag{6.18}$$
$$\gamma_1 + \gamma_e + a_1 = \min(\gamma_1 + a_1, \min(\gamma_{-1}, \gamma_2)) + a_2. \tag{6.19}$$

The quasi-equilibrium approximation corresponds to the case when the reaction constant k_{-1} is much faster than the reaction constant k_2. In terms of orders, this condition reads $\gamma_{-1} < \gamma_2$. In this case, the two tropical equilibration equations (6.18), (6.19) are identical, because the $\min(\gamma_{-1}, \gamma_2) = \gamma_{-1}$. Let $\gamma_m = \gamma_{-1} - \gamma_1$ denote the order of the parameter $K_m = k_{-1}/k_1$. There are two branches of solutions of (6.18), namely $a_2 = \gamma_e, a_1 \leq \gamma_m$ and $a_2 = a_1 + \gamma_e - \gamma_m, a_1 \geq \gamma_m$ corresponding to the $\min(\gamma_1 + a_1, \gamma_{-1}) = \gamma_1 + a_1$ and to the $\min(\gamma_1 + a_1, \gamma_{-1}) = \gamma_{-1}$, respectively. Using the relation between orders and concentrations the first branch of solutions is identified with the saturation regime $x_2 \approx e_0$ (the free enzyme is negligible) and $x_1 >> K_m$ (the substrate has large concentration) and the second branch with the linear regime $x_2 << e_0$ (the concentration of the attached enzyme is negligible) and $x_1 << K_m$ (the substrate has low concentration).

In the linear regime of quasi-equilibrium the fast truncated system (obtained after removing all dominated monomials from (6.17)) reads

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 e_0 + k_{-1} x_2, \\ \dot{x}_2 &= k_1 x_1 e_0 - k_{-1} x_2. \end{aligned} \tag{6.20}$$

The variable $y = x_1 + x_2$ is conserved by the fast truncated system (6.20), but not by the full system (6.17). Therefore, y has to be considered as a new slow variable. By summing the two equations of (6.17) term by term the following is obtained

$$\dot{y} = -k_2 x_2. \tag{6.21}$$

Using the quasi-equilibrium equation $-k_1 x_1 e_0 + k_{-1} x_2 = 0$ x_1, x_2 are eliminated by expressing them as $x_1 = y/(1 + k_1 e_0/k_{-1})$, $x_2 = y/(1 + k_{-1}/(k_1 e_0))$. Finally, the reduced model is obtained for the slow variable y,

$$\dot{y} = -k_2 y/(1 + k_{-1}/(k_1 e_0)) = -V_{max} y/(e_0 + K_m), \tag{6.22}$$

where $V_{max} = k_2 e_0$.

If $\dot{y}$ is expressed as a function of the substrate concentration x_1, then $\dot{y} = -(V_{max}/K_m) x_1$, which is the well known Michaelis-Menten reaction rate in the linear regime.

In the saturated quasi-equilibrium regime, the fast truncated system reads

$$\begin{aligned} \dot{x}_1 &= -k_1 x_1 (e_0 - x_2), \\ \dot{x}_2 &= k_1 x_1 (e_0 - x_2). \end{aligned} \tag{6.23}$$

From (6.23) the quasi-equilibrium the equation $x_2 = e_0$ and further, using (6.21), the reduced model is obtained as shown below

$$\dot{y} = -V_{max}. \tag{6.24}$$

The tropical method also allows us to test that variables x_1, x_2 are faster than y, which means that the reductions are consistent (fast variables are eliminated and the reduced model is written in the slow variables only). In terms of ν orders defined by (6.9), one has to check that $\nu_1 < \nu_y$ and $\nu_2 < \nu_y$. Using (6.9) together with the quasi-equilibrium condition, it is found

that $\nu_2 = \gamma_{-1}$ in the linear regime and $\nu_2 = \gamma_1 + a_1$ in the saturated regime. Furthermore, $\nu_1 \leq \gamma_{-1} + a_2 - a_1$, $\nu_y = \gamma_2 + a_2 - \min(a_1, a_2)$ for both regimes. The condition $\nu_1 < \nu_y$ is satisfied because $\gamma_2 > \gamma_{-1}$. $\nu_2 < \nu_y$ is satisfied in the linear regime because $\gamma_2 > \gamma_{-1}$. The same condition is satisfied also in the saturated regime because $a_1 \leq \gamma_m = \gamma_{-1} - \gamma_1$ in this regime.

To summarize, the unique condition for quasi-equilibrium is $\gamma_2 > \gamma_{-1}$. In particular, this approximation does not depend on the initial data because γ_e does not occur in the above condition.

The quasi-steady state approximation corresponds to the situation when x_2 is equilibrated and faster than x_1. In this case one has to combine (6.19) with the condition $\nu_2 < \nu_1$. Let us denote by $\gamma_m = \min(\gamma_{-1}, \gamma_2) - \gamma_1$ the order of the parameter $K_m = (k_{-1} + k_2)/k_1$. (6.19) alone has two branches of solutions. The first branch is defined by $a_1 \leq \gamma_m$, $a_2 = \gamma_e$ and corresponds to the saturated regime of the quasi-steady state. The second branch is defined by $a_1 \geq \gamma_m$, $a_2 = a_1 + \gamma_e - \gamma_m$ and corresponds to the linear regime. From (6.17) it is found that $\nu_1 = \min(\gamma_1 + a_1 + \gamma_e, \gamma_1 + a_1 + a_2, \gamma_{-1} + a_2) - a_1$ and $\nu_2 = \gamma_1 + a_1 + \gamma_e - a_2$. By elementary inequality algebra it follows that the condition $\nu_2 < \nu_1$ is equivalent to $a_1 < \gamma_e$ at saturation and to $\gamma_m < \gamma_e$ in the linear regime.

Summarizing, the conditions for quasi-steady state are $a_1 < \min(\gamma_m, \gamma_e)$ (saturation) or $\gamma_m < \min(a_1, \gamma_e)$ (linear regime). In contrast to quasi-equilibrium, quasi-steady state depends on the initial conditions.

The quasi-steady state equations at saturation are $k_1 x_1(e_0 - x_2) = 0$, leading to $x_2 = e_0$. In the linear regime one has $k_1 x_1 e_0 - (k_{-1} + k_2)x_2 = 0$, leading to $x_2 = e_0 x_1/K_m$. Using (6.21) the well known expressions $\dot{y} = -k_2 e_0 = -V_{max}$ and $\dot{y} = -V_{max} x_1/K_m$ are obtained representing the reaction rate in the saturated and linear regimes, respectively.

The timescales of variables and the validity of quasi-steady state for the Michaelis-Menten irreversible kinetics were previously derived by Segel (Segel, 1988; Segel and Slemrod, 1989). The time scales and conditions using the tropical method are compatible with the ones of Segel on pieces, i.e. in the linear and in the saturated regime of quasi-steady state. For instance, like in (Segel, 1988), the conditions here imply that quasi-steady state can be valid for small γ_e (large enzyme) provided that γ_m is smaller (very large K_m).

6.4.2 The cell cycle model

This model describes the interaction between cyclin and cyclin-dependent kinase cdc2 during the progression of the eukaryotic cell cycle (see Fig. 6.1). Cyclin (variable x_5) is synthesized during interphase stage of the cycle (reaction of constant k_6). Newly synthesized cyclin forms a complex with the phosphorylated kinase cdc2 (cdc2 is the variable x_2 and the complex formation reaction has constant k_4). The resulting complex (variable x_4) is called inactive or pre-maturation promoter factor (pre-MPF). pre-MPF needs to be activated to enter into mitosis in order to phosphorylate many substrates controlling processes essential for nuclear and cellular division. The active form of MPF (variable x_3) is produced from pre-MPF either by a non-regulated transformation (reaction of constant k_{10}) or by an autocatalytic process (reaction of constant k_9). At the end of mitosis the active complex dissociates (reaction of constant k_1), resulting in the phosphorylated cyclin (variable x_6) that is degraded (reaction of constant k_8) and the de-phosphorylated kinase cdc2 (variable x_1). The kinase is equilibrated with its phosphorylated form (variable x_2) by phosphorylation and dephosphorylation reactions (of constants k_2 and k_3 respectively).

The full model has a stable periodic attractor, a limit cycle. The stable limit cycle oscillations correspond to the periodic succession of interphase and mitosis phases of the cell cycle.

The corresponding system of differential equations along with conservation laws for the above model can be described as

$$\begin{aligned}
\dot{x}_1 &= k_1x_3 - k_2x_1 + k_3x_2,\ \dot{x}_2 = k_2x_1 - k_3x_2 - k_4x_2x_5,\\
\dot{x}_3 &= k_{10}x_4 - k_1x_3 + k_9x_3^2x_4,\\
\dot{x}_4 &= k_4x_2x_5 - k_{10}x_4 - k_9x_3^2x_4,\ \dot{x}_5 = k_6 - k_4x_2x_5,\\
\dot{x}_6 &= k_1x_3 - k_8x_6,\ x_1 + x_2 + x_3 + x_4 = 1.
\end{aligned} \tag{6.25}$$

The value of the conservation law $x_1 + x_2 + x_3 + x_4 = 1$ follows from the initial conditions $\boldsymbol{x}(0) = (0, 0.75, 0, 0.25, 0, 0)$ that were taken from (Tyson, 1991). Other initial conditions with the same value of the conservation law would lead to the same tropical equilibration solutions.

Computing the tropical equilibration defined in Section 4.1 and (4.6) to this model, the following set of inequalities are obtained :

$$\begin{aligned}
\min(a_3 + \gamma_1, a_2 + \gamma_3) &= a_1 + \gamma_2,\\
a_1 + \gamma_2 &= \min(a_2 + \gamma_3, a_2 + a_5 + \gamma_4),\\
\min(a_4 + \gamma_{10}, 2a_3 + a_4 + \gamma_9) &= a_3 + \gamma_1,\\
a_2 + a_5 + \gamma_4 &= \min(a_4 + \gamma_{10}, 2a_3 + a_4 + \gamma_9),\\
\gamma_6 &= a_2 + a_5 + \gamma_4,\\
a_3 + \gamma_1 = a_6 + \gamma_8, &\min(a_1, a_2, a_3, a_4) = 0.
\end{aligned} \tag{6.26}$$

Using the numerical values of the parameters from the original paper we find, for $\varepsilon = 1/9$, $\gamma_1 = 0$, $\gamma_2 = -6$, $\gamma_3 = -3$, $\gamma_4 = -2$, $\gamma_6 = 2$, $\gamma_8 = 0$, $\gamma_9 = -2$, $\gamma_{10} = 2$ (cf. Section 4.1).

Remark: One may notice that the orders γ depend on which units were used for the parameters. However, if the parameter units are changed, the set of tropical equilibrations is transformed into an equivalent one. Indeed, the model equations should be invariant with respect to units conversion. In particular, if units of second order reaction constants (i.e. coefficients of second order monomial rates) are multiplied by k, one should subtract $\log(k)/\log(\varepsilon)$ from the parameter orders and add the same quantity to the concentration orders. This will generate an equivalent set of solutions, up to rounding errors.

Using the algorithm (cf. Algorithm 1 in Sections 4.3.4) three tropical equilibrations for this system are obtained, namely $\boldsymbol{a}_1 = (8, 5, 2, 0, -1, 2)$, $\boldsymbol{a}_2 = (5, 2, 2, 0, 2, 2)$, $\boldsymbol{a}_3 = (3, 0, 2, 0, 4, 2)$.

The rescaled truncated system for the solution $\boldsymbol{a}_3$ reads

$$\begin{aligned}
\dot{\bar{x}}_1 = \varepsilon^{-6}(\bar{k}_3\bar{x}_2 - \bar{k}_2\bar{x}_1),\ \dot{\bar{x}}_2 &= \varepsilon^{-3}(\bar{k}_2\bar{x}_1 - \bar{k}_3\bar{x}_2),\\
\dot{\bar{x}}_3 &= \bar{k}_{10}\bar{x}_4 - \bar{k}_1\bar{x}_3 + \bar{k}_9\bar{x}_3^2\bar{x}_4,\\
\dot{\bar{x}}_4 &= \varepsilon^2(-\bar{k}_{10}\bar{x}_4 + \bar{k}_4\bar{x}_2\bar{x}_5 - \bar{k}_9\bar{x}_3^2\bar{x}_4),\\
\dot{\bar{x}}_5 = \varepsilon^{-2}(\bar{k}_6 - \bar{k}_4\bar{x}_2\bar{x}_5),\ \dot{\bar{x}}_6 &= \bar{k}_1\bar{x}_3 - \bar{k}_8\bar{x}_6.
\end{aligned} \tag{6.27}$$

It appears clearly that the variables x_1, x_2, x_5 are fast. More precisely, their characteristic times are $\nu_1^{-1} = \varepsilon^6$, $\nu_2^{-1} = \varepsilon^3$, $\nu_5^{-1} = \varepsilon^2$, respectively. The largest of these timescales is here approximately 0.01 (in minutes which are the time units of the model). The remaining slow variables have characteristic times from ε^0 to ε^{-2}, i.e. approximately from 1 to 100 min. Therefore, the timescales of slow and fast species are separated by a gap, and the singular perturbation small parameter (cf. Section 6.2) is $\eta = t_{fast}/t_{slow} \sim \varepsilon^2$ (the power 2 arises as the difference between $\nu_6 = \nu_3 = 0$, coming from the fastest slow species and $\nu_5 = -2$, coming from the slowest fast species).

The fast truncated system reads

$$\dot{x}_1 = k_3x_2 - k_2x_1, \dot{x}_2 = k_2x_1 - k_3x_2,$$
$$\dot{x}_5 = k_6 - k_4x_2x_5. \tag{6.28}$$

and has a single conservation law $C_1 = x_1 + x_2$ that provides a new slow variable. This conservation law, not conserved by the full system (6.25), indicates the presence of a fast cycle in the model. It is the rapid phosphorylation/dephosphorylation cycle transforming the cyclin x_1 into its phosphorylated form x_2 and back. The fast variables are eliminated from the system obtained by adding to (6.28) the definition of the fast conservation law cf. Section 6.3.2:

$$k_3x_2 - k_2x_1 = 0, k_6 - k_4x_2x_5 = 0, y = x_1 + x_2. \tag{6.29}$$

The differential equation for y is obtained by adding the first two equations of the full system (6.25), and thus restoring the terms k_1x_3 and $k_4x_2x_5$, that have order ε^2 and were pruned in the first step.

Finally, the following reduced model is obtained

$$\dot{x}_3 = k_{10}x_4 - k_1x_3 + k_9x_3^2x_4, \tag{6.30}$$
$$\dot{x}_4 = -k_{10}x_4 + k_6 - k_9x_3^2x_4, \tag{6.31}$$
$$\dot{x}_6 = k_1x_3 - k_8x_6, \dot{y} = k_1x_3 - k_6. \tag{6.32}$$

and the slaved fast variables are given by $x_1 = yk_2/(k_2+k_3) \approx yk_2/k_3$, $x_2 = yk_3/(k_2+k_3) \approx y$, $x_5 = k_6(k_2+k_3)/(k_4k_3y) \approx k_6/(k_4y)$, where we have used (6.29) and the fact that $k_2 \ll k_3$.

Let us note that the variable y has the same order as x_2 ($a_y = a_3 = 2$), it is tropically equilibrated ($\gamma_1 + a_3 = \gamma_6 = 2$ in (6.32)), and has $\nu_y = \gamma_6 - a_y = 0$ meaning that it is slow.

Let us call this four variables model reduced model 1. Note that in this model the dynamics of the variables x_3, x_4 is decoupled from the two others. It can therefore be concluded that by the tropical based approach a two dimensional minimal cell cycle model is obtained.

Repeating the procedure for the equilibrations $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ two other rescaled truncated systems were obtained.

The rescaled truncated system for the solution $\boldsymbol{a}_1$ reads

$$\dot{\bar{x}}_1 = \varepsilon^{-6}(\bar{k}_3\bar{x}_2 - \bar{k}_2\bar{x}_1 + k_1x_3), \dot{\bar{x}}_2 = \varepsilon^{-3}(\bar{k}_2\bar{x}_1 - \bar{k}_3\bar{x}_2),$$
$$\dot{\bar{x}}_3 = \bar{k}_{10}\bar{x}_4 - \bar{k}_1\bar{x}_3 + \bar{k}_9\bar{x}_3^2\bar{x}_4,$$
$$\dot{\bar{x}}_4 = \varepsilon^2(-\bar{k}_{10}\bar{x}_4 + \bar{k}_4\bar{x}_2\bar{x}_5 - \bar{k}_9\bar{x}_3^2\bar{x}_4),$$
$$\dot{\bar{x}}_5 = \varepsilon^3(\bar{k}_6 - \bar{k}_4\bar{x}_2\bar{x}_5), \dot{\bar{x}}_6 = \bar{k}_1\bar{x}_3 - \bar{k}_8\bar{x}_6, \tag{6.33}$$

and for the solution $\boldsymbol{a}_2$ the following rescaled truncated system is obtained

$$\dot{\bar{x}}_1 = \varepsilon^{-6}(\bar{k}_3\bar{x}_2 - \bar{k}_2\bar{x}_1), \dot{\bar{x}}_2 = \varepsilon^{-3}(\bar{k}_2\bar{x}_1 - \bar{k}_3\bar{x}_2),$$
$$\dot{\bar{x}}_3 = \bar{k}_{10}\bar{x}_4 - \bar{k}_1\bar{x}_3 + \bar{k}_9\bar{x}_3^2\bar{x}_4,$$
$$\dot{\bar{x}}_4 = \varepsilon^2(-\bar{k}_{10}\bar{x}_4 + \bar{k}_4\bar{x}_2\bar{x}_5 - \bar{k}_9\bar{x}_3^2\bar{x}_4),$$
$$\dot{\bar{x}}_5 = (\bar{k}_6 - \bar{k}_4\bar{x}_2\bar{x}_5), \dot{\bar{x}}_6 = \bar{k}_1\bar{x}_3 - \bar{k}_8\bar{x}_6. \tag{6.34}$$

In both cases, the variable x_5 is slow, which was not the case for the equilibration $\boldsymbol{a}_1$. This is possible, because for a nonlinear model, the timescale of a variable depends on the concentration range in which the model functions. The equilibrations $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ correspond to very low and low concentrations of phosphorylated kinase x_2 (proportional to ε^5 and ε^2, respectively), meaning slow consumption of the cyclin x_5. The concentration of x_2 is large for the equilibration $\boldsymbol{a}_3$ (proportional to ε^0) leading to rapid consumption of x_5 (see (6.25)).

The two tropical equilibrations $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ lead to the same reduced model, which is called reduced model 2:

$$\begin{aligned} \dot{x}_3 &= k_{10}x_4 - k_1x_3 + k_9x_3^2x4, \\ \dot{x}_4 &= -k_{10}x_4 + k_6 - k_9x_3^2x_4, \\ \dot{x}_5 &= k_6 - k_4yx_5, \\ \dot{x}_6 = k_1x_3 - k_8x_6, \dot{y} &= k_1x_3 - k_6, \end{aligned} \tag{6.35}$$

to be considered together with $x_1 = yk_2/k_3$, $x_2 = y$.

The tropical setting confirms ideas from the theory of nonlinear dynamical systems. The two reduced models are nested. Reduced model 2 has a larger number of slow (relaxing) variables than reduced model 1. This means that the corresponding invariant manifolds are embedded one into another with the lowest dimensional one defined by the reduced model 1 carrying the dynamics on the limit cycle attractor. Starting with initial low concentrations of the phosphorylated kinase corresponding to the equilibration $\boldsymbol{a}_1$ or $\boldsymbol{a}_2$, the system will increase these concentrations to levels corresponding to the equilibration $\boldsymbol{a}_3$ that allow the stable limit cycle oscillations.

One can notice that the reduced model 1 does not contain the parameters k_2, k_3, k_4 of the full model. This means that as long as the phosphorylation and the dephosphorylation of the free kinase, as well as the formation of cyclin kinase complex are fast enough, the actual values of the kinetic constants of these processes are not important.

In his paper, Tyson (Tyson, 1991) also proposes a two variables reduced model:

$$\begin{aligned} \dot{u} &= k_9(v-u)(\alpha+u^2) - k_1u, \\ \dot{v} &= k_6 - k_1u, \end{aligned} \tag{6.36}$$

where $u = x_3$, $v = x_3 + x_4 + x_5$, $\alpha = k_{10}/k_9$.

It can be easily checked that (6.36) are equivalent with (6.30),(6.31), provided that $x_5 \ll x_4$ and $x_5 \ll x_3$. These last conditions, justified by intuitive arguments, were used in the derivation of the reduced model in (Tyson, 1991). In the current approach, the same conditions follow immediately from the orders of the species concentrations. Indeed, for the equilibration $\boldsymbol{a}_3$, $x_5 \sim \varepsilon^4$, $x_3 \sim \varepsilon^2$, $x_4 \sim \varepsilon^0$, therefore $x_5 \ll x_4$ and $x_5 \ll x_3$. To summarize, the advantage of the current approach is that it is automatic and can be applied to larger models that are more difficult or impossible to grasp by simple intuition.

6.5 Benchmarking on Biomodels database

Biological models are selected based on Section 4.3 and the average number of slow variables across Biomodels with different time thresholds are computed to get an estimate of the dimension of the invariant manifold. The number of slow variables are computed based on the rescaled orders i.e. $\mu_i - a_i > \mu_{\text{threshold}}$ and a certain time threshold as explained in (6.11). In addition to slow species, the quasi buffered species are computed which are slow variables with very high time threshold. To compute the timescale threshold is fixed to $100,000$ seconds and the slow species at this threshold are labeled as quasi buffered species. In the model, such species are practically constant and are subtracted from the slow variables. Boxplot showing the compression ratio (i.e. ratio of average number of slow variables /total number of variables) over all the tropical equilibrations for each model with respect to different time thresholds is shown in Fig. 6.4, 6.7 for tropical equilibrations computed Algorithm 1 and Algorithm 2 (computing tropical minimal branches). Additionally for these two algorithms, the average number of slow variables across all models are presented as boxplot in Fig. 6.3, 6.6 respectively.

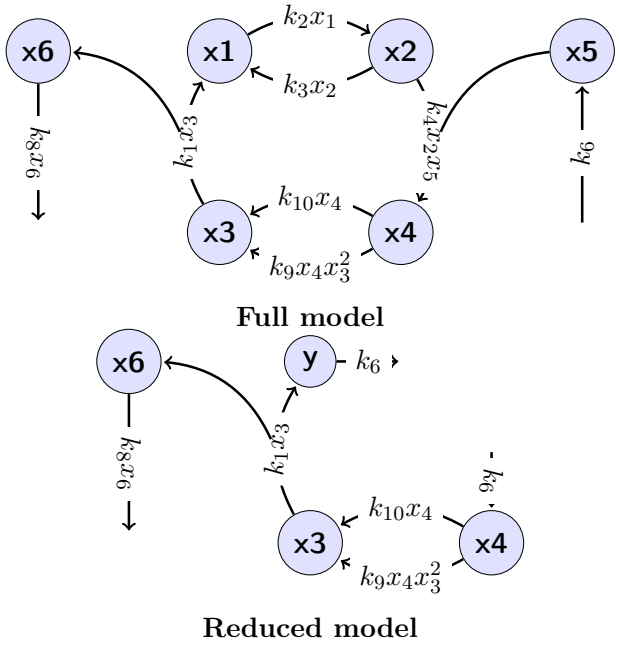


FIGURE 6.1: Graphic representation of the full cell cycle model (Tyson, 1991) (referenced as BIOMD00000005 by Biomodels) and of the reduced model 1. The full model describes the cyclin production and complex formation between the cyclin and the kinase cdc2, the autocatalytic activation (by dephosphorylation), and the dissociation of this complex followed by the destruction of the cyclin. The reduced model represents accurately the same processes, on the invariant manifold containing the periodic attractor. The different variables mean: x_1 : cdc2, x_2 : cdc2-P, x_3 : cdc2 : cyclin-P i.e. active MPF complex, x_4 : P-cdc2 : cyclin-P i.e. pre-MPF complex, x_5 : cyclin, x_6 : cyclin-P, $y = x_1 + x_2$: total free cdc2. Both full and reduced model are biochemical networks with polynomial rate functions.

The slowest timescale is also computed for each model which is defined as the smallest time threshold at which all species become fast (the quasi-buffered species were removed from the model before performing this step). A histogram showing the distribution slowest timescale is presented in Fig. 6.2, 6.5 for Algorithm 1 and Algorithm 2 (computing tropical minimal branches), respectively. To estimate the slowest timescale, the time threshold is varied and the number of slow species are counted, the threshold at which all species become fast is considered to be the slowest timescale of that model. This histogram indicates that the benchmarked models are representative of a wide variety of cellular processes whose timescales range from fractions of seconds to one day.

6.6 Testing the method

In order to test the method the cell cycle model (Tyson, 1991) (referenced as BIOMD00000005 by Biomodels) is considered in more detail. This is the cell cycle model example analysed in Section 6.4.2. Here the detection of slow and fast species and the accuracy of model reduction is tested.

6.6.1 Slowness index

The detection of slow fast species is tested by comparison with a numerical method introduced in (Radulescu et al., 2008). This method consists of simulating the trajectories $x_i(t)$ for each species i of the model and comparing them to the imposed trajectories x_i^* calculated as solutions

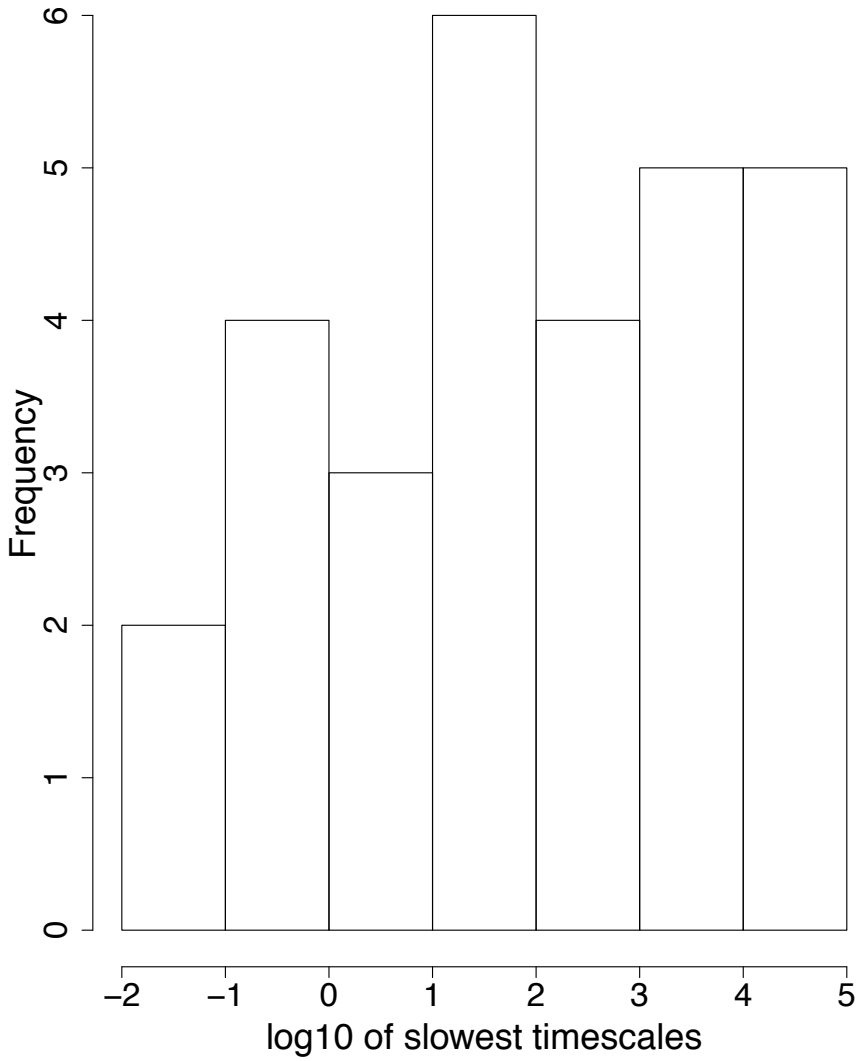


FIGURE 6.2: (A) Histogram showing the distribution of slowest timescales for 35 models corresponding to $\varepsilon = 1/23$. The tropical equilibrations are computed based on Algorithm 1.

of quasi-steady state equations cf. (6.7). Precisely, x_i^* is the solution of $\sum_j k_j S_{ij} \boldsymbol{x}^{\boldsymbol{\alpha}_j} = 0$ in which all species of indices $l \neq i$ are replaced by their simulated values $x_l(t)$. Like in (Radulescu et al., 2008) the slowness index $I_i(t) = |\log_{10}(x_i(t)/x_i^*(t))|$ is used (the base of the logarithm is purely conventional). Fast species obey quasi-steady state conditions (see (6.7) and Section 6.2). Therefore, for fast species, I_i is close to zero. For slow species, the trajectories $x_i(t)$ are different from $x_i^*(t)$ and the index I_i is high. Fig. 6.8 shows the values of this index for all the species in the cell cycle model BIOMD00000005. In the tropical method a species is fast or slow depending how the orders $\nu_i = \mu_i - a_i$ compared to a timescale threshold. For $\varepsilon = 1/9$, three tropical solutions were found which are already discussed in Section 6.4.2. For the solution $\boldsymbol{a}_3$ the species 1, 2, and 5 are fast and the species 3, 4, and 6 are slow (timescales 1 min or slower). This solution leads to the reduced model 1 described in Section 6.4.2. In contrast, species 5 is slow for the two other equilibrations corresponding to the reduced model

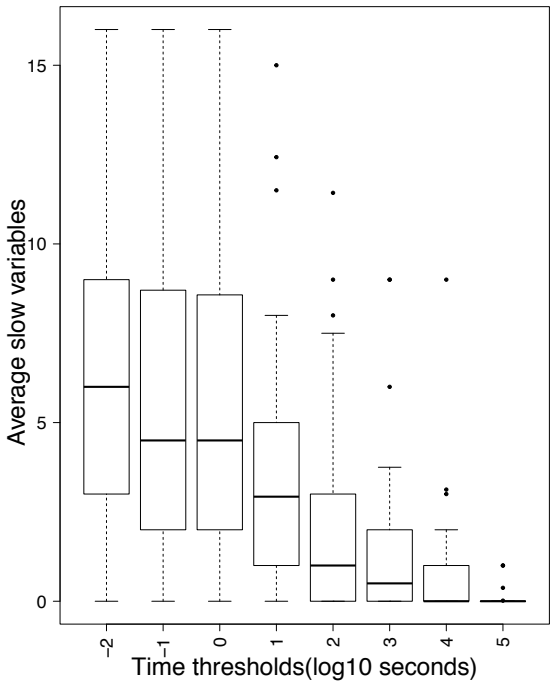


FIGURE 6.3: Boxplots showing the average number of slow variables in the Biomodels database for different values of time threshold θ. Each point represents the compression ratio over all the tropical equilibrations for each model with respect to different time thresholds. The time thresholds -2 to 5 in the plot are the $\log 10_{10}$ transformed values of time thresholds $0.01, 0.1, 1, 10, 100, 1000, 10000, 100000$ in secs. The boxplot corresponds to $\varepsilon = 1/23$. The boxplots of other ε values look similar. The tropical equilibrations are computed based on Algorithm 1.

2. The numerical method based on the slowness index classifies species 1,2, and 5 as fast and is thus compatible with the new method for the tropical solution $\boldsymbol{a}_3$ (Fig. 6.8a)). The reduced model 1 corresponding to the tropical solution $\boldsymbol{a}_3$ reproduces with good accuracy the limit cycle oscillations of the cell cycle model as shown in Fig. 6.8c).

6.6.2 Accuracy of the reduction

A quantitative estimate of reduction accuracy can be based on the L^2 norm of the difference between trajectories $\boldsymbol{x}(t)$, $\boldsymbol{x}_{red}(t)$ simulated with the full and reduced model, respectively. However, because periods are slightly changed by the reduction, the error could be defined as $err = inf_a \|x(t) - x_{red}(at)\|/\|x(t)\|$, where a is a time scaling parameter close to 1. For the trajectories shown in Fig. 6.8c), err is less than 0.01 and the optimal scaling parameter is $a = 1.0002$ (the relative change of the period is 0.0002).

The two other equilibrations lead to the reduced model 2 that is at least as accurate as the reduced model 1 (in short, in the reduced model 2, species 5 is considered slow and is not eliminated). This reduction accurately reproduces the dynamics not only on the limit cycle attractor, but also when initial data is far from this attractor. This is illustrated in Fig. 6.8d). The full model and the two reduced models are simulated starting from several initial data $\boldsymbol{x}_{0i}$, $i = 1, \ldots, 3$. The initial data of the reduced models is obtained by projection on the corresponding invariant manifolds. For example, the reduced model 1 evolves on an invariant manifold whose equations (up to small correcting terms) are given by (6.29) and read $x_5 = k_6/(k_4 x_2)$, $x_1 = k_3 x_2/k_1$. By computing the eigenvalues of the Jacobian of system

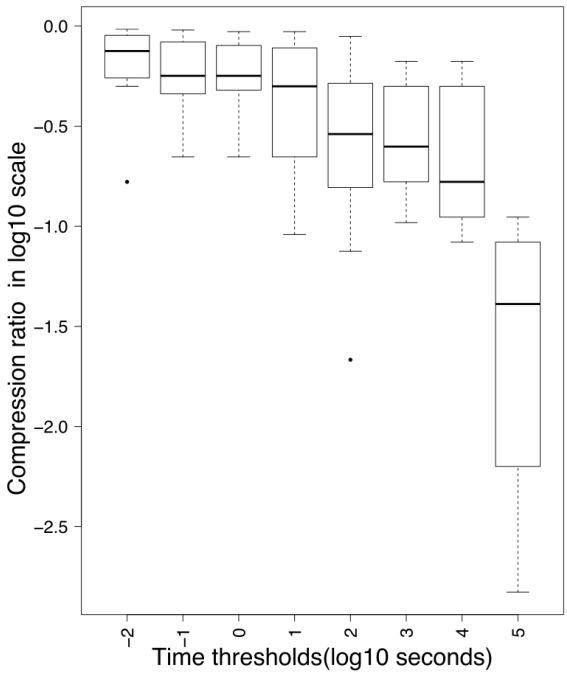


FIGURE 6.4: Boxplots showing compression ratio in the Biomodels database for different values of time threshold θ. The compression ratio is defined as the average number of slow variables / number of variables in the model. Each a point represents the average slow variables. The number of slow variables were computed based on rescaled orders (see (6.11)) and certain time threshold in seconds. The time thresholds -2 to 5 in the plot are the $\log 10_{10}$ transformed values of time thresholds $0.01, 0.1, 1, 10, 100, 1000, 10000, 100000$ in secs. The boxplot corresponds to $\varepsilon = 1/23$. The boxplots of other ε values look similar. The tropical equilibrations are computed based on Algorithm 1.

(6.25) it was found that this invariant manifold has an attractive, stable region (all eigenvalues, except the zero ones corresponding to exact conservation laws, have negative real parts) and an unstable region (where there are eigenvalues with positive real parts). The initial data vectors $\boldsymbol{x}_{01}$ and $\boldsymbol{x}_{02}$ are close to the unstable region of the invariant manifold. Therefore, the trajectories starting from these initial data first get away from the manifold and after large excursions approach the attractive part of the manifold. Reduced model 2 is able to reproduce these transients but not the reduced model 1 (Fig. 6.8d) because the latter is valid only on the slowest attractive invariant manifold.

6.7 Comparison with COPASI time separation method

The proposed method was tested against the existing tool COPASI (Hoops et al., 2006; Surovtsova et al., 2009). COPASI is a software for simulation and analysis of biochemical networks. This software accepts and generates several model exchange formats including the widely spread SBML format and is very popular in the computational biology community. COPASI is one of major biochemical networks tools that implements time separation of variables. To accomplish this aim COPASI proposes a modified ILDM (intrinsic low dimensional manifold) method. This method computes slow and fast modes which are transformations of species concentrations as described in (Deuflhard and Heroth, 1996; Zobeley et al., 2005; Surovtsova et al., 2009). More precisely, COPASI performs a Schur block decomposition of the Jacobian matrix $\boldsymbol{J}$, consisting

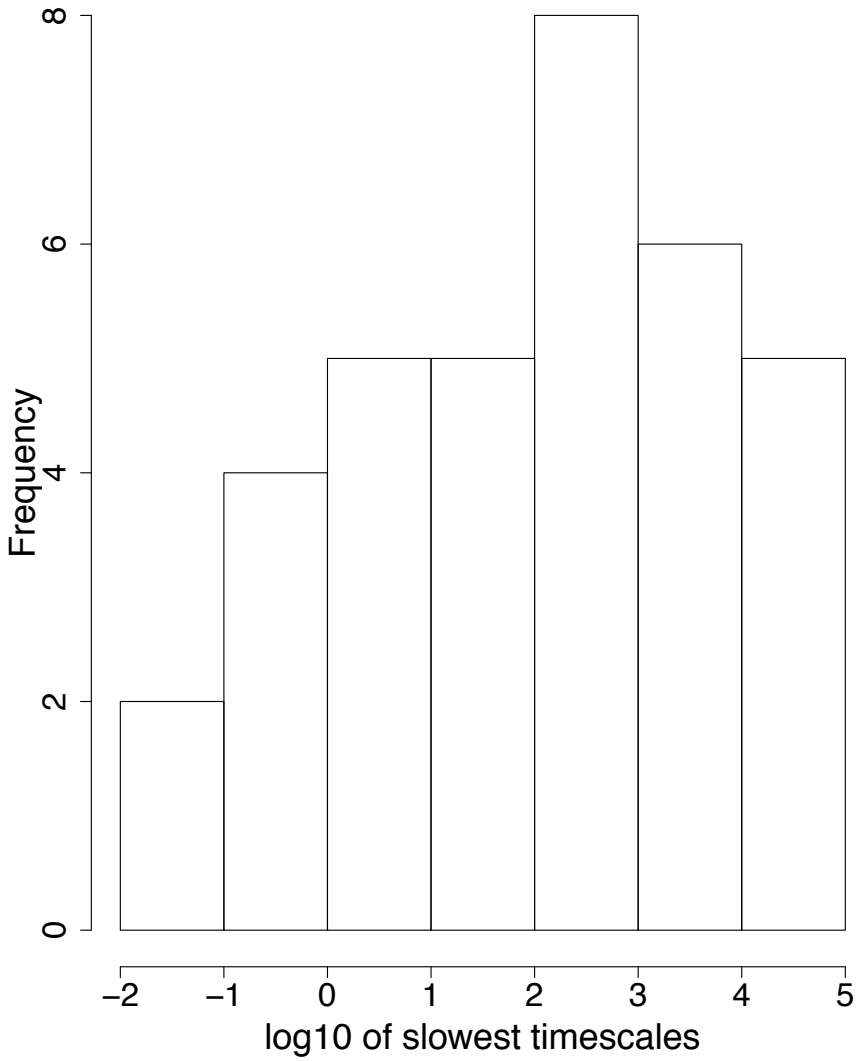


FIGURE 6.5: (A) Histogram showing the distribution of slowest timescales for 35 models corresponding to $\varepsilon = 1/23$. The tropical equilibrations are computed based on Algorithm 2.

in finding a non-singular matrix T such that $T^{-1}JT = \begin{pmatrix} S_{slow} & 0 \\ 0 & S_{fast} \end{pmatrix}$, where S_{slow}, S_{fast} have real Schur form, i.e. they are upper triangular matrices with possibly non-vanishing elements on the first subdiagonal.

The time threshold (needed to separate the slow and fast blocks of the Jacobian matrix) is automatically captured in this method by finding a gap in the spectrum of the Jacobian (cf. Section 6.2).

In order to compare the modified ILDM method against the tropical and slowness index methods, the fast space of the model was computed using COPASI for 100 time steps between 1 and 100 min and the contribution of each species to this fast space was checked. COPASI defines the contribution of a species i to a mode j as the matrix element T^{-1}_{ji}. These contributions of various species to fast modes are provided by COPASI as fractions p_i, where i is the species index, $0 \leq p_i \leq 1$, $\sum_i p_i = 1$. COPASI declare species with largest contribution

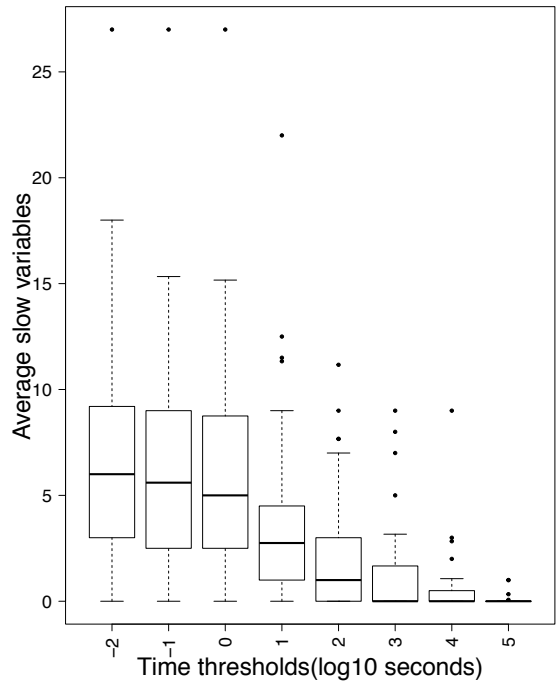


FIGURE 6.6: Boxplots showing the average number of slow variables in the Biomodels database for different values of time threshold θ. Each point represents the compression ratio over all the tropical equilibrations for each model with respect to different time thresholds. The time thresholds -2 to 5 in the plot are the $\log 10_{10}$ transformed values of time thresholds $0.01, 0.1, 1, 10, 100, 1000, 10000, 100000$ in secs. The boxplot corresponds to $\varepsilon = 1/23$. The boxplots of other ε values look similar. The tropical equilibrations are computed based on Algorithm 2.

to the fast space (largest p_i) as fast species. For exactly the same trajectory, the values of the slowness indices $I_i = |\log_{10}(x_i(t)/x_i^*(t))|$ were computed. For fast species, I_i should be close to zero. Fig. 6.8a) and b) summarizes the comparison between the slowness index and the tropical method. The tropical solution $\boldsymbol{a}_3$ leads to the reduced model 1 (cf. Section 6.4.2) and copes with the limit cycle trajectory used in this test. The timescale orders ν_i of the variables for this tropical solutions identify species x_1, x_2, x_5 as fast and species x_3, x_4, x_6 as slow (see also Section 6.4.2). As can been seen in Fig. 6.8a) the slowness index of species x_1, x_2, x_5 is close to zero for all times. The slowness index of species x_3, x_4, x_6 can reach large values. Therefore the tropical method and the slowness index method provide exactly the same timescale decomposition. COPASI time separation can not be compared directly to the tropical method, because it generates a timescale decomposition that changes with time and which is valid for a trajectory. However, it can be compared with the slowness index decomposition. Fig. 6.9 summarizes the comparison between COPASI and the slowness index. It should be noted that the species x_3 is automatically eliminated by COPASI using the single conservation law present in the model. The slowness index and COPASI contribution to fast space should be anticorrelated: when the first one is small the latter should be big and vice versa. Species x_1 and x_5 have high contribution towards the fast space and very low slowness index (see Fig. 6.9b). For these species it can be said that there is consistence between COPASI and slowness index. Species x_2 also has large contribution to fast space except for some intervals where COPASI may classify it as slow. The method unambiguously classifies this species as fast (cf. Fig. 6.9b) its slowness index is very low for all times). Most importantly,

COPASI fails to identify correctly time intervals where species 6 is slow as indicated by

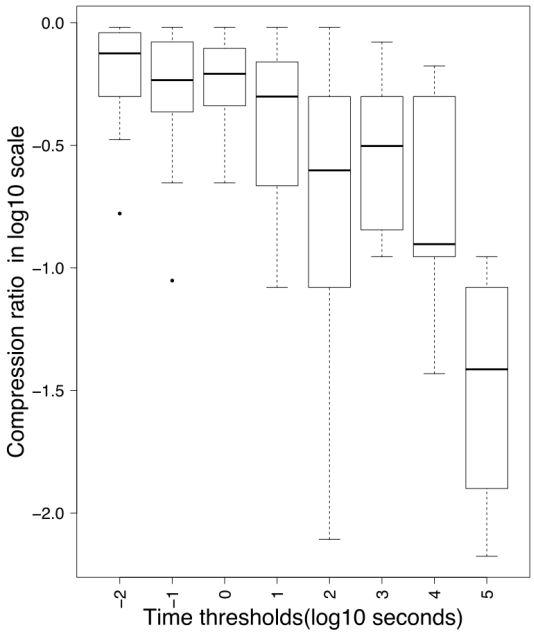


FIGURE 6.7: Boxplots showing compression ratio in the Biomodels database for different values of time threshold θ. The compression ratio is defined as the average number of slow variables / number of variables in the model. Each a point represents the average slow variables. The number of slow variables were computed based on rescaled orders (see (6.11)) and certain time threshold in seconds. The time thresholds -2 to 5 in the plot are the $\log 10_{10}$ transformed values of time thresholds $0.01, 0.1, 1, 10, 100, 1000, 10000, 100000$ in secs. The boxplot corresponds to $\varepsilon = 1/23$. The boxplots of other ε values look similar. The tropical equilibrations are computed based on Algorithm 2.

the large value of the slowness index co-existing with large values of contribution p_i (as large as for species 1 and 2, see Fig. 6.9b). According to COPASI this species is similar to the fast species x_2, whereas the proposed method indicate it similar to the slow species x_4 and x_3. As a matter of fact, COPASI determines slow variables by comparing the values of contributions p_i to the fast and slow modes. Despite the existence of a spectral gap, the differences of the indices p_i between species that are considered slow and fast can be relatively small and therefore this classification is not robust. In contrast, the tropical and slowness indices methods classify species in a robust way. Indeed, timescales to species are directly associated, via the orders ν_i and these timescales are well separated for slow and fast species. Using these orders, it was found that the fastest slow species x_3 and x_6 are 100 times slower than the slower fast species x_5. Furthermore, as shown in Fig.6.8a, the slowness index is very sensitive to differences in timescales. Fast species x_1, x_2, x_5 keep this index low for all times, whereas the corresponding COPASI indices are not always high. Generally, it should not be recommended to use species contributions to modes as indicative of their timescales, as COPASI does. For instance, the sum of two or more species can be a slow mode, even if all these species are fast (cf. Section 6.3.2, this situation is typical for fast cycles). The fast species have in this case higher contributions to a slow mode which may qualify them as slow according to the species contribution criterion.

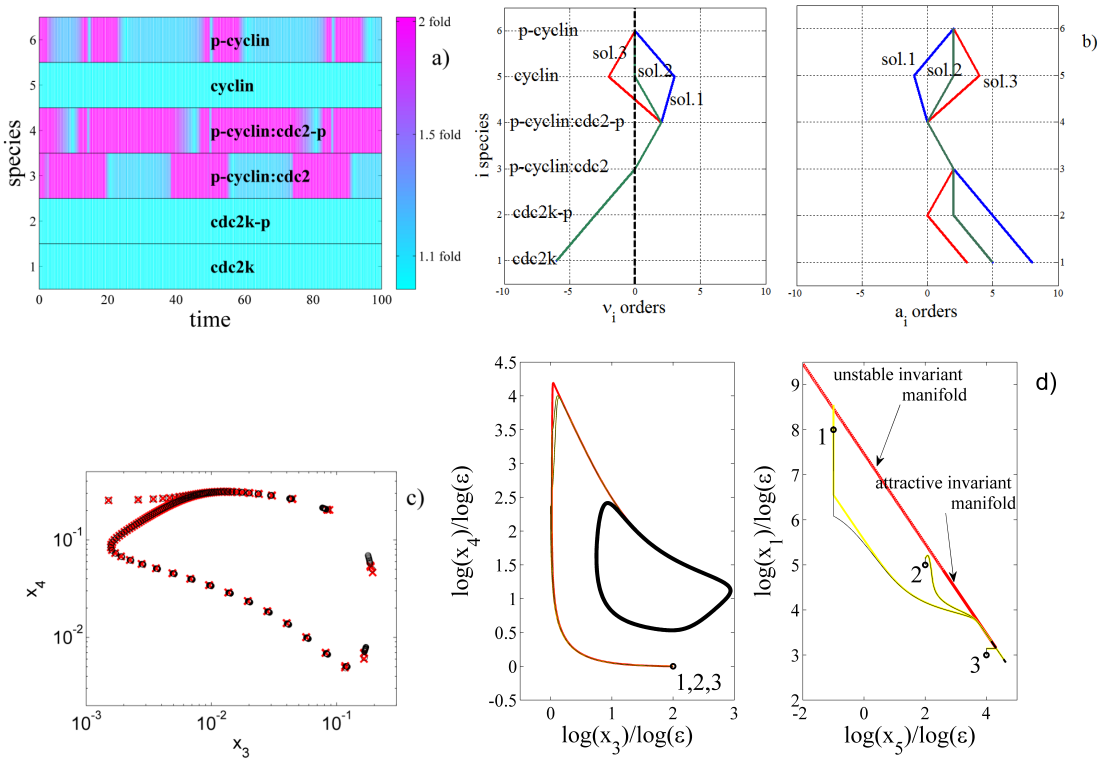


FIGURE 6.8: Testing tropical slow/fast decomposition and accuracy of reduction of BIOMD00000005 (cell cycle model). a) The slowness index is represented as a function of time on trajectories: slow variables have large slowness index (the fold ratio stands for the exponentiated index $exp(\log(x_i(t)/x_i^*(t))|)$); b) Left : The orders $\nu_i = \mu_i - a_i$ are represented for different species and for three tropical solutions. Cf. (6.8) a species i evolves on the timescale $\varepsilon^{-\nu_i}$ and hence lower ν_i mean faster species. The threshold θ separating slow and fast species set to 1 min to satisfy the gap condition (6.10) corresponds to $\mu_{\text{threshold}} = 0$ as defined by (6.11). The threshold order is represented as a dotted line. Fast species have orders below this value, namely species x_1, x_2, x_5 are fast for the tropical solution $\boldsymbol{a}_3$, whereas only species x_1, x_2 are fast for the tropical solutions $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$. Right : Orders a_i for different species and tropical solutions indicate species concentrations. Cf. (4.3) higher a_i mean lower concentration. For all order calculations we have used $\epsilon = 1/9$. c) Comparison of the limit cycle trajectories computed with the full (black circles) and reduced model (red crosses). d) Model trajectories for the full model (black), reduced model 1 (red) and reduced model 2 (yellow), starting from three initial data, corresponding to three different tropical equilibrations. The limit cycle attractor is contained in an invariant manifold. The reduced model 1 provides a good approximation of the dynamics on the invariant manifold (such as starting from initial data 3), but not outside. The reduced model 2 is accurate also outside the invariant manifold (see trajectories starting from equilibrations 1 and 2).

6.8 Discussions

The problem of timescale decomposition and model reduction of biochemical networks is addressed in this application. The approach described here relies on the notion of tropical equilibration. Tropical equilibrations represent a generalization of steady states and correspond to compensation of dominant fluxes acting on species concentrations. The remaining, uncompensated weaker fluxes are responsible for the slow dynamics of the system on attractive invariant manifolds. The correspondence between the tropical equilibrations and attractive invariant manifolds has been exploited here to associate, to each tropical equilibration, a reduced model. It is shown elsewhere (Samal et al., 2015b) that when there is an infinity of tropical equilibration solutions, these can be organised into branches and minimal branches, points on the same branch correspond to the same tropically truncated system.

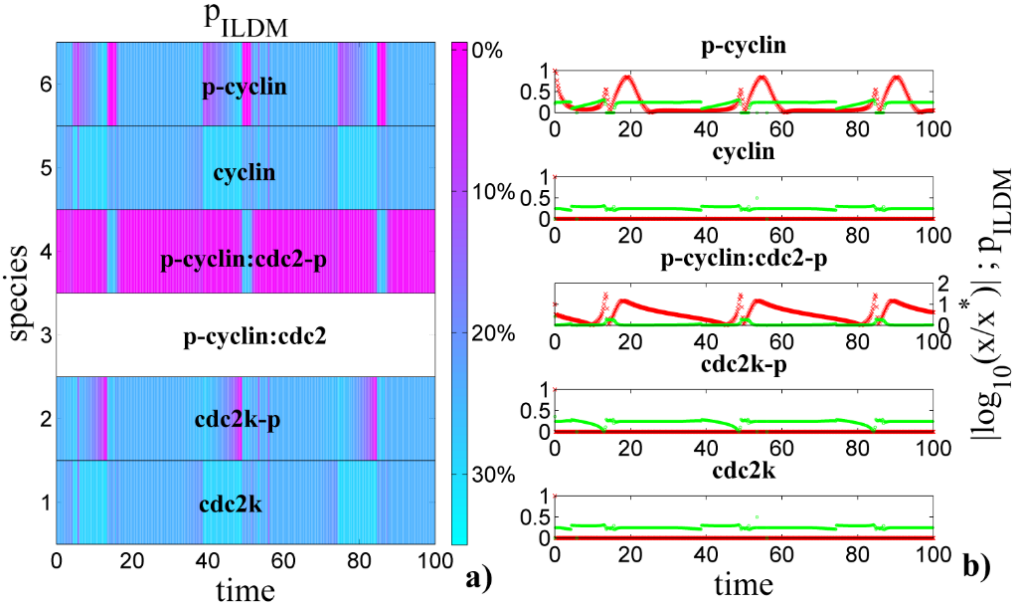


FIGURE 6.9: Summary of the analysis of model BIOMD00000005 using ILDM method and comparison with the method based on the slowness index. The model was simulated in COPASI from 0 to 100 min with default initial concentrations. The species 3 is eliminated using the single conservation law. For all the remaining species we represent the time dependence of their contributions p_i, where $0 \leq p_i \leq 1$, $\sum_i p_i = 1$ to the fast space. The fractions p_i are generated by COPASI and the values p_i are color coded in the left panel a). Chemical species with largest contribution to the fast space (largest p_i) are supposed to be fast species (as explained in (Surovtsova et al., 2009)). Therefore p_i (in green in right panel b)) and the slowness index $I_i = |\log_{10}(x_i/x_i^*)|$ (in red) should be anti-correlated. This is well verified for species $1, 4, 5$ (p_i are relatively high when I_i are close to zero, and close to zero when I_i are relatively high, but it is not well verified for species 6 (p-cyclin) whose slowness index has large peaks in places where the COPASI contribution to fast modes stays constantly high.

An algorithm is proposed to compute tropical equilibrations in Section 4.1, which is used to determine the fast/slow partition of chemical species in a network of biochemical reactions. This is the first and often a critical step in the model reduction algorithms. In particular, the number of slow species provides the size of minimal dynamic models to which complex biochemical networks can be reduced. The validity of the reductions presented above depends on concentration and parameter orders, as well as on initial data. For the simple example of Michaelis-Menten kinetics validity conditions are obtained for various reductions as inequalities among orders of magnitude of concentrations and parameters. These validity conditions define large domains in the concentrations and parameters spaces. Previous work on larger models suggests a larger applicability of this result which implies that the resulting reductions are robust (Radulescu et al., 2008).

The benchmarking of the algorithm on the Biomodels database shows that a significant dimension compression can be performed on cell dynamics models at timescales of $1000s$ and larger. Starting with complex models having more than 30 variables, the minimal models have median numbers of 2 slow variables. This suggests that, at least piece-wise in parameter and phase space, the tasks fulfilled by molecular networks are relatively simple. The need for having complex machineries with many regulators to perform simple tasks (such as relaxation to steady states or limit cycle oscillations) could be justified by the system robustness. A system having a very large number of variables and parameters, multiple timescales and only a few slow degrees of freedom is generically robust with respect to perturbations of variables and parameters (Gorban and Radulescu, 2007).

The described method can be also used to study sensitivity issues and identifiability of parameters from trajectories. A parameter is sensitive if changing its value induces large changes of model trajectories. In the analysis of Tyson's model it is seen that some parameters of the full model are not present in the reduced model. Although the orders of magnitude of these parameters are important (changing them may change the reduction), small changes of their values do not lead to changes of the model trajectories. Parameters of the full model that are not present in the reduced model are therefore insensitive. It may also happen that parameters of the reduced model are combinations (for instance multivariate monomials) of the parameters of the full model. If these combinations are sensitive, then so are the parameters they contain. However, parameters that are parts of such combinations can not be determined independently from the observed trajectories, which leads to parameter non-identifiability issues (Radulescu et al., 2012). Thus, insensitive parameters and parameter lumping resulting from model reduction can be used to asses local identifiability of the system parameters (Radulescu et al., 2012). The idea of relating parameter lumping and parameter identifiability can also be found in other computational algebraic geometry approaches (Meshkat, Eisenberg, and DiStefano, 2009).

Solving the tropical equilibration problem and finding a slow-fast decomposition is the first step for model reduction. The remaining steps consist of elimination of the fast variables by solving systems of algebraic equations. This has been shown for simple examples. In the case of more complex models, the elimination can be performed numerically, or symbolically. Tropical methods can simplify this task by replacing the full systems by tropically truncated systems. In particular, the binomial or toric case when the truncation has only two monomials is particularly interesting because for this case there are rapid algorithms for computing the steady states (Grigoriev and Weber, 2012; Millán et al., 2012). Higher approximation can be provided by Newton-Puiseux expansions (Radulescu, Vakulenko, and Grigoriev, 2015), that encompass tropical solutions in their lowest order. Although the calculations needed for formal reduction could be long, once the model is reduced, it can be used in various applications, such as a part of larger networks, or in models of tissues and organisms where the same biochemical network has to be replicated in several interacting cells. Furthermore, the reductions have a strong geometrical basis. In future work, this property will be exploited to show how to endow the reduced model with a reaction network structure and how to identify inclusion relations among different reduced models.

The slowest part of the dynamics often ends with a stable steady state. Tropical equilibrations that correspond to steady states should be on the tropical variety. The Puiseux series exist in this case and provide higher order approximations to the position of the steady state. However, tropical counterparts of stable invariant manifolds can be parts of prevariety without being in the variety. Only the last invariant manifold, the slowest in the hierarchy and the absorbing state in the automaton of metastable regimes (cf. Chapter 7) may be associated to parts of the variety (Radulescu, Vakulenko, and Grigoriev, 2015).

Several other open questions will be addressed in future work. For instance, the current algorithm finds the tropical equilibrations for fixed values of the parameters. It would be very interesting to formally classify all the possible reductions and phase portraits of a reaction network with a given topology and reaction rates, for all possible values of the parameters. The problem of Michaelis-Menten kinetics is solved by hand. In the future the algorithms will be extended in order to compute the tropical equilibration solutions depending on the parameters. For this purpose, the techniques used for *linear quantifier elimination* shall be extended (Weispfenning, 1988; Dolzmann and Sturm, 1997; Weber, Sturm, and Abdel-Rahman, 2011) and incorporated them into the current framework.

Chapter 7

Metastability, Symbolic Dynamics and Application to TGFβ Signalling

In this chapter the application of tropical minimal branches in determining the MRs is discussed. The MRs of smooth biochemical reaction networks representing the low dimensional regions of the phase space close to which the dynamics is much slower compared to the rest of the phase space. The transitions between the MRs are modelled using finite state automaton. The discrete states in such an automaton correspond to the tropical minimal branches and are referred to as metastable regimes. In addition, it will be shown that for TGFβ pathway distinct metastable regimes are found to be associated with the clinical phenotypes namely epithelium-like (non-aggressive) and mesenchymal-like (aggressive) cell lines.

7.1 Background

In order to cope with the qualitative data, boolean (Kauffman, 1969; Thomas, 1973) or multi-valued networks (Thomas, 1991; Naldi, Thieffry, and Chaouiya, 2007) are often used instead of continuous models. For this reason, many efforts were focused on coarse graining dynamical networks described by ordinary differential equations (ODE) to boolean networks (Davidich and Bornholdt, 2008). However, in spite of their advantages, dynamical properties of large Boolean or multi-valued networks are still difficult to study. The difficulty originates from the number of possible states, which for multi-valued networks with m levels (Boolean networks correspond to $m = 2$) is m^n. Although there are efficient methods for computing attractors of Boolean networks (algorithms based on binary decision diagrams or on satisfiability solvers can handle synchronous networks with hundreds of variables (Dubrova and Teslenko, 2011)), more intricate dynamical properties like the metastable regimes discussed here, ask for comprehensive analysis of the state transition graph which is hard to perform and moreover for analysis of the hierarchy of time scales which is even harder for the Boolean and multi-valued networks. The coarse graining method proposed here leads to a drastic reduction of the number of states. This offers unprecedented possibilities for qualitative analysis of the dynamics. This is illustrated in Fig. 7.1, where we plot the number of minimal branches versus the number of equations in the model. As can be noticed, this number is much lower than the number of states of a Boolean network with the same number of variables, which illustrates the advantage of our coarse graining with respect to other methods that discretize the values of the variables in order to obtain Boolean or multi-value networks (Davidich and Bornholdt, 2008).

Here a new method is proposed for model analysis that uses coarse grained descriptions of continuous dynamics as discrete automata defined on finite states. These states will not be obtained by discretization of network variables, but by identification of a finite number of collective modes describing possible coordinated activity of several variables.

For large networks with ODE based dynamics and multiple timescales it is reasonable to consider the following property: a typical trajectory consists of a succession of qualitatively different slow segments separated by faster transitions. The slow segments, generally called

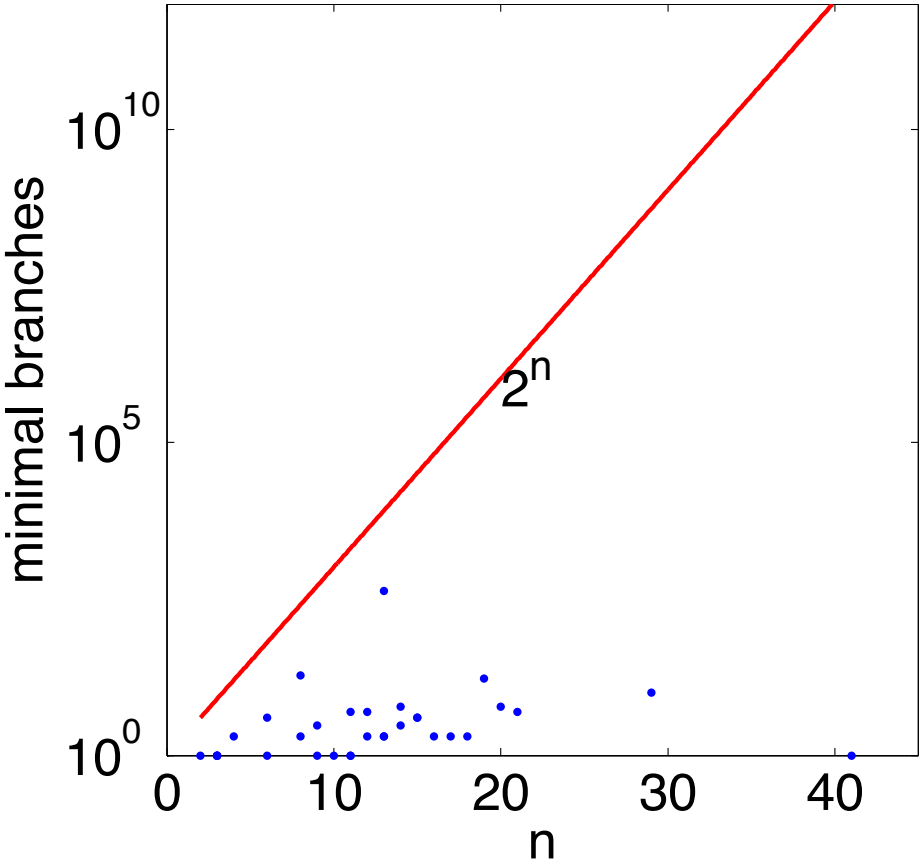


FIGURE 7.1: Semi-log plot of the minimal branches versus the number of equations in the models from Biomodels repository for $\varepsilon = 1/5$. Comparison with a binary network number of states 2^n suggests sub-exponential scaling.

metastable states or regimes, can be of several types such as the attractive invariant manifolds (Gorban and Karlin, 2005), Milnor attractors (Rabinovich et al., 2006) or saddle connections (Rabinovich et al., 2012). The notion of metastability generalizes the notion of attractor. Like in the case of attractors, distant parts of the system can have coordinated activity for metastability. The dynamical states of large networks can be represented as points in a high dimensional space, called phase space. In this representation each coordinate represents the concentration of a molecular species. Coordinated activity means that many of the species concentrations are correlated, which can be geometrically represented by the proximity to a lower dimension hypersurface in the phase space. The application on model reduction discussed in the previous section exemplified the computation of such lower dimensional hypersurface. A system remains in the proximity of an attractor after entering its basin of attraction, but can leave a metastable regime after a relatively long time (much longer than the time needed for transitions between two different regimes). Fig. 1.1 summarizes this geometrical picture. The term *crazy-quilt* was coined to describe such a patchy landscape of multiscale networks dynamics (Gorban and Radulescu, 2008).

This phenomenon, called itinerancy, received particular attention in neuroscience (Tsuda, 1991). Itinerant behaviour is shown by mathematical models with stable heteroclinic sequences (defined as open chains of saddle fixed points connected by one-dimensional separatrices) and was also observed in transient activity of antennal lobe neurons involved in insect olfaction or in the activity of high vocal centers controlling songbird patterns (Rabinovich et al., 2006). It can be believed that similar phenomena occur in molecular biology for chemical reaction networks.

A well studied example sustaining this picture is the biochemical and gene expression dynamics guiding the orderly progression of the cell cycle. The main feature of this dynamics is the sequential activation of cyclin dependent kinase/cyclin complexes. More than 30 years since cyclins were discovered the main cell cycle control events are now well understood and it is agreed that each of them involve the collective action of several regulator molecules. In addition, studies of periodic gene expression in synchronized cell division of yeast show the existence of waves of coordinate expression corresponding to different cell cycle phases or transitions (Rustici et al., 2004). Furthermore, mathematical models of the cell cycle machinery (currently more than 150 published models (Weis et al., 2014)) illustrate the stage dependent coordination of biochemical variables. As an example, the structure of steady state branches of the Wee1-Cdc25 module in yeast lead (Tyson, Csikasz-Nagy, and Novak, 2002) to consider that the exit from mitosis is a collective phenomenon that can be described as a saddle-node bifurcation. The analysis of such models also showed that branches and bifurcations of states occur naturally in the context of cell cycle modelling (Noel et al., 2012). In a more general context, geometric analysis of single-cell expression data from human and mouse tissues showed that gene expression is structured not only in clusters but also in continua of states within polyhedra whose vertices can be understood as specialized key tasks (Korem et al., 2015). These findings were interpreted in terms of multi-objective optimization solutions (Korem et al., 2015), but could also suggest transient behaviour between specialized states. The idea of associating cell lineage commitment to collective behaviour of gene networks variables was used in many other contexts including cancer genomics where it was proposed that cancer cells are trapped in some abnormal attractors (Huang, Ernberg, and Kauffman, 2009).

In an ODE system, when there is only one dominant term or when the dominant terms have all the same sign, the dynamics is fast and the system tends rapidly towards a region in phase space where at least two dominant terms of opposite signs are equilibrated. Here, such regions will be referred to as metastable regimes and the machinery of tropical geometry methods will be utilised to compute them. Specifically, the tropical minimal branches will be treated as proxies for the metastable regimes and will be used to describe the qualitative network dynamics as a sequence of transitions from one minimal branch to another. The complexity of the qualitative dynamics depends on the number of branches. In order to test it, the minimal branches are computed for a selection of models from Biomodels database (cf. Table 4.2). Although there are theoretical results suggesting that the number of branches should be small, these results are valid in the average in the probabilistic space of all the models. Furthermore, the biological significance of metastable states in the context of TGFβ signalling in cancer cell lines is also investigated.

7.2 Approach

For a given biochemical reaction network, one can try to reconstruct the transitions happening between the metastable regimes as illustrated in Fig. 1.1 by computing the tropical minimal branches (cf. computed using the algorithm explained in Section 4.1). The minimal branches are treated as proxies for the metastable regimes. By abstraction, it can be considered that the metastable regimes (patches in the "crazy-quilt" picture in Fig.1.1) are represented as nodes of a graph. Two nodes in the graph are connected if and only if there is at least one transition from one regime to the other. This can be viewed as an abstraction of a finite state machine, because the number of regimes is finite. However, given one regime, there may be several different possibilities to leave this regime, each leading to a different metastable regime. The choosing of a particular transition depends on the initial data. Thus, although the initial model is deterministic the finite state machine abstraction is generally stochastic. From one node, several transitions can leave each one having a different probability per unit time. Because the

initial system spends a long time on the metastable regime and little time on the transitions, it is natural to expect that the memory of previous transitions is lost and that the stochastic process is Markovian. In the following a method is proposed to learn this stochastic, Markov process, from many simulations of the full model, each one starting from a different, randomly chosen, initial state. The method is similar to the Markov state models (Bowman et al., 2009) used to coarse grain phase space for protein folding molecular dynamics. Contrary to energy landscape in protein folding, the proposed "crazy-quilt" here is not that rugged as it consists of a small number of patches. Therefore, standard Monte Carlo procedures function well in this case.

In order to construct the finite state machine, the phase space of the continuous model needs to be mapped to a finite set of branches. First, minimal branches of tropical equilibrations were computed as subsets of the euclidian space $\mathbb{R}^n$ where n is the number of variables. Tropical equilibrations minimal branches are stored as matrices $A_1, A_2, \ldots, A_b$, whose lines are tropical solutions within the same branch. Here b is the number of minimal branches.

As the method computes numerical approximations of the tropical prevariety, given a value of ε, this approximation is better when the denominator bound d is high. At fixed d, the dependence of the precision on ε follows more intricate rules dictated by Diophantine approximations. For this reason, the number b is systematically tested and the truncated systems corresponding to minimal branches are robust when changing the value of ε.

Trajectories $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))$ of the smooth dynamical system are generated with different initial conditions, chosen uniformly. For each time t, the Euclidian distance $d_i(t) = \min_{\boldsymbol{y}\in A_i} \|\boldsymbol{y} - \log_\varepsilon(\boldsymbol{x}(t))\|$, is computed where $\|*\|$ denotes the Euclidean norm and $\log_\varepsilon(\boldsymbol{x}) = (\log x_1/\log(\varepsilon), \ldots, \log x_n/\log(\varepsilon))$. This distance classifies all points of the trajectory as belonging to a tropical minimal branch. The result is a symbolic trajectory $s_1, s_2, \ldots$ where the symbols s_i belong to the set of minimal branches. In order to include the possibility of transition regions an unique symbol t is included to represent the situations when the minimal distance is larger than a fixed threshold. The choice of this threshold is robust (see discussion in Section 7.3 and Fig. 7.3). The residence times $\tau_1, \tau_2, \ldots$ that represent the time spent in each of the state are also stored.

Special care should be taken when the model has a number of conservations laws. A conservation law is a linear combination of species concentrations that is kept constant during the dynamics, in other words an equation of the type $c_1^i x_1 + c_2^i x_2 + \ldots + c_n^i x_n = K_i$. It is supposed that the conservation laws are semi-positive (all c_i are positive or nought). Then, several such equations together with the positivity conditions $x_i \geq 0$ define a polyhedron. The initial conditions for the trajectories $x(t)$ uniformly in such a polyhedron needs to be picked.

To start, each component j can take its initial concentration in the range as shown below

$$[0, \min(K_i/c_j^i, \text{ such that } c_j^i > 0)] \tag{7.1}$$

but as initial concentration of one component is set, the range available for components which are involved in the same constraint is reduced. The last component picked for each constraint must take the maximal value. Thus the sampling depends on the order in which initial concentrations are selected. To avoid introducing a bias related to this order, a random ordering of components is selected for each random initial state. Furthermore, if a component is picked as last in a constraint, its value is enforced (it must take all what remains). Conflicts may arise if this component is part of other constraints as well. To avoid this, it is ensured that the last assigned item is specific to the constraint (not involved in any other constraint). This step may create problems with highly interdependent sets of constraints or constraints with less than two specific components. For the models considered here, this sampling works well as constraints have more specific components than overlapping ones.

By this method, N symbolic trajectories of length M defining the vectors of successive states are generated $(s_1^j, s_2^j, \ldots, s_M^j)$ and residence times $(\tau_1^j, \tau_2^j, \ldots, \tau_M^j)$, where $j \in [1, N]$.

The stochastic automaton is learned as a homogenous, finite states, continuous time Markov process, defined by the lifetime (mean sojourn time) of each state T_i, $1 \leq i \leq b$ and by the transition probabilities $p_{i,j}$ from a state i to another state j. Following estimators for the lifetimes and for the transition probabilities are used:

$$T_i = (\sum_{n=1}^{N}\sum_{m=1}^{M} \tau_n \mathbb{1}_{s_m^n=i})/(\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{s_m^n=i}) \tag{7.2}$$

$$p_{i,j} = (\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{s_m^n=i, s_{m+1}^n=j})/(\sum_{n=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{s_m^n=i}),\ i \neq j, \tag{7.3}$$

where $\mathbb{1}_C$ stands for the indicator function equal to one if condition C is fulfilled or else equal to nought.

7.3 Biological significance of metastable branches for TGF-β signalling

As a case study, a nonlinear model of dynamic regulation of Transforming Growth Factor beta TGF-β signalling pathway that is recently described by (Andrieux et al., 2012) is considered. TGF-β signalling occurs through association of the ligand with TGF-beta type I (TGFBR1) and type II (TGFBR2) serine/threonine kinase receptors. TGF-β binding to TGFBR2 induces recruitment and phosphorylation of TGFBR1, which in turn transmits the signal through phosphorylation of SMAD2 transcription factor. Once phosphorylated, the SMAD2 hetero-dimerizes with SMAD4 and the complexes then migrate to the nucleus, where they regulate the transcription of TGF-β-target genes. In that context, the Transcriptional Intermediary Factor 1, TIF1-γ have been shown to function either as a transcriptional repressor or as an alternative transcription factor that promote TGF-β signalling. The apparent controversial effect of TIF1-γ on regulation of the SMAD-dependent TGF-β signalling was solved by a model integrating a ternary complex associating TIF1-γ with SMAD2 and SMAD4 complexes. This model has a dynamics defined by $n = 18$ polynomial differential equations and 25 biochemical reactions (the full set of ordinary differential equations can be found in the Section A.1 in Appendix A, the reaction scheme can be found in (Andrieux et al., 2012)). The computation of the tropical equilibrations for this model shows that there are 9 minimal branches. The connectivity graph of these branches (defined in Section 4.1) and the learned finite-state automaton (obtained with the method of Section 7.2) are shown in Fig. 7.2. The Table 7.1 illustrates the convergence of the estimated transition probabilities when N, the number of Monte-Carlo samples, is increased. In order to illustrate the robustness of the classification of the states in the trajectories for each branch B_i, $i \in [1, 9]$ we computed the distribution of probability of Euclidian distances between randomly chosen states of the model, compatible with the conservation laws, and the branches. In Fig. 7.3 these distributions are compared with the distribution of minimal distances used to classify various states on the model's trajectories. The latter distances are smaller and clearly separated from the random states distances.

The transition probabilities of the automaton are coarse grained properties of the statistical ensemble of trajectories for different initial conditions (cf. Section 7.2). Given a state and a minimal branch close to it, it will depend on the actual trajectory to which other branch the system will be close to next. However, when initial data and the full trajectory are not known, the automaton will provide estimates of where we go next and with which probability. For the example studied and for nominal parameter values, the branch B1 is a globally attractive

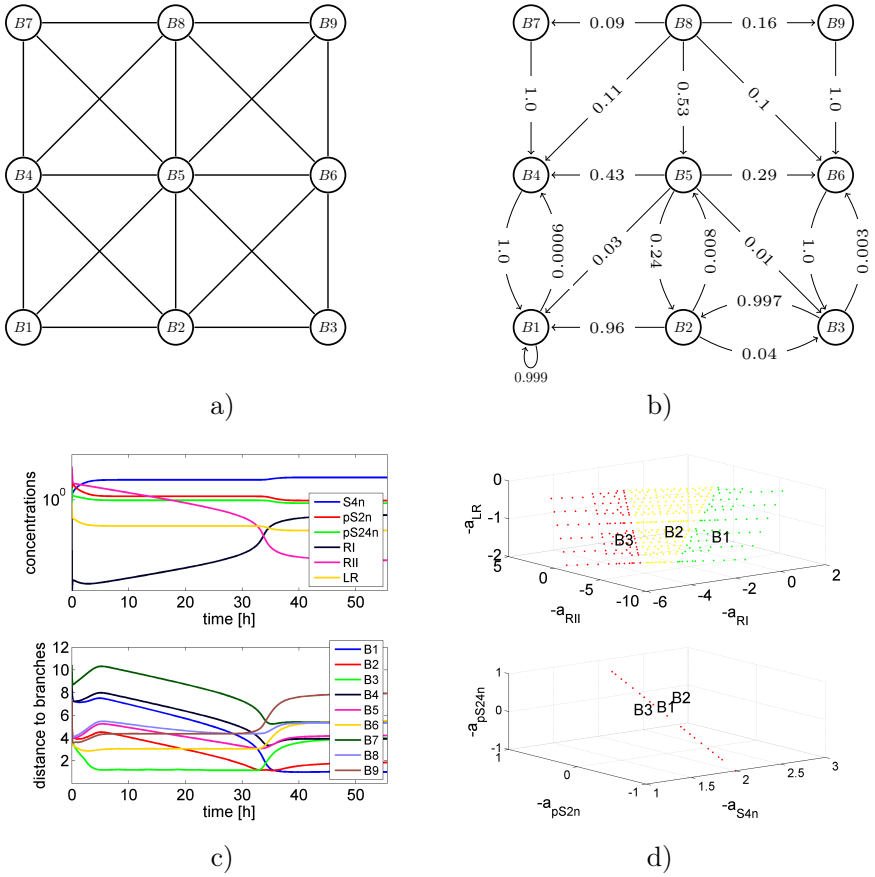


FIGURE 7.2: Summary of tropical geometry analysis of the TGFβ model. a) Connectivity graph of tropical minimal branches; b) finite state automaton; c) a single trajectory of the system (starting from initial data chosen randomly close to the branch B_3) is represented by plotting the concentration of different species vs. time (upper sub-figure); the distances to different branches of solutions vs. time (lower sub-figure) shows that the sequence of branches for this trajectory is B_3, B_2, B_1 (all points of the trajectory are close to one of these three branches and significantly more distant to the other branches). d) The different branches of solutions are defined by allowed concentrations of different variables, represented here by orders of magnitudes a_i; the opposite concentration orders $-a_i$ are proportional to the logarithms of concentrations $-a_i \sim \log(x_i)$. All the order calculations were performed using $\varepsilon = 1/11$. The most used branches B_1, B_2, B_3 are shown in projection onto sets of three variables. The variables RI, RII, LR are plasma membrane receptors and ligand-receptor complex (signalling input layers), whereas pS2n, S4n, pS24n are nuclear transcription factors and complexes (effectors). The structure of tropical branches shows that composition of input layers is more flexible (varies on planar domains that are disjoint for different branches) than the concentrations of effectors (vary on linear intervals that overlap for different branches).

sink: starting from anywhere, the automaton will reach B1 with probability one. This branch contains the unique stable steady state of the initial model. This calculation illustrates the basic properties of minimal branches of equilibrations. Trajectories of the dynamical system can be decomposed into segments that remain close to minimal branches. Furthermore, all the observed transitions between the branches are contained in the connectivity graph resulting from the polyhedral complex of the tropical equilibration branches. This result proves the solidity of the tropical approach, because the geometric connectivity was not enforced to constrain the possible transitions; the fact that it is really the case that emerges from the analysis of the trajectories. A change of parameter values can have several consequences: change the connectivity graph, change of the probabilities of transitions and change of the attractor position.

TABLE 7.1: Estimation of the transition probabilities between branches of the TGFβ model, for different values of N, the number of Monte-Carlo samples.

N	1→1	1→4	2→1	2→3	2→5	3→2	3→6	4→1	5→1	5→2	5→3
100	1.0000	0.0000	0.9836	0.0164	0.0000	1.0000	0.0000	1.0000	0.0227	0.2727	0.0227
200	1.0000	0.0000	0.9675	0.0325	0.0000	1.0000	0.0000	1.0000	0.0217	0.2391	0.0109
500	1.0000	0.0000	0.9614	0.0386	0.0000	1.0000	0.0000	1.0000	0.0216	0.2381	0.0173
1000	1.0000	0.0000	0.9703	0.0264	0.0033	1.0000	0.0000	1.0000	0.0280	0.2258	0.0086
2000	1.0000	0.0000	0.9585	0.0365	0.0050	1.0000	0.0000	1.0000	0.0246	0.2173	0.0086
3000	1.0000	0.0000	0.9561	0.0389	0.0050	1.0000	0.0000	1.0000	0.0295	0.2342	0.0084
5000	0.9996	0.0004	0.9542	0.0399	0.0059	0.9996	0.0004	1.0000	0.0340	0.2460	0.0093
10000	0.9994	0.0006	0.9556	0.0363	0.0081	0.9973	0.0027	1.0000	0.0344	0.2370	0.0110

N	5→4	5→6	6→3	7→4	8→4	8→5	8→6	8→7	8→9	9→6
100	0.4091	0.2727	1.0000	1.0000	0.1373	0.4902	0.0784	0.0784	0.2157	1.0000
200	0.4130	0.3152	1.0000	1.0000	0.1300	0.5200	0.0600	0.0600	0.2300	1.0000
500	0.4156	0.3074	1.0000	1.0000	0.1098	0.5294	0.0784	0.0980	0.1843	1.0000
1000	0.4280	0.3097	1.0000	1.0000	0.1051	0.5333	0.0949	0.0909	0.1758	1.0000
2000	0.4497	0.2998	1.0000	1.0000	0.1121	0.5354	0.1020	0.0859	0.1646	1.0000
3000	0.4439	0.2840	1.0000	1.0000	0.1122	0.5315	0.0984	0.0951	0.1627	1.0000
5000	0.4227	0.2880	1.0000	1.0000	0.1160	0.5272	0.0969	0.0965	0.1633	1.0000
10000	0.4304	0.2872	1.0000	1.0000	0.1132	0.5344	0.0997	0.0932	0.1595	1.0000

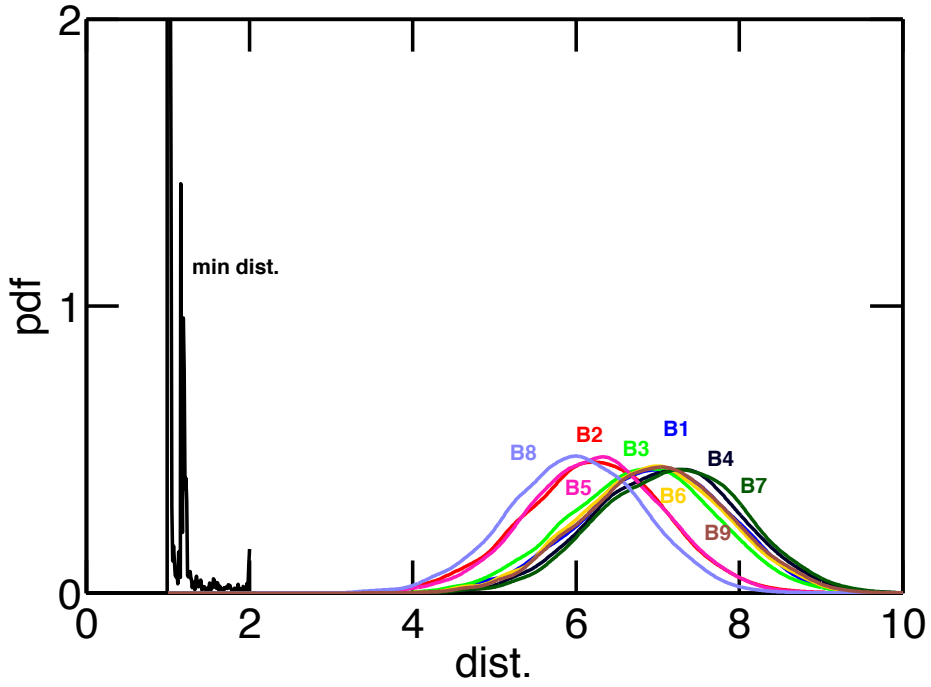


FIGURE 7.3: Distribution of distances to tropical branches of the TGFβ model. For each branch B_i, $i \in [1,9]$ we represent the probability density function of Euclidian distances between randomly chosen states of the model, compatible with the conservation laws, and the branches ($dist = \min_{i\in[1,9]} d_i$). The minimal distances computed for states on model's trajectories ($mindist$) are smaller and clearly separated from the random states distances. In order to compute the transitions represented in Fig.7.2b) we have used the threshold $dist = 3$ (states with $dist > 3$ were declared transient, whereas states with $dist < 3$ were classified as belonging to some branch.)

In order to understand the significance of the minimal branches and their relation with dynamic and physiologic properties of the network an analytic study of the tropical equilibration solutions is performed. It is shown (see also the Section A.2 in Appendix A) that the most

important cause of the multiplicity of branches is the dynamics of the TGFBR1 and TGFBR2 receptors whose internalization and trafficking regulates TGF-β signalling (Le Roy and Wrana, 2005). These two receptors belong to a ligand-receptor module of 6 variables and 12 reactions that is decoupled from the rest of the network. More precisely, the ligand-receptor module activates the SMAD transcription factors but receives no feed-back (see Fig. 7.4) and can be studied independently from the rest of the variables. This module has been used with little variation in many models of TGF-β signalling (Vilar, Jansen, and Sander, 2006; Zi and Klipp, 2007; Chung et al., 2009).

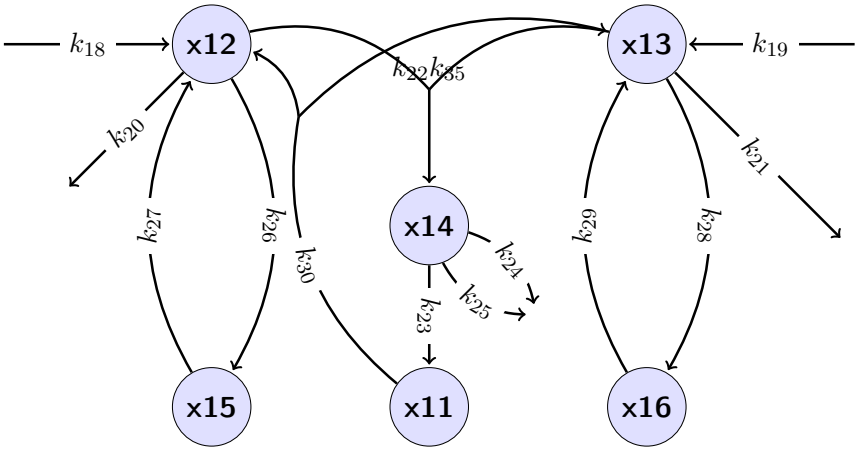


FIGURE 7.4: Graphic representation of the ligand-receptor module of the TGF-β full model. The different variables mean: x_{12} : RI (TGBR1), x_{13} : RII (TGFBR2), x_{14} : LR (ligand-receptor complex), x_{15} : RIe (TGFBR1 in endosome), x_{16} : RIIe (TGFBR2 in endosome), x_{11} : LRe (LR in endosome).

It is shown in the Section A.2 in Appendix A that the tropical equilibration of the ligand-receptor module form a two dimensional polyhedron conveniently parametrized by the concentration orders a_{12} and a_{13} of the receptors TGFBR1 and TGFBR2 respectively. The branches can be calculated analytically ((A.9) and (A.13) in Section A.2 in Appendix A). For the nominal values of the model parameters one of these branches is empty and the three remaining branches correspond to B_1, B_2 and B_3. The two other triplets of branches (B_4, B_5, B_6) and (B_7, B_8, B_9) correspond to the same mutual relations of variables in the ligand-receptor module. They are distinguished by the values of the remaining variables (the transcription factors module). As the computation of the automaton showed that the branches B_i, $i \in [4, 9]$ are practically inaccessible from states in branches B_i, $i \in [1, 3]$, therefore we will not discuss them here.

Symbolic computation is used to determine the steady states of the ligand-receptor module. This module has an unique steady state corresponding to concentrations orders that can be placed inside the polyhedron of tropical solutions using the (4.3). The minimal branch containing the steady state is a sink of the coarse grained dynamics. The polyhedron of tropical solutions, its decomposition into minimal branches, and the position of the steady states inside it, depend on the model parameters. Among model parameters two are important: k_{18} and k_{19} representing the production rate of the protein receptors TGFBR1 and TGFBR2, respectively. Consequently, these two parameters are correlated to gene expression and account for possible variability in mRNA levels of the two types of receptors. Fig. 7.5 shows the tropical equilibration branches of the ligand-receptor modules for various parameters k_{19} corresponding to various TGFBR2 expression levels. For the nominal parameters used in the model, the branch $B1$ is a sink i.e. an attractor (the coarse-grained dynamics shows that the probability to leave this state is negligible), and the branches $B2$ and $B3$ are transient i.e. metastable. This means that starting in the branch $B2$ or $B3$ the receptor module will reach the branch $B1$ after a

certain time and will remain there. However, over-expression of TGFBR2 modelled by changing the parameter k_{19} and illustrated in the Fig. 7.5a-c can tilt the balance in favor of large concentrations of receptor of type 2 corresponding to branches $B2$, $B3$ of tropical solutions in the model. Interestingly, this change occurs by a displacement of the position of the steady state from $B1$ to $B2$ and $B3$ and not by a change of the concentration values allowed for these branches.

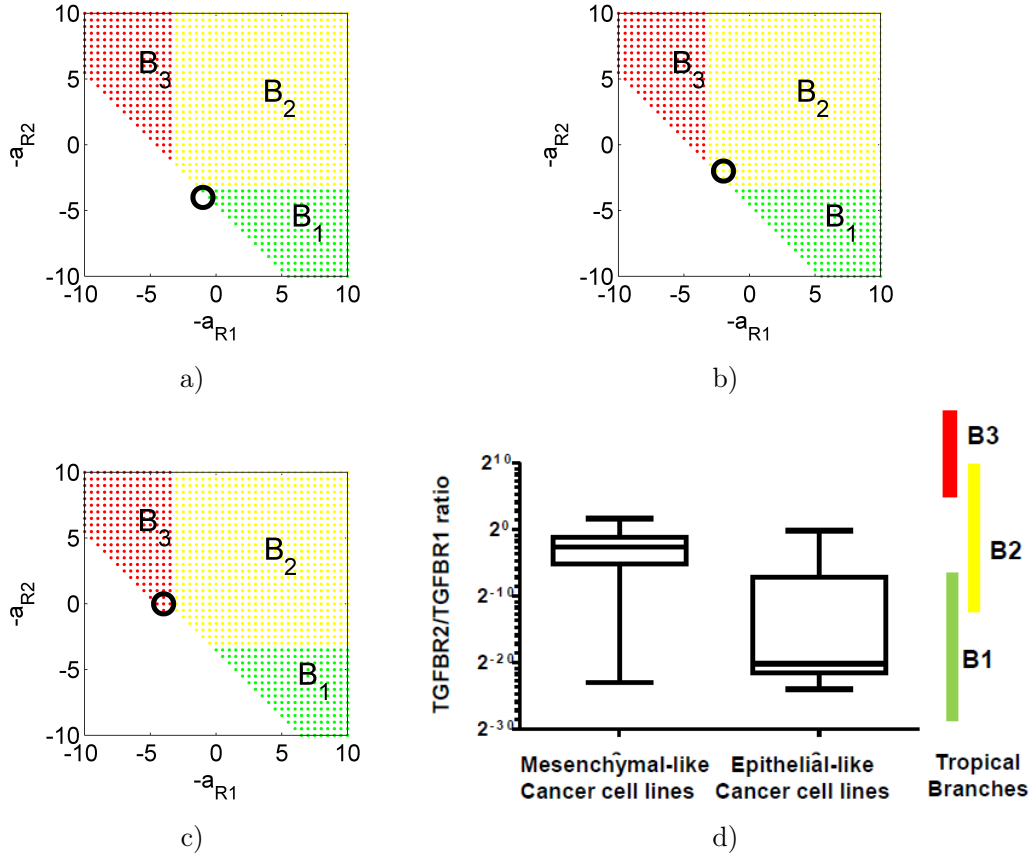


FIGURE 7.5: Tropical equilibrations of the ligand-receptor modules for various values of TGFBR2 (R2) gene expression are represented in projection in the plane (a_{R1}, a_{R2}) (a point in this plane provides the orders of concentrations of the protein receptors) (a,b, and c) and comparison with proteomics data from (Gholami et al., 2013) (d). Branches of tropical equilibrations are calculated for a) nominal value k_{19} (TGFBR2 expression) (this is the same as Fig. 7.2d in projection onto the plane (a_{R1}, a_{R2})), b) $\times 2$ TGFBR2 overexpression , and c) $\times 10$ TGFBR2 overexpression. The circle represents the position of the stable steady state and the branch containing is an attractor of the finite-state automaton. Like in Fig. 7.2 large opposite concentration orders $-a_i$ indicate large concentrations. All order calculations were performed using $\varepsilon = 1/11$. d) Proteomic data from NCI-60 cancer cell lines. Aggressive lines cover ratios of receptor concentrations intervals (indicated as bars at the right side of the sub-figure) corresponding to branches B3 (red) and B2 (yellow), whereas non-aggressive lines correspond to low expression of TGFBR2 in branch B1 (green). The receptors concentration ratios are well separated in the two classes (Mann-Whitney test, p-value 0.0006).

While Vilar et al. (Vilar, Jansen, and Sander, 2006) have speculated that the ligand-receptor module is responsible for the versatility of the response of the TGF-β pathway, no experimental evidence support this hypothesis. Here, it is demonstrate that there are correlations between dynamical specificity characterized by membership to a particular branch of equilibration and cell phenotype. Such a comparison for the NCI-60 panel of cancer cell lines is illustrated here, a

well established tool for tumor comparison and drug screening provided by the National Cancer Institute. Based on microarray analysis, these cell lines were found to cluster into two classes: epithelium-like (non-aggressive) and mesenchymal-like (aggressive) cell lines (Ross et al., 2000).

Using the global proteome analysis of this NCI-60 panel (Gholami et al., 2013), the protein expression levels of TGFBR1 and TGFBR2 are extracted which showed that the mesenchymal-like (aggressive) cell lines can be distinguished from the epithelial-like (non-aggressive) cell lines by the increased level of TGFBR2 (Fig 7.5d.

The proteome data was compared to membership to tropical branches. According to (4.3) there is a linear relation between opposite concentration orders $-a_i$ and logarithms of concentrations, $-a_i \approx b\log(x_i)$, $i = 1, \ldots, n$ $(b = -1/\log(\varepsilon) > 0)$. Opposite concentration orders $-a_i$ were used instead of a_i because they change in the same direction as the concentrations (small opposite orders mean small concentrations and large opposite orders mean large concentrations). Therefore, in Fig 7.5a-c the relation $TGFBR1 = TGFBR2$ is verified on the bissector of the first quadrant, whereas $TGFBR2 > TGFBR1$ and $TGFBR2 < TGFBR1$ are valid above and below the bissector, respectively.

When compared the proteome results with the membership to a particular branch of equilibration, it is found that the distribution of concentration orders in branches place non-aggressive cancer cell lines in a range covered by branch 1, whereas the aggressive cancer cell lines are placed in a range covered by branches 2 and 3 (Fig 7.5d). Indeed, the ratio $TGFBR2/TGFBR1$ is small for the branch B_1 and in non-aggressive cancer cell lines, and is much larger for B_2, B_3 and in aggressive cell lines.

Furthermore, the association of up-regulation of TGFBR2 with mesenchymal-like appearance in an independent dataset of 51 breast cancer cell lines (Neve et al., 2006) is validated. In a recent work (Ruff et al., 2015), comparative analyses between Basal B cell lines with mesenchymal-like phenotype and Basal A and Luminal cell lines with epithelial morphology permitted to identify more than 600 differentially expressed genes that include TGFBR2. Gene expression data were now extracted for TGFBR1 and TGFBR2 and we showed that TGFBR2 gene expression is significantly induced in mesenchymal-like cell lines while TGFBR1 did not vary (Supplementary Fig.1). In accordance with to observation discussed here, Parker et al. (Parker et al., 2007) have previously reported the association of low TGFBR2 expression with a lower aggressive tumor phenotype.

In summary, the tropical geometry analysis of the TGFβ signalling model first shows that the multiplicity of branches is due to the dynamics of TGFβ receptors. The more important parameters in this ligand-receptor module are the concentration of TGFBR1 and TGFBR2 and three main tropical branches are distinguished by the value of TGFBR2/TGFBR1 ratio (small in B1, intermediate in B2, large in B3). Importantly, we showed that the TGFBR2/TGFBR1 ratio is associated with tumor cell lines phenotype (high and low TGFBR2/TGFBR1 ratio for aggressive and non aggressive cell lines, respectively). Together these data demonstrated that tropical geometry analysis permits to discriminate between cellular states based on the evaluation of TGFβ receptors concentration. The importance of such up-regulation of TGFBR2 in aggressive cancer cell lines might be related to its implication in SMAD-independent signalling that includes PI3K-Akt, JNK, p38MAPK and Rho-like GTPases and which highly contribute to epithelial-mesenchymal transition (Zhang, 2009; Moustakas and Heldin, 2012).

Together these observations suggest that metastable regimes defined by branches of minimal tropical equilibrations are associated with cell phenotypes. The idea of associating tropical minimal branches with clinical phenotype is similar to the idea of cancer attractors (Huang, Ernberg, and Kauffman, 2009) where the idea is that cancer cells are trapped in some abnormal attractors.

7.4 Discussions

A method is presented to coarse grain the dynamics of a smooth biochemical reaction network to a discrete symbolic dynamics of a finite state automaton. The coarse graining was obtained using a tropical geometry approach to compute the states. These states correspond to metastable dynamic regimes and to relatively slow segments of the system trajectories. The coarse grained model can be used for studying statistical properties of biochemical networks such as occurrence and stability of temporal patterns, recurrence, periodicity and attainability problems.

Further improvement and evolution is possible for this approach. First, the coarse graining can be performed in a hierarchical way. For the nonlinear example studied in the paper only the full tropical equilibrations are computed. These stand for the lowest order in the hierarchy (coarsest model). As discussed in Section 6.3.1 partial equilibrations can be considered when slow variables are not equilibrated and thus refining the automaton. Generally, there are more partial equilibrations than total equilibrations and learning an automaton on the augmented state set will produce refinements. Second, and most importantly, the dynamics within a branch could also be described. As shown in Chapter 6, reductions of the systems of ordinary differential equations are valid locally close to tropical equilibrations. Furthermore, the same reduction is valid for all the equilibrations in a branch. This suggests that a hybrid approach, combining reduced ODE dynamics within branch with discrete transitions between branches is feasible. The transitions can be autonomously and deterministically commanded by crossing the boundaries between branches that are perfectly determined by the current approach.

The most important result of this application is the extension of the notion of attractor to metastable regimes of chemical reaction networks and the proposition of a practical recipe to compute metastability. Metastable regimes correspond to low-dimensional hypersurfaces of the phase space, along which the dynamics is relatively slower. Most likely, metastable regimes have biological importance because the network spends most of its time in these states. The itinerancy of the network, described as the possibility of transitions from one metastable regime to another is paramount to the way neural networks compute, retrieve and use information (Tsuda, 1991) and can have similar role in biochemical networks. The approach based on tropical geometry provides an algorithmic method to test these ideas further. The extension of this approach i.e. making use of statistical methods to compute the association of the tropical minimal branches with clinical phenotypes based on "-omics" data remains a topic of future research.

Chapter 8

Pathway based modelling of Biochemical Reactions Networks

In this chapter, the ECs are associated with different clinical phenotypes or outcomes. A method based on statistical modelling is discussed which integrates the gene expression profile of a biological system with the ECs. ECs represent the pathways of a biochemical network model.

8.1 Background

Metabolic networks are usually represented as a collection of enzyme catalysed biochemical reactions. There are numerous databases with such information (Kanehisa and Goto, 2000; Croft et al., 2011; Le Novére et al., 2006). One way to study such a network is by decomposing the network into sub-networks or pathways in an unbiased manner. Such pathways represent different metabolic routes for the production of given metabolites along with the essential enzymes. Extreme Currents (ECs) (Clarke, 1988), Elementary Flux Modes (EFMs) (Schuster, Fell, and Dandekar, 2000) and Extreme Pathways(EPs) (Schilling, Letscher, and Palsson, 2000) are three widely used techniques in this context. The common assumption is that the underlying biochemical reaction system is in a steady state, and with additional constraints on reaction fluxes, the solution space can be represented as a polyhedron. The vertices of such a polyhedron have the biochemical interpretation of being pathways in the network. A comparison between different metabolic pathway techniques can be found in (Llaneras and Picó, 2010). One major drawback of pathway enumeration is that the number of pathways can explode in a combinatorial fashion with the size of the network (Klamt and Stelling, 2002). Hence, for larger networks optimization techniques are frequently used. For an overview (Rezola et al., 2015) is recommended.

In short, the decomposed pathways are basically steady state reaction-flux distributions capturing wide range of behaviour that the network is capable to display. Essentially, these pathways are invariants of the network and hence do not require kinetic rate parameters to be known or estimated, which is often very difficult. In biological systems the functioning of metabolic networks is regulated by gene and protein expression levels which affect the enzyme concentrations. Therefore, out of many decomposed pathways some are more associated with a given phenotype than others.

Given, such a network the ECs are computed, represented as a gene set and the ones that are associated with given clinical or biological phenotypes based on gene expression data are identified. This translated to the question of performing an enrichment of gene sets with respect to gene expression data. For this purpose, there exists several methods in the bioinformatics community. Such methods can be broadly divided into "self-contained" or "competitive" methods. A detailed overview can be found in (Maciejewski, 2014; Nam and Kim, 2008).

The self-contained method considers only the genes mappable to a pathway of interest while computing the association with the phenotype and disregarding the genes not in that gene set.

The null hypothesis in self-contained methods is that no gene in the gene set is associated with the phenotype. A prominent example is the global test (GT) (Goeman et al., 2004). The GT tries to reject the null hypothesis that all coefficients in a generalized linear model linking gene expression and phenotype are exactly zero.

In competitive methods each gene set of interest is viewed together with the whole universe of existing genes. The most prominent example is the GSEA method (Subramanian et al., 2005). The null hypothesis in competitive methods is that genes in the given gene set are "at most as often differentially expressed" between conditions of interest as the genes in the complement of the given gene set.

Existing approaches estimate the association of EFMs with a phenotype based on high-throughout "-omics" data using a hypergeometric test which is a competitive method. In (Schwartz et al., 2007) the enrichment of EFMs computed from KEGG pathways are associated with gene expression data in various stress conditions of *Saccharomyces cerevisiae*. The approach in (Rezola et al., 2013) uses a multivariate hypergeometric test to link EFMs in a genome scale human metabolic network to tissue specific gene expression differences in human. This technique was also applied to study metabolic differences between lung cancer subtypes (Rezola et al., 2014). The methods based on hypergeometric test require a threshold to discretise the genes into two or more classes based on their expression level differences in different biological or clinical outcomes. Therefore, they are sensitive to such a threshold. Moreover, the application of hypergeometric test is limited to discrete phenotypes (e.g. cancer or healthy) and not applicable to continuous clinical outcomes e.g. survival times. Finally, the universe of all genes has to be defined appropriately (Maciejewski, 2014) which could be tricky. A different approach was undertaken by YANA (Schwarz et al., 2005) which uses optimization to minimise the sum of squared error between reaction fluxes in EFMs and the fluxes based on gene or protein expression. A review on existing approaches can be found in (Rezola et al., 2015).

The method proposed here is based on a generalized linear model with l_1 regularized norm called sparse group lasso (SGL) (Simon et al., 2013) to identify phenotype associated ECs based on gene expression data. SGL selects a sparse set of feature groups and also introduces sparsity within each group. Features in the model are clusters of ECs, and feature groups are defined based on correlations among these features. This also suits the application scenario if the number of features exceeds the number of samples. The current approach has similarities to a self-contained gene set test. However, in contrast to other techniques we estimate the probability of a network feature (i.e. EC cluster) to be a predictor of the biological or clinical phenotype. This is done via a bootstrap approach. Using simulations, we show that the current method performs better than the GT, which is a self-contained gene set test.

Furthermore, other than SGL there exists techniques based on regularized linear models to address such a problem e.g. Clustered lasso (She, 2010), Elastic net (Zou and Hastie, 2005) , OSCAR (Bondell and Reich, 2008). These techniques tend to equalise the coefficients of features which are correlated but do not take explicitly the correlation into account. Therefore, to demonstrate the applicability and efficacy of the current approach, in addition to GT, the method is compared with Elastic net approach in the simulation setting. Elastic net performs variable selection e.g. selects a sparse set of features as well as selects group of correlated variables. This grouping of correlated variables is demonstrated in Theorem 1 in (Zou and Hastie, 2005) which essentially means that the coefficient paths between highly correlated variables is almost 0. This is achieved by the elastic net penalty which is a convex combination of l_1 regularized norm as in lasso (Tibshirani, 1996b) and l_2 regularized norm as in ridge regression (Hoerl and Kennard, 1970). Therefore, the grouping of correlated variables is implicitly done by objective function of Elastic net.

The advantage of using ECs instead of EFMs or EPs in this approach is to provide the link to well established techniques from stoichiometric network analysis. For example, ECs have been employed to compute Hopf bifurcations (yielding oscillations) (Errami et al., 2015; Gatermann,

Eiswirth, and Sensse, 2005b) and bistability (Aguda and Clarke, 1987) of biochemical reaction networks. However, the method described in this paper can be readily applied to EFMs and EPs as well.

The major benefits of the method are four fold: Firstly, it is a method to address the typical high overlap of individual ECs because of the network structure as well as correlations at the level of -"omics" data. This issue is addressed by focusing on "clusters of ECs" to define the network features and defining groups (i.e. clusters of network features) in the statistical model based on correlations among these features. Secondly, the approach is flexible enough to analyse different types of phenotypes e.g. categorical (cancer vs healthy), continuous or censored survival times. Thirdly, the approach does not require a discretization of genes into categories (as required by hypergeometric test). Finally, the method estimates a probability of each network feature to be associated with the phenotype, hence appropriately taking into account the noise in observed data.

The method is applied to metabolic models from KEGG database (when these models are available as biochemical reaction networks). ECs then correspond to pathways within a given KEGG model. Although KEGG models are frequently used to analyze and understand human diseases based on gene expression or other -omics data, the reaction network structure of these models is usually ignored. In contrast, the current approach relates changes in specific sub-reaction systems defined as ECs and groups of ECs to observed changes of the phenotype. Importantly, a given KEGG model pathway can have numerous ECs, implying that KEGG models as a whole might not always offer the right granularity to describe and understand a biological mechanism. Additionally, the ECs in a given reaction network can identify the feedback loops (Sensse, Hauser, and Eiswirth, 2006) which are important for dynamics.

8.2 Approach

The ECs are steady state flux distributions. In order to investigate the ECs with respect to the gene expression data, enzymes in the flux carrying reactions need to be mapped to genes. For every EC i.e. row vector of matrix E (cf. Section 3.2 describing the mathematical background on E) denoted by E_j ((3.27)), a row vector g_j, corresponding to the j-th EC, is constructed by mapping the flux carrying reactions denoted by $\{1 \leq i \leq n \mid E_{ji} \neq 0\}$ to those genes, which participate in the reactions as enzymes. More specifically, this is done by constructing an ℓ dimensional indicator vector (where $\ell \leq n$ because the mapping from fluxes to genes can be many to one) with as many entries as there are mappable genes in the overall pathway. Then, entry k' in g_j is set to 1, if gene k' can be mapped to EC E_j. Such a mapping can result in duplicates. Therefore, only the distinct vectors $g_1, \ldots, g_m$ are considered where m denotes the number of gene sets obtained from k ECs and $m \leq k$ is due to the duplicates resulting from the mapping. Finally, the gene set vectors $g_1, \ldots, g_m$ can be joined into a $m \times \ell$ matrix G.

8.2.1 Computation of Network Features

Let us define gene set corresponding to g_j as the support of g_j: $\text{supp}(g_j) = \{i : G_{ji} = 1\}$ (i.e. the non-zero entries in row vector G_j). For matrix G, let S be set of extreme currents represented as gene sets. The ECs having empty gene sets are excluded from S. In practice, the gene sets are often highly overlapping or even identical. The issue of identical gene sets is already addressed in the previous section. Therefore, individual extreme currents and associated gene sets, respectively, will be hard to discriminate based on gene expression data. To address this issue the gene sets in S are grouped into clusters.

Let us denote S' as the partition of S such that S' breaks S into subsets denoted as $S' = \{S'_1, \ldots, S'_{m'}\}$ where $S'_l, 1 \leq l \leq m'$ denotes a cluster and $m' \leq m$ (as ECs with empty gene sets are excluded). It satisfies the condition $S'_1 \cup S'_2 \cup, \ldots \cup S'_{m'} = S'$ and $S'_i \cap S'_j = \emptyset$ for i and

j from $1 \ldots m'$, $i \neq j$. A review on clustering algorithms can be found in (Jain and Dubes, 1988). Here agglomerative hierarchical clustering (Maechler et al., 2015) was used which is a widely accepted technique for cluster analysis. Agglomorative hierarchical clustering works by successively joining data points (which are gene sets i.e. elements in S) based on a defined similarity measure, resulting in a tree like structure (*dendrogram*), which represents similarities of groups of data points (in this case gene sets) in a hierarchical manner. In the current case, the Jaccard index was employed as similarity measure between gene sets and the complete linkage strategy was used to join clusters. A grouping is then found by prescribing a certain number of clusters and cutting the dendrogram at the corresponding height. Here the number of clusters m' is selected by computing the silhouette index (Rousseeuw, 1987), which is an established approach to estimate the quality of a clustering. For this purpose, the number of clusters is varied from 2 to $m' - 1$ and selected the optimal number of clusters according to the maximum silhouette index. The network feature corresponding to $S'_l, 1 \leq l \leq m'$ is then computed by taking the union of the gene sets in S'_l. These network features are used in the subsequent steps to fit the linear model.

8.2.2 Combining Network Features with Gene Expression Data and Phenotypes

The idea behind the current approach is to combine gene expression data, phenotypes and metabolic network features. Gene expression data of interest is generally represented in form of a $l \times q$ matrix D, where rows represent the genes present in the microarray chip and columns represent biological or patient samples. The biological or clinical phenotype is given by a q-dimensional vector y. Entries in y could be categorical (e.g. cancer or healthy) or continuous real numbers (e.g. a patient's quantitative response to a certain treatment).

First, a $q \times m'$ feature matrix X is constructed (recall m' is the number of clusters as defined in the previous section) as:

$$X_{ji} = [pc_1]_{S'_i}[D_j^{S'_i}]_{S'_i}, \tag{8.1}$$

where $[pc_1]_{S'_i}$ is a row vector corresponding to the first principal component of $D^{S'_i}$ (explaining the maximum proportion of variation in the data), where $D^{S'_i}$ is a matrix derived from D with rows representing only those genes that map to the genes in S'_i. $[D_j \hat{} S'_i]_{S'_i}$ is the column vector consisting of genes S'_i in the j-th sample (column) in $D^{S'_i}$. The equation thus projects expression profiles of the S'_i genes to the direction explaining maximum variance in the data. Hence, an effective summarization of expression profiles of multiple genes in set S'_i into one activity score is performed. Notably, this is similar to the method suggested by (Bild et al., 2006).

8.2.3 Modelling Categorical and Continuous Phenotypes

It is assumed that the phenotype y is linked to feature matrix X via a generalized linear model

$$f(y) = X\beta + \epsilon, \tag{8.2}$$

where ϵ the measurement noise and f a link function, e.g. logit in case of categorical phenotypes and the identity function in case of continuous phenotypes (McCullagh and Nelder, 2000).

Importantly, it is supposed that the β to be a sparse vector, i.e. most coefficients are zero. That means the phenotype is determined by a sparse combination of network features, which is to be identified from expression data and observed phenotypes. This addresses the non-identifiability issue arising in high dimensional setting where number of features is greater than number of samples. Another complicating factor in the current context is that the gene expression data typically exhibits a non-trivial correlation structure that may not coincide with network features. For example, certain groups of genes might be activated or deactivated by

common regulatory elements, resulting in a block-like correlation matrix. In order to address this type of correlation, the network features are assigned to groups. Groups should be assigned in such a manner that the distance between the features inside the group (intra-group) should be small whereas the distance to features of other group (inter-group) features should be high. This is done by hierarchical average clustering of network features based on gene expression data where the distance between two features S'_i and S'_j is defined as $1- \mid cor(S'_i, S'_j) \mid$ where *cor* refers to the Pearson correlation of the corresponding gene expression profiles. Such an approach based on clustering to handle correlated variables and cope with multicollinearity in the context of linear models is described in (Bühlmann et al., 2013) and explained below. The optimal number of obtained clusters (estimated via the silhouette index) is denoted as g. The grouping of features addresses the issue of near linear dependency between group of features due to high correlation. Such group membership information can enter the above linear model as explained in the next section. It should be pointed out that, clustering is performed two times, first to generate the network features from ECs and second to cluster the network features based on the correlation in the gene expression data. An overview of the current approach is represented in Fig. 8.1.

8.2.4 Fitting the Model via Sparse Group Lasso

Following suggestions in (Bühlmann et al., 2013) the generalized linear model (8.2) under the above described restrictions, namely sparsity and organisation of network features into correlated groups, can be fitted via a sparse group lasso (SGL) (Simon et al., 2013). The SGL solves the following optimization problem:

$$\begin{aligned} \min_\beta \frac{1}{2q} \| y - \sum_{l=1}^{g} f^{-1}(X^{(l)}\beta^{(l)}) \|_2^2 \\ +(1-\alpha)\lambda \sum_{l=1}^{g} \sqrt{p_l} \|\beta^{(l)}\|_2 + \alpha\lambda\|\beta\|_1, \end{aligned} \tag{8.3}$$

where q is the dimension of the response vector y, $X^{(l)}$ is the submatrix of X with columns corresponding to the predictors in group l, $\beta^{(l)}$ is the coefficient vector of that group and p_l is the length of $\beta^{(l)}$ and λ a the tuning parameter controlling the sparsity of coefficient vector β: Larger λ implies more sparsity, but less precise fit to the data. Moreover, $\alpha \in [0, 1]$ is a parameter that balances between sparse selection of whole feature groups and sparse selection within each feature group. For the extreme case $\alpha = 0$ we recover the group lasso (Yuan and Lin, 2006), which only performs selection of feature groups of a whole. For $\alpha = 1$ we arrive at the original lasso model (Tibshirani, 1996a), which selects features independent of any group structure.

In practice λ and α are tuning parameters. These are usually estimated via 5-fold cross-validation: Data is split into 5 distinct sets (= folds), and the SGL model is trained on 4 folds while leaving out the rest for testing. The prediction error made by the model on the test set is then recorded. The whole procedure is successively repeated until each fold has been left out once for testing. The average prediction error made during the cross-validation procedure can be used as a quality measure for the SGL model, which can be recorded over a grid of λ and α parameters. Accordingly, a good combination of these parameters can be selected.

As stated in (Bühlmann et al., 2013) an advantage of the proposed approach compared to an elastic net (Zou and Hastie, 2005) is the explicit handling of the feature correlation structure.

8.2.5 Patient Survival Outcomes

The current approach can also be employed for associating network features with patient survival: In case of patient survival data $y = \{(t_i, \delta_i) | i = 1, \ldots, q\}$ where t_i is the observed survival

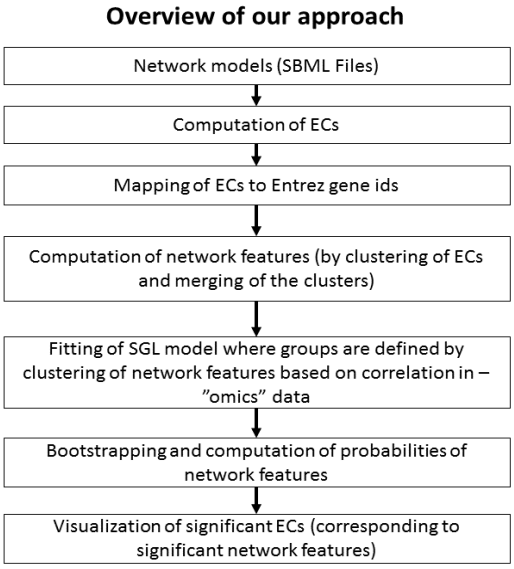


FIGURE 8.1: An overview of the approach

time of patient i and $\delta_i \in \{0, 1\}$ indicates the censoring status. Instead of a generalized linear a, Cox proportional hazard model is assumed, i.e.:

$$f(t \mid X) = f_0(t) \exp(X\beta), \tag{8.4}$$

where $f(t \mid X)$ is the hazard at time t and $f_0(t)$ the baseline hazard. See (McCullagh and Nelder, 2000) for an overview about survival models. Notably, the sparse group lasso can also be employed to fit this type of model.

8.2.6 Bootstrapping

In order to quantify the uncertainty associated with selection of a particular feature a bootstrap test is employed. The uncertainty with respect to the relevance of a particular feature is measured by the probability that a particular feature is associated to the phenotype. The test is performed by randomly sampling the rows in X and the corresponding entries in y (with replacement) and re-estimating coefficients of the SGL model. This is repeated a number of times (100 times during the simulations and 500 times with for real data), and the fraction of times in which a coefficient is non-zero, gives the probability for the association of a specific feature with the phenotype. This probability value is denoted as bootstrap frequency.

8.3 Results

8.4 Metabolic Network Reconstruction

Metabolic networks were obtained from KEGG database in KGML format (Kanehisa and Goto, 2000). The KGML files were converted to SBML using KEGGtranslator (Wrzodek, Dräger, and Zell, 2011). For testing the method models corresponding to Glycolysis / Gluconeogenesis (hsa00010) model is used.

KEGG reactions are extensively annotated with main and side reactant-product pairs called RPAIR (Kotera et al., 2004). For the current purpose, only the main reaction pairs are considered. A single reaction can have multiple main pairs and in that case each main pair is treated

as an independent reaction. It is observed that the number of ECs increase substantially by considering only the main pairs. This happens due to the presence of compounds in side pairs (e.g. water, cofactors, etc) which may not always be in steady state. The benefits of such a decomposition is analysed in (Faust, Croes, and Helden, 2009) for searching of the optimal paths in KEGG networks. Moreover, only main pairs are displayed in the visualization of the pathways on the KEGG website.

Compounds lying at the boundary of the system (i.e. acting as source or sinks) are assumed not to be in steady state condition. Thus, the reaction network for EC computation involves only those species participating in the main reaction pairs and occurring as non-boundary species. In the computation of EFMs these are referred as external metabolites. Those chemical species which are either produced or consumed are considered to be boundary species.

8.5 Computation of ECs and Feature Matrix

KEGG metabolic models are taken and the network features are computed using the steps described above. A benchmark on the CPU running times and a histogram on the distribution of ECs is described in Chapter 5. The mapping ECs to gene sets and clustering them to generate network features results in considerable reduction in the number of features. For example, the models analysed in applications namely, hsa00240 and hsa00260 have 17 and 9 features derived from 5510 and 58 ECs respectively. Essentially, this tells us that a large number of extreme currents in flux space often map to few features defined by gene sets.

8.6 Simulations

A certain proportion of features were to be declared to be relevant, denoted by set F and the remaining set of features were denoted by set F'. The response variables (i.e. censored survival times or categorical phenotypes) were generated from F. A similar approach for simulations was followed in (Simon et al., 2013). A SGL penalized Cox regression model for censored survival times and SGL penalized logistic regression for categorical phenotypes were fitted to the data to in order to identify the relevant features. The performance of SGL was compared with GT based on the Area Under the Curve (AUC) values measuring the performance of a binary classifier computed using the pROC package in R (Robin et al., 2011). The details of the simulations are described below

8.6.1 Survival Outcomes

The simulation was designed as follows:

1. A certain set of features, F was declared to be relevant. Non-relevant features are denoted as F'.

2. Feature matrix X was drawn from a q dimensional multivariate normal distribution $N_q(0, \Sigma)$. The covariance Σ_{ij} between features i and j was defined according to the Jaccard index between gene sets, i.e.:

$$\Sigma_{ij} = \frac{|G'_i \cap G'_j|}{|G'_i \cup G'_j|} \tag{8.5}$$

   This represents the correlation arising from the overlap i.e. outcome of network structure.

3. To simulate the effect of correlation from the expression data, a higher covariance was set i.e. 0.8 between the features comprising of 50% randomly selected elements from F and

50% randomly selected elements from F' denoted by F_{cor} i.e.

$$\Sigma_{ij} = \begin{cases} 0.8, & \text{if } i \neq j \text{ and } i, j \in F_{cor} \\ 1, & i = j. \end{cases} \tag{8.6}$$

4. The response (i.e. survival and censoring time) was generated from relevant features as $y_j = exp(\sum \beta f_j)$ (where $\beta = 3$). The indicator for censoring (i.e. 0) or failure (i.e. 1) was generated independently by sampling the indicators (with replacement) with a probability that 30% of the samples are censored.

5. A sparse group lasso (SGL) penalized Cox regression model was fitted the data, and the fitting procedure was bootstrapped 100 times. This resulted into a probability per network feature.

6. As an alternative to the current method a global test (GT) was performed for each network feature, resulting into one p-value per feature.

7. The performance of both the methods was compared based on their AUC values.

8.6.2 Categorical Phenotypes

The simulation was set up similar to before, the only difference being that the response y_j was generated from relevant features as a Bernoulli distributed random variable with parameter $exp(\sum \beta f_j)/(1 + exp(\sum \beta f_j))$ (where $\beta = 3$). Accordingly, a SGL penalized logistic regression model was fitted to the data. The fitting procedure was bootstrapped as described above.

Simulation results for censored survival times as response variables are shown in Tables 8.1 and S5 - S8 and for categorical phenotypes as response variables are shown in Tables 8.2 and S9 - S12. In both the scenarios, a clear overall benefit of the current approach based on SGL could be observed. Thus, indicating a statistically significant improvement of the current method compared to the GT, in the vast majority of simulation settings, which increased with the number of samples in the dataset. The statistical significance is estimated based on a Wilcoxon signed rank test. Notably, different reconstructed pathways were tested in the simulation, which differed in their number of extractable network features. It was observable that in simulation cases, where the number of network features far exceeded the number of samples in the dataset the performances of both compared methods were comparably low and equal, whereas in all other cases the SGL based method could benefit significantly more than GT.

Interestingly with increasing number of relevant features a decrease in the performance of both GT and SGL was observed. This is most likely due to the increasing number of correlated relevant features, resulting in non-identifiability of the individual network features.

8.7 Application in Prostate Cancer

A comprehensive analysis of pathways in prostate cancer was reported in (Sreekumar et al., 2009) where metabolomic profiles were found to be differentially expressed in benign prostate, clinically localized prostate cancer and metastatic disease. More specifically, Sarcosine was found to be highly elevated in the tumor samples.

8.7.1 Data source

Sarcosine can be found in the glycine, serine and threonine pathway (KEGG ID: hsa00260) in the KEGG database. Normalised gene expression data was mapped from (Brase et al., 2011),

TABLE 8.1: Summary of simulation on Glycolysis / Gluconeogenesis (hsa00010) KEGG pathway for survival time analysis. Avg refers to median value of Area Under the Curve (AUC) values obtained in 50 simulation runs. GT, EN and SGL refer to glocal test, elastic net and sparse group lasso respectively. P-values refer to the null hypothesis of no difference between GT and SGL denoted as P-Value GT-SGL. The P-value between SGL and EN is denoted as P-value EN-SGL.

Samples	Relevant Features	Features	Iterations	Avg(AUC):GT	Avg(AUC):EN	Avg(AUC):SGL	P-value GT-SGL	P-value EN-SGL
10	05%	11	50	1.00±0.02	0.70±0.14	1.00±0.02	0.02	6.41e-10
10	15%	11	50	0.73±0.16	0.61±0.14	0.94±0.12	1.60e-09	7.45e-10
10	25%	11	50	0.78±0.12	0.71±0.12	0.93±0.12	2.27e-09	6.77e-10
20	05%	11	50	1.00±0.04	0.80±0.12	1.00±0.00	0.00024	5.60e-10
20	15%	11	50	0.83±0.15	0.77±0.11	1.00±0.06	5.22e-09	1.10e-09
20	25%	11	50	0.75±0.15	0.83±0.08	1.00±0.08	2.41e-09	2.28e-09
40	05%	11	50	1.00±0.05	0.70±0.13	1.00±0.00	3.98e-05	6.82e-10
40	15%	11	50	0.86±0.08	0.80±0.08	1.00±0.01	8.05e-10	6.98e-10
40	25%	11	50	0.87±0.13	0.60±0.05	1.00±0.00	1.64e-08	6.44e-10
60	05%	11	50	1.00±0.03	0.70±0.12	1.00±0.00	7.49e-05	6.77e-10
60	15%	11	50	0.77±0.06	0.58±0.10	1.XX±0.00	4.23-10	7.32e-10
60	25%	11	50	0.87±0.10	0.62±0.08	1.00±0.00	7.72e-09	7.39e-10

TABLE 8.2: Summary of simulation on Glycolysis / Gluconeogenesis (hsa00010) pathway from KEGG for classification analysis. Avg refers to median value of Area Under the Curve (AUC) values obtained in 50 simulation runs. GT, EN and SGL refer to glocal test, elastic net and sparse group lasso respectively. P-values refer to the null hypothesis of no difference between GT and SGL denoted as P-Value GT-SGL. The P-value between SGL and EN is denoted as P-value EN-SGL. NA* means that the EN method failed the linear model for the given simulation parameters. NA** means that the p-value computation is not application as the list of AUC values corresponding to both the methods are exactly the same.

Samples	Relevant Features	Features	Iterations	Avg(AUC):GT	Avg(AUC):EN	Avg(AUC):SGL	P-value GT-SGL	P-value EN-SGL
8	05%	11	50	0.95±0.15	NA*	1.00±0.15	0.003	NA*
8	15%	11	50	0.73±0.16	NA*	0.73±0.14	1.49e-07	NA*
8	25%	11	50	0.68±0.13	NA*	0.70±0.14	0.31	NA*
20	05%	11	50	1.00±0.04	1.00±0.02	1.00±0.01	0.0003	0.34
20	15%	11	50	0.75±0.12	0.83±0.17	0.88±0.14	9.55e-10	3.42e-07
20	25%	11	50	0.64±0.11	0.75±0.14	0.79±0.15	6.08e-09	5.76e-05
40	05%	11	50	1.00±0.02	1.00±0.00	1.00±0.00	8.90e-05	NA**
40	15%	11	50	0.86±0.14	1.00±0.09	1.00±0.06	3.37e-08	0.0004
40	25%	11	50	0.81±0.13	1.00±0.08	1.00±0.06	4.98e-09	2.71e-05
60	05%	11	50	1.00±0.03	1.00±0.00	1.00±0.00	0.03	NA**
60	15%	11	50	0.77±0.11	0.94±0.16	1.00±0.08	6.48e-10	1.13e-06
60	25%	11	50	0.81±0.10	1.00±0.09	1.00±0.06	6.91e-10	0.0025

comprising 47 prostate tumor tissue samples along with 48 normal prostate tissue samples (GSE29079, Affymetrix Human Exon 1.0 ST).

8.7.2 Computation of ECs and Integration of expression data

The SGL based Cox regression model was fitted with 9 network features. The number of permutations in the bootstrap test was fixed at 500.

8.7.3 Results

The histogram of bootstrap frequencies shown in Fig. 8.2. A confidence cut-off value of 0.8 is chosen because of gap in the histrogram before this value. It resulted in 6 features out of 9 as significant. Upon investigation, out of these 6 features, 2 features included glycine-N-methyltransferase (GNMT) (Fig. 8.3). GNMT has been associated with elevated Sarcosine levels in (Sreekumar et al., 2009) and prostate cancer progression in general (Song et al., 2011). The right most EC in Fig. 8.3 shows the reversible reaction of Sarcosine to Glycine catalysed by different set of enzymes for forward and backward reaction components. The left most EC links

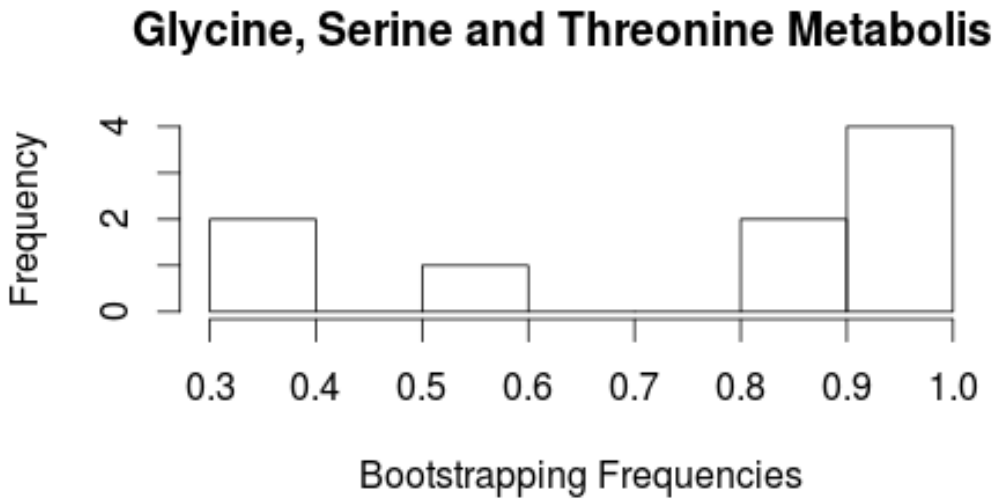


FIGURE 8.2: Histogram of bootstrap frequencies of 9 network features in Glycine, serine and threonine metabolism for prostate cancer data

Choline with 5-Aminolevulinate as well as Methanal. As visualized in Fig. 8.3 (left) Sarcosine can be converted into Methanal (formaldehyde) via PIPOX. It has been reported that prostate cancer patients show increased formaldehyde concentrations in their urine (Španěl et al., 1999).

8.8 Application in Glioblastoma multiforme (GBM)

In (Wolf, Agnihotri, and Guha, 2010) the authors discussed the role of nucleotide metabolism in Glioblastoma Multiforme. Based on that it was decided to investigate pyrimidine metabolism in more depth, which is a sub-category of nucleotide metabolism.

8.8.1 Data source

Normalised expression profiles from 342 patients were downloaded and combined from The Cancer Genome Atlas (TCGA) database (al, 2008) (HG-U133A Affymetrix Array platform). To correct for possible batch effects the ComBat method (Johnson, Li, and Rabinovic, 2007) was applied. Expression data were mapped to the pyrimidine metabolism KEGG pathway (ID: hsa00240). In addition to gene expression clinical data were retrieved from TCGA. The median, minimum and maximum survival times were 357, 26 and 3880 days respectively. 27.19% of the patients had censored outcomes.

8.8.2 Computation of ECs and Integration of expression data

The SGL based Cox regression model was fitted with 17 network features. The number of permutations in the bootstrap test was fixed at 500.

8.8.3 Results

The histogram of bootstrap frequencies is shown in Fig. 8.4. A confidence cut-off value of 0.8 is chosen because of gap in the histrogram before this value. It resulted in 3 features out of 17 as significant. It was found that 1 out of these 3 relevant features contained enzymes previously

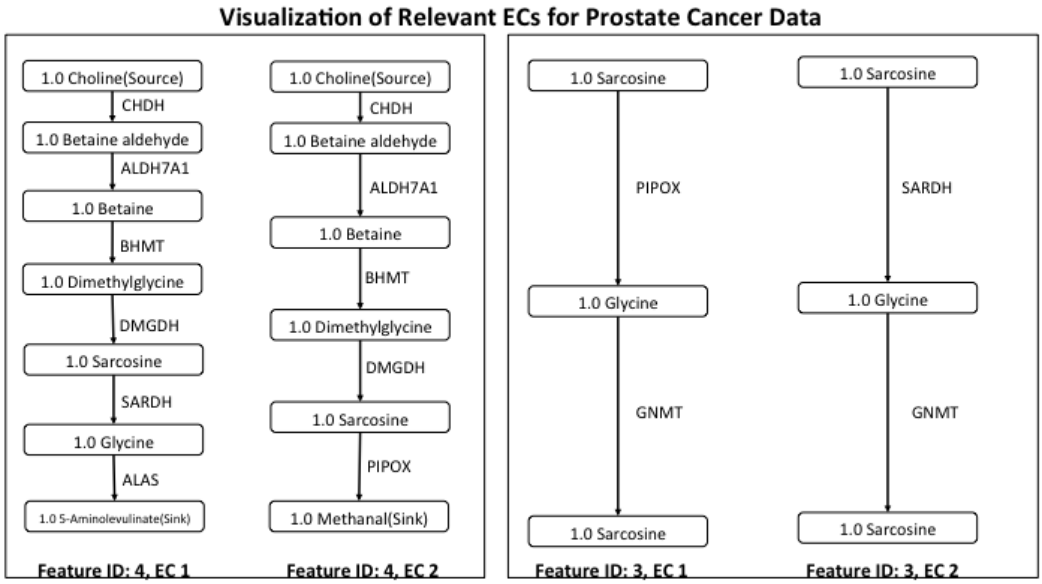


FIGURE 8.3: Two network features from the Glycine, Serine and Threonine metabolism identified to discriminate prostate cancer and normal patients with high probability. The stoichiometric values are added as prefix to the names of chemical species. Each EC is presented as a separate directed graph within each network feature (indicated via boxes). Vertices are chemical species (as they appear in the corresponding KEGG file) and directed edges refer to chemical reactions with reactant at the tail and product at the head of the edge. Edge labels indicate enzymes.

reported as reported as dysregulated in GBM (Wolf, Agnihotri, and Guha, 2010; Bardot et al., 1994), namely thymidylate synthase (TYMS) and thymidine phosphorylase (TYMP). As seen from Fig. 8.5 TYMS catalyses the reaction from dUMP to dTMP. The other enzymes catalyze several reactions. The current method unravels these multiple reaction pathways.

8.9 Discussions

A method to associate features (namely reaction pathways based on ECs) of a metabolic network to clinical or biological phenotypes with the help of gene expression data was described. Extraction of relevant reaction pathways from a metabolic network has previously found application in drug target identification and network robustness analysis (see (Papin et al., 2003) for an overview). ECs are invariants of the reaction network structure and enumerate possible network behaviours. However, under real biological conditions not all of these possible behaviours are realized. Gene expression data reflect constraints on concentrations of the enzymes, which regulate the functioning of the metabolic network. Gene expression data can thus be employed to identify, which of the possible network behaviours and reaction pathways, respectively, in reality contribute to a particular biological or clinical phenotype. Notably, this is not only possible in categorical sense for e.g. discriminating normal from cancer cells, but with respect to more complex clinical outcome measurements, such as censored survival times. However, a major difficulty arising in that context is that gene expression data shows a non-trivial correlation structure, which is not necessarily in agreement with defined network features. The current method uses a sparse group lasso based approach to address this issue. Simulation studies demonstrated the superiority of the current method compared to a conventional GT as well as Elastic net. In essence the current method combines algebraic analysis of the metabolic network structure with an unbiased view on the transcriptional level.

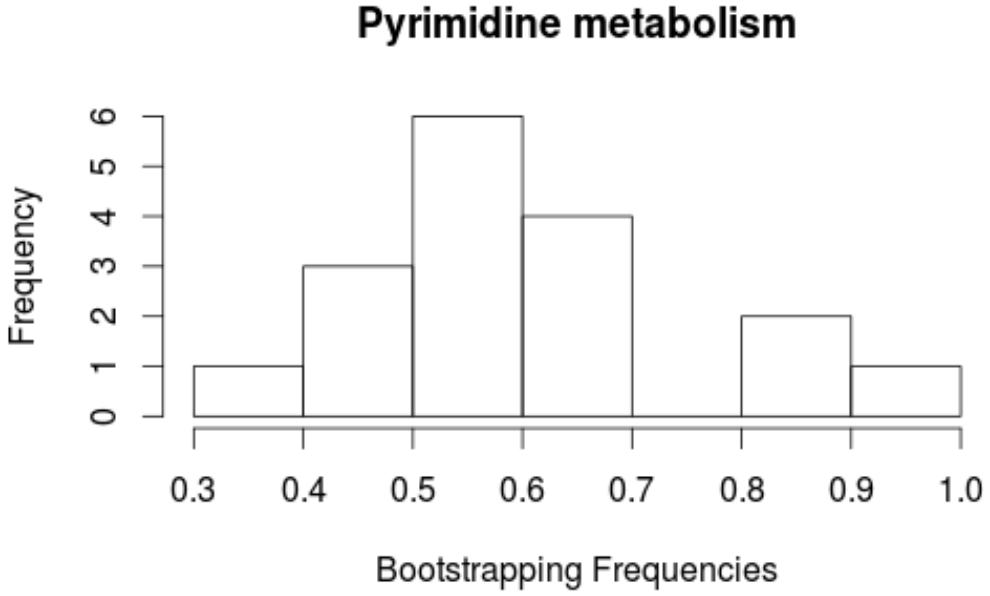


FIGURE 8.4: Histogram of bootstrap frequencies of 17 network features in Pyrimidine metabolism for GBM

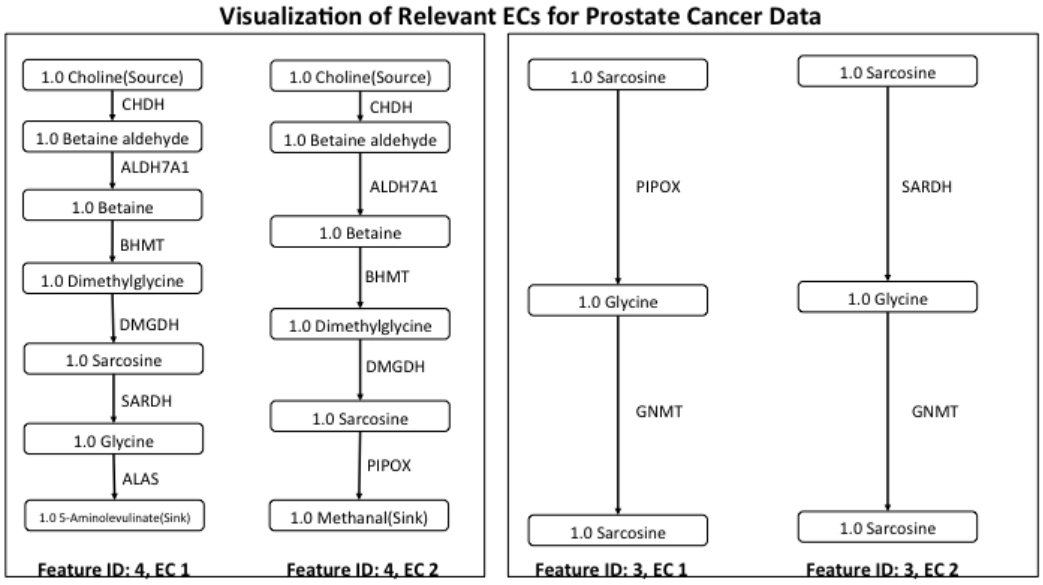


FIGURE 8.5: The single network features from the Pyrimidine metabolism is identified which discriminate prostate cancer and normal patients with high probability. The stoichiometric values are added as prefix to the names of chemical species. Each EC is presented as a separate directed graph within each network feature (indicated via boxes). Vertices are chemical species (as they appear in the corresponding KEGG file) and directed edges refer to chemical reactions with reactant at the tail and product at the head of the edge. Edge labels indicate enzymes.

The application of SGL method was demonstrated by investigating the role of the metabolite Sarcosine in prostate cancer and pyrimidine metabolism in GBM. In both cases statistically stable identified network features could be linked to findings from the medical literature, supporting the relevance of the current method. Furthermore, the distribution of bootstrapping frequencies as shown in the histograms showed a gap before 0.8, which lead to the choice of selecting this value as the cut-off for declaring significant features. The source code along with simulation results on additional KEGG models are available at http://www.abi.bit.uni-bonn.de/index.php?id=17.

The current approach was made to be computationally practical via parallel computing. The computation time for the GBM example was less than 2 hours and for the prostate cancer application it took less than 30 minutes on a server with 64 cores and 512 MB RAM. This was achievable, due to the focus on individual pathway models. However, the ultimate goal would be a genome-scale analysis of metabolic networks, which currently is impossible with the current method due to the computational complexity of EC computation. Nevertheless, the method would be in principle also extendable to such a situation in the future, by using network decomposition into tractable submodels like in (Hunt et al., 2014), or genetic algorithm combined with linear programming solutions from (Kaleta et al., 2009).

Appendix A

Appendix: TGFβ Signalling

A.1 Description of the TGFb model used in the thesis.

The model is described by the following system of differential equations

$$
\begin{aligned}
\frac{\mathrm{d}x_1}{\mathrm{d}t} &= k_2x_2 - k_1x_1 - k_{16}x_1x_{11} \\
\frac{\mathrm{d}x_2}{\mathrm{d}t} &= k_1x_1 - k_2x_2 + k_{17}k_{34}x_6 \\
\frac{\mathrm{d}x_3}{\mathrm{d}t} &= k_3x_4 - k_3x_3 + k_7x_7 + k_{33}k_{37}x_{18} - k_6x_3x_5 \\
\frac{\mathrm{d}x_4}{\mathrm{d}t} &= k_3x_3 - k_3x_4 + k_9x_8 - k_8x_4x_6 \\
\frac{\mathrm{d}x_5}{\mathrm{d}t} &= k_5x_6 - k_4x_5 + k_7x_7 + 2k_{11}x_9 - 2k_{10}x_5^2 - k_6x_3x_5 + k_{16}x_1x_{11} \\
\frac{\mathrm{d}x_6}{\mathrm{d}t} &= k_4x_5 - k_5x_6 + k_9x_8 + 2k_{13}x_{10} - 2k_{12}x_6^2 - k_{17}k_{34}x_6 + k_{31}k_{36}x_8 - k_8x_4x_6 \\
\frac{\mathrm{d}x_7}{\mathrm{d}t} &= k_6x_3x_5 - x_7(k_7 + k_{14}) \\
\frac{\mathrm{d}x_8}{\mathrm{d}t} &= k_{14}x_7 - k_9x_8 - k_{31}k_{36}x_8 + k_8x_4x_6 \\
\frac{\mathrm{d}x_9}{\mathrm{d}t} &= k_{10}x_5^2 - x_9(k_{11} + k_{15}) \\
\frac{\mathrm{d}x_{10}}{\mathrm{d}t} &= k_{15}x_9 - k_{13}x_{10} + k_{12}x_6^2 \\
\frac{\mathrm{d}x_{11}}{\mathrm{d}t} &= k_{23}x_{14} - k_{30}x_{11} \\
\frac{\mathrm{d}x_{12}}{\mathrm{d}t} &= k_{18} - x_{12}(k_{20} + k_{26}) + k_{30}x_{11} + k_{27}x_{15} - k_{22}k_{35}x_{12}x_{13} \\
\frac{\mathrm{d}x_{13}}{\mathrm{d}t} &= k_{19} - x_{13}(k_{21} + k_{28}) + k_{30}x_{11} + k_{29}x_{16} - k_{22}k_{35}x_{12}x_{13} \\
\frac{\mathrm{d}x_{14}}{\mathrm{d}t} &= k_{22}k_{35}x_{12}x_{13} - x_{14}(k_{23} + k_{24} + k_{25}) \\
\frac{\mathrm{d}x_{15}}{\mathrm{d}t} &= k_{26}x_{12} - k_{27}x_{15} \\
\frac{\mathrm{d}x_{16}}{\mathrm{d}t} &= k_{28}x_{13} - k_{29}x_{16} \\
\frac{\mathrm{d}x_{17}}{\mathrm{d}t} &= k_{31}k_{36}x_8 - k_{32}x_{17} \\
\frac{\mathrm{d}x_{18}}{\mathrm{d}t} &= k_{32}x_{17} - k_{33}k_{37}x_{18}
\end{aligned}
\tag{A.1}
$$

These variables are as follows:

- Receptors on plasma membrane: x_{12} = RI (receptor 1), x_{13} = RII (receptor 2), x_{14} = LR (ligand-receptor complex).
- Receptors in the endosome: x_{11} = LRe, x_{15} = RIe, x_{16} = RIIe.
- Transcription factors and complexes in cytosol: x_1 = S2c, x_3 = S4c, x_5 = pS2c, x_7 = pS24c, x_9 = pS22c, x_{18} = S4ubc.
- Transcription factors and complexes in the nucleus: x_2 = S2n, x_4 = S4n, x_6 = pS2n, x_8 = pS24n, x_{10} = pS22n, x_{17} = S4ubn.

A.2 Calculation of tropical equilibration branches for the ligand-receptor module.

Tropical equilibration solutions for the variables x_{11},x_{12}, x_{13}, x_{14}, x_{15}, x_{16} (the submodel in Fig. 7.4) can be computed independently from the rest of the variables of the TGFβ model. The ordinary differential equations for these variables form a subsystem that is decoupled (receives no feed-back) from the rest of the equations.

We can reduce the system of 6 tropical equations to a simplified system of 3 tropical equations using the following two general properties.

Property 1 (binomial species)**.** *Y is a binomial species if the ordinary differential equation defining its rate of variation contains only one positive monomial term and only one negative monomial term*

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = M_1(\boldsymbol{X})Y^{n_1} - M_2(\boldsymbol{X})Y^{n_2},$$

where $\boldsymbol{X}$ denotes the other variables. We further assume that $n_1 < n_2$. Then, the species Y can be eliminated and the resulting simplified tropical system has the same tropical equilibration solutions as the full system. The simplification is performed by eliminating the equation for Y and replacing everywhere Y by $(M_1/M_2)^{1/(n_2-n_1)}$.

Proof. The proof follows from the fact that the tropical equation for the order a of Y has the unique solution $a = \frac{1}{(n_2-n_1)}(\mu_1 - \mu_2)$. □

Property 2 (dominated first order reactions)**.** *If a species Y is consumed by several first order reactions of kinetic constants $k_1, k_2, \ldots, k_r$ and if $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_p < \gamma_{p+1} \leq \gamma_{p+2} \leq \ldots \leq \gamma_r$, then the reactions $k_{p+1}, \ldots, k_r$ can be eliminated and the resulting simplified tropical system has the same tropical equilibration solutions as the full system.*

Proof. The proof follows from the following obvious property of the min operation as shown below

$$\min(\gamma_1, \ldots, \gamma_p, \ldots, \gamma_r) = \min(\gamma_1, \ldots, \gamma_p) \tag{A.2}$$

. □

Using $\gamma_{26} < \gamma_{20}$, $\gamma_{28} < \gamma_{21}$ (a condition satisfied by the nominal model parameters and meaning that internalization is more rapid than degradation for both receptors 1 and 2) and the Properties 1,2 we can justify the reduction illustrated in Fig. A.1. Because the reduced model has the same tropical solutions as the full, larger model, it is enough to solve the tropical equilibration problem for the reduced model. This reads

$$\min(\gamma_{18}, a_{14} + \gamma_{23}, a_{12} + \gamma_{26}) = \min(a_{12} + a_{13} + \gamma_{22} + \gamma_{35}, a_{12} + \gamma_{26}) \tag{A.3}$$

$$\min(\gamma_{19}, a_{14} + \gamma_{23}, a_{13} + \gamma_{28}) = \min(a_{12} + a_{13} + \gamma_{22} + \gamma_{35}, \gamma_{28} + a_{13}) \tag{A.4}$$

$$\min(\gamma_{24}, \gamma_{25}, \gamma_{23}) + a_{14} = a_{12} + a_{13} + \gamma_{22} + \gamma_{35} \tag{A.5}$$

Suppose now that the following condition is true

$$\min(\gamma_{24}, \gamma_{25}, \gamma_{23}) = \gamma_{23}. \tag{A.6}$$

This condition is satisfied by the nominal parameters and, like the previous condition, means that receptors have relatively large life-times. Then from (A.5) we got $a_{14} = a_{12} + a_{13} + \gamma_{22} + \gamma_{35} - \gamma_{23}$ and the equations (A.3),(A.4) become

$$\begin{aligned} \min(\gamma_{18}, a_{14} + \gamma_{23}, a_{12} + \gamma_{26}) &= \min(a_{14} + \gamma_{23}, a_{12} + \gamma_{26}) && \text{(A.7)} \\ \min(\gamma_{19}, a_{14} + \gamma_{23}, a_{13} + \gamma_{28}) &= \min(a_{14} + \gamma_{23}, a_{13} + \gamma_{28}) && \text{(A.8)} \end{aligned}$$

The solutions of (A.7), (A.8) can be easily found and form the following polyhedron

$$\begin{aligned} (\{a_{12} + a_{13} + \gamma_{22} + \gamma_{35} \leq \gamma_{18}\} \quad &\cup \quad \{\gamma_{26} + a_{12} \leq \gamma_{18}\}) \cap \\ (\{a_{12} + a_{13} + \gamma_{22} + \gamma_{35} \leq \gamma_{19}\} \quad &\cup \quad \{\gamma_{28} + a_{13} \leq \gamma_{19}\}), \\ a_{14} \quad &= \quad a_{12} + a_{13} + \gamma_{22} + \gamma_{35} - \gamma_{23}. \end{aligned} \tag{A.9}$$

The orders of the remaining variables can be found as indicated in Prop. 1:

$$\begin{aligned} a_{15} &= a_{12} + \gamma_{26} - \gamma_{27}, && \text{(A.10)} \\ a_{16} &= a_{13} + \gamma_{28} - \gamma_{29}, && \text{(A.11)} \\ a_{11} &= a_{12} + a_{13} + \gamma_{22} + \gamma_{35} - \gamma_{30}. && \text{(A.12)} \end{aligned}$$

The polyhedron of tropical solutions defined by Eq.(A.9) can be partitioned into minimal branches (also polyhedra). This can be done by checking which term is dominant in the ordinary differential equations for the variables x_{12}, x_{13} and x_{14} (see Eqs. (A.1)). The result is that there are at most four minimal branches defined by one of the conditions

$$\begin{aligned} &\{a_{12} + \gamma_{26} < a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \cap \{a_{13} + \gamma_{28} < a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \\ &\{a_{12} + \gamma_{26} < a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \cap \{a_{13} + \gamma_{28} > a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \\ &\{a_{12} + \gamma_{26} > a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \cap \{a_{13} + \gamma_{28} < a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \\ &\{a_{12} + \gamma_{26} > a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \cap \{a_{13} + \gamma_{28} > a_{12} + a_{13} + \gamma_{22} + \gamma_{35}\} \end{aligned} \tag{A.13}$$

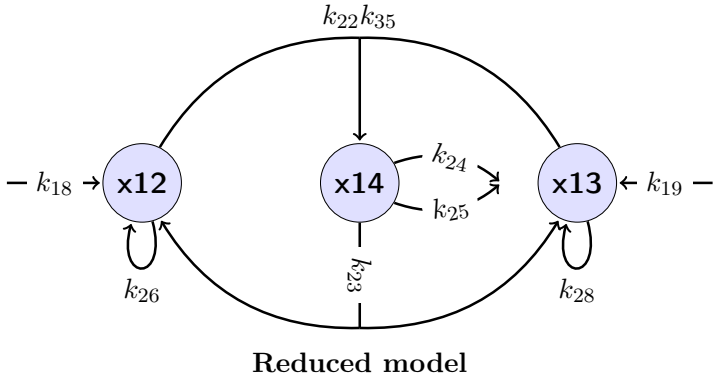


FIGURE A.1: In order to compute the branches of tropical equilibration of the ligand-receptor module we use a reduced model. The reduced model is not necessarily a good approximation of the full dynamics, but has exactly the same tropical solutions as the full model. The different variables mean: x_{12} : RI (TGBR1), x_{13} : RII (TGFBR2), x_{14} : LR (ligand-receptor complex), x_{15} : RIe (TGFBR1 in endosome), x_{16} : RIIe (TGFBR2 in endosome), x_{11} : LRe (LR in endosome).

Bibliography

Aguda, Baltazar D. and Bruce L. Clarke (1987). "Bistability in chemical reaction networks: Theory and application to the peroxidase-oxidase reaction". In: *The Journal of Chemical Physics* 87.6, pp. 3461–3470. DOI: http://dx.doi.org/10.1063/1.452991. URL: http://scitation.aip.org/content/aip/journal/jcp/87/6/10.1063/1.452991.

al, Roger McLendon et (2008). "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". In: *Nature* 455.7216, pp. 1061–1068.

Albert, Réka (2005). "Scale-free networks in cell biology". In: *Journal of Cell Science* 118.21, pp. 4947–4957. ISSN: 0021-9533. DOI: 10.1242/jcs.02714. eprint: http://jcs.biologists.org/content/118/21/4947.full.pdf. URL: http://jcs.biologists.org/content/118/21/4947.

Andrieux, Geoffroy et al. (2012). "Dynamic regulation of Tgf-B signaling by Tif1γ: a computational approach". In: *PloS One* 7.3, e33761.

Aubin, Jean-Pierre (2010). "Macroscopic traffic models: Shifting from densities to "Celerities"". In: *Applied Mathematics and Computation* 217.3. Special Issue in Honor of George Leitman on his 85th Birth year, pp. 963 –971. ISSN: 0096-3003. DOI: http://dx.doi.org/10.1016/j.amc.2010.02.032. URL: http://www.sciencedirect.com/science/article/pii/S0096300310002006.

Barabasi, Albert-Laszlo and Zoltan N Oltvai (2004). "Network biology: understanding the cell's functional organization". In: *Nat Rev Genet* 5.2, pp. 101–113. ISSN: 1471-0056. URL: http://dx.doi.org/10.1038/nrg1272.

Barbano, Paolo E et al. (2007). "A mathematical tool for exploring the dynamics of biological networks". In: *Proceedings of the National Academy of Sciences* 104.49, pp. 19169–19174.

Bardot, V et al. (1994). "Purine and pyrimidine metabolism in human gliomas: relation to chromosomal aberrations". In: *British journal of cancer* 70.2, pp. 212–218. ISSN: 0007-0920. DOI: 10.1038/bjc.1994.282. URL: http://europepmc.org/articles/PMC2033491.

Barillot, Emmanuel et al. (2012). *Computational systems biology of cancer*. CRC Press. ISBN: 9781439831441.

Bertalanffy, L. von (1968). *General System Theory*. George Braziller Publisher.

Bhalla, U S and R Iyengar (1999). "Emergent properties of networks of biological signaling pathways." en. In: *Science* 283.5400, pp. 381–7. ISSN: 0036-8075. DOI: 10.1126/science.283.5400.381. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{\&}db=PubMed{\&}dopt=Citation{\&}list{_}uids=9888852$\backslash$nhttp://www.ncbi.nlm.nih.gov/pubmed/9888852.

Bihan, Frédéric (2014). "Irrational mixed decomposition and sharp fewnomial bounds for tropical polynomial systems". In: *arXiv preprint arXiv:1410.7905*.

Bild, Andrea H et al. (2006). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." In: *Nature* 439.7074, pp. 353–357. DOI: 10.1038/nature04296. URL: http://dx.doi.org/10.1038/nature04296.

Bogart, Tristram et al. (2007). "Computing tropical varieties". In: *Journal of Symbolic Computation* 42.1, pp. 54–73.

Bondell, Howard D. and Brian J. Reich (2008). "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR". In: *Biometrics* 64.1, pp. 115–123. ISSN: 1541-0420. DOI: 10.1111/j.1541-0420.2007.00843.x. URL: http://dx.doi.org/10.1111/j.1541-0420.2007.00843.x.

Bovier, Anton and Frank den Hollander (2015). *Metastability. A Potential-Theoretic Approach*. Springer International Publishing, p. 581. ISBN: 978-3-319-24775-5.

Bowman, Gregory R et al. (2009). "Progress and challenges in the automated construction of Markov state models for full protein systems". In: *The Journal of Chemical Physics* 131.12, p. 124101.

Brase, Jan C. et al. (2011). "TMPRSS2-ERG -specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling". In: *BMC Cancer* 11.1, pp. 1–8. ISSN: 1471-2407. DOI: 10.1186/1471-2407-11-507. URL: http://dx.doi.org/10.1186/1471-2407-11-507.

Brehme, Marc et al. (2016). "Combined Population Dynamics and Entropy Modelling Supports Patient Stratification in Chronic Myeloid Leukemia". In: *Scientific Reports* 6, 24057 EP –. URL: http://dx.doi.org/10.1038/srep24057.

Bruno, A D (2000). *Power Geometry in Algebraic and Differential Equations*. San Diego, CA: Elsevier. URL: http://cds.cern.ch/record/1086406.

Bühlmann, Peter et al. (2013). "Correlated variables in regression: Clustering and sparse estimation". In: *Journal of Statistical Planning and Inference* 143.11, pp. 1835–1858. ISSN: 0378-3758. DOI: http://dx.doi.org/10.1016/j.jspi.2013.05.019. URL: http://www.sciencedirect.com/science/article/pii/S0378375813001225.

Capra, F (1997). *The Web of Life: A New Scientific Understanding of Living Systems*. Knopf Doubleday Publishing Group. ISBN: 9780385476768. URL: http://books.google.de/books?id=3OAIaUZQhM0C.

Chen, Pei et al. (2016). "Detecting critical state before phase transition of complex biological systems by hidden Markov model". In: *Bioinformatics* 32.March, btw154. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw154. URL: http://bioinformatics.oxfordjournals.org/content/early/2016/03/18/bioinformatics.btw154.short?rss=1.

Chiavazzo, Eliodoro and Ilya Karlin (2011). "Adaptive simplification of complex multiscale systems". In: *Physical Review E* 83.3, p. 036706.

Chung, Seung-Wook et al. (2009). "Quantitative modeling and analysis of the transforming growth factor β signaling pathway". In: *Biophysical Journal* 96.5, pp. 1733–1750.

Clarke, BruceL. (1988). "Stoichiometric network analysis". English. In: *Cell Biophysics* 12.1, pp. 237–253. ISSN: 0163-4992. DOI: 10.1007/BF02918360. URL: http://dx.doi.org/10.1007/BF02918360.

Clarke, Edmund M, Orna Grumberg, and Doron Peled (1999). *Model checking*. MIT press.

Craver, Carl F. (2006). "When mechanistic models explain". In: *Synthese* 153.3, pp. 355–376. ISSN: 1573-0964. DOI: 10.1007/s11229-006-9097-x. URL: http://dx.doi.org/10.1007/s11229-006-9097-x.

Croft, David et al. (2011). "Reactome: a database of reactions, pathways and biological processes". In: *Nucleic Acids Research* 39.suppl 1, pp. D691–D697. DOI: 10.1093/nar/gkq1018. eprint: http://nar.oxfordjournals.org/content/39/suppl_1/D691.full.pdf+html. URL: http://nar.oxfordjournals.org/content/39/suppl_1/D691.abstract.

Dassow, George von et al. (2000). "The segment polarity network is a robust developmental module". In: *Nature* 406.6792, pp. 188–192.

Davidich, Maria and Stefan Bornholdt (2008). "The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model". In: *Journal of Theoretical Biology* 255.3, pp. 269–277.

Davydow, Alex and Dima Grigoriev (2015). "Bounds on the number of connected components for tropical prevarieties". In: *arXiv preprint arXiv:1511.06609*.

De Moura, Leonardo and Nikolaj Bjørner (2008). "Z3: An Efficient SMT Solver". In: *Proceedings of the Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. TACAS'08/ETAPS'08. Budapest,

Hungary: Springer-Verlag, pp. 337–340. ISBN: 3-540-78799-2, 978-3-540-78799-0. URL: http://dl.acm.org/citation.cfm?id=1792734.1792766.

Demetrius, Lloyd and Thomas Manke (2005). "Robustness and network evolution—an entropic principle". In: *Physica A: Statistical Mechanics and its Applications* 346.3, pp. 682–696.

Deuflhard, Peter and J. Heroth (1996). *Dynamic dimension reduction in ODE models*. Springer.

Dolzmann, Andreas and Thomas Sturm (1997). "REDLOG: Computer Algebra Meets Computer Logic". In: *ACM SIGSAM Bulletin* 31.2, pp. 2–9. ISSN: 0163-5824.

Dräger, Andreas et al. (2015). "SBMLsqueezer 2: context-sensitive creation of kinetic equations in biochemical networks". In: *BMC Systems Biology* 9.1, pp. 1–17. ISSN: 1752-0509. DOI: 10.1186/s12918-015-0212-9. URL: http://dx.doi.org/10.1186/s12918-015-0212-9.

Dubrova, Elena and Maxim Teslenko (2011). "A SAT-based algorithm for finding attractors in synchronous boolean networks". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 8.5, pp. 1393–1399.

Einsiedler, Manfred, Mikhail Kapranov, and Douglas Lind (2006). "Non-archimedean amoebas and tropical varieties". In: *Journal für die reine und angewandte Mathematik (Crelles Journal)* 2006.601, pp. 139–157.

Emiris, Ioannis Z. and John F. Canny (1995). "Efficient Incremental Algorithms for the Sparse Resultant and the Mixed Volume". In: *Journal of Symbolic Computation* 20.2, pp. 117 – 149. ISSN: 0747-7171. DOI: http://dx.doi.org/10.1006/jsco.1995.1041. URL: http://www.sciencedirect.com/science/article/pii/S0747717185710413.

Errami, Hassan et al. (2015). "Detection of Hopf bifurcations in chemical reaction networks using convex coordinates". In: *Journal of Computational Physics* 291, pp. 279–302. ISSN: 0021-9991. DOI: http://dx.doi.org/10.1016/j.jcp.2015.02.050. URL: http://www.sciencedirect.com/science/article/pii/S0021999115001400.

Faust, Karoline, Didier Croes, and Jacques van Helden (2009). "Metabolic Pathfinding Using {RPAIR} Annotation". In: *Journal of Molecular Biology* 388.2, pp. 390–414. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1016/j.jmb.2009.03.006. URL: http://www.sciencedirect.com/science/article/pii/S0022283609002654.

Feinberg, Martin (1987). "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, pp. 2229 –2268. ISSN: 0009-2509. DOI: http://dx.doi.org/10.1016/0009-2509(87)80099-4. URL: http://www.sciencedirect.com/science/article/pii/0009250987800994.

Fenichel, Neil (1979). "Geometric singular perturbation theory for ordinary differential equations". In: *Journal of Differential Equations* 31.1, pp. 53 –98. ISSN: 0022-0396. DOI: http://dx.doi.org/10.1016/0022-0396(79)90152-9. URL: http://www.sciencedirect.com/science/article/pii/0022039679901529.

Feret, J et al. (2009). "Internal coarse-graining of molecular systems". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6453–6458. DOI: 10.1073/pnas.0809908106. eprint: http://www.pnas.org/content/106/16/6453.full.pdf+html. URL: http://www.pnas.org/content/106/16/6453.abstract.

Fukuda, Komei and Alain Prodon (1996). "Double description method revisited". In: *Combinatorics and Computer Science: 8th Franco-Japanese and 4th Franco-Chinese Conference Brest, France, July 3–5, 1995 Selected Papers*. Ed. by Michel Deza, Reinhardt Euler, and Ioannis Manoussakis. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 91–111. ISBN: 978-3-540-70627-4. DOI: 10.1007/3-540-61576-8_77. URL: http://dx.doi.org/10.1007/3-540-61576-8_77.

Gatermann, Karin, Markus Eiswirth, and Anke Sensse (2005a). "Toric ideals and graph theory to analyze Hopf bifurcations in mass action systems". In: *Journal of Symbolic Computation* 40.6, pp. 1361 –1382. ISSN: 0747-7171. DOI: 10.1016/j.jsc.2005.07.002. URL: http://www.sciencedirect.com/science/article/pii/S0747717105001215.

Gatermann, Karin, Markus Eiswirth, and Anke Sensse (2005b). "Toric ideals and graph theory to analyze Hopf bifurcations in mass action systems". In: *Journal of Symbolic Computation* 40.6, pp. 1361–1382. ISSN: 0747-7171. DOI: http://dx.doi.org/10.1016/j.jsc.2005.07.002. URL: http://www.sciencedirect.com/science/article/pii/S0747717105001215.

Gatermann, Karin and Birkett Huber (2002). "A Family of Sparse Polynomial Systems Arising in Chemical Reaction Systems". In: *Journal of Symbolic Computation* 33.3, pp. 275 –305. ISSN: 0747-7171. DOI: http://dx.doi.org/10.1006/jsco.2001.0512. URL: http://www.sciencedirect.com/science/article/pii/S0747717101905127.

Gawrilow, E and M Joswig (2000). "Polymake: a framework for analyzing convex polytopes". In: *Polytopes Combinatorics and Computation*. URL: http://link.springer.com/chapter/10.1007/978-3-0348-8438-9_2.

Gholami, Amin Moghaddas et al. (2013). "Global proteome analysis of the NCI-60 cell line panel". In: *Cell Reports* 4.3, pp. 609–620.

Goeman, Jelle J. et al. (2004). "A global test for groups of genes: testing association with a clinical outcome". In: *Bioinformatics* 20.1, pp. 93–99. DOI: 10.1093/bioinformatics/btg382. eprint: http://bioinformatics.oxfordjournals.org/content/20/1/93.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/20/1/93.abstract.

Gorban, A N and I V Karlin (2005). *Invariant manifolds for physical and chemical kinetics*. Vol. 660. Lect. Notes Phys. Springer.

Gorban, A. N. and O. Radulescu (2008). "Dynamic and static limitation in reaction networks, revisited". In: *Advances in Chemical Engineering – Mathematics in Chemical Kinetics and Engineering*. Ed. by David West Guy B. Marin and Gregory S. Yablonsky. Vol. 34. Advances in Chemical Engineering. Elsevier, pp. 103–173. DOI: 10.1016/S0065-2377(08)00002-1.

Gorban, Alexander N. and Iliya V. Karlin (2003). "Method of invariant manifold for chemical kinetics". In: *Chemical Engineering Science* 58.21. International Symposium on Mathematics in Chemical Kinetics and Engineering, pp. 4751 –4768. ISSN: 0009-2509. DOI: http://dx.doi.org/10.1016/j.ces.2002.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0009250903003774.

Gorban, Alexander N and Ovidiu Radulescu (2007). "Dynamical robustness of biological networks with hierarchical distribution of time scales". In: *IET Systems Biology* 1.4, pp. 238–246.

Gorban, Alexander N., Ovidiu Radulescu, and Andrei Y. Zinovyev (2010). "Asymptotology of chemical reaction networks". In: *Chemical Engineering Science* 65.7. International Symposium on Mathematics in Chemical Kinetics and Engineering, pp. 2310–2324. ISSN: 0009-2509. DOI: 10.1016/j.ces.2009.09.005.

Gorban, Alexander N and Muhammad Shahzad (2011). "The Michaelis-Menten-Stueckelberg theorem". In: *Entropy* 13.5, pp. 966–1019.

Grigor'ev, D. Yu. and M. F. Singer (1991). "Solving Ordinary Differential Equations in Terms of Series with Real Exponents". English. In: *Transactions of the American Mathematical Society* 327.1, pp. 329–351. ISSN: 00029947. URL: http://www.jstor.org/stable/2001845.

Grigoriev, Dima and Vladimir V. Podolskii (2013). "Complexity of Tropical and Min-plus Linear Prevarieties". In: *computational complexity* 24.1, pp. 31–64. ISSN: 1420-8954. DOI: 10.1007/s00037-013-0077-5. URL: http://dx.doi.org/10.1007/s00037-013-0077-5.

Grigoriev, Dima and Andreas Weber (2012). "Complexity of Solving Systems with Few Independent Monomials and Applications to Mass-Action Kinetics". English. In: *Computer Algebra in Scientific Computing*. Ed. by Vladimir P. Gerdt et al. Vol. 7442. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 143–154. ISBN: 978-3-642-32972-2. DOI: 10.1007/978-3-642-32973-9_12. URL: http://dx.doi.org/10.1007/978-3-642-32973-9_12.

Gross, Charles G (1998). "Claude Bernard and the Constancy of the Internal Environment". In: *The Neuroscientist* 4.5, pp. 380–385. DOI: 10.1177/107385849800400520. URL: http://nro.sagepub.com/content/4/5/380.abstract.

Gunawardena, J. (2003). *Chemical reaction network theory for in-silico biologists*. Technical Report. http://vcp.med.harvard.edu/papers/crnt.pdf.

Gurobi Optimization (2012). *Gurobi optimizer reference manual*. URL: URL : http://www.gurobi.com.

Gutenkunst, Ryan N et al. (2007). "Universally sloppy parameter sensitivities in systems biology models". In: *PLoS computational biology* 3.10, e189.

Haller, George and Themistoklis Sapsis (2010). "Localized instability and attraction along invariant manifolds". In: *SIAM Journal on Applied Dynamical Systems* 9.2, pp. 611–633.

Heidergott, Bernd, Geert Jan Olsder, and Jacob W. van der Woude (2006). *Max Plus at work : modeling and analysis of synchronized systems : a course on Max-Plus algebra and its applications*. Princeton series in applied mathematics. Princeton (N.J.): Princeton University Press. ISBN: 0-691-11763-2. URL: http://opac.inria.fr/record=b1119405.

Henk, Martin, Jürgen Richter-Gebert, and Günter M Ziegler (2004). "16 Basic Properties of Convex Polytopes". In: *Handbook of discrete and computational geometry*, p. 355.

Hept, Kerstin and Thorsten Theobald (2009). "Tropical bases by regular projections". In: *Proceedings of the American Mathematical Society* 137.7, pp. 2233–2241.

Hirsch, Morris W, Stephen Smale, and Robert L Devaney (2012). *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

Hoerl, Arthur E. and Robert W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems". In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634. eprint: http://amstat.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634. URL: http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

Hoops, Stefan et al. (2006). "COPASI – a complex pathway simulator". In: *Bioinformatics* 22.24, pp. 3067–3074. DOI: 10.1093/bioinformatics/btl485. eprint: http://bioinformatics.oxfordjournals.org/content/22/24/3067.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/22/24/3067.abstract.

Huang, Sui, Ingemar Ernberg, and Stuart Kauffman (2009). "Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective". In: *Seminars in Cell & Developmental Biology* 20.7, pp. 869–876. ISSN: 10849521. DOI: 10.1016/j.semcdb.2009.07.003. URL: http://linkinghub.elsevier.com/retrieve/pii/S1084952109001499.

Hunt, Kristopher A et al. (2014). "Complete enumeration of elementary flux modes through scalable, demand-based subnetwork definition". In: *Bioinformatics*, btu021.

Jain, Anil K. and Richard C. Dubes (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN: 0-13-022278-X.

Jensen, Anders N. (2006). "A Presentation of the Gfan Software". In: *Proceedings of the Second International Conference on Mathematical Software*. ICMS'06. Castro Urdiales, Spain: Springer-Verlag, pp. 222–224. ISBN: 3-540-38084-1, 978-3-540-38084-9. DOI: 10.1007/11832225_21. URL: http://dx.doi.org/10.1007/11832225_21.

Jirsa, Viktor K. et al. (2014). "On the nature of seizure dynamics". In: *Brain* 137.8, pp. 2210–2230. ISSN: 0006-8950. DOI: 10.1093/brain/awu133. eprint: http://brain.oxfordjournals.org/content/137/8/2210.full.pdf. URL: http://brain.oxfordjournals.org/content/137/8/2210.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127. DOI: 10.1093/biostatistics/kxj037. eprint: http://biostatistics.oxfordjournals.org/

content/8/1/118.full.pdf+html. URL: http://biostatistics.oxfordjournals.org/content/8/1/118.abstract.

Kaleta, Christoph et al. (2009). "EFMEvolver: computing elementary flux modes in genome-scale metabolic networks". In: *Lect. Notes Inform.* Pp. 179–89.

Kanehisa, Minoru and Susumu Goto (2000). "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1, pp. 27–30. DOI: 10.1093/nar/28.1.27. eprint: http://nar.oxfordjournals.org/content/28/1/27.full.pdf+html. URL: http://nar.oxfordjournals.org/content/28/1/27.abstract.

Kauffman, Stuart (2004). "A proposal for using the ensemble approach to understand genetic regulatory networks". In: *Journal of theoretical biology* 230.4, pp. 581–590.

Kauffman, Stuart A (1969). "Metabolic stability and epigenesis in randomly constructed genetic nets". In: *Journal of Theoretical Biology* 22.3, pp. 437–467.

Kitano, H. (2002a). "Systems Biology: A Brief Overview". In: *Science* 295.5560, pp. 1662–1664. ISSN: 00368075. DOI: 10.1126/science.1069492. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1069492.

Kitano, Hiroaki (2002b). "Computational systems biology". In: *Nature* 420.6912, pp. 206–210. URL: http://dx.doi.org/10.1038/nature01254.

Klamt, Steffen and J00F6rg Stelling (2002). "Combinatorial Complexity of Pathway Analysis in Metabolic Networks". English. In: *Molecular Biology Reports* 29.1-2, pp. 233–236. ISSN: 0301-4851. DOI: 10.1023/A:1020390132244. URL: http://dx.doi.org/10.1023/A%3A1020390132244.

Korem, Yael et al. (2015). "Geometry of the Gene Expression Space of Individual Cells". In: *PLoS Comput Biol* 11.7, e1004224.

Kotera, Masaaki et al. (2004). "RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions". In: *Genome Informatics* 15, P062.

Krivulin, Nikolai (2014). "Tropical optimization problems". In: *arXiv preprint arXiv:1408.0313*.

Lam, SH and DA Goussis (1994). "The CSP method for simplifying kinetics". In: *International Journal of Chemical Kinetics* 26.4, pp. 461–486.

Le Novére, Nicolas et al. (2006). "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems". In: *Nucleic Acids Research* 34.suppl 1, pp. D689–D691. DOI: 10.1093/nar/gkj092. eprint: http://nar.oxfordjournals.org/content/34/suppl_1/D689.full.pdf+html. URL: http://nar.oxfordjournals.org/content/34/suppl_1/D689.abstract.

Le Novere, Nicolas et al. (2006). "BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems". In: *Nucleic Acids Research* 34.suppl 1, pp. D689–D691. DOI: 10.1093/nar/gkj092. eprint: http://nar.oxfordjournals.org/content/34/suppl_1/D689.full.pdf+html. URL: http://nar.oxfordjournals.org/content/34/suppl_1/D689.abstract.

Le Roy, Christine and Jeffrey L Wrana (2005). "Clathrin-and non-clathrin-mediated endocytic regulation of cell signalling". In: *Nature Reviews Molecular Cell Biology* 6.2, pp. 112–126.

Lin, C. and L. Segel (1988). *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611971347. eprint: http://epubs.siam.org/doi/pdf/10.1137/1.9781611971347. URL: http://epubs.siam.org/doi/abs/10.1137/1.9781611971347.

Litvinov, GL (2007). "Maslov dequantization, idempotent and tropical mathematics: a brief introduction". In: *Journal of Mathematical Sciences* 140.3, pp. 426–444.

Litvinov, G.L., V.P. Maslov, and G.B. Shpiz (2001). "Idempotent functional analysis: an algebraic approach". In: *Mathematical Notes* 69.5, pp. 696–729.

Llaneras, Francisco and Jesús Picó (2010). "Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal

generators". In: *Journal of Biomedicine and Biotechnology* vol. 2010, Article ID 753904, 13 pages. DOI: 10.1155/2010/753904.

Maas, Ulrich and Stephen B Pope (1992). "Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space". In: *Combustion and Flame* 88.3, pp. 239–264.

MacArthur, Ben D., Avi Ma'ayan, and Ihor R. Lemischka (2009). "Systems biology of stem cell fate and cellular reprogramming". In: *Nat Rev Mol Cell Biol* 10.10, pp. 672–681. ISSN: 1471-0072. DOI: 10.1038/nrm2766. URL: http://dx.doi.org/10.1038/nrm2766.

Machado, Daniel et al. (2011). "Modeling formalisms in Systems Biology". In: *AMB Express* 1.1, pp. 1–14. ISSN: 2191-0855. DOI: 10.1186/2191-0855-1-45. URL: http://dx.doi.org/10.1186/2191-0855-1-45.

Maciejewski, Henryk (2014). "Gene set analysis methods: statistical models and methodological differences". In: *Briefings in Bioinformatics* 15.4, pp. 504–518. DOI: 10.1093/bib/bbt002. eprint: http://bib.oxfordjournals.org/content/15/4/504.full.pdf+html. URL: http://bib.oxfordjournals.org/content/15/4/504.abstract.

Maclagan, Diane and Bernd Sturmfels (2015). "Introduction to tropical geometry". In: vol. 161. Graduate Studies in Mathematics. American Mathematical Society, RI.

Maechler, Martin et al. (2015). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3.

Maslov, Victor P and Vasiliy N Kolokoltsov (1994). *Idempotent analysis and its applications to optimal control theory*.

McCullagh, P. and J. A. Nelder (2000). *Generalized Linear Models*. Chapman & Hall/CRC. Chap. Chapter 13: Models for Survival Data.

Meiske, Wolfgang (1978). "An approximate solution of the Michaelis-Menten mechanism for quasi-steady and state quasi-equilibrium". In: *Mathematical Biosciences* 42.1, pp. 63–71.

Meshkat, Nicolette, Marisa Eisenberg, and Joseph J DiStefano (2009). "An algorithm for finding globally identifiable parameter combinations of nonlinear ODE models using Gröbner Bases". In: *Mathematical biosciences* 222.2, pp. 61–72.

Millán, Mercedes Pérez et al. (2012). "Chemical reaction systems with toric steady states". In: *Bulletin of mathematical biology* 74.5, pp. 1027–1065.

Milnor, John (1985). "On the concept of attractor". In: *Communications in Mathematical Physics* 99.2, pp. 177–195. ISSN: 0010-3616. DOI: 10.1007/BF01212280. URL: http://link.springer.com/10.1007/BF01212280.

Moustakas, Aristidis and Carl-Henrik Heldin (2012). "Induction of epithelial–mesenchymal transition by transforming growth factor β". In: *Seminars in Cancer Biology*. Vol. 22. 5. Elsevier, pp. 446–454.

Nabli, Faten et al. (2016). "On enumerating minimal siphons in Petri nets using CLP and SAT solvers: theoretical and practical complexity". In: *Constraints* 21.2, pp. 251–276. ISSN: 1572-9354. DOI: 10.1007/s10601-015-9190-1. URL: http://dx.doi.org/10.1007/s10601-015-9190-1.

Naldi, Aurélien, Denis Thieffry, and Claudine Chaouiya (2007). "Decision diagrams for the representation and analysis of logical models of genetic networks". In: *Computational Methods in Systems Biology*. Springer, pp. 233–247.

Nam, Dougu and Seon-Young Kim (2008). "Gene-set approach for expression pattern analysis". In: *Briefings in Bioinformatics* 9.3, pp. 189–197. DOI: 10.1093/bib/bbn001. eprint: http://bib.oxfordjournals.org/content/9/3/189.full.pdf+html. URL: http://bib.oxfordjournals.org/content/9/3/189.abstract.

Neve, Richard M et al. (2006). "A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes". In: *Cancer Cell* 10.6, pp. 515–527.

Noble, Denis (2008). "Claude Bernard, the first systems biologist, and the future of physiology." In: *Experimental physiology* 93.1, pp. 16–26. ISSN: 0958-0670. DOI: 10.1113/expphysiol.2007.038695. URL: http://www.ncbi.nlm.nih.gov/pubmed/17951329.

Noël, Vincent (2012). "Modèles réduits et hybrides de réseaux de réactions biochimiques: applications à la modélisation du cycle cellulaire". PhD thesis. Université Rennes 1.

Noel, Vincent et al. (2012). "Tropical Geometries and Dynamics of Biochemical Networks Application to Hybrid Cell Cycle Models". In: *Proceedings of the 2nd International Workshop on Static Analysis and Systems Biology (SASB 2011)*. Ed. by Jérôme Feret and Andre Levchenko. Vol. 284. Electronic Notes in Theoretical Computer Science. Elsevier, pp. 75–91.

— (2014). "Topical and Idempotent Mathematics and Applications". In: ed. by G Litvinov and S Sergeev. Vol. 616. American Mathematical Society. Chap. Tropicalization and tropical equilibration of chemical reactions.

Okino, Miles S., and Michael L. Mavrovouniotis (1998). "Simplification of Mathematical Models of Chemical Reaction Systems". In: *Chemical Reviews* 98.2. PMID: 11848905, pp. 391–408. DOI: 10.1021/cr9502231. eprint: http://dx.doi.org/10.1021/cr9502231. URL: http://dx.doi.org/10.1021/cr9502231.

Pachter, Lior and Bernd Sturmfels (2004). "Tropical geometry of statistical models". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.46, pp. 16132–16137. DOI: 10.1073/pnas.0406010101. eprint: http://www.pnas.org/content/101/46/16132.full.pdf. URL: http://www.pnas.org/content/101/46/16132.abstract.

Papin, Jason A. et al. (2003). "Metabolic pathways in the post-genome era". In: *Trends in Biochemical Sciences* 28.5, pp. 250–258. ISSN: 0968-0004. DOI: http://dx.doi.org/10.1016/S0968-0004(03)00064-1. URL: http://www.sciencedirect.com/science/article/pii/S0968000403000641.

Parker, Alexander Scott et al. (2007). "Lower expression levels of the transforming growth factor beta receptor type II protein are associated with a less aggressive tumor phenotype and improved survival among patients with clear cell renal cell carcinoma". In: *Human Pathology* 38.3, pp. 453–461.

Qian, H. and D. A. Beard (2006). "Metabolic futile cycles and their functions: a systems analysis of energy and control". In: *IEE Proceedings - Systems Biology* 153.4, pp. 192–200. ISSN: 1741-2471. DOI: 10.1049/ip-syb:20050086.

Rabinovich, Mikhail I et al. (2006). "Dynamical principles in neuroscience". In: *Reviews of modern physics* 78.4, p. 1213.

Rabinovich, Mikhail I et al. (2012). "Information flow dynamics in the brain". In: *Physics of Life Reviews* 9.1, pp. 51–73.

Radulescu, O., S. Vakulenko, and D. Grigoriev (2015). "Model Reduction of Biochemical Reactions Networks by Tropical Analysis Methods". In: *Math. Model. Nat. Phenom.* 10.3, pp. 124–138. DOI: 10.1051/mmnp/201510310. URL: http://dx.doi.org/10.1051/mmnp/201510310.

Radulescu, O, A Zinovyev, and A Lilienbaum (2007). "Model Reduction and Model Comparison for NFκB Signalling". In: *Proceedings of 2nd Conference Foundations of Systems Biology in Engineering (FOSBE-2007)*. Suttgart, Germany.

Radulescu, Ovidiu et al. (2008). "Robust simplifications of multiscale biochemical networks." En. In: *BMC systems biology* 2.1, p. 86. ISSN: 1752-0509. DOI: 10.1186/1752-0509-2-86. URL: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-2-86.

Radulescu, Ovidiu et al. (2012). "Reduction of dynamical biochemical reactions networks in computational biology". In: *Frontiers in Genetics* 3.131. ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00131.

Radulescu, Ovidiu et al. (2015). "Symbolic dynamics of biochemical pathways as finite states machines". In: *Computational Methods in Systems Biology - 13th International Conference (CMSB 2015)*. Ed. by Olivier F Roux and Jérémie Bourdon. Vol. 9308. Lecture Notes in Computer Science. Springer, pp. 104–120. DOI: 10.1007/978-3-319-23401-4{_}10.

Rand, David A (2008). "Mapping global sensitivity of cellular network dynamics: sensitivity heat maps and a global summation law". In: *Journal of The Royal Society Interface* 5.Suppl 1, S59–S69.

Reder, Christine (1988). "Metabolic control theory: a structural approach". In: *Journal of Theoretical Biology* 135.2, pp. 175–201.

Rezola, Alberto et al. (2013). "Selection of human tissue-specific elementary flux modes using gene expression data". In: *Bioinformatics* 29.16, pp. 2009–2016. DOI: 10.1093/bioinformatics/btt328. eprint: http://bioinformatics.oxfordjournals.org/content/29/16/2009.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/29/16/2009.abstract.

Rezola, Alberto et al. (2014). "*In-Silico* Prediction of Key Metabolic Differences between Two Non-Small Cell Lung Cancer Subtypes". In: *PLoS ONE* 9.8, e103998. DOI: 10.1371/journal.pone.0103998. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0103998.

Rezola, Alberto et al. (2015). "Advances in network-based metabolic pathway analysis and gene expression data integration". In: *Briefings in Bioinformatics* 16.2, pp. 265–279. DOI: 10.1093/bib/bbu009. eprint: http://bib.oxfordjournals.org/content/16/2/265.full.pdf+html. URL: http://bib.oxfordjournals.org/content/16/2/265.abstract.

Richter-Gebert, Jurgen, Bernd Sturmfels, and Thorsten Theobald (2005). "First steps in tropical geometry". In: *Contemporary Mathematics* 377, pp. 289–318.

Robin, Xavier et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12.1, p. 77. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-77. URL: http://www.biomedcentral.com/1471-2105/12/77.

Rojas, J Maurice (2002). "Why polyhedra matter in non-linear equation solving". In: *arXiv preprint math/0212309*.

Ross, Douglas T et al. (2000). "Systematic variation in gene expression patterns in human cancer cell lines". In: *Nature Genetics* 24.3, pp. 227–235.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: http://dx.doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

Roussel, Marc R. (1997). "Classroom Note:An Analytic Center Manifold For a Simple Epidemiological Model". In: *SIAM Review* 39.1, pp. 106–109. DOI: 10.1137/S0036144594293035. eprint: http://dx.doi.org/10.1137/S0036144594293035. URL: http://dx.doi.org/10.1137/S0036144594293035.

Roussel, Marc R (2005). *Lecture Notes: Invariant manifolds*. Accessed: 2016-04-28. URL: http://people.uleth.ca/~roussel/nld/manifolds.pdf.

Rowley, Clarence W and Jerrold E Marsden (2000). "Reconstruction equations and the Karhunen–Loève expansion for systems with symmetry". In: *Physica D: Nonlinear Phenomena* 142.1, pp. 1–19.

Ruff, Michaël et al. (2015). "The Disintegrin and Metalloprotease ADAM12 Is Associated with TGF-β-Induced Epithelial to Mesenchymal Transition". In: *PloS One* 10.9, e0139179.

Rustici, Gabriella et al. (2004). "Periodic gene expression program of the fission yeast cell cycle". In: *Nature Genetics* 36.8, pp. 809–817.

Samal, Satya Swarup, Hassan Errami, and Andreas Weber (2012). "PoCaB: A Software Infrastructure to Explore Algebraic Methods for Bio-chemical Reaction Networks". English. In: *Computer Algebra in Scientific Computing*. Ed. by VladimirP. Gerdt et al. Vol. 7442. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 294–307. ISBN: 978-3-642-32972-2. DOI: 10.1007/978-3-642-32973-9_25. URL: http://dx.doi.org/10.1007/978-3-642-32973-9_25.

Samal, Satya Swarup et al. (2015a). "A Geometric Method for Model Reduction of Biochemical Networks with Polynomial Rate Functions." In: *Bulletin of mathematical biology* 77.12,

pp. 2180–211. ISSN: 1522-9602. DOI: 10.1007/s11538-015-0118-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/26597097.

Samal, Satya Swarup et al. (2015b). "Analysis of Reaction Network Systems Using Tropical Geometry". In: *Computer Algebra in Scientific Computing – 17th International Workshop (CASC 2015)*. Ed. by Vladimir P. Gerdt et al. Vol. 9301. Lecture Notes in Computer Science. Aachen, Germany: Springer, pp. 422–437. DOI: 10.1007/978-3-319-24021-3_31.

— (2015c). "Computer Algebra in Scientific Computing: 17th International Workshop, CASC 2015, Aachen, Germany, September 14-18, 2015, Proceedings". In: ed. by P. Vladimir Gerdt et al. Cham: Springer International Publishing. Chap. Analysis of Reaction Network Systems Using Tropical Geometry, pp. 424–439. ISBN: 978-3-319-24021-3. DOI: 10.1007/978-3-319-24021-3_31. URL: http://dx.doi.org/10.1007/978-3-319-24021-3_31.

Samal, Satya Swarup et al. (2016). "Geometric analysis of pathways dynamics: Application to versatility of TGF-β receptors". In: *Biosystems*, pp. –. ISSN: 0303-2647. DOI: http://dx.doi.org/10.1016/j.biosystems.2016.07.004. URL: http://www.sciencedirect.com/science/article/pii/S0303264716301174.

Savageau, M. A. and E. O. Voit (1987). "Recasting nonlinear differential equations as S-systems: a canonical nonlinear form". In: *Mathematical biosciences* 87.1, pp. 83–115.

Savageau, Michael A et al. (2009). "Phenotypes and tolerances in the design space of biochemical systems". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6435–6440.

Schilling, Christophe H, David Letscher, and Bernhard Ø Palsson (2000). "Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective". In: *Journal of Theoretical Biology* 203.3, pp. 229–248. ISSN: 0022-5193. DOI: http://dx.doi.org/10.1006/jtbi.2000.1073. URL: http://www.sciencedirect.com/science/article/pii/S0022519300910737.

Schnell, S. and P.K. Maini (2002). "Enzyme kinetics far from the standard quasi-steady-state and equilibrium approximations". In: *Mathematical and Computer Modelling* 35.1–2, pp. 137–144. ISSN: 0895-7177. DOI: http://dx.doi.org/10.1016/S0895-7177(01)00156-X. URL: http://www.sciencedirect.com/science/article/pii/S089571770100156X.

Schuster, Stefan, David A Fell, and Thomas Dandekar (2000). "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks". In: *Nature biotechnology* 18.3, pp. 326–332.

Schwartz, Jean-Marc et al. (2007). "Observing metabolic functions at the genome scale". In: *Genome Biology* 8.6, R123. ISSN: 1465-6906. DOI: 10.1186/gb-2007-8-6-r123. URL: http://genomebiology.com/2007/8/6/R123.

Schwarz, Roland et al. (2005). "YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities". In: *BMC Bioinformatics* 6.1, p. 135. ISSN: 1471-2105. DOI: 10.1186/1471-2105-6-135. URL: http://www.biomedcentral.com/1471-2105/6/135.

Segel, Lee A (1988). "On the validity of the steady state assumption of enzyme kinetics". In: *Bulletin of mathematical biology* 50.6, pp. 579–593.

Segel, Lee A and Marshall Slemrod (1989). "The quasi-steady-state assumption: a case study in perturbation". In: *SIAM review* 31.3, pp. 446–477.

Sensse, Anke, Marcus J. B. Hauser, and Markus Eiswirth (2006). "Feedback loops for Shil'nikov chaos: The peroxidase-oxidase reaction". In: *The Journal of Chemical Physics* 125.1, 014901, p. 014901. DOI: http://dx.doi.org/10.1063/1.2207140. URL: http://scitation.aip.org/content/aip/journal/jcp/125/1/10.1063/1.2207140.

Sharma, PV and Bhagwan Dash (1981). "Charaka Samhita, Vol. I". In: *Chaukhambha Orientalia, Varanasi, India*.

She, Yiyuan (2010). "Sparse regression with exact clustering". In: *Electron. J. Statist.* 4, pp. 1055–1096. DOI: 10.1214/10-EJS578. URL: http://dx.doi.org/10.1214/10-EJS578.

Simon, Noah et al. (2013). "A Sparse-Group Lasso". In: *Journal of Computational and Graphical Statistics* 22.2, pp. 231–245. DOI: 10.1080/10618600.2012.681250. eprint: http://dx.

doi.org/10.1080/10618600.2012.681250. URL: http://dx.doi.org/10.1080/10618600.2012.681250.

Soliman, Sylvain, François Fages, and Ovidiu Radulescu (2014). "A constraint solving approach to model reduction by tropical equilibration". In: *Algorithms for Molecular Biology* 9.1, p. 24.

Sommars, Jeff and Jan Verschelde (2016). "Pruning Algorithms for Pretropisms of Newton Polytopes". In: *Computer Algebra in Scientific Computing: 18th International Workshop, CASC 2016, Bucharest, Romania, September 19-23, 2016, Proceedings*. Ed. by P. Vladimir Gerdt et al. Cham: Springer International Publishing, pp. 489–503. ISBN: 978-3-319-45641-6. DOI: 10.1007/978-3-319-45641-6_31. URL: http://dx.doi.org/10.1007/978-3-319-45641-6_31.

Song, Yoo Hyun et al. (2011). "The important role of glycine N-methyltransferase in the carcinogenesis and progression of prostate cancer". In: *Mod Pathol* 24.9, pp. 1272–1280. ISSN: 0893-3952. URL: http://dx.doi.org/10.1038/modpathol.2011.76.

Sreekumar, Arun et al. (2009). "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression". In: *Nature* 457.7231, pp. 910–914. ISSN: 0028-0836. URL: http://dx.doi.org/10.1038/nature07762http://www.nature.com/nature/journal/v457/n7231/suppinfo/nature07762{_}S1.html.

Strogatz, Steven H (2014). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview press.

Sturmfels, Bernd (2002). *Solving systems of polynomial equations*. Vol. 97. CBMS Regional Conference Series in Math. American Mathematical Society, Providence, RI.

Subramanian, Aravind et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. DOI: 10.1073/pnas.0506580102. eprint: http://www.pnas.org/content/102/43/15545.full.pdf. URL: http://www.pnas.org/content/102/43/15545.abstract.

Surovtsova, Irina et al. (2009). "Accessible methods for the dynamic time-scale decomposition of biochemical systems". In: *Bioinformatics* 25.21, pp. 2816–2823. DOI: 10.1093/bioinformatics/btp451. eprint: http://bioinformatics.oxfordjournals.org/content/25/21/2816.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/25/21/2816.abstract.

Terzer, M. (2009). *Large scale methods to enumerate extreme rays and elementary modes*. Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), Zurich.

Theobald, Thorsten (2006). "On the frontiers of polynomial computations in tropical geometry". In: *Journal of Symbolic Computation* 41.12, pp. 1360–1375.

Thomas, René (1973). "Boolean formalization of genetic control circuits". In: *Journal of theoretical biology* 42.3, pp. 563–585.

— (1991). "Regulatory networks seen as asynchronous automata: a logical description". In: *Journal of Theoretical Biology* 153.1, pp. 1–23.

Thomas, René and Richard d'Ari (1990). *Biological feedback*. CRC press.

Tibshirani, R. (1996a). "Regression shrinkage and selection via the lasso". In: *J. Royal. Statist. Soc B.* 58.1, pp. 267–288.

Tibshirani, Robert (1996b). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178.

Tikhonov, Andrei Nikolaevich (1952). "Systems of differential equations containing small parameters in the derivatives". In: *Matematicheskii sbornik* 73.3, pp. 575–586.

Tsuda, Ichiro (1991). "Chaotic itinerancy as a dynamical basis of hermeneutics in brain and mind". In: *World Futures: Journal of General Evolution* 32.2-3, pp. 167–184.

Turányi, T., T. Bérces, and S. Vajda (1989). "Reaction rate analysis of complex kinetic systems". In: *International Journal of Chemical Kinetics* 21.2, pp. 83–99. ISSN: 1097-4601. DOI: 10.1002/kin.550210203. URL: http://dx.doi.org/10.1002/kin.550210203.

Tyson, John J (1991). "Modeling the cell division cycle: cdc2 and cyclin interactions". In: *Proceedings of the National Academy of Sciences* 88.16, pp. 7328–7332.

Tyson, John J, Attila Csikasz-Nagy, and Bela Novak (2002). "The dynamics of cell cycle regulation". In: *Bioessays* 24.12, pp. 1095–1109.

Vicsek, Tamas (2001a). "A question of scale". In: *Nature* 411.6836, pp. 421–421. URL: http://dx.doi.org/10.1038/35078161.

Vicsek, Tamás (2001b). *Fluctuations and scaling in biology*. Oxford University Press New York.

Vilar, JM, Ronald Jansen, and Chris Sander (2006). "Signal processing in the TGF-beta superfamily ligand-receptor network". In: *PLoS Comput Biol* 2.1, e3.

Von Neumann, John (1956). "Probabilistic logics and the synthesis of reliable organisms from unreliable components". In: *Automata studies* 34, pp. 43–98.

Španěl, Patrik et al. (1999). "Analysis of formaldehyde in the headspace of urine from bladder and prostate cancer patients using selected ion flow tube mass spectrometry". In: *Rapid Communications in Mass Spectrometry* 13.14, pp. 1354–1359. ISSN: 1097-0231. DOI: 10.1002/(SICI)1097-0231(19990730)13:14<1354::AID-RCM641>3.0.CO;2-J. URL: http://dx.doi.org/10.1002/(SICI)1097-0231(19990730)13:14<1354::AID-RCM641>3.0.CO;2-J.

Waddington, Conrad Hal (1940). *Organisers and genes*. English. Cambridge Biological Studies. University Press, Cambridge. URL: http://www.cabdirect.org/abstracts/19401601308.html$\backslash$nhttp://krishikosh.egranth.ac.in/bitstream/1/21239/1/25218.pdf.

— (1957). *The Strategy of the Genes*. London: Allen and Unwin.

Wagner, Clemens and Robert Urbanczik (2005). "The Geometry of the Flux Cone of a Metabolic Network". In: *Biophysical Journal* 89.6, pp. 3837–3845. ISSN: 0006-3495. DOI: http://dx.doi.org/10.1529/biophysj.104.055129. URL: http://www.sciencedirect.com/science/article/pii/S0006349505730265.

Weber, Andreas, Thomas Sturm, and Essam O. Abdel-Rahman (2011). "Algorithmic Global Criteria for Excluding Oscillations". English. In: *Bulletin of Mathematical Biology* 73.4, pp. 899–916. ISSN: 0092-8240. DOI: 10.1007/s11538-010-9618-0. URL: http://dx.doi.org/10.1007/s11538-010-9618-0.

Weis, Michael C et al. (2014). "A data-driven, mathematical model of mammalian cell cycle regulation". In: *PloS One* 9.5, e97130.

Weispfenning, Volker (1988). "The Complexity of Linear Problems in Fields". In: *Journal of Symbolic Computation* 5.1&2, pp. 3–27.

Westerhoff, Hans V and Bernhard O Palsson (2004). "The evolution of molecular biology into systems biology". In: *Nat Biotech* 22.10, pp. 1249–1252. ISSN: 1087-0156. URL: http://dx.doi.org/10.1038/nbt1020.

Wolf, Amparo, Sameer Agnihotri, and Abhijit Guha (2010). "Targeting Metabolic Remodeling in Glioblastoma Multiforme". In: *Oncotarget* 1.7. ISSN: 1949-2553. URL: http://www.impactjournals.com/oncotarget/index.php?journal=oncotarget&page=article&op=view&path%5B%5D=190.

Woodward, James (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Wrzodek, Clemens, Andreas Dräger, and Andreas Zell (2011). "KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats". In: *Bioinformatics* 27.16, pp. 2314–2315. DOI: 10.1093/bioinformatics/btr377. eprint: http://bioinformatics.oxfordjournals.org/content/27/16/2314.full.pdf+html. URL: http://bioinformatics.oxfordjournals.org/content/27/16/2314.abstract.

Yuan, Ming and Yi Lin (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zhang, Ying E (2009). "Non-Smad pathways in TGF-β signaling". In: *Cell Research* 19.1, pp. 128–139.

Zi, Zhike and Edda Klipp (2007). "Constraint-based modeling and kinetic analysis of the Smad dependent TGF-beta signaling pathway". In: *PLoS One* 2.9, e936–e936.

Zobeley, Jürgen et al. (2005). "A new time-dependent complexity reduction method for biochemical systems". In: *Transactions on Computational Systems Biology I*. Springer, pp. 90–110.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320. ISSN: 1467-9868. DOI: 10.1111/j.1467-9868.2005.00503.x. URL: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

List of Publications

Peer Reviewed Journals

S. S. Samal, D. Grigoriev, H. Fröhlich, A. Weber, and O. Radulescu. A geometric method for model reduction of biochemical networks with polynomial rate functions. Bulletin of Mathematical Biology, 77(12):2180–2211, 2015. DOI: 10.1007/s11538-015-0118-0.

S. S. Samal, A. Naldi, D. Grigoriev, A. Weber, N. Théret, O. Radulescu, Geometric analysis of pathways dynamics: Application to versatility of TGF-β receptors, Biosystems, Available online 21 July 2016, ISSN 0303-2647. DOI: http://dx.doi.org/10.1016/j.biosystems.2016.07.004.

I. K. Aybar, O. O. Aybar, B. Ferčec, V. G. Romanovski, S. S. Samal, and A. Weber. Investigation of Invariants of a Chemical Reaction System with Algorithms of Computer Algebra. MATCH Communications in Mathematical and in Computer Chemistry, 74(3):465–480, 2015.

D. Grigoriev, S. S. Samal, S. Vakulenko, and A. Weber. Algorithms to study large metabolic network dynamics. Math. Model. Nat. Phenom., 10(5):100– 118, 2015. DOI: 10.1051/mmnp/201510507.

Conference Proceedings

S. S. Samal, D. Grigoriev, H. Fröhlich, and O. Radulescu. Analysis of reaction network systems using tropical geometry. In Vladimir P. Gerdt, Wolfram Koepf, Werner M. Seiler, and Evgenii V. Vorozhtsov, editors, Computer Algebra in Scientific Computing, volume 9301 of Lecture Notes in Computer Science, pages 424–439. Springer International Publishing, 2015. DOI: 10.1007/978-3-319-24021-3_31.

O. Radulescu, S. S. Samal, A. Naldi, D. Grigoriev, and A. Weber. Symbolic dynamics of biochemical pathways as finite states machines. In Olivier F. Roux and Jérémie Bourdon, editors, Computational Methods in Systems Biology - 13th International Conference (CMSB 2015), volume 9308 of Lecture Notes in Computer Science, pages 104–120. Springer, 2015. DOI: 10.1007/978-3-319-23401-4_10.

S. S. Samal, O. Radulescu, D. Grigoriev, H. Fröhlich, and A. Weber. A tropical method based on newton polygon approach for algebraic analysis of biochemical reaction networks. In 9th European Conference on Mathematical and Theoretical Biology, 2014.

S. S. Samal, H. Errami, and A. Weber. Pocab: A software infrastructure to explore algebraic methods for bio-chemical reaction networks. In Vladimir P. Gerdt, Wolfram Koepf, Ernst W. Mayr, and Evgenii V. Vorozhtsov, editors, Computer Algebra in Scientific Computing, volume 7442 of Lecture Notes in Computer Science, pages 294–307. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-32973-9_25.

S. S. Samal, A. Mishra, S. Samal, J.K. Pattnaik, and P. Bhuyan. First advisory and real-time health surveillance to reduce maternal mortality using mobile technology. In Tomasz Janowski and Hrushikesha Mohanty, editors, Distributed Computing and Internet Technology, volume 5966 of Lecture Notes in Computer Science, pages 279–281. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-11659-9_30.

Posters

Satya Swarup Samal, Andreas Weber, and Holger Fröhlich. Functional Analysis of Metabolic Networks using Sparse Group Lasso. Poster abstracts of GCB 2015: 12.

Andreas Till, Anja Wieland, Roman Reinartz, Niklas Schäfer, Sied Kebir, Franziska Lorbeer, Sabine Normann, Mihaela Keller, Heike Höfer, Satya Samal Swarup, Ashar Ahmad, Joao Dinis, Matthias Simon, Holger Fröhlich, Martin Glas and Björn Scheffler. Dissection of Glioblastoma Heterogeneity by using High Content Imaging of Primary Tumor Cells. Poster abstracts of NOA Bonn 2014.

Niklas Schäfer, Holger Fröhlich, Matthias Simon, Roman Reinartz, Elke Hattingen, Anja Wieland, Björn Scheffler, Andreas Till, Martin Glas, Satya Swarup, Joao Dinis, Ulrich Herrlinger. Tracking relapse-inducing cells in human glioblastoma. Poster abstracts of 8th International Meeting des Kompetenznetzwerks für Stammzellforschung NRW fand vom 21. - 22. April 2015.