LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Helmholtz Zentrum München
Institut für Bioinformatik und Systembiologie

Masterarbeit
in Bioinformatik

Stochastic models of the genetic toggle switch

Michael Strasser

Aufgabensteller:	Prof. Fabian Theis
Betreuer:	Dr. Carsten Marr
Abgabedatum:	29.07.2010

Ich versichere, dass ich diese Masterarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

29.07.2010 ______________________________

Michael Strasser

Abstract

The toggle switch motif consists of two genes that mutually repress their expression and exhibits two stable states where one gene is upregulated and the other is repressed. The toggle switch motif frequently occurs in regulatory networks of differentiating cells and is thought to act as a cellular memory unit that chooses and maintains cell fate decisions. Recently, the dynamics of the toggle switch have been studied using deterministic models, which are justified for systems with large molecular abundances. However, due to low copy numbers of DNA and mRNA, stochastic effects can alter the system's dynamics, for example introducing random transitions between the two stable states of the toggle switch. To account for this intrinsic noise, we introduce different probabilistic models of the toggle switch. In the one-stage model of the toggle switch gene expression is condensed into a single synthesis reaction. In the two-stage model mRNA is introduced and finally, we include autoactivation into our model. For each model, we define three dynamical regimes of the toggle switch, where either of the two genes dominates or the switch is in an undecided state. From simulations we determine the prevalence of the regimes and determine the parameter dependence of systemic features. We find that increasing the mean protein level of the system also increases the switching time between the regimes. Inclusion of mRNA and autoactivation increases the switching time even more. Furthermore, we investigate the robustness of the toggle switch against parameter asymmetries with respect to the switching bias. Parameter asymmetries occur due to different protein properties, changing chromatin states or asymmetric cell division. Intriguingly, a maximally robust system is achieved by mRNA levels on the order of tens and protein levels on the order of ten thousands, granting unbiased lineage decision over a wide range of asymmetries. Autoactivation further increases the robustness of the system. In this work, we show that the study of a probabilistic toggle switch model allows the reconciliation of principles of motif dynamics with measured molecular abundances.

Zusammenfassung

Der Toggle Switch, ein regulatorisches Motif bestehend aus zwei sich gegenseitig inhibierenden Genen, ist ein häufiger Bestandteil regulatorischer Netze in sich differenzierenden Zellen. Dieses regulatorische Motif fungiert als Informationsspeicher in der Zelle. Man geht davon aus, dass die Differenzierungsentscheidung einer Zelle durch den Toggle Switch gefällt und gespeichert wird. In den letzten Jahren wurde die Dynamik des Toggle Switch anhand verschiedener deterministischer Modelle beschrieben, die allerdings nur im Falle großer Molekülzahlen anwendbar sind. Augrund der geringen Anzahl von mRNA- und DNA-Molekülen in der Zelle treten stochastische Effekte auf, die die Dynamik des System grundlegend ändern. In dieser Arbeit werden verschiedene probabilistische Modelle des Toggle Switch vorgestellt und der Einfluss der stochastischen Effekte analysiert. Im sogenannten one-stage Modell werden Transkription und Translation zu einer einzigen Synthesereaktion zusammengefasst, im two-stage Modell werden beide Prozess getrennt modelliert. Zuletzt werden die Modelle durch zusätzliche Autoaktivierung erweitert. Für jedes Modell werden drei dynamische Regime definiert: Entweder ist jeweils eines der beiden Gene hochreguliert, während das andere inhibiert ist, oder das System befindet sich in einem noch nicht entschiedenen Zustand. Anhand simulierter Zeitverläufe wird überprüft, ob die Aufenthaltswahrscheinlichkeit des Systems für die Regime ausgeglichen ist. Weiterhin wird anaylsiert, wie weitere Systemeigenschaften von den gewählten Parametern abhängen. Höhere Proteinzahlen verlängern die Zeit, die das System benötigt um das Regime zu wechseln. Durch die Hinzunahme von mRNA und Autoaktivierung wird diese Zeit zusätzlich erhöht. Die simulierten Zeiten werden mit einer analytischen Näherung verglichen. Des weiteren wird die Robustheit des Toggle Switch bezüglich des Switching Bias bei asymmetrisch gewählten Parametern untersucht. Asymmetrische Parameter werden zum Beispiel durch unterschiedliche Proteineigenschaften, sich ändernden Chromatinzustand oder asymmetrische Zellteilung verursacht. Ein maximale Robustheit wird durch mRNAs in der Größenordnung von zehn sowie durch Proteine in der Größenordnung von mehreren zehntausend Molekülen erzielt. Die Robustheit des System wird durch Autoregulation zusätzlich erhöht. Diese Arbeit zeigt, dass ein probabilistisches Modell des Toggle Switch es ermöglicht, generelle Prinzipen dieses regulatorischen Motif mit experimentell gemessenen Proteinzahlen in Verbindung zu bringen.

Acknowledgements

Thanks to Carsten Marr for being the best supervisor one could wish to have. His constant feedback and motivation made this work much easier. Thanks to Fabian Theis for giving me the opportunity to do this work in his group, allowing me to explore an interesting field of research. Thanks to all members of the CMB group for the inspiring working atmosphere.

Special thanks to Flori, my friends and family for the abundance of patience and support throughout this work.

Contents

Chapter 1

Introduction

Cellular phenotypes are the result of the well-orchestrated interplay of thousands of genes and their products. One level of regulation is mediated by proteins known as transcription factors controlling the expression of genes. Other levels of regulation, for example microRNAs or post-translational modifications act on top of this layer of transcriptional regulation. As the transcription factors themselves are gene products, they are regulated as well, leading to a complex network of regulatory interactions. Advances in high throughput experiments have unraveled large parts of these networks.

1.1 Regulatory networks in development

The developmental program of an organism is encoded in its regulatory network. Studies of large transcription factor networks mostly analyze the topology of the network. Interesting properties like scale freeness [4], modularity [42] or the small world property [63] emerged and seem to constitute basic principles in biological networks. However, the networks are only available for some organisms, e.g. *Escherichia coli* and *Saccharomyces cerevisiae* and due to the complexity of these large networks, the dynamics – giving insight to what actually happens in the cell – remain unclear. Therefore, subsets of the complete regulatory network are investigated.

Meso-scale models contain only genes involved in a certain process of interest, for example the formation of the mid-hindbrain boundary [64] or hematopoesis [29]. These networks are often modeled using boolean dynamics [18]: Genes have only two states (they are either ON or OFF), interactions between genes are modeled as boolean functions. Simulating the system with boolean dynamics leads to attractors that can be interpreted as stable configurations of gene expression. A common idea is that these attractors correspond to cellular phenotypes, e.g. different cell types [24]. An intuitive picture to illustrate the attractor concept was published by Waddington [60] (see Fig. 1.1), describing the differentiation process of a cell as a marble rolling down a hilly landscape (termed the epigenetic landscape). During differentiation the cell (marble) arrives at junctions and has to choose a direction. Ultimately it will come to rest at a stable state, corresponding to a cell type. The shape of the landscape is determined by the network topology and the parameters of regulatory interactions.

Small-scale models describe small regulatory motifs occurring much more frequently in the whole network then expected [36]. They typically consisting of two to four genes. Alon

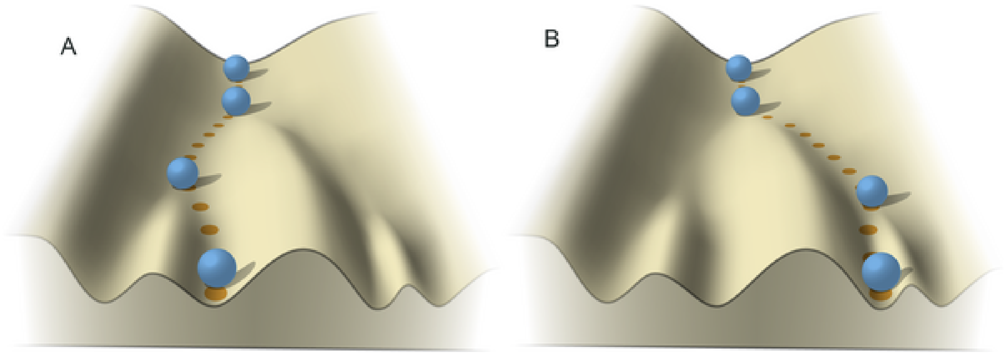

Figure 1.1: Illustration of Waddington's epigenetic landscape adopted from Mitchell [37]. The cell moves in a hilly landscape ultimately coming to rest at a stable state corresponding to a certain phenotype.

[3] suggested that these network motifs carry out specific functions, which can be studied in isolation from the complete network.

A network motif known as the toggle switch is of special interest in this work. It consists of two genes A and B that mutually repress their expression (see Fig. 1.2 A). Intuitively, this motif can be in three states: Either A is active, repressing the expression of B, or vice versa. In the third state both genes are repressed and the system is in a deadlock state. This is often referred to as priming. Gardner et al. [12] constructed a synthetic toggle switch in the bacterium *E. coli* and showed that this motif can act as a cellular memory unit in vivo. In eucaryotes this motif frequently occurs in developmental networks that control cell fate decisions. One of the best studied stem cell systems is hematopoesis, the formation of blood cells. Here, mature blood cells are derived in a hierarchical manner from the hematopoetic stem cells. Multiple occurrences of the toggle switch motif have been found [44], which are thought to control the cell fate, promoting either one or the other fate, but making them mutually exclusive (Fig. 1.2 B). One well studied example is the toggle switch of the transcription factors PU.1 and GATA-1 presumably controlling the cell fate decision of common myeloid progenitors. The two possible states of the toggle switch resemble the respective lineage choice: PU.1 is highly expressed in granulocyte/macrophage progenitors (GMPs), while GATA-1 is repressed. In megakaryocyte/erythroid progenitors (MEPs) GATA-1 is expressed, but not PU.1. In the common myeloid progenitor itself, PU.1 and GATA-1 are only expressed at a basal level, corresponding to the primed state. Another example can be found in megakaryocyte/erythroid progenitors, where the toggle switch between EKLF and Fli controls differentiation into either red blood cells or megakaryocytes. Again, the toggle switch motif occurs in cell fate decision of granulocyte/macrophage progenitors. Therefore, to understand the process of hematopoesis on a quantitative dynamical level a comprehensive analysis of the toggle switch motif is necessary.

Recently, many different models of the PU.1/GATA-1 toggle switch in myloid differentiation have be proposed. They all use ordinary differential equations to deterministically describe the dynamics of involved molecular species. Roeder and Glauche [49] introduced a model of the toggle switch including self- and crossactivation of PU.1 and GATA-1 and a PU.1/GATA-1 heterodimer that acts as an inhibitor. The system is capable of two stable states, corresponding to the two possible lineage decisions. However, to achive bistability, assumptions on molecular details are made which could not be confirmed experimentally.

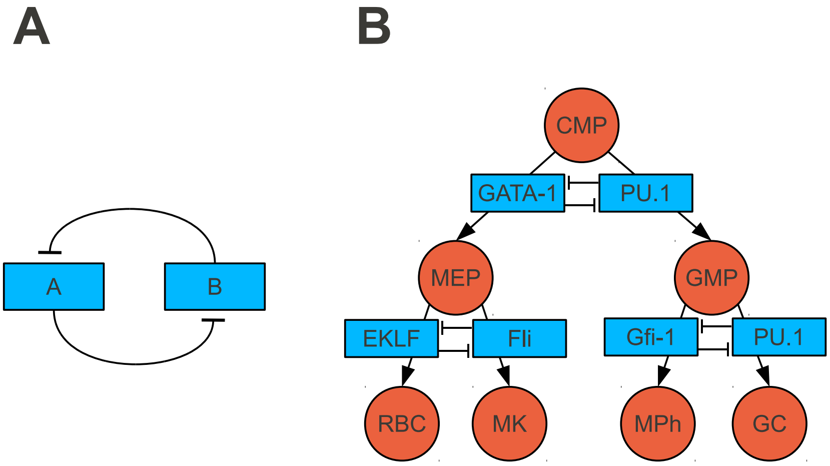

Figure 1.2: A: Scheme of the toggle switch motif. B: Repeated occurrence of the toggle switch in myeloid differentiation. Cell types are shown in red, regulatory proteins are shown in green. CMP: common myeloid progenitor, MEP: megakaryocyte/erythroid progenitor, GMP: granulocyte/macrophage progenitor, RBC: red blood cell, MK: megakaryocyte, MPh: macrophage, GC: granulocyte.

Huang et al. [21] modeled the mutual inhibition and autoactivation of the PU.1/GATA-1 toggle switch as Hill functions. The model exhibits two or three steady states respectively, depending on the parameters and thus can serve as a conceptual framework to understand commitment of differentiation cells. However, this model requires a binding cooperativity of at least two – a feature that was not experimentally validated – to achive multistability. Chickarmane et al. [8] proposed a PU.1/GATA-1 model involving an additional unknown player to exhibit bistability without binding cooperativity. Krumsiek [29] introduced an additional cosuppressor (pRB) in the PU.1/GATA-1 toggle switch, resulting in an asymmetric system. It was suggested that pRB plays an important role in the lineage decision mechanism and could actually control the switching.

All these models rely on deterministic dynamics. Speaking in terms of the epigenetic landscape: If one places the cell at the same location in the epigenetic landscape multiple times, it will always end up in the same attractor state. During the last years, the fact that cellular systems are inherently noisy became popular, for a review, see e.g. [23]. In a seminal work Elowitz et al. [10] showed that gene expression is an inherently noisy process due to the low numbers and discreteness of involved players. The cell moves in the epigenetic landscape, but is subject to random fluctuations, which e.g. can push the cell over mountain ridges in the landscape.

Probabilistic models of the toggle switch have been studied in theory by Loinger et al. [33], who examined different realizations of the toggle switch with respect to their steady state solution. Loinger et al. determined the range of parameters where bistability is possible using numerical integration and stochastic simulation and studied the effect of cooperative binding, protein-protein interaction and degradation of bound repressors on the dynamics of the switch. Schultz et al. [54] provided an analytical solution of the probabilistic toggle switch using linear noise and fast transition approximations based on the work of Kepler and Elston [26], and investigated the influence of parameters on the shape of the steady state distribution. to describe the Sasai and Wolynes [53] mapped the toggle switch to a quantum many body problem and solved the system using the Hartree approximation, also showing the influence of the parameters on the steady state distribution. A detailed discussion of existing literature will be given in section 5.1.2.

We extend these studies on the probabilistic toggle switch by

1. expanding the toggle switch model by introducing mRNA. All previously mentioned studies have considered a simple model of gene expression where transcription and translation are condensed into a single synthesis reaction of proteins.

2. analyzing the effect of different protein copy numbers on the systems dynamics.

3. investigating the robustness of the system against asymmetric parameters.

1.2 Robustness of cellular systems

Robustness is thought to be a major underlying principle of all biological systems [27]. Robust systems can maintain specific function in the presence of internal or external perturbations. Internal perturbation are e.g. fluctuations due to inherent noise, for example in gene expression, which drives the cell out of equilibrium. External perturbations are often related to environmental changes, such as changes in energy sources or stress. A classical example of a robust system is the fate decision in λ-phage. The fate decision mechanism is not the result of a set of fine-tuned systems parameters (like binding affinities between repressors and promoter sequences), but results from the structure of the regulatory circuits. Little et al. [32] showed that the system is robust against promoter point mutations in the sense that it still can maintain its function as a fate decision unit.

Following the ideas of Kitano [27], robustness in a dynamical system can be realized in two ways:

1. The system, driven out of its current state through a perturbation, falls back into the original state (attractor) and restores its function.

2. The system does not return to the current state, but instead moves to another attractor, which has a different configuration (e.g. in terms of molecular expression profile), but the same function.

According to Kitano [27], three concepts contribute to a robust system: First, system control in terms of positive and negative feedback loops allows for a robust response to perturbations. Second, redundancy and diversity allow the system to maintain its function. Even if one component fails, a second one can rescue the system. Third, modularity minimizes the spread of local perturbation across the whole system.

During this work we will investigate the robustness of the toggle switch motif against asymmetries in system parameters. We require that the system is able to maintain its function as a cellular memory unit and provides equal probability for both possible decisions. Instead of using the general definition of robustness by Kitano [27], we use an approach more similar to classical sensitivity analysis [59], quantifying the influence of the perturbation on the system's output. However, one can redefine these internal into external parameter perturbations. For example all parameters of our genetic system are related to the DNA sequence – most obviously DNA binding affinities or promoter strength, but even protein 3D structure –, and this DNA sequence is subject to external perturbations, such as UV-light, induced mutations etc.

1.3 Overview

In this work, we investigate the dynamics of different toggle switch models and examine the robustness of the toggle switch against parameter asymmetries.

In chapter 2 we introduce the general formalisms and methods used to describe and work with our models. We represent the models deterministically with ordinary differential equations (ODEs), semi-deterministically using stochastic differential equations (SDEs) and in a probabilistic manner using the chemical master equation (CME). Furthermore we introduce methods to approximate the CME either numerically or through kinetic Monte Carlo simulation. At the end of this chapter, we present the general modeling assumptions used throughout this work.

In chapter 3 we study the simplest model of the toggle switch, where gene expression is a one-stage process, neglecting the intermediate mRNA stage. We give a quantitative description of its dynamics and define the three major dynamical regimes of the toggle switch. We derive an analytical expression for the switching time of the system and find a strong dependence of the switching time on the overall protein level in the system. Finally, we focus on how the dynamics change when asymmetric parameters are introduced and find that increased protein levels enhance the robustness of the system.

In chapter 4 we extend the one-stage toggle switch by introducing an mRNA stage and elaborate how this additional complexity influences the system's behavior. We analyze the dynamics of this two-stage switch in terms of transition times. We extend the analytical expression of the switching time for the two-stage switch and find a similar dependence of the switching time on the protein level of the system. Furthermore, we assess the robustness of the system against asymmetric parameters and find an increased robustness for switches with higher protein levels.

In chapter 5 we compare both toggle switch models and find that the two-stage switches have longer switching times and are more robust than the one-stage systems. We extend the two-stage switch by autoregulation and show its influence on the dynamics. Autoregulation increases switching times and robustness even more. Finally, we give a biological interpretation of the results and line out possible extensions of the present work.

Chapter 2

Modeling framework

In this chapter we introduce the formalisms used to describe the dynamical systems, show the derivation of the stochastic simulation algorithm and discuss general modeling assumptions.

2.1 Formalisms

We start by first describing a general model in terms of biochemical reactions. We then show how one can assess the dynamics of this model using deterministic and probabilistic approaches.

2.1.1 Reactions

A general chemical reaction model consists of a set of N molecular species $X_1, \dots X_N$ and a set of M biochemical reactions with reaction rates $k_1 \dots k_m$, specifying the speed of the reactions. Each reaction μ can be written in stoichiometric form:

$$e_{\mu 1} X_1 + e_{\mu 2} X_2 \dots \xrightarrow{k_\mu} p_{\mu 1} X_1 + p_{\mu 2} X_2 \dots \quad . \tag{2.1}$$

Thereby, $e_{\mu i}$ $(p_{\mu i}) \in \mathbb{N}_0$ is the number of X_i molecules consumed (produced) by reaction μ (possibly 0 if the species does not take part in the reaction). We can represent the whole system by an educt matrix

$$E = e_{\mu i}, E \in \mathbb{N}^{(MxN)},$$

a product matrix

$$P = p_{\mu i}, P \in \mathbb{N}^{(MxN)},$$

and a vector of reaction rate constants $k = (k_\mu) \in \mathbb{R}^M$. The stoichiometric matrix is defined as $V = -E + P$. The row vector $V_\mu, \mu = 1, ..., M$ of V corresponds to the change in molecule numbers if reaction μ occurs.

2.1.2 Deterministic description: Reaction rate equations

We can treat the model deterministically by converting the biochemical reactions in equation (2.1) into a set of ordinary differential equations (ODEs) by applying the law of mass action. The ODEs describe the change of molecule species over time. We get one ODE for each molecule species in the system, which describes the time-rate change of the molecular abundance of this species in terms of the abundances of other species. Reaction constants are treated as reaction rates. This results in the following N-dimensional ODE system, which consists of reaction rate equations

$$\frac{dX_i(t)}{dt} = f_i(X_1(t) \dots X_N(t)) \ , \tag{2.2}$$

where $X_i(t)$ denotes the abundance of species X_i at time t and the functions f_i are determined by E, P and the reaction constants k:

$$f_i(X) = \sum_\mu p_{\mu i} k_\mu \prod_j X_j^{e_{\mu i}} - \sum_\mu e_{\mu i} k_\mu \prod_j X_j^{p_{\mu i}}$$

Solving the system of equations (2.2) gives the evolution of the continuous molecular abundances over time. Note that the solutions of ODEs derived from biochemical reaction networks are always non-negative, if initial conditions are non-negative [38].

The reaction rate equation approach is only valid if the molecule numbers in the system are large, so that continuous molecular abundances are an adequate description of the system and random fluctuations of species numbers can be neglected. In section 2.2 we will derive the exact conditions under which the reaction rate equation approach is justified.

2.1.3 Probabilistic description given large molecular numbers: Stochastic differential equations

Ordinary differential equations provide an easy way to assess cellular dynamics, but assume a noise free processes, which is unrealistic in a cellular environment, which is inherently noisy. An intuitive way to incorporate noise into the deterministic reaction rate equations is the addition of a white noise term $\xi(t)$ to the right hand side of the ODE:

$$\frac{dX_i(t)}{dt} = f_i(X(t)) + g_i(X(t)) \cdot \xi_i(t) \ , \tag{2.3}$$

$X_i(t)$ is now a stochastic process and g_i are functions that determine how the noise term $\xi_i(t)$ acts upon the species. These functions are not defined for stochastic differential equations in general, but we will see in section 2.2.2 that one can derive them from molecular physics in case of biochemical reaction systems. Note that $\frac{dW}{dt} = \xi(t)$, where W is called the Wiener process, which fulfills

$$\begin{aligned} W(0) &= 0 \\ E[W(t)] &= 0 \\ W(t) - W(s) &\sim N(0, t-s) \ , \end{aligned} \tag{2.4}$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The Wiener process is e.g. used to describe Brownian motion. We can rewrite equation (2.3) in terms of the Wiener process to obtain the stochastic differential equation:

$$dX_i(t) = f_i(X(t))dt + g_i(X(t)) \cdot dW(t) \ . \tag{2.5}$$

Given an stochastic differential equation of the form (2.5), the Euler-Maruyama method approximates the solution over the interval $[0, T]$ by dividing the interval in b subintervals of width $\delta = T/b$ and iterative calculation of

$$\begin{aligned} X(t+\delta) &= X(t) + f(X(t)) + g(X(t))[W(t+\delta) - W(t)] \\ &\overset{(2.4)}{=} X(t) + f(X(t)) + g(X(t))N(0,\delta) \\ &= X(t) + f(X(t)) + g(X(t))\sqrt{\delta}N(0,1) \ , \end{aligned}$$

with $X(0) = x_0$ and $0 \leq t \leq T$. This scheme allows an easy way to integrate stochastic differential equations, but has some disadvantages: Mass conservation is not guaranteed ad hoc due to the noise term unless one includes correlated noise for players which are interconnected via mass conservation.

It is not entirely clear how to select the number of integration steps b, which is the most crucial point for ODE-solvers. Large b lead to decreased performance, small b will lead to inaccurate solutions. More sophisticated algorithms that use dynamical step sizes exist, e.g. the Milstein method [28], but are not subject of this work.

Stochastic differential equations do account for fluctuations, however the inclusion of a noise term is artificial and lacks solid physical foundation. For example it is unclear how to define the function $g(X(t))$ in equation (2.5). However, in section 2.2 will will show how this term can be derived from molecular physics and we will provide the conditions under which the stochastic differential equation approach is justified.

Furthermore, when we transfer a chemical reaction model to a set of stochastic differential equations it is not guaranteed that its solution will always be non-negative (opposed to the ordinary differential equations). This is most likely when low molecule numbers are present in the system. Here, the noise term can drive the species abundance into negative amounts. Since we study a system with inherently small molecular numbers (e.g. one DNA molecule) later on, we do not apply the stochastic differential equation approach.

2.1.4 Probabilistic description: Chemical Master Equation

Often processes in genetic systems involve molecular species that are present in only very small amounts, introducing fluctuations. In these cases deterministic approaches are expected to give misleading results (examples are given in [52]).

The chemical master equation (CME) provides a fully probabilistic description of the model and is derived from underlying molecular processes. Later we show its connection to the stochastic differential equations.

We define the state of the system at time t as a vector $x(t) \in S, S \subseteq \mathbb{N}_0^N$ (S is only a subspace of $\mathbb{N}_0^N$ if mass conservation is included in the system), where S is called the model's state space and N is the number of molecule species. $x_i(t), i = 1, ...N$ is the number of molecules of species i at time t. Note that we neglect spatial position and velocities of the particles here, because we assume the system to be well stirred. As biological systems are typically not well stirred and even more, gradients in the cytoplasm are important for e.g. cell division, this assumption is questionable, but has to be applied for the sake of simplicity.

The reaction constants are interpreted as reaction probabilities per unit time. Therefore, we define the propensity $a_\mu(x)$ of a reaction μ as

$$\begin{aligned} a_\mu(x)\delta t :=& \text{Probability that, given the system is in state } x \text{ at time } t, \\ & \text{reaction } \mu \text{ will occur in the infinitesimal interval } [t, t+\delta t] \end{aligned}$$

The propensity of reaction μ is calculated as

$$a_\mu(x) = k_\mu \cdot \prod_{i=1}^{N} \binom{x_i}{e_{\mu i}} ,$$

where k_μ is the reaction constant of reaction μ and $\binom{n}{k}$ is the binomial coefficient. This expressions is based on molecular physics, taking into account the frequency of educt collision and the chance for a reaction once educts collided.

Given some initial state the system will jump from one state to the next by executing one of the possible reactions. Which reaction is executed only depends on the propensities of the system. As argued by Gillespie [14] the process $x(t)$ is described by a time-continuous Markov chain which has state dependent state-transition rates. Often $x(t)$ is also referred to as a Markov jump process. We can describe how the probability $P(x,t)$ of being in a certain state x at time t changes over time by using the chemical master equation (CME):

$$\frac{\partial P(x,t)}{\partial t} = \sum_{\mu=1}^{M} \left[a_\mu(x - V_\mu, t)P(x - V_\mu, t) - a_\mu(x,t)P(x,t)\right] \tag{2.6}$$

V is the stoichiometric matrix of the system. The first term on the right hand side describes how the system can reach state x from other states which are exactly one reaction μ away from x. Thereby the probability of state x increases. The second term takes into account that the system can leave the current state via reaction μ, which reduces the probability for state x.

We can also write the CME in matrix form if we assume an arbitrary enumeration X of the state space:

$$\frac{\partial P(X,t)}{\partial t} = Q \cdot P(X,t) \tag{2.7}$$

where the matrix Q is defined as

$$Q_{x,y} = \begin{cases} -\sum_\mu a_\mu(x) & x = y \\ a_\mu(x) & y = x + V_\mu \\ 0 & \text{otherwise} . \end{cases}$$

x and y are elements of the state space, thus the dimension of Q is equal to the size of the possibly infinite state space. Q has the following properties: It is independent of t. All its off-diagonal elements are non-negative, all of its diagonal elements are non-positive and all its columns sum to exactly zero. Note that equation (2.7) defines a system of linear differential equations.

The CME is a system of coupled linear differential equations where the number of equations is equal to the size of the state space. The solution of the CME fully describes the distribution of system states over time and hence the dynamics of the system. Often the state space is huge or infinite even for simple systems (exponential in the number of species) and analytical solutions to the CME are possible only in some cases (in particular if all propensities are linear, e.g. if the system involves only unimolecular reactions). Numerical methods do exist, but are often computationally expensive for larger systems. Therefore stochastic differential equations are often used in case of large systems. However, stochastic differential equations fail if the system exhibits large numbers of one species (which are well described by stochastic differential equations) and small numbers of another species (which are only poorly approximated using stochastic differential equations).

2.1.5 Solutions to the CME

In a few cases approximate analytical solutions to the CME can be obtained using the generating function formalism [22, 55] or methods from quantum physics [53]. The success of these method depends on the properties of the system and adequate approximations and cannot be applied to any system in general. Thus in this section we focus on a numerical approximations of the CME, called the finite state projection [40].

As shown in equation (2.7) we can express the CME in matrix form. This system of linear differential equations has the general solution

$$P(X,t) = e^{Q\cdot t} \cdot P(X,0) \; , \tag{2.8}$$

where e^{Q} is the matrix exponential of Q and $P(X,0)$ is a vector of initial probabilities.

For systems that can only reach a finite number of states due to mass conservation, equation (2.8) provides an exact solution to the CME. However for systems with large state space the matrix exponential is hard to compute. Even tough the matrix exponential is still defined for infinite matrices (corresponding to an infinite state space) – it is a mapping between infinite Banach spaces – it is obviously not possible to calculate it on a computer. For infinite state spaces it is not defined at all. To overcome this problem, one can truncate the state space in a useful manner, thereby reducing the complexity of the problem enormously. If the truncation is chosen wisely, the solution to the finite subsystem will be a good approximation to the true solution.

Subject of all finite state projection algorithms is to find a suitable truncation of the state space. We will not describe advanced algorithms (the interested reader is referred to [39, 40, 41]), but use a very basic version of finite state projection, choosing a static projection according to prior knowledge on the systems' dynamics. Suppose we find a valid truncation $S_T \subset S$ of the state space S. For this truncation we can calculate the matrix Q_T, which describes how the probabilities change inside S_T over time. Note that this system will loose probability mass: Consider a state that is a the edge of the state space. Transitions out of S_T are still possible, but transitions into the truncated space are not, resulting in an overall loss of probability in the projected system. We can capture the probability flow out of the system by introducing an artificial absorbing state $\emptyset$ and redirecting reactions that lead out of the projected space into this absorbing state. An example is shown in Fig. 2.1. By including $\emptyset$ in the matrix Q_T we get a new matrix $Q_{T'}$ where lost probability will accumulate in the absorbing state $\emptyset$. We can find the solution of the projected system by $P_{T'}(X,t) = e^{Q_{T'}\cdot t} \cdot P_{T'}(X,0)$, where $P_{T'}(\emptyset,t)$ is the probability that the system has left the finite state space. The error $\epsilon = P_{T'}(\emptyset,t)$ is introduced through the projection and provides an easy way to control the accuracy of the finite state projection. If the states that are removed by projection cannot

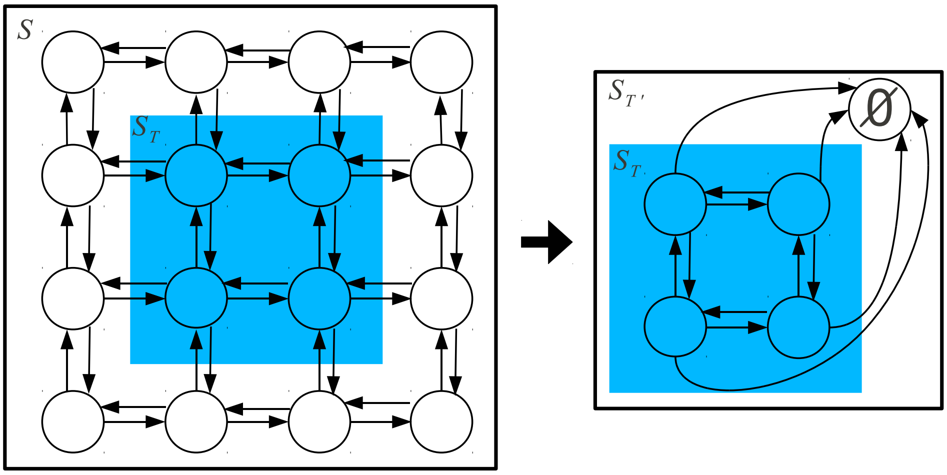

Figure 2.1: Scheme of the finite state projection. Each node corresponds to a state x of the complete state space S, arrows indicate possible state transitions. The subspace S_T is highlighted in blue. Left: state space before projection. Right: state space after projection, transitions leading out of the subspace are redirected into the absorbing state $\emptyset$.

be reached at all from the chosen initial conditions the error ϵ will be 0. In a similar way, removing states that are very unlikely to be reached will contribute only little to the overall error, whereas removing a state that has high probability to be reached will introduce large error. Therefore the general strategy of finite state projection is the identification and removal of states that have low or 0 probability to be reached from initial conditions.

Note however that this approach reaches its limits for systems that have an underlying distribution that has no distinct peaks, but where probability mass is spread equally across the state space. Here, reduction of the state space is not possible without introducing significant error.

Barzel's method for estimation of transition times

Often one is not interested in the complete probability distribution of the system over time, but only in the time it takes the system being in a certain subspace $S_1 \subset S$ to reach another subspace $S_2 \subset S$. Barzel and Biham [5] proposed a general method to calculate transition times from any subset S_1 of the state space to another subset S_2 of the state space for stochastic systems.

This method is based on a linear equation system which defines for each state x of the system the time $T(x)$ to get to a predefined set S_2 of the state space:

$$\tau(x) = M \cdot T(x) \ , \tag{2.9}$$

where

$$\tau(x) = \begin{cases} \frac{1}{\sum_m a_m(x)} & x \notin S_2 \\ 0 & x \in S_2 \end{cases} ,$$

and $M = I - N$ with

$$N_{x,y} = \begin{cases} \frac{a_m(x)}{\sum_i a_i(x)} & x \notin S_2 \wedge y = x + v_m \\ 0 & x \in S_2 \end{cases} .$$

$\tau(x)$ is the average time the system stays in state x, also called waiting time. We can solve for $T(x)$ using standard linear algebra techniques. Note that the dimension of the equation system is equal to the size of the state space and the same problems occur as for the solution of the CME itself. However, we can work on a useful projection of the state space in a similar way to the finite state projection algorithm. Beyond this boundary are only states that are very unlikely to be reached at all and are therefore excluded.

$T(x)$ gives the average time it takes to reach any state within S_2 from state x. To obtain the average time it takes to get from the subspace S_1 to S_2 we have to calculate the weighted average of $T(x), x \in S_1$. The weights are the probabilities of the states within S_1 itself, which are for example be obtained from the solution of the CME or set to user specified values.

2.2 Kinetic Monte-Carlo simulation methods

In this section we derive Gillespie's stochastic simulation algorithm [17] and show its connection to stochastic differential equations and ordinary differential equations.

2.2.1 Gillespie's algorithm

Although methods like the finite state projection can approximate the CME solution in theory (despite the difficulty to find suitable truncations of the state space), often the truncated state space is still too big to be handled, e.g. if the entries of the matrix Q exceed computer memory, which can happen when dealing with a large number of molecules in the system.

Gillespie [17] proposed the stochastic simulation algorithm (SSA) to generate timecourses that satisfy the CME, which corresponds to drawing samples from the process $x(t)$. If we can draw a sufficient number of samples we can approximate the underlying distribution which is the solution to the CME.

The algorithm to simulated trajectories in a stochastic system evolves around the probability $p(\mu, \tau | x, t)d\tau$ of only the reaction μ occurring at the time interval $[t+\tau, t+\tau+d\tau]$ given the system is in state x at time t.

The system is at time t, during τ nothing happens and during the infinitesimal small interval $d\tau$ reaction μ happens.

To derive the joint probability $p(\mu, \tau | x, t)d\tau$ of reaction μ happening in $[t+\tau, t+\tau+d\tau]$, we partition the time interval $[t, t+\tau+d\tau]$ (from the current time of the system until the next reaction happens) into $k+1$ subintervals. The first k intervals have length $\frac{\tau}{k}$, the last interval has length $d\tau$. The probability of exactly one reaction μ happening in $[t, t+\tau+d\tau]$, but not in $[t, t+\tau]$ is equal to the probability of no reaction happening in the first k subintervals ($P_0(x,t)$) multiplied by the probability of reaction μ happening in the last subinterval ($P_\mu(x,t)$):

$$\begin{aligned} P_0(x,t) &= \left[1 - a_0(x,t)\frac{\tau}{k}\right]^k \\ P_\mu(x,t) &= a_\mu(x,t) \cdot d\tau \ , \end{aligned}$$

where $a_0(x,t) = \sum_\mu a_\mu(x,t)$.

Algorithm 1: The stochastic simulation algorithm.

```
Input: Initial conditions $x_0$, maximal simulation time $t_{max}$, propensity functions $a_i$
       and stoichiometric matrix $V$
Output: Timecourse (x,t) of species abundances
$j = 0$;
$t_0 = 0$;
while $t < t_{max}$ do
    $a_0 = \sum_\mu a_\mu(x_j)$;
    $r_1$ = Random(0,1), $r_2$ = Random(0,1);
    $\tau = -\frac{\log(r_1)}{a_0}$;
    $sum = 0$;
    $\mu = 0$;
    while $sum < r_2 a_0$ do
        $\mu = \mu + 1$;
        $sum = sum + a_\mu(x_j)$;
    end
    $t_{j+1} = t_j + \tau$;
    $x_{j+1} = x_j + V_\mu$;
    $j = j + 1$
end
```

We can derive the joint probability

$$\begin{aligned} p(\mu,\tau|x,t)d\tau &= \lim_{k\to\infty} P_0(x,t)\cdot P_\mu(x,t) \\ &= \lim_{k\to\infty}\left[1 - a_0(x,t)\frac{\tau}{k}\right]^k \cdot a_\mu(x,t)\cdot d\tau \\ p(\mu,\tau|x,t) &= \lim_{k\to\infty}\left[1 - a_0(x,t)\frac{\tau}{k}\right]^k \cdot a_\mu(x,t) \\ &= e^{-a_0(x,t)\tau}\cdot a_\mu(x,t)\ . \end{aligned}$$

Samples (μ, τ) from this joint distribution can be drawn by

$$\tau = -\frac{\log(r_1)}{a_0(x,t)} \tag{2.10}$$

and μ being the smallest integer to satisfy

$$\sum_{i=1}^{\mu} a_i(x,t) > r_2 a_0(x,t)\ , \tag{2.11}$$

where r_1, r_2 are samples from a uniform distribution on the interval $[0, 1]$. This is known as the inversion method for generating exponential and uniform random numbers.

Based on this, the stochastic simulation algorithm runs by iteratively drawing τ and j according to equations (2.10) and (2.11), and updating the system's state and time (see Algorithm 1).

This results in trajectories of the system state x over time that are guaranteed to satisfy the CME. Simulating infinitely many trajectories would correspond to the solution of the CME.

2.2.2 τ-leaping

Many improvements have been made to speed up the algorithm (see e.g. [13, 17, 34]). They all are as exact as the original algorithm, but involve some twists on storing and precalculating the propensities. Often the SSA and its refinements are too slow for practical applications, because every single reaction happening in the system is simulated. If molecule numbers are large, the propensities and the sum of propensities a_0 get large. Increasing a_0 leads to smaller time steps τ in equation (2.10). To simulate an arbitrary time interval T in a system with high molecule numbers now takes many more iterations and hence computation time compared to a system with small molecule numbers, because the time steps taken are much smaller. To overcome this limitation of the SSA an approximation can be made that allows to simulate several reactions in one iteration, not only a single one. This method is known as τ-leaping [16]. Assume the system is in state x at time t. Further assume there exists a time τ so that the propensity functions do not change much during $[t, t+\tau]$:

$$a(x) \approx a(x') \;, \tag{2.12}$$

where x' is the system state at $t+\tau$.

Under these conditions the number n_μ of occurrences of a reaction μ during $[t, t+\tau]$ can be approximated by a Poisson distribution:

$$n_\mu \sim \text{Poisson}(a_\mu(x)\tau) \;.$$

The Poisson distribution expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.

So instead of simulating every reaction of μ on its own during $[t, t+\tau]$ several reactions are simulated at once by drawing a sample n_μ from a Poisson distribution for each reaction species μ and firing that reaction n_μ times. The state of the system now changes according to:

$$x(t+\tau) = x(t) + \sum_\mu \text{Poisson}(a_\mu(x)\tau) V_\mu \tag{2.13}$$

Finding a τ that satisfies the leaping condition (2.12) is not obvious, but several methods [7, 16] are available that guarantee to find such τ. All these approaches rely on finding the maximal τ by bounding the expected change in propensities during τ. Also one has to take care that the system is not driven into negative population numbers (which is possible because of the unboundedness of the Poisson distribution and lack of coordination between reactions during τ).

Moreover, τ-leaping is only advantageous if the mean number of reactions that are simulated together (the mean of the Poisson distribution, $a_\mu(x)\tau$) is large, otherwise the performance is similar to the SSA.

Chemical Langevin Equation

To see the connection to the stochastic differential equations and reaction rate equations from sections 2.1.3 and 2.1.2, in the following we show how τ-leaping can be related to stochastic differential equations and ultimately to the deterministic reaction rate equations.

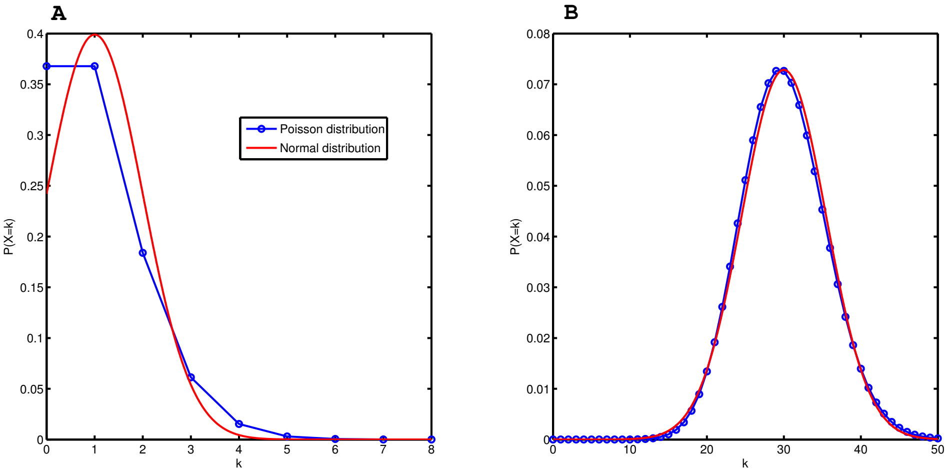

Figure 2.2: Approximation of a Poisson distribution with parameter λ through a Normal distribution with $\mu = \sigma^2 = \lambda$. A: Probability density functions for $\lambda = 2$. Both distributions differ significantly. B: Probability density functions for $\lambda = 30$. Both distribution are similar.

For τ-leaping to be useful,

$$a_\mu(x)\tau \gg 1 \; . \tag{2.14}$$

In this case the Poisson distribution is well approximated by a normal distribution (see Fig. 2.2):

$$\text{Poisson}(a_\mu(x)\tau) = \text{N}(a_\mu(x)\tau, a_\mu(x)\tau) = a_\mu(x)\tau + \sqrt{a_\mu(x)\tau}\text{N}(0,1) \; ,$$

because $\text{N}(\mu, \sigma^2) = \mu + \sigma \cdot \text{N}(0,1)$. Plugging this into the τ-leaping formula (2.13) gives the Langevin leaping formula consisting of a drift and a diffusion term:

$$x(t+\tau) = x + \sum_\mu \underbrace{V_\mu a_\mu(x)\tau}_{\text{Drift}} + \sum_\mu \underbrace{V_\mu \sqrt{a_\mu(x)} N(0,1)\sqrt{\tau}}_{\text{Diffusion}}$$

Note that by substituting the discrete Poisson distribution by the continuous Normal distribution, the discreteness of the system is lost. This can also be written as a stochastic differential equation known as the Chemical Langevin Equation [15]:

$$\frac{dx(t)}{dt} = \sum_\mu V_\mu a_\mu(x(t)) + \sum_\mu V_\mu \sqrt{a_\mu(x(t))}\Gamma_\mu(t) \; , \tag{2.15}$$

where $\Gamma_j(t)$ are independent white noise processes. The Chemical Langevin Equation is a stochastic differential equation and can be seen as the link between stochastic and deterministic descriptions of the same model: In the thermodynamic limit (species abundances x_i and system volume Ω approach infinity, with x_i/Ω being constant) the rightmost term of equation (2.15) becomes neglectable and the Chemical Langevin Equation reduces to the deterministic reaction rate equation (2.2).

We have hereby derived the conditions under which the stochastic differential equations in section 2.1.3 and reaction rate equations in section 2.1.2 are justified. Stochastic differential equation describe the dynamics with sufficient accuracy if assumptions (2.14) and (2.12) are fulfilled. Reaction rate equations can be used if the system is close to or at the thermodynamic limit.

Note that the noise term $g_i(X(t)) \cdot \xi(t)$ of the stochastic differential equation (2.15) could be derived using findings from molecular physics and is not an artificial extension of ODEs to account for fluctuations.

2.2.3 Hybrid algorithms

Both τ-leaping and the exact SSA will become inefficient when applied to stiff systems. Stiffness has no unique mathematical definition, but the main idea is that the system can be separated into slow and fast dynamical modes, which are stable. In the context of stochastic biochemical reaction systems this means that at a certain time some of the reactions will be very fast whereas the remaining reactions are slow. This separation of scales leads to inefficient simulations: The fast modes will equilibrate rather quickly, followed by a long phase where dynamics are dominated by the slow modes. Although dynamics are determined by the slow modes, the SSA has to simulate every reaction happening in the system, thus spending most of the time on the now uninteresting fast modes. The τ-leaping algorithm is also constrained to select τ according to the fast modes.

Hybrid simulation algorithms can overcome this issues. They are called hybrid, because not the whole system is treated in a stochastic manner but some parts are approximated deterministically.

One of the first hybrid algorithms for stochastic systems was introduced by Haseltine and Rawlings [19]. The mathematical derivation of this approach will not be given here and the interested reader is referred to [19, 20]. Here we only show the general idea behind the algorithm as it also describes the basic ideas for the more complex hybrid algorithms (e.g. [6]). The key idea of this hybrid algorithm is to divide the reactions into fast and slow subsets. The speed of the reaction is expressed in the size of its propensity. The algorithm applies a different simulation method to each partition: The slow reactions y are simulated by the exact SSA, whereas the fast reactions z are simulated either completely deterministically via ODEs or via stochastic differential equations. As the system of interest might not be in a state were one can distinguish between slow and fast reactions, the SSA simulation is used until slow and fast subsets emerge. Here, one switches to the hybrid simulation approach.

Depending on both the propensities of the slow and fast subset, τ is selected, which is the time step taken. The fast reactions are now simulated by integrating the corresponding ordinary or stochastic differential equations over τ. This gives some intermediate state of the system, where the fast reactions have already occurred but no slow reaction. Based on this intermediate state the propensities of the slow subset (which have changed due to the change in the fast reaction) are recalculated and one slow reaction to be executed is chosen according to the standard SSA procedure from equation (2.11).

Using this scheme the algorithm does not need to simulate every reaction in the fast subset on its own and will run much more efficiently. However, the main issue of all hybrid algorithms is the partition of reactions into a fast and a slow subset. Often previous knowledge on the system's dynamics must be available to determine fast and slow reactions. Moreover, the algorithm of Haseltine and Rawlings [19] relies on a static assignment of these sets. However, in systems like the toggle switch such a static assignment is not possible since the speed of the reactions heavily depends on the systems current state.

2.3 Modeling assumptions

In the following we give an overview of the assumptions and parameter choices used to model the toggle switch in the next chapters.

2.3.1 Level of detail

Throughout this work we use a simple model of gene expression. Basic transcription and translation reactions create mRNA or proteins. The decay of mRNA and protein are also modeled as a single reactions. Furthermore proteins can interact with DNA through binding and dissociation reactions.

Of course, reducing these complex processes to single reactions is questionable. Each individual process involves several distinct steps. For example, let us consider transcription in eucaryotes: Assuming the chromatin structure is in a state were the gene of interest is accessible, the polymerase has to bind the gene's promoter. This is only possible if several supporting proteins, so called general transcription factors have to assemble at the gene's promoter: TFIID (transcription factor for polymerase II D) binds the promoter's TATA-box, causing structural chances in the DNA that are recognized by several other general transcription factors such as TFIIA and TFIIB. The polymerase now can bind this assembly, completing the so called transcription initiation complex. With help of the general transcription factors the DNA is partially unwound in the promoter region so that the polymerase can start synthesizing the a few basepairs. Afterwards, the polymerase undergoes a conformational change, being released from the initiation complex, recruiting new cofactors and starting the mRNA elongation phase of transcription. Even at this state the polymerase does not proceed along the DNA smoothly but often stops and reaccelerates. Additionally the polymerase has to cope with superhelical tension in the DNA introduced by the unwinding. The polymerase stops when it reaches a stop codon, releases the mRNA and dissociates. However even at this stage the mRNA is not completed. mRNA postprocessing occurs, involving capping and splicing. As a last step, the mRNA has to be exported from the nucleus into the cytoplasm.

We see that this process is highly complex and condensing it into a single transcription reaction will certainly not account for that. For the translation process and mRNA/protein degradation the situation is similar. As we want to find general principles of the toggle switch mechanism, we do not aim for a model that reflects molecular biology in such detail in this work, but reduce it to the essential mechanism of gene expression. However we would like to note that there are several publications that aim to find a balance between total model reductionism and details that indeed are important for general behavior of the model: For work on the gene expression process, see e.g. the work of Roussel and Zhu [50, 51] and Swain et al. [57].

2.3.2 Parameters

First we derive upper boundaries on the transcription and translation rates. Transcription of DNA into mRNA is accomplished by the RNA-polymerase. One polymerase can process about 10-20 nucleotides per second in eucaryotes [1, 9, 56]. However, Alon [2] finds 80 nucleotids per seconds in *E. coli* and a mean transcription time of one minute per mRNA in procaryotes and 30 minutes per mRNA in mammals. As described by Alberts et al. [1] the newly elongated

Reaction	Symbol	Parameter value
Synthesis (one-stage model)	α	$0.5\ s^{-1}$
Transcription	α	$0.05 s^{-1}$
Translation	β	$0.05\ s^{-1}\text{mRNA}^{-1}$
mRNA degradation	γ	$0.005\ s^{-1}\text{mRNA}^{-1}$
Protein degradation	δ	$5 \cdot 10^{-3}$ to $5 \cdot 10^{-6}\ s^{-1}\text{Protein}^{-1}$, adjusted to the desired protein level
repressor-DNA binding	τ^{+}	$1\ s^{-1}\text{Protein}^{-1}$
activator-DNA binding	π^{+}	$1\ s^{-1}\text{Protein}^{-1}$
repressor-DNA dissociation	τ^{-}	$0.1\ s^{-1}$
activator-DNA dissociation	π^{-}	$0.1\ s^{-1}$

Table 2.1: Parameter of the toggle switch used throughout this work. Protein degradation is chosen according to the desired protein level μ. If not mentioned otherwise, all simulations and plots are based on this set of parameters.

RNA fragment is immediately released from the DNA, which enables other polymerases to follow up even before the first RNA has been completed. The distance between polymerases is estimated to be around 100 nucleotides [25]. The rate of transcription is independent of the sequence length n, since the longer the gene, the more polymerases can process it in parallel. Altogether we find a maximal transcription rate of

$$\begin{aligned}\alpha &= \frac{\text{Speed}}{\text{Sequence length}} \cdot \#\text{ of transcribing polymerases}\\ &= \frac{10nt\ s^{-1}}{n} \cdot \frac{n}{100nt}\\ &= 0.1s^{-1}\end{aligned}$$

in case that enough polymerases and nucleotides are present.

The maximal translation rate can be inferred in a similar way: Ribosomes, large complexes of proteins and rRNAs that translate mRNA into polypeptides, proceed with a speed of 2 codons (=6 nucleotides =2 amino acids) per second in eucaryotes [1]. One mRNA can be processed by many ribosomes (polyribosomes) at the same time [1]. The average space between two ribosomes is 80 nucleotides or 27 amino acids [1]. Therefore the overall translation rate for one mRNA of length n is

$$\begin{aligned}\beta &= \frac{\text{Speed}}{\text{Sequence length}} \cdot \#\text{ of transcribing polymerases}\\ &= \frac{2AA\ s^{-1}}{n} \cdot \frac{n}{27AA}\\ &= 0.074s^{-1}\ ,\end{aligned}$$

again independent of the length n of the mRNA in terms of codons. This corresponds to the maximally possible translation rate. The actual rate will be smaller when not enough ribosomes or other involved molecules (tRNA, amino acids) are present. This result is in good agreement with literature, where the time needed for one translation is said to be between 20 seconds and several minutes [1] (we estimate the minimal translation time as $\frac{1}{0.074s^{-1}} = 13.5s$).

However we have to stress here that these numbers found in literature show large variation. Often measurements of these rates are only available in procaryotes, which seem to have different (faster) kinetics than eucaryotes [2], and therefore cannot be transfered to eucaryotes. Overall these estimates of transcription and translation rates should be used to get a general idea of the timescale of these processes.

We use a transcription and translation rate of $\alpha = \beta = 0.05\ s^{-1}$, corresponding to an average time of 20 seconds per product, which seems to be reasonable in the context of the above considerations. The fact that both rates are equal is not expected to have influences on the results.

Interactions between proteins and DNA are mediated by specific regions of the proteins, called DNA-binding domains, which on the one hand can recognize specific DNA sequences and on the other hand maintain the interaction between DNA and protein. Zinc Fingers, Leucine Zippers or Helix-Turn-Helix motifs are prominent examples of DNA binding domains [1]. The binding between DNA and protein is maintained by hydrogen bonds, ionic bonds, and hydrophobic interactions. Single interactions are weak, but as many bonds are formed the binding between DNA and protein becomes stronger. The binding rates are very fast compared to transcription and translation processes and according to Alon [2] in the range of $1s^{-1}$ in *E. coli*. The unbinding rate depends on the strength of the interaction and is assumed to be 10 times smaller ($0.1s^{-1}$) in our model, leading to strong binding of the protein to the DNA. All reaction rates of the models used in this work are summarized in table 2.1.

As we showed above, the transcription and translation rate have upper bounds. The only way for a cellular system to further increase the abundance of proteins is to modulate the degradation rates of mRNA or proteins, giving longer lifetime to mRNA and proteins. Thus, during this work we manipulate the degradations rates to adjust the system's protein level to a desired steady state.

Throughout chapter 4, we assume that in the system a mean number of 10 mRNAs is present. This surprisingly low copy number is motivated by findings from Warren et al. [62], who showed that the mRNA number of PU.1 is in the range of 10-20. The transcription factor PU.1 together with GATA-1 forms a toggle switch motif, that is thought to determine cell fate decision in myeloid differentiation. Therefore we adopt the low mRNA copy number in our model.

In sections 3.2 and 4.2 we assume that the main source of asymmetry in a biological toggle switch originates from the individual protein DNA interaction strength of the two players. Both promoter sequence and 3D-structure of the transcription factors can be very different and may result in different interaction parameters. Also the transcription and translation rates can be non-symmetric. Although the actual transcription rate is independent of the sequence, the rate limiting step is the binding of the polymerase to the DNA, which depends on the promoter sequence. Core parts of promoter sequence have been conserved during evolution (e.g. the TATA-box), but the remaining promoter sequence has large variability from gene to gene. Similar arguments can be made for the translation, that mainly depends on the rate of ribosomes binding the ribosomal binding site of the mRNA. However we choose to focus on asymmetries originating in the protein-DNA interactions for the sake of simplicity.

In section 5.1.5 we do not allow simultaneous binding of inhibitor and activator, because binding sites are assumed to be in close distance, leading to sterical hindrance. This assumption is not valid in general, since binding sites for transcription factors can be several kilobases upstream of the core promoter of the gene. This exclusive binding of either activator or inhibitor has a big influence on the switching time, since the activator can protect its own promoter from inhibition in this model. Independent binding of activator and repressor will remove this effect and will be investigate in later work, however the transcriptional potency of this double bound gene is unclear.

Chapter 3

One-stage toggle switch

In this chapter we investigate the dynamics of a one-stage toggle switch where transcription and translation is condensed into a single process.

3.1 System description

At first we describe the model qualitatively and quantitatively using the frameworks provided in the last chapter.

3.1.1 Gene expression and regulatory interactions

The central dogma in molecular biology claims that genetic information is transcribed from genes into mRNA by RNA-polymerase in the nucleus. The mRNA gets exported into the cytoplasm and is translated into proteins by ribosomes. The transcription and translation itself are highly complex processes including binding of various proteins to the nucleic acids, recruitment of cofactors and different stages of the actual synthesis process.

In this chapter we condense all processes involved in transcription and translation into a single first order biochemical reaction, that produces proteins from DNA with a rate α:

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

As this reaction is a mixture of the original transcription and translation process we will call it the synthesis reaction.

Protein degradation is also modeled as a single reaction:

$$\text{Protein} \xrightarrow{\delta} \emptyset$$

As we effectively removed the mRNA stage of gene expression, we call this model a one-stage model of gene expression, following the nomenclature of Shahrezaei and Swain [55].

To establish a toggle switch, we need to incorporate the mutual inhibitory interactions between proteins and genes into the model. Transcriptional regulation is mediated through proteins known as transcription factors that can bind the promoters of a gene, thereby influencing its transcription. This binding is reversible and is maintained through highly conserved parts of the protein known as DNA binding domains (e.g. the zinc finger domain or the GATA domain) that recognize and bind to specific DNA sequences. An inhibitory transcription factor binds and occupies the promoter of its target gene, thereby blocking important parts of the transcriptional machinery from accessing essential DNA sequences. This shuts down transcription completely because no transcription initiation complex can be formed.

As there is no elementary transcription process in our one-stage model, all regulatory interactions will instead have influence on the synthesis reaction. One way of including regulatory interactions is to express the synthesis reaction rate not as a constant but as a function of protein numbers. The Hill function is a widely used example [64, 65], leading to a gradual response.

The more accurate way of modeling this protein-DNA interaction is to introduce binding and unbinding reactions and a new species representing protein-bound DNA:

$$\begin{aligned} \text{Protein} + \text{DNA} &\xrightarrow{\tau^+} \text{DNA}^{\text{bound}} \\ \text{DNA}^{\text{bound}} &\xrightarrow{\tau^-} \text{Protein} + \text{DNA}\ . \end{aligned} \tag{3.1}$$

This bound DNA has a synthesis rate of 0 if the transcription factor works as an inhibitor. To model the mutual inhibition of two players, from now on called A and B we simply define a reaction where Protein_A binds DNA_B, leading to arrest of synthesis, and vice versa (see Fig. 3.1 for a graphical representation of the model). This approach of additional binding/unbinding reactions is advantageous because all reactions of our system are still based on the law of mass action, whereas reactions with rates as functions of species are not. Note that we do not allow degradation of DNA-bound proteins in our model. We assume that the protease – a large protein complex that degrades proteins – cannot access proteins in close distance to the DNA due to sterical hindrance. Using one-stage gene expression and mutual inhibition through protein-DNA interaction we can express the model as the set of biochemical reactions listed in table 3.1.

3.1.2 Mathematical representation

Deterministic model

Under the assumption of large species abundances, we can apply the deterministic framework – introduced in section 2.1.2 – to the model, resulting in the following set of ODEs (for simplicity we do not explicitly write the time dependence of the variables):

$$\begin{aligned} \frac{d}{dt}d_\text{A} &= \tau_\text{B}^-(1-d_\text{A}) - \tau_\text{B}^+ \cdot d_\text{A} \cdot p_\text{B} \\ \frac{d}{dt}d_\text{B} &= \tau_\text{A}^-(1-d_\text{B}) - \tau_\text{A}^+ \cdot d_\text{B} \cdot p_\text{A} \\ \frac{d}{dt}p_\text{A} &= \alpha_\text{A} \cdot d_\text{A} - \delta_\text{A} \cdot p_\text{A} + \tau_\text{A}^-(1-d_\text{B}) - \tau_\text{A}^+ \cdot d_\text{B} \cdot p_\text{A} \\ \frac{d}{dt}p_\text{B} &= \alpha_\text{B} \cdot d_\text{B} - \delta_\text{B} \cdot p_\text{B} + \tau_\text{B}^-(1-d_\text{A}) - \tau_\text{B}^+ \cdot d_\text{A} \cdot p_\text{B}\ , \end{aligned} \tag{3.2}$$

where p_A and p_B denote the abundances of Protein_A and Protein_B, respectively. The abundance of DNA_A and DNA_B is denoted by d_A and d_B.

Here we do not include the bound DNA state explicitly but express it in terms of the unbound state due to mass conservation: $(d_\text{A} + {d_\text{A}}^{\text{bound}} = 1)$. This reduces the size of the ODE system by two. We can solve for the steady state values of the players by setting

$$\frac{d}{dt}d_\text{A} = \frac{d}{dt}d_\text{B} = \frac{d}{dt}p_\text{A} = \frac{d}{dt}p_\text{B} = 0\ . \tag{3.3}$$

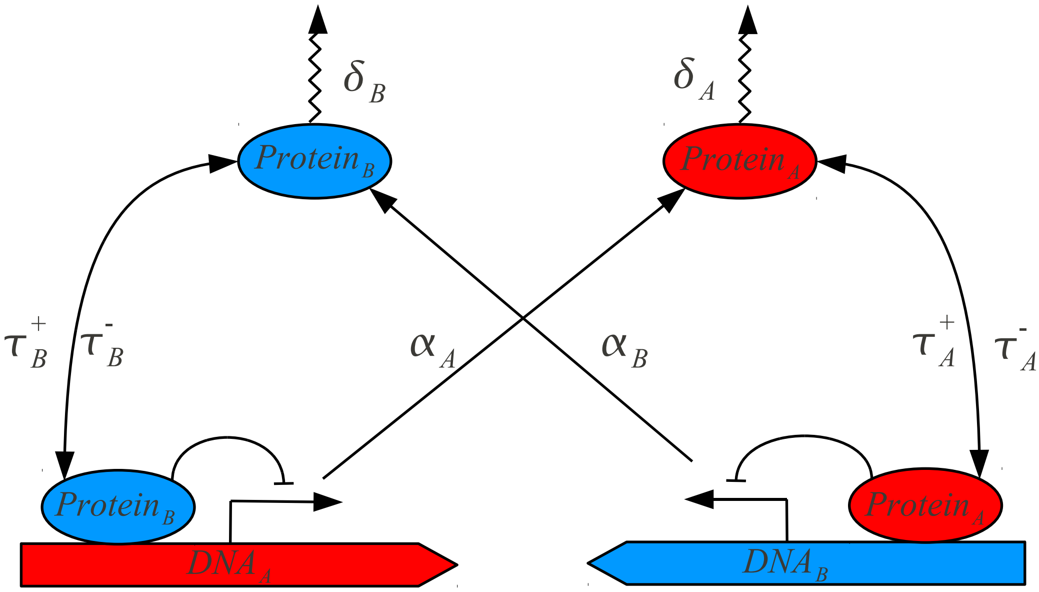

Figure 3.1: Scheme of the one-stage toggle switch where the complex processes of transcription and translation are condensed to a single synthesis reaction. Species associated with the player A are shown in red, species associated with the player B are shown in blue. Reactions/Interactions are indicated as arrows, jagged arrows indicate degradation reactions. Protein$_\mathrm{A}$ is synthesized from DNA$_\mathrm{A}$ with rate α_A. It decays with rate δ_A. Moreover it binds (unbinds) the promoter of DNA$_\mathrm{B}$ with with rate τ_A^+ (τ_A^-). Protein bound promoters lead to synthesis arrest.

$$
\begin{aligned}
\mathrm{DNA_B} &\xrightarrow{\alpha_\mathrm{B}} \mathrm{DNA_B} + \mathrm{Protein_B} & \qquad \mathrm{DNA_A} &\xrightarrow{\alpha_\mathrm{A}} \mathrm{DNA_A} + \mathrm{Protein_A} \\
\mathrm{Protein_B} &\xrightarrow{\delta_\mathrm{B}} \emptyset & \qquad \mathrm{Protein_A} &\xrightarrow{\delta_\mathrm{A}} \emptyset \\
\mathrm{Protein_B} + \mathrm{DNA_A} &\xrightarrow{\tau_\mathrm{B}^+} \mathrm{DNA_A^{bound}} & \qquad \mathrm{Protein_A} + \mathrm{DNA_B} &\xrightarrow{\tau_\mathrm{A}^+} \mathrm{DNA_B^{bound}} \\
\mathrm{DNA_A^{bound}} &\xrightarrow{\tau_\mathrm{B}^-} \mathrm{Protein_B} + \mathrm{DNA_A} & \qquad \mathrm{DNA_B^{bound}} &\xrightarrow{\tau_\mathrm{A}^-} \mathrm{Protein_A} + \mathrm{DNA_B}
\end{aligned}
$$

Table 3.1: List of reactions for the one-stage toggle switch. The model consists of protein synthesis, protein degradation and protein-DNA binding/dissociation. Degraded proteins are absorbed by the protein sink $\emptyset$.

There exist two non-trivial solutions to equations (3.2) and (3.3), resulting from the solution to a second order polynomial. One solution involves negative protein numbers, thus being biologically meaningless.

As the solution terms are quite complicated (see Appendix A.1), we focus on the solution obtained by an numerical ODE-solver. Fig. 3.2 shows the timecourse of the ODE using two different initial conditions and two sets of parameters. Fig. 3.3 shows the timecourses of many different initial conditions projected onto the p/g phase plane. For symmetric parameters p_{A} and p_{B} evolve exactly identically (see Fig. 3.2 A, B and 3.3 A), asymmetric parameters lead to preference of one species (Fig. 3.2 C, D and 3.3 B).

Bistability is not possible for this system: Given non-negative initial conditions the solution of the ODE will also always be non-negative (see section 2.1.2). Since only one positive steady state exists the solution can only converge towards this single steady state. This can also be seen in the phase portraits in Fig. 3.3.

Probabilistic model

Using the probabilistic approach from section 2.1.4, we can fully describe the system through the CME. Instead of writing down one large master equation we split the CME up into four coupled equations, each corresponding to one DNA configuration. This is possible since the numbers of DNA molecules are discrete in the probabilistic framework ($\mathrm{DNA}_{P/G} \in \{0,1\}$).

$$P_{ij}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) := P(\mathrm{Protein}_{\mathrm{A}} = p_{\mathrm{B}}, \mathrm{Protein}_{\mathrm{B}} = p_{\mathrm{B}}, \mathrm{DNA}_{\mathrm{A}}^{\mathrm{bound}} = i, \mathrm{DNA}_{\mathrm{B}}^{\mathrm{bound}} = j, t)$$

is the probability of finding the system at time t in a state with p_{A} and p_{B} copies of $\mathrm{Protein}_{\mathrm{A}}$ and $\mathrm{Protein}_{\mathrm{B}}$ and the corresponding promoter state. For example if $i = 0, j = 0$ both promoters are unbound, allowing synthesis of both proteins. Correspondingly, if $i = 1, j = 0$ $\mathrm{DNA}_{\mathrm{A}}$ is bound and synthesis of $\mathrm{Protein}_{\mathrm{A}}$ is inhibited. Synthesis of $\mathrm{Protein}_{\mathrm{B}}$ is possible since $\mathrm{DNA}_{\mathrm{B}}$ is unbound. Probability mass is exchanged between the four equations only due to binding and unbinding reactions, that is, changes in the promoter state.

$$\begin{aligned}
\frac{d}{dt}P_{00}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) &= \tau_{\mathrm{A}}^{-} P_{01}(p_{\mathrm{A}} - 1, p_{\mathrm{B}}, t) + \tau_{\mathrm{B}}^{-} P_{10}(p_{\mathrm{A}}, p_{\mathrm{B}} - 1, t) \\
&\quad + [-\tau_{\mathrm{B}}^{+} p_{\mathrm{B}} - \tau_{\mathrm{A}}^{+} p_{\mathrm{A}} + \alpha_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{-} - 1) + \alpha_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{-} - 1) \\
&\qquad + \delta_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{+} - 1) \cdot p_{\mathrm{A}} + \delta_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{+} - 1) \cdot p_{\mathrm{B}}] \\
&\quad P_{00}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) \\
\frac{d}{dt}P_{11}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) &= \tau_{\mathrm{A}}^{+}(p_{\mathrm{A}} + 1) P_{10}(p_{\mathrm{A}} + 1, p_{\mathrm{B}}, t) + \tau_{\mathrm{B}}^{+}(p_{\mathrm{B}} + 1) P_{01}(p_{\mathrm{A}}, p_{\mathrm{B}} + 1, t) \\
&\quad + \big[-\tau_{\mathrm{A}}^{-} - \tau_{\mathrm{B}}^{-} + \delta_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{+} - 1) \cdot p_{\mathrm{A}} + \delta_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{+} - 1) \cdot p_{\mathrm{B}}\big] \\
&\quad P_{11}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) \\
\frac{d}{dt}P_{10}(p_{\mathrm{A}}, p_{\mathrm{B}}, t) &= \tau_{\mathrm{A}}^{-} P_{11}(p_{\mathrm{A}} - 1, p_{\mathrm{B}}, t) + \tau_{\mathrm{B}}^{+}(p_{\mathrm{B}} + 1) P_{00}(p_{\mathrm{A}}, p_{\mathrm{B}} + 1, t) \\
&\quad + \big[-\tau_{\mathrm{B}}^{-} - \tau_{\mathrm{A}}^{+} p_{\mathrm{A}} + \alpha_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{-} - 1) + \delta_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{+} - 1) \cdot p_{\mathrm{A}} + \delta_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{+} - 1) \cdot p_{\mathrm{B}}\big] \\
&\quad P_{10}(p_{\mathrm{A}}, p_{\mathrm{B}}, t)
\end{aligned} \tag{3.4}$$

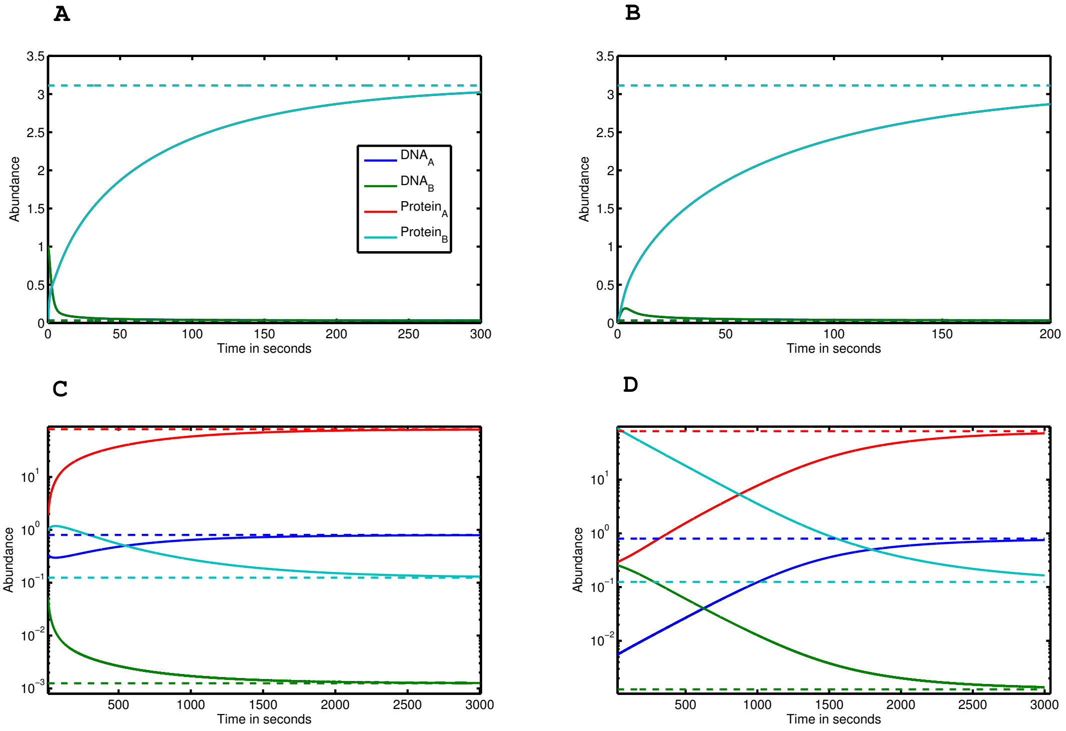

Figure 3.2: Numerical solution of the reaction rate equations in the one-stage model for different parameters and initial conditions. Dashed lines indicate steady state values. A,B: Symmetric parameters for both A and B (see Tab. 2.1), leading to identical time evolution of the corresponding species. Initial conditions were set to $p_{\mathrm{A}} = p_{\mathrm{B}} = 0$ and $d_{\mathrm{A}} = d_{\mathrm{B}} = 1$ (A) and $d_{\mathrm{A}} = d_{\mathrm{B}} = 0$ (B). C,D: Asymmetrical binding constants $\tau_{\mathrm{A}}^{+} = 1, \tau_{\mathrm{B}}^{+} = 0.5$, leading to A-dominance. Initial conditions are set to $p_{\mathrm{A}} = p_{\mathrm{B}} = 0,\ d_{\mathrm{A}} = d_{\mathrm{B}} = 1$ (C) and $p_{\mathrm{A}} = 0, p_{\mathrm{B}} = 100,\ d_{\mathrm{A}} = 0,\ d_{\mathrm{B}} = 1$ (D).

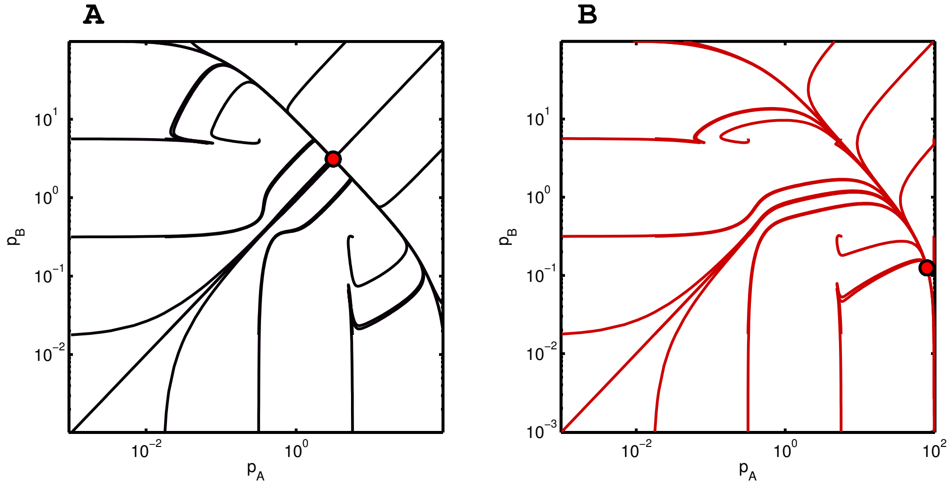

Figure 3.3: Phase portrait of the one-stage model. This plot shows the ODE-system's trajectories in the $p_{\mathrm{A}}/p_{\mathrm{B}}$-phaseplane starting from different initial conditions. The steady state solution is indicated by a red dot. A: completely symmetric parameters for both players lead to a steady state lying on the symmetry axis $p = g$. B: Asymmetric binding constants $\tau_{\mathrm{A}}^{+} = 1, \tau_{\mathrm{B}}^{+} = 0.5$, leading to a steady state shifted below the symmetry axis. Other parameters were set according to Tab. 2.1. Note that trajectories do intersect in this plot, since it shows a projection of a four-dimensional solution onto a two-dimensional phaseplane.

$$\begin{aligned}\frac{d}{dt}P_{01}(p_{\mathrm{A}},p_{\mathrm{B}},t) = {} & \tau_{\mathrm{A}}^{+}(p_{\mathrm{A}}+1)P_{00}(p_{\mathrm{A}}+1,p_{\mathrm{B}},t)+\tau_{\mathrm{B}}^{-}P_{11}(p_{\mathrm{A}},p_{\mathrm{B}}-1,t)\\ & +\left[-\tau_{\mathrm{B}}^{+}p_{\mathrm{B}}-\tau_{\mathrm{A}}^{-}+\alpha_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{-}-1)+\delta_{\mathrm{A}}(E_{p_{\mathrm{A}}}^{+}-1)\cdot p_{\mathrm{A}}+\delta_{\mathrm{B}}(E_{p_{\mathrm{B}}}^{+}-1)\cdot p_{\mathrm{B}}\right]\\ & P_{00}(p_{\mathrm{A}},p_{\mathrm{B}},t)\end{aligned}$$

The shift operators E_x^+ and E_x^- increase or decrease the function argument x by one, i.e.

$$\begin{aligned}E_x^+ f(x,y) &= f(x+1,y)\\ E_x^- f(x,y) &= f(x-1,y)\ .\end{aligned}$$

The first equation in (3.4) describes how the probability of the state with p_{A} molecules of Protein$_{\mathrm{A}}$ and p_{B} molecules of Protein$_{\mathrm{B}}$ and both promoters unbound changes over time: The first term describes how the system, that has a bound DNA$_{\mathrm{B}}$ and an unbound DNA$_{\mathrm{A}}$ (P_{01}) and one less Protein$_{\mathrm{A}}$ molecule previously, enters the current state, where both promoters are unbound. This happens by Protein$_{\mathrm{A}}$ dissociating from the promoter of DNA$_{\mathrm{B}}$ (with rate τ_{A}^{-}), changing the promoter state to 00 and increasing the number of Protein$_{\mathrm{A}}$ molecules by 1 from $p_{\mathrm{A}}-1$ to p_{A}. The second term is analogously the unbinding of a Protein$_{\mathrm{B}}$ from a former bound promoter of DNA$_{\mathrm{A}}$. The first two terms in the bracket account for the loss of probability mass in the current state due to transitions into the two other possible promoter configurations. The third term in brackets accounts for Protein$_{\mathrm{A}}$ production: The system can enter the current state $P_{00}(p_{\mathrm{A}},p_{\mathrm{B}},t)$ from the state with one less Protein$_{\mathrm{A}}$ molecule ($P_{00}(p_{\mathrm{A}}-1,p_{\mathrm{B}},t)$) – the shift operator expresses this missing Protein$_{\mathrm{A}}$ molecule –, or the system can leave the current state (accounted for by the constant factor -1) by synthesizing an additional Protein$_{\mathrm{A}}$, moving it to the state $P_{00}(p_{\mathrm{A}}+1,p_{\mathrm{B}},t)$. The synthesis of Protein$_{\mathrm{B}}$ is similar. The next term expresses the influence of the degradation of Protein$_{\mathrm{A}}$ on the state probability. The system can move into the current state $P_{00}(p_{\mathrm{A}},p_{\mathrm{B}},t)$ from the state that has one additional Protein$_{\mathrm{A}}$ ($P_{00}(p_{\mathrm{A}}+1,p_{\mathrm{B}},t)$) by degradation of Protein$_{\mathrm{A}}$. Therefore, $P_{00}(p_{\mathrm{A}},p_{\mathrm{B}},t)$ increases. Note that this propensity is now linear in the protein number present. The system also can leave the current state by degrading one Protein$_{\mathrm{A}}$ and therefore the current state loses probability. In a similar way the three other equations can be derived.

3.1.3 Regimes

First we want to qualitatively describe the system's dynamics obtained from simulated data using Gillespie's algorithm. The main feature of the mutual inhibition regulatory motif is its bistability, meaning that the system can – given adequate parameters and initial conditions – adopt two different regimes. As both proteins are fighting each other eventually one will win and dominate the other by inhibiting its expression and backing up its own dominance. Therefore the system will be either in a state where Protein$_{\mathrm{A}}$ dominates Protein$_{\mathrm{B}}$ or vice versa.

Looking at simulations, we find the following: For an initial state where proteins are present ($p_{\mathrm{A}}=0$, $p_{\mathrm{B}}=0$), after some time protein synthesis of, lets say Protein$_{\mathrm{A}}$, will occur. Two different things can happen now:

1. The newly synthesized Protein$_{\mathrm{A}}$ binds the antagonistic promoter quickly blocking the synthesis of Protein$_{\mathrm{B}}$ and keeping its expression level at 0. This will be followed by further synthesis of Protein$_{\mathrm{A}}$, backing up its dominance. This will lead to the two regimes mentioned above (as both proteins can dominate depending which one got synthesized first).

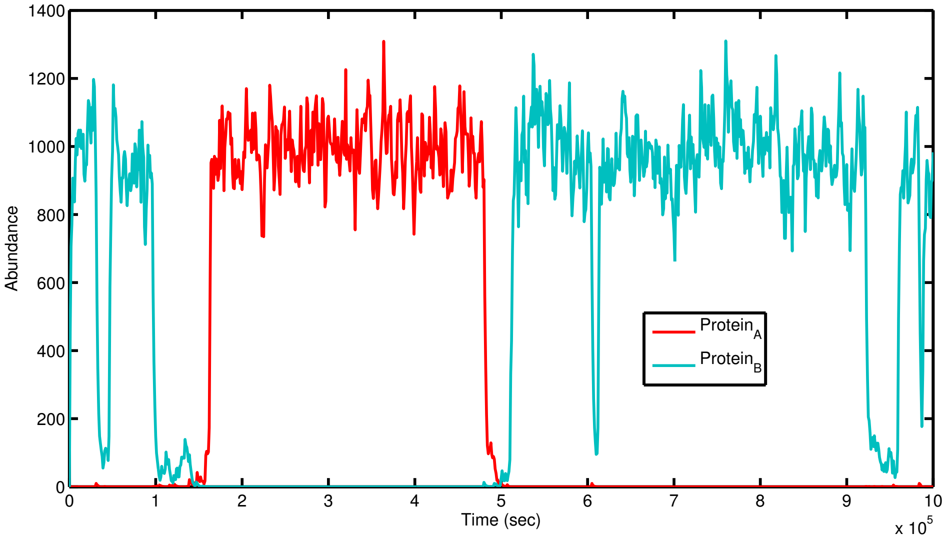

Figure 3.4: An exemplary timecourse of the one-stage toggle switch obtained by stochastic simulation. The number of proteins is plotted against time. The bistability is clearly visible: Either Protein_A is upregulated and Protein_B is repressed or vive versa.

2. In the second scenario an antagonistic Protein_B will be synthesized before Protein_A can bind and inhibit DNA_B. This results in a state where both Protein_A and Protein_B are present. As both proteins eventually bind to the other promoter the system will end up in a state were both promoters are occupied, shutting down the synthesis of the whole system. We will refer to this regime as a deadlock.

However none of these scenarios is terminal. With a certain probability the system leaves the current regime and will be driven to another one. In order to understand how the system can leave the regime where one protein dominates, we have to take a closer look at the promoter state. Consider the regime where player A dominates. Although Protein_A is abundant and the promoter will be in a bound state most of the time there is a constant probability > 0 that an unbinding reaction will occur. This unbinding reaction will result in a short time window were Protein_B can be synthesized. If this happens the system will be driven out of the dominating regime into the deadlock regime as proteins of both species are present now and both promoters get bound. To leave the deadlock regime, one protein species must be fully degraded to enable synthesis of the other protein again.

In order to give a quantitative description of the systems behavior we need to define the features mentioned above. We start by defining the three regimes (Protein_A dominating, Protein_B dominating, deadlock) as subsets of the complete state space S. As the whole system is stochastic there are no sharp boundaries between regimes. However, we find that, while the system is for example in the regime, where A dominates, the synthesis of Protein_B is shut down whereas the synthesis of Protein_A is similar to an unregulated gene. Therefore, we can approximate the distribution within these two regimes by using results from Thattai and van Oudenaarden [58], who showed that for unregulated genes the distribution of the protein numbers is Poissonian and the mean and variance of protein x's copy number during steady state obeys:

$$\mu_x = \sigma_x^2 = \frac{\alpha_x}{\delta_x} \ . \tag{3.5}$$

This corresponds to a simple birth/death process.

These statistics are not completely correct for our model. The mean is expected to be smaller as one protein is bound to repress the antagonistic promoter, reducing the mean number of free proteins by one. The variance is expected to be larger since random binding/unbinding event interrupt the synthesis frequently. This indicates that the underlying distribution is not Poissonian after all, since mean and variance should be equal. Nevertheless, we can use these statistics to approximate the protein distribution of the system while it is either in the Protein$_\mathrm{A}$-dominating or Protein$_\mathrm{B}$-dominating state and use them to define the boundaries of the regimes:

$$\begin{aligned} S_\mathrm{A} &= \{s \in S | p_\mathrm{A} > \phi_\mathrm{A} \wedge p_\mathrm{B} < \phi_\mathrm{B}\} \\ S_\mathrm{B} &= \{s \in S | p_\mathrm{A} < \phi_\mathrm{A} \wedge p_\mathrm{B} > \phi_\mathrm{B}\} \\ S_0 &= \{s \in S | p_\mathrm{A} < \phi_\mathrm{A} \wedge p_\mathrm{B} < \phi_\mathrm{B}\} \ . \end{aligned} \tag{3.6}$$

The regimes and boundaries are illustrated in Fig. 3.5. The regime S_A (S_B) is the subspaces of S where Protein$_\mathrm{A}$ (Protein$_\mathrm{B}$) is dominating, regime S_0 is the subspace where the system is in a deadlock situation. The regime boundary ϕ_x – chosen here at the lower boundary of protein x's distribution assuring that $(1 - \alpha')\%$ percent of the distribution lie beyond the lower bound – is defined as

$$\phi_x = F^{-1}_{\mu_x}(\alpha') \ , \tag{3.7}$$

with F^{-1} being the inverse cumulative distribution function of the Poisson-distribution. For example using $\alpha' = 0.001$ assures that the regime contains 99.9% of the probability distribution.

3.1.4 A numerical approximation of the CME

Using the results from above, we can apply the finite state projection method described in section 2.1.5 to solve the CME of the one-stage system (3.4) in case of small protein numbers.

To obtain a suitable subspace of the six-dimensional state space S (the six dimensions correspond to the six molecular species involved), the states which do not obey DNA mass conservation are removed. Furthermore we have to truncate the state space using an upper bound of the protein distribution (3.5).

$$\phi_x^+ = F^{-1}_{\mu_x}(0.999) \ .$$

All states with $p_\mathrm{A} > \phi_\mathrm{A}^+$ or $p_\mathrm{B} > \phi_\mathrm{B}^+$ are removed, making the subspace finite. As the system will only enter states that lie beyond these bounds with probability $1 - 0.999 = 10^{-3}$ – we assured that using the far edges ϕ_x^+ of the protein distributions –, removing those states will introduce an error of $\epsilon = 10^{-3}$, which is a good approximation of the exact solution.

As an example we will use a one-stage system with a mean number of proteins $\mu_\mathrm{A} = \mu_\mathrm{B} = 25$. Parameters are chosen symmetric for both genes : $\alpha = 0.5$, $\delta = 0.02$, $\tau^+ = 1$, $\tau^- = 0.1$).

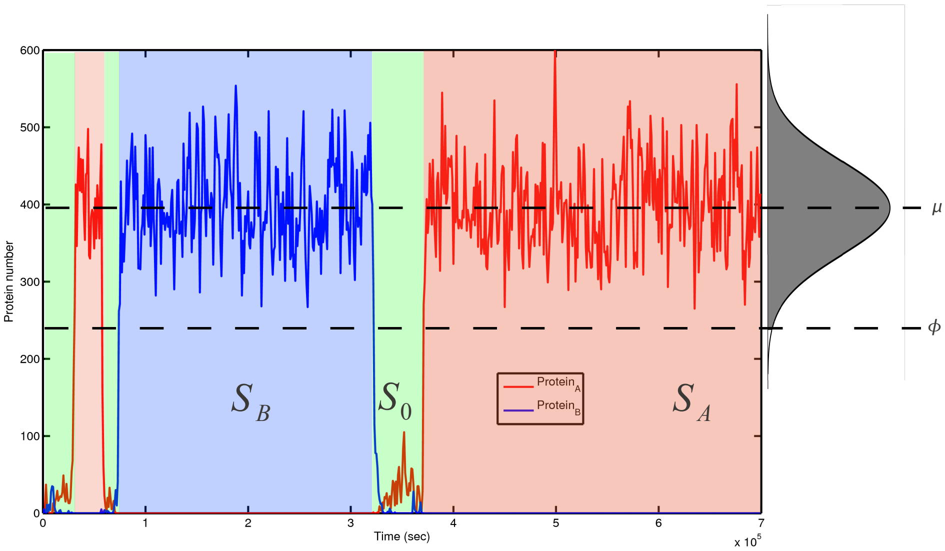

Figure 3.5: Timecourse of the one-stage toggle switch passing through different regimes. S_A is highlighted in red, S_B in blue and S_0 in green. Additionally, the distribution of proteins in S_A (S_B) is shown on the right, illustrating the mean protein level μ and the boundary ϕ between regimes. The gray area under the distribution corresponds to 99.9% probability mass.

Using $\alpha' = 0.001$, we find an upper bound of the protein distribution at $\phi^+ = 42$. Applying these constraints on the state space results in a subspace containing $42^2 \cdot 4 = 7056$ states. In order to solve the CME for this subspace one has to evaluate equation (2.8) for a 7056×7056 matrix Q, which is sparse, but still contains $\approx 7056 \cdot 8$ non-zero entries (8 corresponding to the number of possible reactions in each state). Note that the state space could be further restricted using the fact the toggle switch mechanism should prevent the system from going into regions of the state space where both proteins are highly expressed.

Setting a desired initial condition p_0, we can solve equation (2.8) for times t, resulting in the evolution of probability distribution over the system's state space. Some exemplary snapshots of the evolving probability distribution using different initial conditions are shown in Fig. 3.6. The approximate solution of the CME is accurate ($\epsilon \approx 10^{-3}$), suggesting that the projection was suitable. Further increasing the state space would decrease the error even more. We see that the time dependent solutions for all three initial conditions converge to a similar steady state solution. The shape of this distribution clearly reveals the bistable nature of the system, consisting of the three regime S_A, S_B and S_0. For example in Fig. 3.6 C we see that systems where both protein levels are high run into a deadlock and both proteins get degraded. The probability distribution approaches the origin, corresponding to S_0. From here the distribution partitions into three distinct peaks, representing S_0, S_A and S_B.

3.1.5 Switching time

How long will the toggle switch stay in a specific regime? The regulatory motif of a toggle switch is thought to be a mechanism of cells to make and maintain a decision [12], for example during cell fate decision in hematopoesis [30]: Once the cell is driven into one lineage it should be irreversibly committed to that lineage. Since our system is stochastic the probability of moving out of one regime can be very small but will never equal zero (unless e.g. the

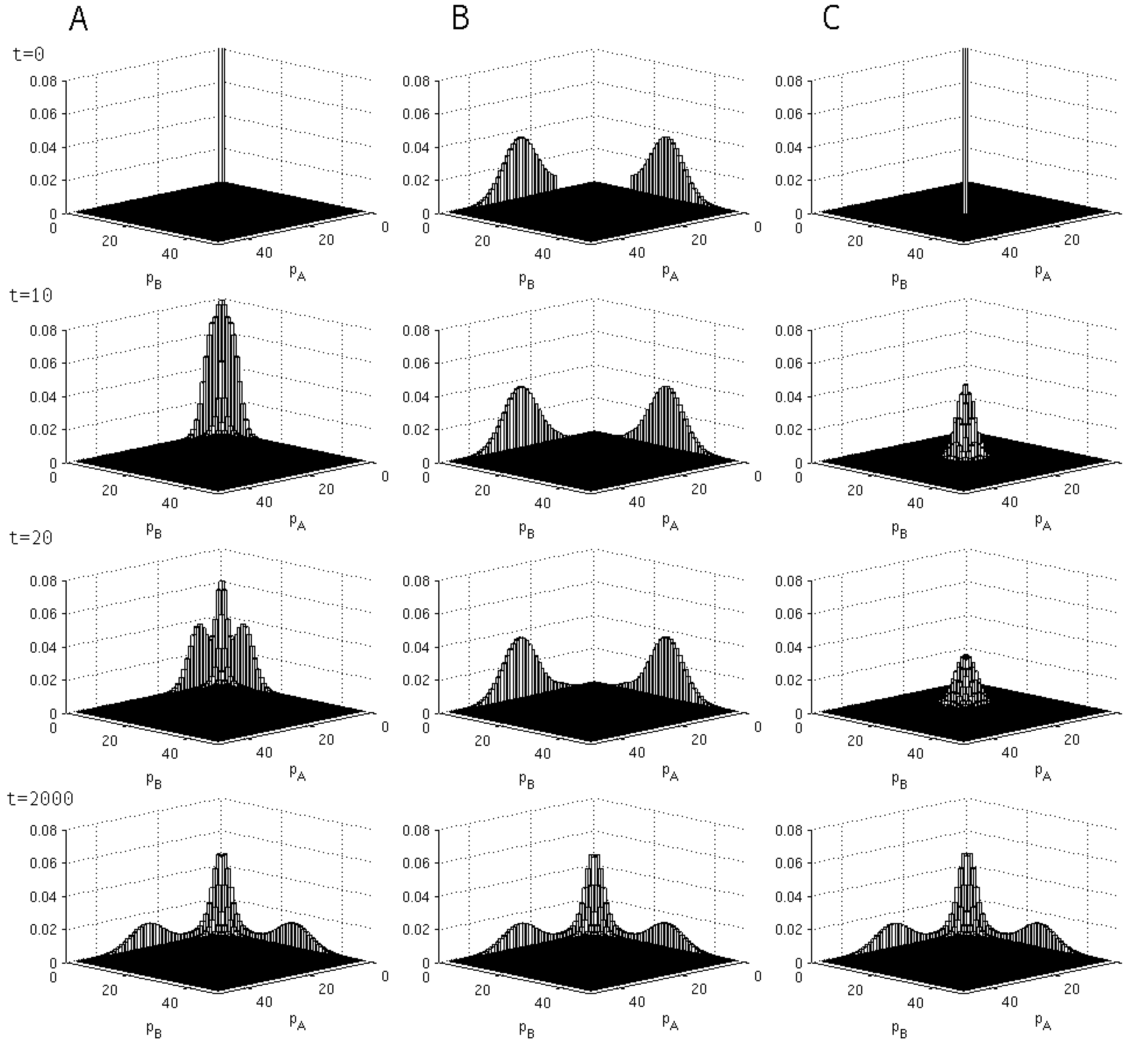

Figure 3.6: Time evolution of the system's probability distribution obtained by finite state projection. As the complete distribution is 6-dimensional, only a projection onto the p_A/p_B plane is shown here. Errors of the approximation are below 10^{-3}. Parameters are symmetrical for both genes: $\alpha = 0.5$, $\delta = 0.02$, $\tau^+ = 1$, $\tau^- = 0.1$, leading to a mean protein number $\mu = 25$. Each column corresponds to the time evolution of the distribution given different initial conditions. A: Initial conditions set to $p_A = p_B = 0$. B: Initial conditions set to truncated normal distributions. C: Initial conditions set to $p_A = p_B = 25$.

unbinding rate $\tau^- = 0$), meaning that it is certain that after a long time the system will eventually change its regime. In the case of S_A and S_B we will call the time it takes to switch the regime *switching time*. For a biological switch this time should be longer than all other relevant processes of the cell – especially longer than cell lifetime or cell cycle time – in order to maintain the decision. In the following an analytical derivation of the switching time is shown and compared to the method introduced in section 2.1.5 and to estimates obtained from stochastic simulation. We are interested in the time it takes to leave the chosen regime. Without loss of generality, let us assume the system is in S_A. So we have the promoter of DNA$_\mathrm{B}$ bound by Protein$_\mathrm{A}$ and the promoter of DNA$_\mathrm{A}$ unbound. This results in a mean Protein$_\mathrm{A}$-level of $\mu_\mathrm{A} = \alpha_\mathrm{A}/\delta_\mathrm{A}$ and $\mu_\mathrm{B} = 0$ (see equation (3.5)). We call this initial state x_0.

In order to leave S_A it is crucial that one Protein$_\mathrm{B}$ is synthesized, which then can bind the promoter of DNA$_\mathrm{A}$, thus shutting down synthesis of Protein$_\mathrm{A}$ and ultimately driving the system into S_0. This trajectory involves the following events:

1. Unbinding of Protein$_\mathrm{A}$ from DNA$_\mathrm{B}$
2. Synthesis of Protein$_\mathrm{B}$ during the unbound phase
3. Binding of Protein$_\mathrm{B}$ to the promoter of DNA$_\mathrm{A}$ before Protein$_\mathrm{B}$ is degraded

First we describe the unbinding of DNA$_\mathrm{B}$: Due to stochasticity even if the system is in S_A, various times Protein$_\mathrm{A}$ dissociates, leaving the the promoter of DNA$_\mathrm{B}$ unbound for a certain time t_u. The average time the promoter stays unbound is equal to the average time until a binding reaction occurs, which is

$$t_u = \frac{1}{\tau_\mathrm{A}^+ \cdot \mu_\mathrm{A}} \; . \tag{3.8}$$

The time time the promoter stays unbound is a random variable itself, but for simplicity we approximate it be its mean value. Note that t_u depends, somewhat counterintuitively, on τ^+ and not on τ^-. The above mentioned synthesis of Protein$_\mathrm{B}$ has to take place during t_u. The probability of k synthesis reactions to happen during t_u is

$$P(K = k) = \frac{(\alpha_\mathrm{B} \cdot t_u)^k}{k!} \cdot \exp(-\alpha_\mathrm{B} \cdot t_u) \; ,$$

as the number of synthesis reactions K during t_u is Poisson-distributed with mean $\alpha_\mathrm{B} \cdot t_u$ (see section 2.2.2). Thus, the probability of one or more synthesis reactions happening during the unbound phase is

$$q = 1 - P(K = 0) = 1 - \exp(-\alpha_\mathrm{B} \cdot t_u) \; .$$

However not one but several unbound phases may occur, each of them giving the chance of successful synthesis. The number N of unbound phases until a successful synthesis of Protein$_\mathrm{B}$ is geometrically distributed with parameter q:

$$N \sim \mathrm{Geom}(q) \tag{3.9}$$

with the probability density function of the geometric distribution as $P_\mathrm{Geom}(N = n) = (1 - q)^{n-1} \cdot q$.

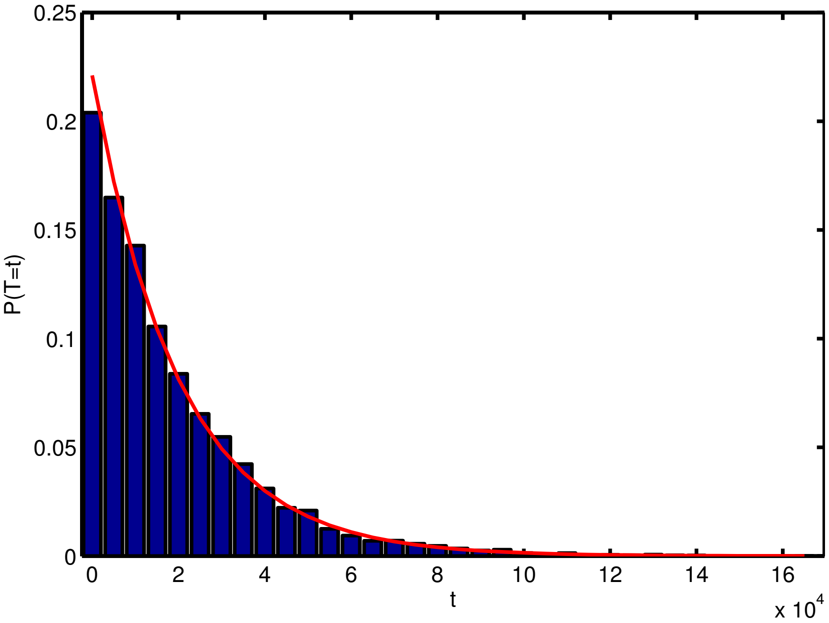

Figure 3.7: Histogram of switching times from simulations (blue) compared to the analytically derived geometric distribution (red).

We can convert this number N of unbound phases until successful synthesis into time: The average number of unbound phases during a time interval Δt is $\tau_{\mathrm{A}}^{-} \cdot \Delta t$. Therefore we can apply a linear transformation to the random variable N:

$$T = \frac{N}{\tau_{\mathrm{A}}^{-}} \ ,$$

relating the number of intervals to actual time by a factor $\frac{1}{\tau_{\mathrm{A}}^{-}}$. T gives not the numbers of intervals, but the actual time until until successful synthesis of Protein$_{\mathrm{B}}$.

The mean of T, which we call *switching time* T_{Switch} and its variance $\sigma^2_{\mathrm{Switch}}$ are then given by:

$$\begin{aligned} T_{\mathrm{Switch}} &= \mathrm{Mean}(T) = \frac{\mathrm{Mean}(N)}{\tau_{\mathrm{A}}^{-}} = \frac{1}{q \cdot \tau_{\mathrm{A}}^{-}} \\ &= \frac{1}{1 - \exp\left(-\frac{\alpha_{\mathrm{B}} \cdot \delta_{\mathrm{A}}}{\tau_{\mathrm{A}}^{+} \cdot \alpha_{\mathrm{A}}}\right)} \cdot \frac{1}{\tau_{\mathrm{A}}^{-}} \\ \sigma^2_{\mathrm{Switch}} &= \mathrm{Var}(T) = \left(\frac{1}{\tau_{\mathrm{A}}^{-}}\right)^2 \cdot \mathrm{Var}(N) = \left(\frac{1}{\tau_{\mathrm{A}}^{-}}\right)^2 \cdot \frac{1-q}{q^2} = \frac{1-q}{\left(\tau_{\mathrm{A}}^{-} \cdot q\right)^2} \ . \end{aligned} \tag{3.10}$$

Note that the mean is almost equal to the standard deviation in case of small q.

We now ask if the synthesis of one protein is enough to drive the system into another regime. Intuitively, only the synthesis of Protein$_{\mathrm{B}}$ is not sufficient to move the system out of S_{A}. Additionally this Protein$_{\mathrm{B}}$ must bind to the antagonistic promoter before it is degraded. Therefore we extend the results from equation (3.10) by taking into account the required binding of Protein$_\mathrm{B}$ to DNA$_\mathrm{A}$. This has to take place before the protein is degraded. The average time until degradation for one Protein$_\mathrm{B}$ is (using the mean of the underlying distribution)

$$t_\mathrm{deg} = \frac{1}{\delta_\mathrm{B}} \; . \tag{3.11}$$

Analogous to the previous approach we find that the probability of having one or more binding reactions during t_{deg} is

$$q_\mathrm{deg} = 1 - \exp(-\tau_\mathrm{B}^{+} \cdot t_\mathrm{deg}) \; . \tag{3.12}$$

The switching time, where we require synthesis of Protein$_\mathrm{B}$ followed by binding for a regime change can be derived in a similar way as above. The only difference is that N is now geometrically distributed with parameter $\lambda = q \cdot q_\mathrm{deg}$ instead of $\lambda = q$, giving

$$\begin{aligned} T_\mathrm{Switch} &= \mathrm{Mean}(T) = \frac{\mathrm{Mean}(N)}{\tau_\mathrm{A}^{-}} = \frac{1}{q \cdot q_\mathrm{deg} \cdot \tau_\mathrm{A}^{-}} \\ \sigma^2_\mathrm{Switch} &= \mathrm{Var}(T) = \left(\frac{1}{\tau_\mathrm{A}^{-}}\right)^2 \cdot \mathrm{Var}(N) = \left(\frac{1}{\tau_\mathrm{A}^{-}}\right)^2 \cdot \frac{1 - q \cdot q_\mathrm{deg}}{(q \cdot q_\mathrm{deg})^2} = \frac{1 - q \cdot q_\mathrm{deg}}{\left(\tau_\mathrm{A}^{-} \cdot q \cdot q_\mathrm{deg}\right)^2} \; . \end{aligned} \tag{3.13}$$

Considering biologically relevant parameter choices, where $\tau_\mathrm{B}^{+} \gg \delta_\mathrm{B}$ (for a discussion on parameters, see section 2.3), gives $q_{deg} \approx 1$. Therefore equations (3.13) reduce to (3.10).

One can use Barzel's method introduced in section 2.1.5 to obtain estimates of the switching time by defining $S_2 = \{x | x \notin S_\mathrm{A}\}$ and solving equation (2.9) for $T(x)$. With x_0 being the initial state in S_A as defined before, $T(x_0)$ should give us a similar switching time as obtained by the analytical solution.

Additionally we estimate the switching time by running 10000 SSA simulations using the Stochkit [31] software package. Initial conditions were set to x_0 and timecourses of 10^6 seconds were simulated. We calculate the simulated mean switching time as

$$T^\mathrm{sim}_\mathrm{Switch} = \frac{1}{10000} \sum_{e=1}^{10000} t_\mathrm{A}^{(1,e)},$$

where $t_R^{(i,e)}$ is defined as the length of the i-th time interval in timecourse e where regime S_R prevails ($R \in \{A, B, 0\}$). We simply look for the time until the simulated system leaves S_A for the first time ($i = 1$) and take the mean over these times. Note that this estimate is expected to be biased since the maximal switching time during simulation is bounded by $10^6 s$, the time of simulation. For systems that exhibit even longer switching times or have a broad distribution of switching time ranging beyond $10^6 s$ this estimate will be to small.

The three estimation methods of switching times are applied to the one-stage model of the toggle switch, using different values for the protein degradation rate to archive different protein levels. Due to the computational expense the Barzel method could only be applied to one-stage system with mean protein level < 1000. Results are shown in Fig. 3.8:

First we find that the switching time increases when the protein degradation rate δ decreases and the protein level thereby increases (see Fig. 3.8 A). Smaller degradation rates lead to prolonged protein life time, thus once a protein has been synthesized, it is more probable to bind the antagonistic promoter if its lifetime is longer (see equation (3.12)). However this has only minor influence on the switching time since even for larger degradation rates the

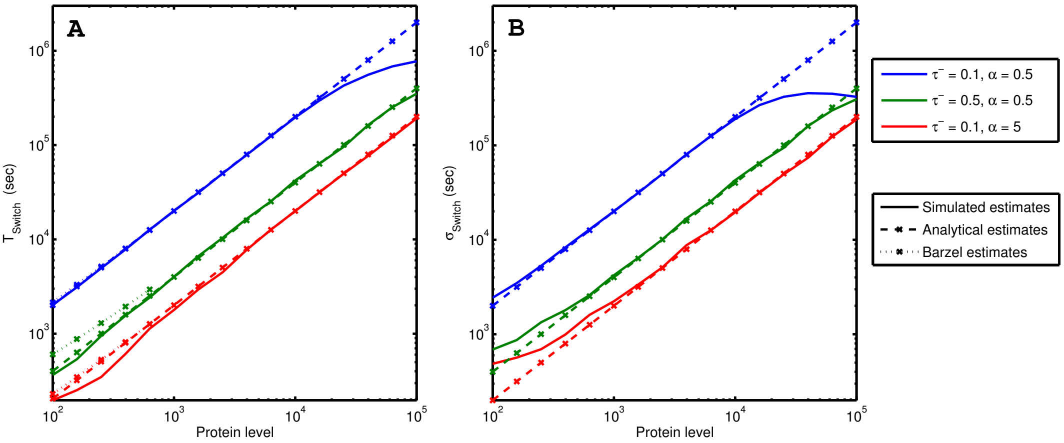

Figure 3.8: Switching times for the one-stage models. Estimates for the switching time over various degradation rates δ resulting in different mean protein levels. Estimates for three different binding/synthesis parameters(blue, red, green) are shown for different estimation methods. Remaining parameters are set according to Table 2.1. Solid lines: SSA estimates. Dashed lines: Analytical estimates. Fine dashed lines: Barzel estimates. A: Estimates of the mean switching time. B: Estimates of the standard deviation of switching time (Barzel estimates are missing, because this method only gives the mean value). All estimates are in good agreement, revealing the the switching time increases as the mean protein level increases.

probability for a protein binding to the promoter is close to 1. Instead, decreasing the protein degradation rate leads to higher amounts of proteins of the dominating species. The more dominating proteins are present, the shorter are the intervals during which the antagonistic promoter is unbound. The number of these interval remains unchanged (depending only on the dissociation rate), but the length of the intervals decreases. Thus the synthesis reaction driving the system out of the current regime is much less probable if many dominating proteins are present.

By comparing the different estimation methods, we find very similar values for all systems considered. For systems with very small protein levels, simulated estimates for the mean are smaller than analytical estimates and Barzel estimates (Fig. 3.8 A). This is due to the configuration of the SSA algorithm, that only returns the systems state in intervals of 1000 seconds. For systems that have switching time < 1000 seconds, the phase during which the system will stay in the committed regime will not be visible in the simulated trajectory. Therefore, the simulated estimates underestimate the real switching time. For medium protein levels, the switching times predicted by all methods agree well. For system where the switching time gets closer to the overall simulated time of 10^6 seconds, simulated estimates are dampened and switching time is underestimated. Small deviations between analytical and simulated data exist, because the time it takes to degrade the former dominating protein was not included into formula (3.10). It only estimates the time until a regime change is certain. One could extend the formula by adding this degradation term, but as seen, the difference is neglectable.

Additionally, we compare the standard deviations of the switching times obtained by the analytical method and the simulation (Fig. 3.8 B). Interestingly, the analytical estimates of the standard deviation are very close to the values obtained by simulation (despite the bias introduced by the simulation as mentioned above), even though we made some mean field approximations during the derivation of the analytical expression (in equations (3.8) (3.11)),

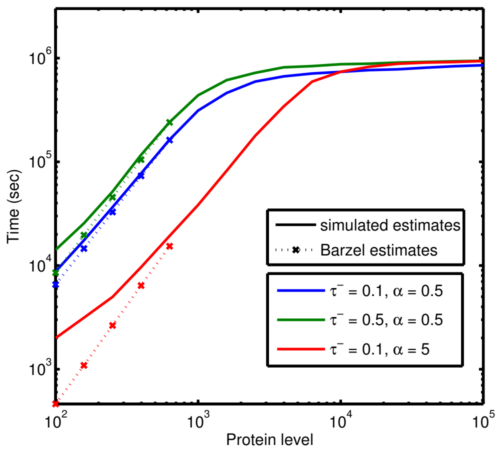

Figure 3.9: Priming times for the one-stage models. Comparison of the simulated mean priming time (solid lines) to estimates obtained by Barzel's approach (dashed lines). Priming times of systems with different mean protein levels μ are plotted for three parameter sets, showing the dependence of the primining time on the protein level. Remaining parameters are set according to Tab. 2.1.

which are expected reduce the variation. This suggests that our mean field assumptions are valid and the main variability comes from the geometric distribution of the number of trials until successfully switching.

We draw the following conclusions from this analysis: The switching time is mainly influenced by the time a promoter is unbound (t_u). This quantity mainly depends on the average protein number μ during S_A or S_B. Further, the switching time depends on the rate of protein synthesis and on the rate of protein dissociation, controlling the number of unbinding phases.

3.1.6 Priming time

In the previous section we were interested in the time until the systems leaves S_A or S_B. In this section we do the same analysis for the time it takes to leave S_0, which will be called priming time.

Deriving an analytical approximation of the priming time is a more challenging task than the derivation of the switching (since it cannot be tracked down to a single event such as protein dissociation) time and will not be given here. Instead we assess the priming time of the model using only simulated and Barzel's estimates.

Analogous to the switching times, we use stochastic simulation and Barzel's approach to estimate the time it takes a one-stage system to leave S_0. We set the initial conditions to both promoters bound by repressors and no free proteins in the system and estimate the time until the system enters either S_A or S_B. Protein degradations rates are different in each system, resulting in different mean protein levels μ.

Results are depicted in Fig. 3.9, showing that also the priming time of the systems depends on the protein level (or degradation rate) of the system. Contrary to the switching time, where this dependence is due to high protein numbers effectively shutting down the antagonistic promoter, for the priming time the degradation rate itself is the major source of dependence. Smaller degradation rates make the degradation of unbound proteins much more improbable, so that they can bind again, keeping the system in S_0.

Comparing simulation data to results of the Barzel approach shows that the general trend of increasing priming times at decreasing degradation rates is also captured by the Barzel estimates. However, for all three sets of parameters there are differences between the estimates especially at very slow protein levels. Simulated data show deviations from the otherwise linear (on loglog scale) relationship between priming time and degradation rate δ. The source of this difference is unclear and could be within the cutoff ϕ to separate the regimes. For small protein levels of separation of S_A, S_B and S_0 is harder, as ϕ will approach 0.

3.2 Switching bias and robustness analysis

We are interested in how the stochastic model of the toggle switch behaves in case of asymmetries between the two genes. As shown before in section 3.1.2 one major drawback of deterministic system is that they are unable to capture bistability and will always converge to the same regime if parameters are slightly non-symmetric. Therefore, we evaluate how the switching decision is influenced by the introduction of asymmetric parameters for the two genes.

3.2.1 Promoter binding

We assume that the main source of asymmetry comes from differences in protein-DNA interaction. For a detailed discussion on this assumption, see section 2.3. Despite this asymmetries, the toggle switch should still be able to reach both committed regimes in vivo and should even more do this with about equal probability for each regime. A differentiating cell should be able to create both lineages and both with equal probability.

As a consequence, we focus our analysis on asymmetries in the Protein-DNA binding reaction and on their influence on the switching decision for different levels of proteins in the system.

We use the stochastic simulation to obtain estimates of the overall probability that the system is in a certain regime R:

$$P_R = \frac{1}{L} \sum_{i,e} t_R^{(i,e)} \ ,$$

where L is the overall simulated time. The switching bias Bias_A of the system is defined as

$$\mathrm{Bias}_\mathrm{A} = \frac{P_\mathrm{A}}{P_\mathrm{A} + P_\mathrm{B}} \ .$$

The bias quantifies if the system has a preference for one of the two regimes. In case of a completely symmetric system the bias is 0.5, meaning that the probability of the system being in S_A is equal to that of being in S_B. In case of a fully tilted switch, the bias would be 0 or 1. Note that this quantity does not include the probability for being in S_0. Calculating the bias analytically is not feasible at this point as we need to know the distribution of the underlying Markov process completely, which corresponds to the solution of the CME. Numerical solutions to the CME could give insights about the bias, but are currently impossible to obtain for systems with large state space.

We obtain systems with different amounts of proteins in S_A and S_B by changing the protein degradation rate. Additionally, we vary the binding parameters in the range $0.5 < \tau_\mathrm{B}^+/\tau_\mathrm{A}^+ < 2$, thus giving one gene up to a two-fold increased binding strength over the other.

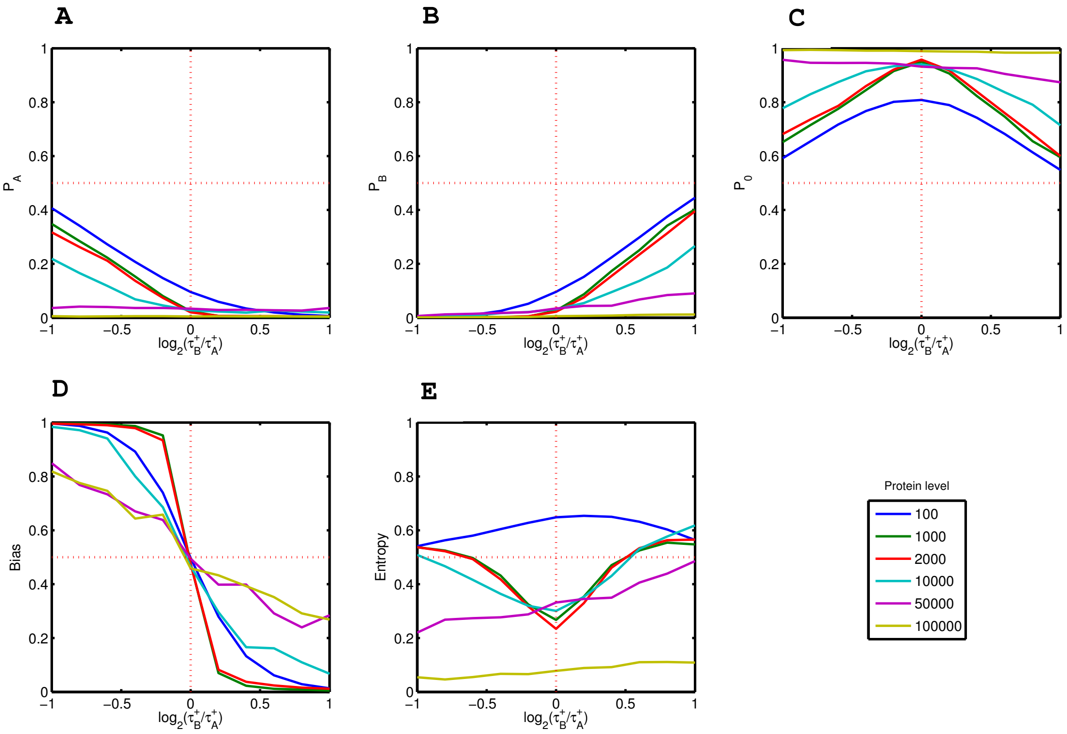

Figure 3.10: Robustness of the one-stage model with respect to binding constants. Estimates of P_{A}, P_{B}, P_0, the bias and the entropy are plotted against the binding asymmetry for six different protein mean values μ. Remaining parameters are set according to Table 2.1. Symmetric parameters give a bias $B_{\mathrm{A}} = 0.5$, whereas asymmetric parameters result in $B_{\mathrm{A}} \neq 0.5$ tilting the switch in to either S_{A} or S_{B}.

For each system, 1000 simulations were computed. Using 1000 trajectories estimates the regime probabilities with good accuracy (see Appendix B). Initial conditions were set to both promoters unbound and no proteins present. For each set of parameters we evaluate the 1000 simulated timecourses with respect to the switching bias. Results are shown in Fig. 3.10. Note that the plots are non-symmetric since for $\log_2(\tau^+_{\mathrm{B}}/\tau^+_{\mathrm{A}}) < 0$ we have decreased one parameter two-fold compared to the other, whereas for $\log_2(\tau^+_{\mathrm{B}}/\tau^+_{\mathrm{A}}) > 0$ we have increased one parameter two-fold compared to the other resulting in different absolute parameter values.

For all protein levels μ between 10^2 and 10^5 we observe as expected that giving advantage to one gene through higher binding rates of its protein tilts the switch into this direction leading to $B_{\mathrm{A}} \neq 0.5$ (Fig 3.10 A,B and D). For the symmetric case $\tau^+_{\mathrm{B}}/\tau^+_{\mathrm{A}} = 1$ both S_{A} and S_{B} are equally probable as expected, leading to a bias of 0.5. For systems with higher protein levels/small degradation rates the bias is less influenced by the parameter asymmetry.

However this effect is time dependent. Using longer simulation time, we find that also systems with small degradation rate become more biased in case of asymmetric parameters. (data not shown). This can be explained if we look at process of decision making. If there is almost no protein in the system yet, the regime choice will only depend on which kind of protein is synthesized first, driving the system into this direction. As synthesis parameters are assumed to be equal for both players, this event is always unbiased (each player has 50% probability of first synthesis). If the system instead runs into a deadlock, (few) proteins of both species are present. This increases the influence of the parameter asymmetry as one of the two protein species has higher DNA affinity, meaning that the other one is more likely to be degraded. The decision will be biased again. In other words, there are two different types of switching decisions. If no proteins are present, the system will decide unbiased, if proteins are present the decision will be biased. This also explains the smaller bias of systems with high protein level μ. Systems with high μ will first enter either S_{A} or S_{B} with equal probability, and due to long switching time (see section 3.1.5), they will keep this unbiased decision over a long time. Following regime changes, which are biased, become less relevant.

Systems with high protein levels also have increased probability of being in S_0, which is in agreement with the findings that these systems have long priming times. Interestingly, introducing asymmetry in binding parameters decreases the probability P_0: If non of both players has an advantage over the other the system is longer locked in S_0 as neither Protein$_{\mathrm{A}}$ or Protein$_{\mathrm{B}}$ can significantly overwhelm the other (see Fig. 3.10 C).

To get not only a qualitative impression on the systems robustness with respect to parameters but a quantitative measure we utilize the Shannon entropy:

$$E = -P_{\mathrm{A}} \cdot \log_2(P_{\mathrm{A}}) - P_{\mathrm{B}} \cdot \log_2(P_{\mathrm{B}})$$

The Shannon entropy is a measure commonly used in information theory and describes the uncertainty of a random variable. For example consider the random variable describing the result of a coin flip. If the coin is fair, the the uncertainty of the coin toss is maximal and the entropy is maximal with $E = 1$. If the coin is unfair, preferring e.g. tail, the uncertainty is less, resulting in $E < 1$. For our model, the entropy is expected to be maximal where the probability of both regimes is equal but also decision time (time in S_0) is not too long.

Calculating the entropy reveals, that even though systems with high protein numbers are less biased in their switching, they spend most of their time in S_0. Therefore, the whole system looses its biological relevance, since it is very improbable that it will establishes a decision at all.

Systems with low protein numbers have higher probability to be in S_{A} and S_{B}, but have the major drawback that they are strongly biased at asymmetric parameters. Additionally we know from section 3.1.5 that these systems can only maintain a decision for a short time, questioning their relevance in biological systems as well.

3.2.2 Initial conditions

Now we analyze the influence of asymmetric initial conditions on the switching bias. A common opinion of how a toggle switch is activated is the shift of the cell from a coexpression state to a switching state [21, 49]. During the coexpression state, the mutual inhibition in the toggle switch motif is thought to be very weak or non-existent, leading to similar expression of both players. This could for example be achieved by the chromatin structure of the DNA, making the inhibitory operator sites unaccessible. In our model this is realized by setting $\tau^+ \ll \tau^-$ leading to only very short periods where promoters are inhibited, thus shutting down mutual inhibition. By an external signal the cell is now moved into the switching state, where $\tau^+ \gg \tau^-$ leading to strong mutual inhibition. Such change could be the result of a change in the DNA structure, making the repressor elements of the genes accessible. As this change involves only the one or two molecules of DNA in the cell, we assume this event is discrete and not a continuous gradual change, meaning that the transition between those two states is rapid.

Therefore, after that transition occurred, the amount of Protein$_\mathrm{A}$ and Protein$_\mathrm{B}$ correspond to the distribution during the coexpressed state. As there is hardly any interaction between both species in the coexpression state, we assume that their protein distributions are equal to the protein distribution of two single genes. Their means and variances obey equation (3.5). It is obvious that, at the time of the transition from coexpression to switching, the numbers of Protein$_\mathrm{A}$ and Protein$_\mathrm{B}$ will not necessarily be the same but fluctuate around the mean values μ_A and μ_B. So even if the binding parameters for both species are symmetric, due to the spontaneous transition from coexpression to strong mutual inhibition, the initial conditions in this switching state will be asymmetrical by chance.

Analogous to the derivation of the regime boundaries in equation (3.7), we now use the Poisson distribution of single genes to get an impression of the possible asymmetries regarding initial protein levels.

We define the minimal/maximal effective number of protein x by

$$\begin{aligned} \phi_x^- &= F_{\mu_x}^{-1}(0.001) \\ \phi_x^+ &= F_{\mu_x}^{-1}(0.999)\ . \end{aligned} \tag{3.14}$$

Setting the initial protein number of one protein to the minimal number and the other one to the maximal number gives the maximal expected asymmetry in initial protein amounts. Note however that the fold change in protein amount is not constant but decreases as the protein steady state increases. Using the boundaries of the probability distribution overcomes this problem.

By varying the ratio of both initial protein amounts from the minimal/maximal bounds towards the mean steady state value μ, we evaluate the influence of this asymmetry on the switching behavior. Contrary to the analysis in section 3.2.2, we are now only interested in how the asymmetry influences the system's first decision. The initial conditions are expected to have only influence on the first decision, leading either into S_A or S_B. Subsequent flipping of the toggle switch is independent of the initial conditions. Therefore, systems that change their regimes frequently during the simulation time will always look unbiased, since they could compensate the possible bias in the first decision by the overwhelming amount of unbiased decision afterwards. Instead of estimating the overall probabilities to be in one or the other regime (P_A and P_B), we estimate the probability that the system will enter S_A or S_B at the first possible choice using N timecourses simulated with the SSA. Switching events after the first one are neglected. These probabilities the relative frequencies of systems first entering S_A or S_B:

$$\begin{aligned} P_\mathrm{A}^{(1)} &= \frac{\#\text{ of timecourses entering } S_\mathrm{A} \text{ first}}{N} \\ P_\mathrm{B}^{(1)} &= \frac{\#\text{ of timecourses entering } S_\mathrm{B} \text{ first}}{N} \\ P_0^{(1)} &= 1 - (P_\mathrm{A}^{(1)} + P_\mathrm{B}^{(1)})\ . \end{aligned}$$

$P_0^{(1)}$ accounts for the systems that do not decide at all during simulation. Fig. 3.11 shows the results obtained, using $N = 1000$ simulated timecourses.

Asymmetry in the initial conditions seems to have no influence on the systems and each system has about the same probability for first entering S_A or S_B. Systems with small degradation rate spend most time in S_0. To understand this one has to take a closer look on the actual timecourses.

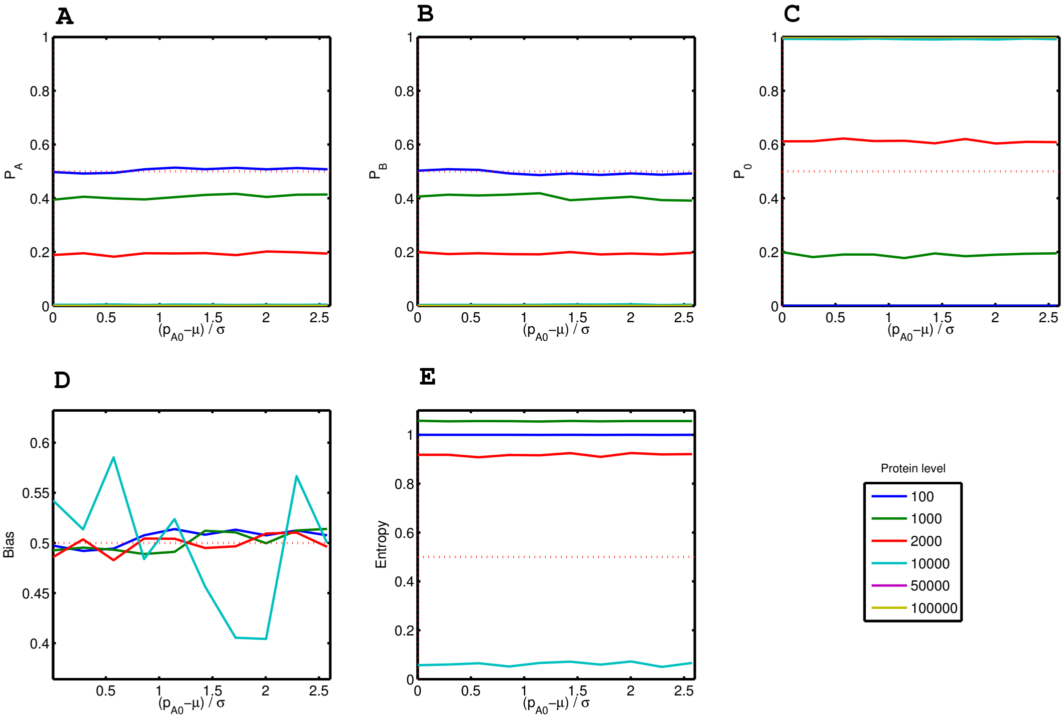

Figure 3.11: Robustness with respect to switching bias and initial conditions. Parameters are chosen from Table 2.1. The initial conditions of the systems were chosen within the bounds ϕ^+ and ϕ^- of the protein level distribution (see equation (3.14)). On the x-axis the distance of the initial number of Protein_A and Protein_B from the distribution mean is given in standard deviations. For systems with distance 0, initial conditions for both proteins were chosen from the center of the distribution, resulting in equal amounts. System with large distance have initial conditions chosen from the opposing edges of the distribution. All systems are robust to asymmetric initial conditions, the switching bias is close to 0.5 even for strong asymmetries. Strong fluctuations of the bias of the 10000-protein system occur, because this system spends most of the time in S_0, leading to very small probabilities of S_A and S_B.

Starting with initial conditions where both proteins are highly expressed, almost immediately both promoters are bound, resulting in a deadlock situation. This is followed by an almost deterministic phase of degradation, where both proteins decay. This phase will be longer if the degradation rate is smaller. The deadlock will be resolved if one of the two species is fully degraded. As both proteins are almost fully degraded, the initial dominance of one species is lost, leading to an unbiased decision afterwards, as observed for 100-10000 protein level system. Since the length of the degradation phase grows with decreasing degradation rate, beyond some degradation rate there will be no decision made during the simulated time (most time the system is in S_0). Systems with a mean of 10000 proteins are close to this boundary: Sometimes all proteins of one species are degraded, so that the system can escape the deadlock, but most of the time this degradation takes longer than the simulated time of 10^6s. This is the reason for the strong fluctuations observed for the switching bias of this systems. Overall, these results show that the one-stage system is robust to asymmetries in the initial conditions.

Chapter 4

Two-stage toggle switch

In the previous chapter, we used a one-stage model of gene expression, where for simplicity transcription and translation were collapsed into a single synthesis reaction. In this chapter we include the mRNA level of gene expression into our model and analyze this model in a similar fashion to the previous chapter. A comparison of both models will be given in chapter 5.

4.1 System description

As in the previous chapter we first define our model as a set of biochemical reactions and then study the system's dynamics using the deterministic and probabilistic framework.

4.1.1 Gene expression and regulatory interactions

Based on the one-stage model, we add a second stage in gene expression by introducing mRNA as additional species and replacing the former synthesis reaction by separate transcription and translation reactions:

$$\begin{aligned}\text{DNA} &\xrightarrow{\alpha} \text{DNA} + \text{mRNA}\\ \text{mRNA} &\xrightarrow{\beta} \text{mRNA} + \text{Protein}\end{aligned}$$

Separate degradation reactions are introduced for mRNA and proteins:

$$\begin{aligned}\text{mRNA} &\xrightarrow{\gamma} \emptyset\\ \text{Protein} &\xrightarrow{\delta} \emptyset\end{aligned}$$

We refer to this as a two-stage model of gene expression.

For the two-stage toggle switch,we include an additional species representing protein-bound DNA, and the corresponding binding/dissociation reactions as before in equation (3.1). This bound DNA cannot be transcribed into mRNA, leading to transcriptional inhibition through protein binding. However, the translation process is independent of the DNA state:

Even if the promoter of the gene is inhibited, as long as mRNA from a previous transcription event is present, translation can occur.

Altogether the two-stage toggle switch can be expressed by the set of reactions listed in Table 4.1. Fig. 4.1 shows a schematic of the system.

4.1.2 Mathematical representation

Deterministic model

The dynamics of the system can be expressed deterministically by the reaction rate equations, where d_{A} (d_{B}) is the abundance of $\mathrm{DNA_A}$ ($\mathrm{DNA_B}$), m_{A} (m_{B}) is the abundance of $\mathrm{mRNA_A}$ ($\mathrm{mRNA_B}$) and p_{A} (p_{B}) is the abundance of $\mathrm{Protein_A}$ ($\mathrm{Protein_B}$):

$$\begin{aligned}
\frac{d}{dt}d_{\mathrm{A}} &= \tau_{\mathrm{B}}^{-} \cdot (1 - d_{\mathrm{A}}) - \tau_{\mathrm{B}}^{+} \cdot d_{\mathrm{A}} \cdot p_{\mathrm{B}} \qquad (4.1)\\
\frac{d}{dt}d_{\mathrm{B}} &= \tau_{\mathrm{A}}^{-} \cdot (1 - d_{\mathrm{B}}) - \tau_{\mathrm{A}}^{+} \cdot d_{\mathrm{B}} \cdot p_{\mathrm{A}}\\
\frac{d}{dt}m_{\mathrm{A}} &= \alpha_{\mathrm{A}} \cdot d_{\mathrm{A}} - \gamma_{\mathrm{A}} \cdot m_{\mathrm{A}}\\
\frac{d}{dt}m_{\mathrm{B}} &= \alpha_{\mathrm{B}} \cdot d_{\mathrm{B}} - \gamma_{\mathrm{B}} \cdot m_{\mathrm{B}}\\
\frac{d}{dt}p_{\mathrm{A}} &= \beta_{\mathrm{A}} \cdot m_{\mathrm{A}} - \delta_{\mathrm{A}} \cdot p_{\mathrm{A}} + \tau_{\mathrm{A}}^{-} \cdot (1 - d_{\mathrm{B}}) - \tau_{\mathrm{A}}^{+} \cdot d_{\mathrm{B}} \cdot p_{\mathrm{A}}\\
\frac{d}{dt}p_{\mathrm{B}} &= \beta_{\mathrm{B}} \cdot m_{\mathrm{B}} - \delta_{\mathrm{B}} \cdot p_{\mathrm{B}} + \tau_{\mathrm{B}}^{-} \cdot (1 - d_{\mathrm{A}}) - \tau_{\mathrm{B}}^{+} \cdot d_{\mathrm{A}} \cdot p_{\mathrm{B}}
\end{aligned}$$

Bound DNA is expressed in terms of unbound DNA due to mass conservation ($\mathrm{DNA} + \mathrm{DNA}^{\mathrm{bound}} = 1$).

We obtain the steady state solution by setting all time derivatives to 0 and find two non-trivial solutions, one biologically irrelevant due to its negative species abundances (see Appendix A.2). Numerical solutions with different parameters and different initial conditions are shown in Fig. 4.2 and a portrait of the $p_{\mathrm{A}}/p_{\mathrm{B}}$ phase-space is depicted in Fig. 4.3. Introducing the mRNA stage can lead to dampened oscillations, due to the fact that regulation is not carried out by mRNA but involves the translation process creating a delay in the regulatory response.

Still this system is not capable of bistability: Given non-negative initial conditions, the solution of the reaction rate equations will always be non-negative (see section 2.1.2) and will therefore converge to the single positive steady state.

Probabilistic model

The stochastic model of the two-stage toggle switch is completely defined by the CME. As before

$$\begin{aligned}
P_{ij}(m_{\mathrm{A}}, m_{\mathrm{B}}, p_{\mathrm{A}}, p_{\mathrm{B}}, t) =& P(\mathrm{mRNA_A} = m_{\mathrm{A}}, \mathrm{mRNA_B} = m_{\mathrm{B}},\\
& \mathrm{Protein_A} = p_{\mathrm{A}}, \mathrm{Protein_B} = p_{\mathrm{B}},\\
& \mathrm{DNA}_{\mathrm{A}}^{\mathrm{bound}} = i, \mathrm{DNA}_{\mathrm{B}}^{\mathrm{bound}} = j, t)
\end{aligned}$$

is the probability to have m_{A} copies of $\mathrm{mRNA_A}$, m_{B} copies of $\mathrm{mRNA_B}$, p_{A} copies of $\mathrm{Protein_A}$, p_{B} copies of $\mathrm{Protein_B}$, and the corresponding promoter configuration ij. This results in

$$
\begin{array}{rlrl}
\mathrm{DNA_B} & \xrightarrow{\alpha_\mathrm{B}} \mathrm{DNA_B} + \mathrm{mRNA_B} & \mathrm{DNA_A} & \xrightarrow{\alpha_\mathrm{A}} \mathrm{DNA_A} + \mathrm{mRNA_A} \\
\mathrm{mRNA_B} & \xrightarrow{\gamma_\mathrm{B}} \emptyset & \mathrm{mRNA_A} & \xrightarrow{\gamma_\mathrm{A}} \emptyset \\
\mathrm{mRNA_B} & \xrightarrow{\beta_\mathrm{B}} \mathrm{mRNA_B} + \mathrm{Protein_B} & \mathrm{mRNA_A} & \xrightarrow{\beta_\mathrm{A}} \mathrm{mRNA_A} + \mathrm{Protein_A} \\
\mathrm{Protein_B} & \xrightarrow{\delta_\mathrm{B}} \emptyset & \mathrm{Protein_A} & \xrightarrow{\delta_\mathrm{A}} \emptyset \\
\mathrm{Protein_B} + \mathrm{DNA_A} & \xrightarrow{\tau_\mathrm{B}^+} \mathrm{DNA_A^{bound}} & \mathrm{Protein_A} + \mathrm{DNA_B} & \xrightarrow{\tau_\mathrm{A}^+} \mathrm{DNA_B^{bound}} \\
\mathrm{DNA_A^{bound}} & \xrightarrow{\tau_\mathrm{B}^-} \mathrm{Protein_B} + \mathrm{DNA_A} & \mathrm{DNA_B^{bound}} & \xrightarrow{\tau_\mathrm{A}^-} \mathrm{Protein_A} + \mathrm{DNA_B}
\end{array}
$$

Table 4.1: List of reactions for the two-stage toggle switch, consisting of transcription, mRNA degradation, translation, protein degradation and binding/dissociation of the repressor. $\emptyset$ denotes a sink state that absorbs degraded mRNAs and proteins.

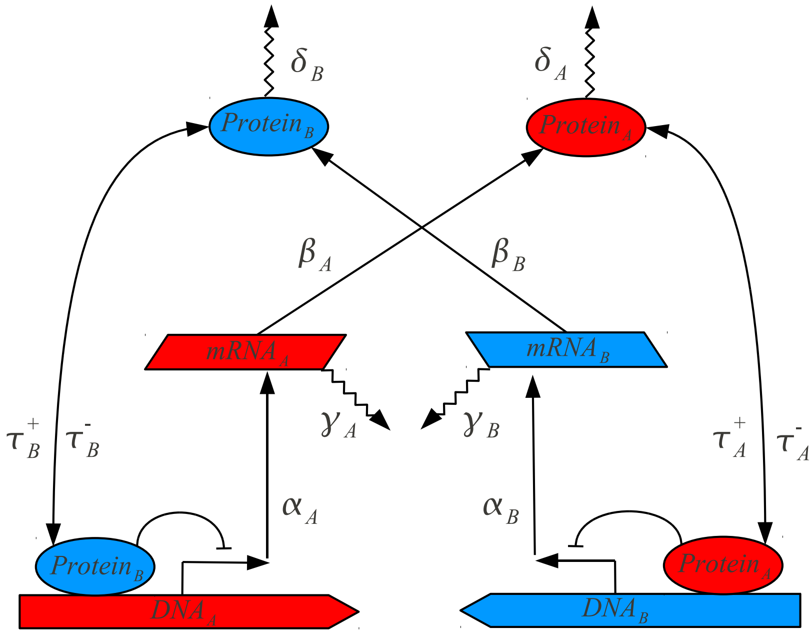

Figure 4.1: Scheme of the two-stage toggle switch. Species associated with player A are shown in red, species associated with B are shown in blue. Reactions/Interactions are indicated as arrows, jagged arrows indicate degradation reactions. $\mathrm{mRNA_A}$ is transcribed from $\mathrm{DNA_A}$ with rate α_A. $\mathrm{mRNA_A}$ decays with rate γ_A and is translated into $\mathrm{Protein_A}$ with rate β_A. The protein decays with rate δ_A and can bind (unbind) $\mathrm{DNA_B}$ with rate τ_A^+ (τ_A^-). Protein-bound DNA lead to transcriptional arrest.

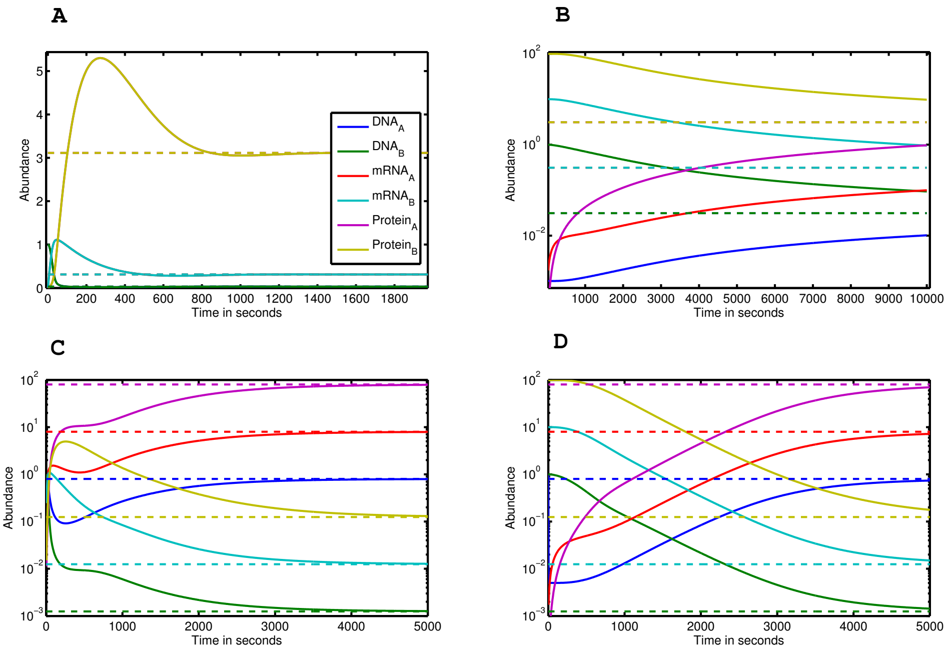

Figure 4.2: Numerical solution of the reaction rate equations in the two-stage toggle switch for different parameters and initial conditions. Dashed lines indicate steady state values. A,B: Symmetrical parameters for both A and B (see Tab. 2.1), leading to identical time evolution of the corresponding species. Initial conditions are $m_{\rm A} = m_{\rm B} = p_{\rm A} = p_{\rm B} = 0, d_{\rm A} = d_{\rm B} = 1$ (A) and $m_{\rm A} = p_{\rm A} = 0, m_{\rm B} = 10, p_{\rm B} = 100, d_{\rm A} = d_{\rm B} = 1$ (B). C,D: Asymmetrical binding constants $\tau_{\rm A}^{+} = 1, \tau_{\rm B}^{+} = 0.5$ leading to A-dominance. Initial conditions are $m_{\rm A} = m_{\rm B} = p_{\rm A} = p_{\rm B} = 0, d_{\rm A} = d_{\rm B} = 1$ (C) and $m_{\rm A} = p_{\rm A} = 0, m_{\rm B} = 10, p_{\rm B} = 100, d_{\rm A} = d_{\rm B} = 1$ (D).

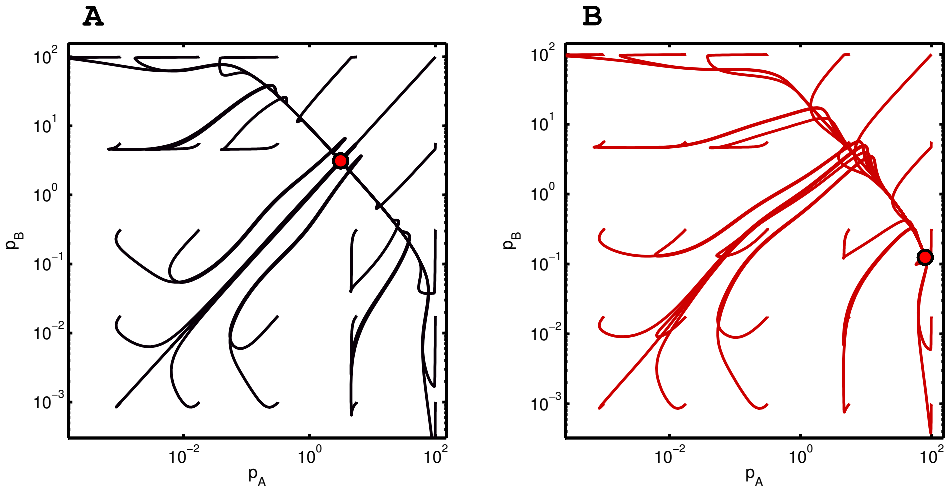

Figure 4.3: Phase portrait of the two-stage toggle switch. This plot shows the ODE-system's trajectories in the $p_{\rm A}/p_{\rm B}$ -phaseplane starting from different initial conditions. The steady state solution is indicated by a red dot. A: completely symmetric parameters for both players lead to a steady state lying on the symmetry axis $p_{\rm A} = p_{\rm B}$. B: $\tau_{\rm A}^{+} > \tau_{\rm B}^{+}$, leading to a steady state shifted below the symmetry axis. Other parameters were set according to Tab. 2.1. Note that trajectories do intersect in this plot, since it shows a projection of a six-dimensional solution onto a two-dimensional phaseplane.

the following CME, that is split up into four coupled equations, corresponding to the four promoter states:

$$
\begin{aligned}
\frac{d}{dt}P_{00}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) = {} & \tau_\mathrm{A}^- P_{01}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A}-1,p_\mathrm{B},t) + \tau_\mathrm{B}^- P_{10}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B}-1,t) \\
& + [-\tau_\mathrm{B}^+ p_\mathrm{B} - \tau_\mathrm{A}^+ p_\mathrm{A} + \alpha_\mathrm{A}(E^-_{m_\mathrm{A}}-1) + \alpha_\mathrm{B}(E^-_{m_\mathrm{B}}-1) \\
& \quad + \gamma_\mathrm{A}(E^+_{m_\mathrm{A}}-1)\cdot m_\mathrm{A} + \gamma_\mathrm{B}(E^+_{m_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \beta_\mathrm{A}(E^-_{p_\mathrm{A}}-1)\cdot m_\mathrm{A} + \beta_\mathrm{B}(E^-_{p_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \delta_\mathrm{A}(E^+_{p_\mathrm{A}}-1)\cdot p_\mathrm{A} + \delta_\mathrm{B}(E^+_{p_\mathrm{B}}-1)\cdot p_\mathrm{B}] \\
& P_{00}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) \\
\frac{d}{dt}P_{11}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) = {} & \tau_\mathrm{A}^+(p_\mathrm{A}+1)P_{10}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A}+1,p_\mathrm{B},t) \\
& + \tau_\mathrm{B}^+(p_\mathrm{B}+1)P_{01}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B}+1,t) \\
& + [-\tau_\mathrm{A}^- - \tau_\mathrm{B}^- + \gamma_\mathrm{A}(E^+_{m_\mathrm{A}}-1)\cdot m_\mathrm{A} + \gamma_\mathrm{B}(E^+_{m_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \beta_\mathrm{A}(E^-_{p_\mathrm{A}}-1)\cdot m_\mathrm{A} + \beta_\mathrm{B}(E^-_{p_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \delta_\mathrm{A}(E^+_{p_\mathrm{A}}-1)\cdot p_\mathrm{A} + \delta_\mathrm{B}(E^+_{p_\mathrm{B}}-1)\cdot p_\mathrm{B}] \\
& P_{11}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) \\
\frac{d}{dt}P_{10}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) = {} & \tau_\mathrm{A}^- P_{11}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A}-1,p_\mathrm{B},t) + \tau_\mathrm{B}^+(p_\mathrm{B}+1)P_{00}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B}+1,t) \\
& + [-\tau_\mathrm{B}^- - \tau_\mathrm{A}^+ p_\mathrm{A} + \alpha_\mathrm{B}(E^-_{m_\mathrm{B}}-1) + \gamma_\mathrm{A}(E^+_{m_\mathrm{A}}-1)\cdot m_\mathrm{A} + \gamma_\mathrm{B}(E^+_{m_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \beta_\mathrm{A}(E^-_{p_\mathrm{A}}-1)\cdot m_\mathrm{A} + \beta_\mathrm{B}(E^-_{p_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \delta_\mathrm{A}(E^+_{p_\mathrm{A}}-1)\cdot p_\mathrm{A} + \delta_\mathrm{B}(E^+_{p_\mathrm{B}}-1)\cdot p_\mathrm{B}] \\
& P_{10}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) \\
\frac{d}{dt}P_{01}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t) = {} & \tau_\mathrm{A}^+(p_\mathrm{A}+1)P_{00}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A}+1,p_\mathrm{B},t) + \tau_\mathrm{B}^- P_{11}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B}-1,t) \\
& + [-\tau_\mathrm{B}^+ p_\mathrm{B} - \tau_\mathrm{A}^- + \alpha_\mathrm{A}(E^-_{m_\mathrm{A}}-1) + \gamma_\mathrm{A}(E^+_{m_\mathrm{A}}-1)\cdot m_\mathrm{A} + \gamma_\mathrm{B}(E^+_{m_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \beta_\mathrm{A}(E^-_{p_\mathrm{A}}-1)\cdot m_\mathrm{A} + \beta_\mathrm{B}(E^-_{p_\mathrm{B}}-1)\cdot m_\mathrm{B} \\
& \quad + \delta_\mathrm{A}(E^+_{p_\mathrm{A}}-1)\cdot p_\mathrm{A} + \delta_\mathrm{B}(E^+_{p_\mathrm{B}}-1)\cdot p_\mathrm{B}] \\
& P_{01}(m_\mathrm{A},m_\mathrm{B},p_\mathrm{A},p_\mathrm{B},t)\ .
\end{aligned}
$$

The shift operators E_x^+ and E_x^- were used to simplify the expression analogous to section 3.1.2. To our knowledge no results have yet been published on the solution of stochastic two-stage toggle switches.

Approaches to solve this CME applying the naive finite state projection method used for the one-stage systems in section 3.1.4 are condemned to failure, since the additional mRNA players increase the state space drastically. Consider a case analogous to the one-stage system, where we chose to cut the state space for $p_\mathrm{A} > 42$, $p_\mathrm{B} > 42$. Additionally we have to restrict the state space in the mRNA dimensions, e.g. at $m_\mathrm{A} > 20$, $m_\mathrm{B} > 20$ (valid for systems with avg. of 10 mRNAs). Due to mass conservation we only need to consider four different combinations of DNA states ($DNA_x + DNA_x^{\mathrm{bound}} = 1$). Overall this will result in a truncated state space S_T of size

$$|S_T| = 42^2 \cdot 20^2 \cdot 4 = 7.05 \cdot 10^5$$

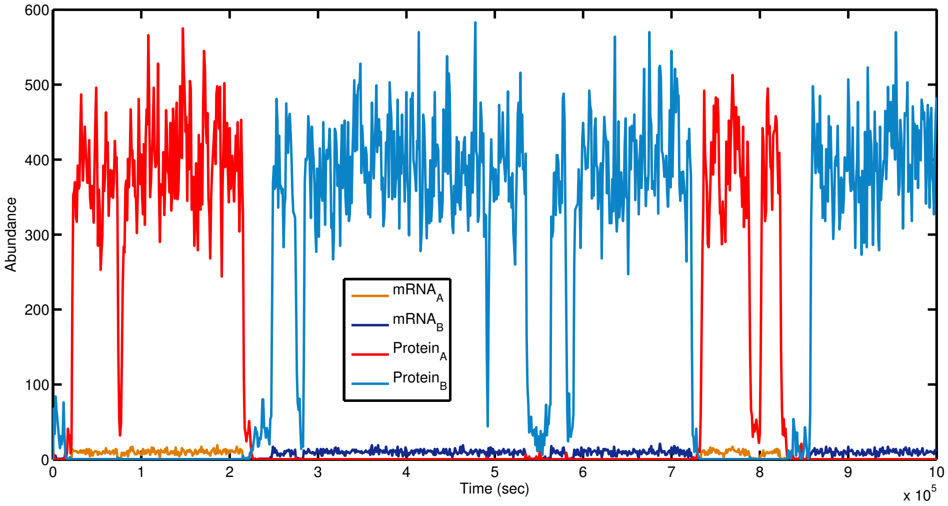

Figure 4.4: Exemplary timecourse of a two-stage toggle switch showing similar dynamical regimes as the one stage switches.

The system's matrix Q (see equation (2.7)) will contain $|S_T| \cdot 12 = 8.47 \cdot 10^6$ non-zero entries. Solving equation (2.8) for such huge matrices is not efficient and not possible for the desired accuracy. For small protein and mRNA levels a more elaborate iterative projection algorithm might be able to solve this CME. However, as the finite state projection is not the main focus of this work, we instead use the stochastic simulation algorithm to analyze the two stage system.

4.1.3 Regimes

The principal dynamics of the two-stage system are similar to the one-stage toggle switch (see Fig. 4.4), resulting in three regimes: In the deadlock regime both promoters are occupied by inhibitors, shutting down the transcription completely. In the A-regime, Protein$_\mathrm{A}$ is abundant and suppresses synthesis of Protein$_\mathrm{B}$ through binding and inhibition of the promoter of DNA$_\mathrm{B}$. The promoter of DNA$_\mathrm{A}$ is unbound. In the B-regime, Protein$_\mathrm{A}$ is dominated by large numbers of Protein$_\mathrm{B}$, inhibiting the promoter of DNA$_\mathrm{A}$. DNA$_\mathrm{B}$ is unbound.

Differences in the dynamics are expected due to the intermediary mRNA state. The regulation of the promoter does not directly depend the promoter's products (mRNA), but involves an additional step of translation, creating a certain delay of promoter response. More importantly the protein distribution itself will be different, as noise from the mRNA level is further amplified through translation.

As before we aim for a mathematical definition of the three regimes of the system. We can again approximate the protein number distribution in the A/B regimes by results from Thattai and van Oudenaarden [58], who showed that for a simple two-state expression model, the mean and variance of protein numbers obey

$$\mu_x = \frac{\alpha_x \beta_x}{\gamma_x \delta_x} \tag{4.2}$$

$$\sigma_x^2 = \frac{\beta_x^2 \alpha_x}{\gamma_x^2 \delta_x + \delta_x^2 \gamma_x} \ . \tag{4.3}$$

The corresponding distribution (shown to be negative binomial by Shahrezaei and Swain [55] in case of $\gamma/\delta \gg 1$) is much broader than for the one-stage model, because transcriptional noise is further amplified. The mRNA distributions are Poissonian with mean and variance of $\frac{\alpha}{\gamma}$, in analogy to the protein distributions in the one-stage model. We define the regimes in the same way as for the one-stage toggle switch (see equations (3.6))

The regime S_A (S_B) is the subspaces of the state space S, where Protein$_\mathrm{A}$ (Protein$_\mathrm{B}$) is dominating, regime S_0 is the subspace where the system is in a deadlock situation. Note that we do not include the mRNA numbers into our regime definitions, since they are only an intermediate and not the effectors of the systems. In a similar way to section 3.1.3, we define the boundary ϕ_x of regime x using a normal approximation of the dominating protein's distribution as

$$\phi_x = \mu_x - Z_{\alpha'} \cdot \sigma_x \ . \tag{4.4}$$

$Z_{\alpha'}$ is the α'% quantil of the standard normal distribution. For example using $\alpha' = 0.001$ assures that 99.9% percent probability mass of the distribution lie beyond the lower boundary. Therefore we are certain to capture all relevant protein numbers belonging to S_A and S_B.

4.1.4 Switching time

Using the results obtained for the one-stage switch in section 3.1.5, the extension of the switching time for two-stage switches is straight forward: Without loss of generality we assume that the system is in S_A. Instead of requiring only synthesis of one Protein$_\mathrm{B}$ followed by a binding reaction, now there needs to be a transcription reaction creating mRNA$_\mathrm{B}$ when DNA$_\mathrm{B}$ is unbound, followed by translation during mRNA lifetime – note that during this time the promoter can be bound by the repressor again – and a binding reaction during the protein lifetime of Protein$_\mathrm{B}$. So we only have to change the parameter of the geometric distribution to account for these events: The probability of a transcription during the unbound phase is (analogously to the synthesis in the one-stage model)

$$q_s = 1 - \exp\left(-\frac{\alpha_\mathrm{B}}{\tau_\mathrm{A}^+ \cdot \mu_\mathrm{A}}\right) \ . \tag{4.5}$$

The probability of translation during average mRNA lifetime is

$$q_t = 1 - \exp\left(-\frac{\beta_\mathrm{B}}{\gamma_\mathrm{B}}\right) \ .$$

q_t will be close to one for biological relevant parameters as $\beta_\mathrm{B} > \gamma_\mathrm{B}$ (see section 2.3), meaning that once mRNA has been synthesized it is quite certain that it will be translated at least once. Finally the probability of a binding reaction during protein lifetime is

$$q_b = 1 - \exp\left(-\frac{\tau_\mathrm{B}^+}{\delta_\mathrm{B}}\right) \ .$$

As for the one-stage system this is expected to be close to 1 for biologically relevant parameters since $\tau_\mathrm{B}^+ \gg \delta_\mathrm{B}$ (see section 2.3).

Altogether we get

$$N \sim \mathrm{Geom}(q_s \cdot q_t \cdot q_b) \tag{4.6}$$

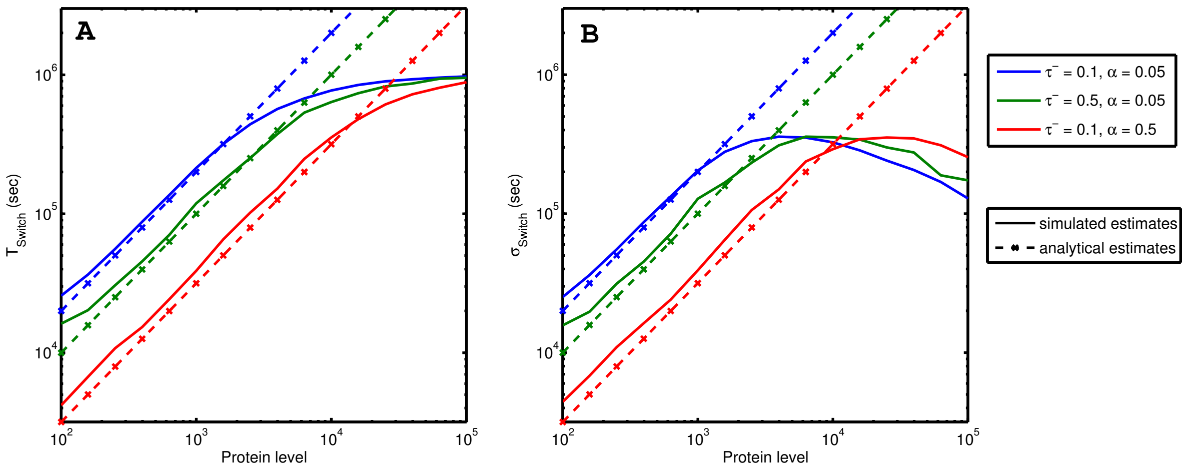

Figure 4.5: Switching times for the two-stage models. Estimates for the switching time over various protein degradation rates δ and three different binding/transcription parameters(blue, red, green) obtained by different methods are shown. Remaining parameters are set according to Table 2.1. Solid lines: SSA estimates. Dashed lines: Analytical estimates. A: Estimates of the mean switching time. B: Estimates of the standard deviation of switching time. The analytical estimates resembles the simulated results. The simulated estimates deviate from analytical estimates for higher protein levels due to the maximal simulation time of 10^6 seconds.

Using the linear transformation $T = \frac{N}{\tau_{\mathrm{A}}^-}$ we end up with the mean and the variance of the switching time as

$$T_{\mathrm{Switch}} = \mathrm{Mean}(T) = \frac{1}{\tau_{\mathrm{A}}^- \cdot q_s q_t q_b}$$
$$\sigma_T^2 = \mathrm{Var}(T) = \frac{1}{(\tau_{\mathrm{A}}^-)^2} \cdot \frac{1 - q_s q_t q_b}{(q_s q_t q_b)^2} \ .$$

Now we compare the the analytical results to estimates obtained by stochastic simulation for models with different protein degradation rates, leading to different protein levels in S_{A} and S_{B}. Barzel's method could not be applied here since the state space is too large even for systems with low species numbers (as shown in section 4.1.2). We simulate 10000 timecourses of these two-stage systems and estimate the mean switching time analogous to the one-stage model in section 3.1.5. Results are shown in Fig. 4.5.

Similar to the one stage model, we find that mean switching time strongly depends on the protein degradation rate (Fig. 4.5 A), which leads to different amounts of proteins during S_{A} and S_{B}. The higher the amount of dominating proteins, the shorter are the phases, where the antagonistic promoter is unbound, thereby reducing the chance of switching.

Analytic means are in good agreement with simulated estimates (Fig. 4.5 A), as long as the switching time of the system is not to close to the maximum simulated time of 10^6 seconds (where simulation results get biased). Small deviations might result from the fact, that in the simulation we estimate the time until the protein levels drop below the threshold ϕ whereas the analytical estimate does not account for this protein degradation phase. Simulated estimates are therefore expected to be slightly higher than analytical estimates. This scenario is very likely since the slope of the analytical curve matches the simulated data quite well, only the y-intercept deviates slightly.

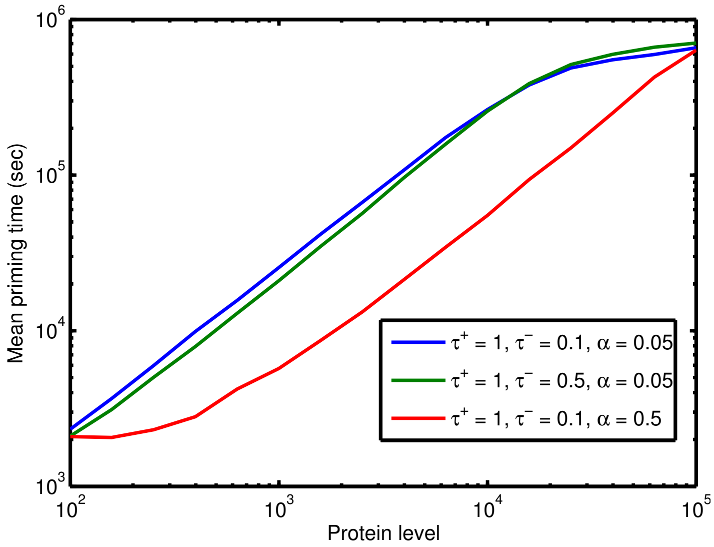

Figure 4.6: Priming times for the two-stage switch. Estimates for the switching time for 3 different parameter sets and various protein degradation rates δ obtained by stochastic simulation. The plot shows the general dependence of the priming time on the protein level μ of the system. Increased transcription rate α also influences the priming time (red line), whereas modified dissociating rates show less influence (green and blue line).

Standard deviations show almost the same characteristics (Fig. 4.5 B), since mean and standard deviation of a geometric distribution become very similar in case of a large distribution parameter. Simulated standard deviations are also dampened for systems with switching times close to 10^6 seconds, ultimately decreasing as most estimates are equal to 10^6 seconds.

4.1.5 Priming time

The priming time is defined as the average time it takes the system to leave S_0 and to reach either S_A or S_B. In case of the two-stage model this involves the degradation of all present proteins and mRNAs of one species. Even if all proteins including the single DNA bound protein are degraded, a remaining mRNA still has the potential to create a burst of proteins again, which will keep the system in S_0.

An analytical solution will not be given here, since we expect it to be even more complex than for the one-stage model. Barzel's approach becomes computationally infeasible, because the state space of two-stage systems is too large as described in section 4.1.2. Simulation results of systems with different protein degradation δ are shown in figure 4.6.

We find that these results are similar to the one-stage model: Increasing protein levels/decreasing protein degradation rates lead to increasing priming time, as it takes much longer to degrade the repressor proteins if their degradation rate is small. We observe that also the mRNA synthesis rate α has a major influence on the priming time. Systems with higher α move out of S_0 quicker, since the chance of transcription during the unbound phase is increased, driving the system towards S_A or S_B. Increasing the unbound phase t_u using larger τ^- does not show much influence, because the t_u term is dominated by the average protein number μ.

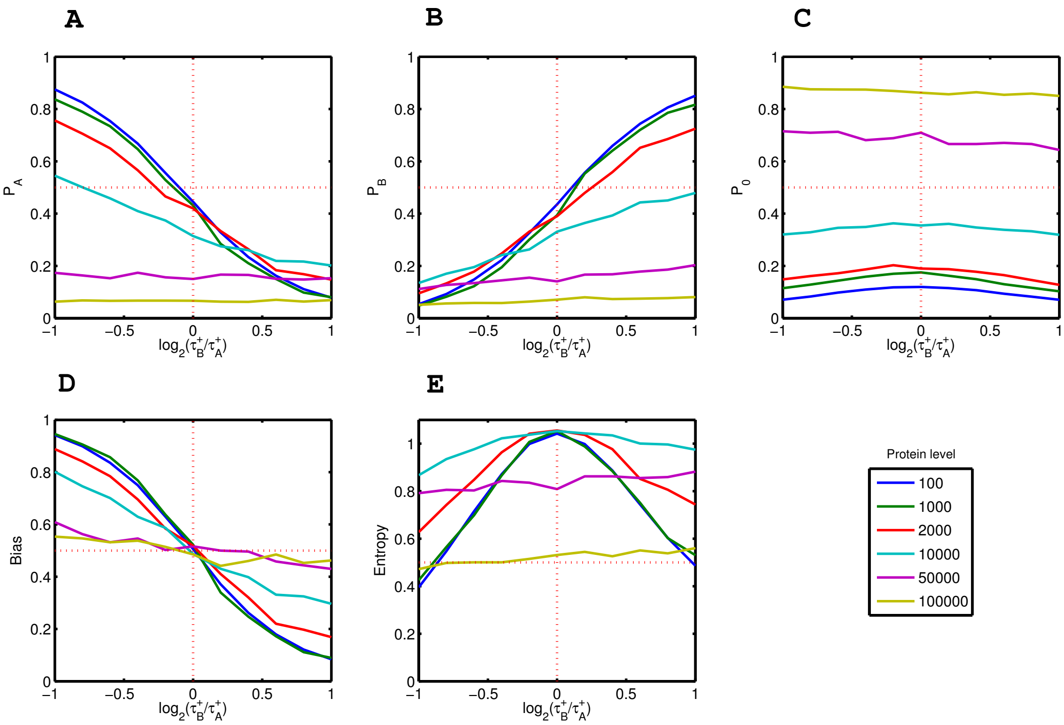

Figure 4.7: Robustness of the two-stage switch with respect to binding constants. Estimates of P_{A}, P_{B}, P_0, the bias and the entropy are plotted against the binding asymmetry for six different protein mean values μ. Symmetric parameters give a bias $\mathrm{Bias}_{\mathrm{A}} = 0.5$, whereas asymmetric parameters result in $\mathrm{Bias}_{\mathrm{A}} \neq 0.5$, tilting the switch either into S_{A} or S_{B}. Remaining parameters are chosen from Table 2.1.

4.2 Switching bias and robustness analysis

In the following sections we analyze the influence of different sources of asymmetry in the system on the switching bias.

4.2.1 Promoter binding

Here we examine the influence of asymmetric binding parameters τ^+ on the switching bias, as defined in equation (3.2.1). The ratio of binding constants is varied in the range $0.5 < \tau_{\mathrm{B}}^+/\tau_{\mathrm{A}}^+ < 2$. Additionally the protein degradation rates are changed, which leads to different average protein copy numbers μ_x in S_{A} and S_{B}. Transcription and mRNA degradation are kept constant, resulting in an overall mean of 10 mRNAs in S_{A} and S_{B}. Results of the simulation are shown in Fig. 4.7.

We see similar results to the one-stage switch: Asymmetry tilts the switching decision into one direction, S_{A} gets more probable than S_{B} or vice versa (Fig. 4.7 A,B). This bias becomes stronger as the parameter asymmetry increases. Systems with higher protein numbers are less influenced and show less bias (Fig. 4.7 D). Evaluating the entropy of the systems shows that it is maximal for systems with symmetric parameters as before. However if the systems become asymmetric, we see a rapid drop of the entropy for systems with small protein numbers, whereas systems with increasing protein number tend towards constant entropy independent of the asymmetry (Fig. 4.7 E). The entropy describes the trade-off between switching bias and the system's ability to decide at all. Low protein number systems have the advantage that they decide very rapidly, but their decision is strongly biased. High protein number systems take longer to decide, but their decision is almost unbiased.

Fig. 4.8 A reveal that, given an asymmetry of $\tau_{\mathrm{B}}^{+}/\tau_{\mathrm{A}}^{+} = 2$, a protein number between 10^4 and 10^5 result in a maximally robust switch according to the entropy: Decisions do not take too long and the decisions are reasonable unbiased. Fig. 4.8 B shows that this maximum is conserved across different asymmetries in binding parameters.

How does the number of mRNAs influence the switch performance? As an extension we now not only vary the protein level μ by changing the protein degradation rates, but also manipulate the mRNA level by changing the mRNA degradation rates γ. The binding asymmetry was fixed at $\tau_{\mathrm{B}}^{+}/\tau_{\mathrm{A}}^{+} = 2$. The results are plotted in Fig. 4.9 and show that different mRNA levels have only moderate influence on the bias, the entropy curves are very similar showing maximal values at protein levels between 10^4 and 10^5.

4.2.2 mRNA synthesis and degradation

To further investigate the role of mRNAs in the system's dynamics, we apply an analogous approach by varying the mRNA degradation rate γ of both genes, which could for example be induced through microRNA, targeting the mRNA of one species, leading to increased degradation of the mRNA.

Different mRNA degradation rates will lead to different protein distributions of both species. Therefore, we investigate not only the influence of asymmetric mRNA degradation rates but also the influence of different protein mean levels. As in section 4.2.1, the protein degradation rates δ are changed symmetrically for both genes to obtain systems with different mean protein numbers μ.

The results shown in Fig. 4.10 look very similar to the results on the binding asymmetries from section 4.2.1. Asymmetries lead to biased switching (4.10 D), but the switching bias of systems with higher average protein levels μ_x is less influenced by the asymmetry.

Finally, the transcription rates α are varied, accounting for e.g. different promoter strengths. Fig. 4.11 shows a qualitatively similar picture as Fig. 4.10. Asymmetries give advantage to one of the genes, tilting the switch towards one regime. However, compared to the binding and mRNA degradation asymmetries, the influence of transcriptional asymmetries on the switching bias seems to be stronger. Consider for example the 10000-protein systems in Fig. 4.11 and Fig. 4.10. Whereas the entropy of the system is close to one for strong degradation asymmetries, the entropy is only 0.7 for strong transcriptional asymmetries. The system is therefore less robust to transcriptional asymmetries.

Nevertheless, we find a general trend that, although transcription, mRNA degradation and protein-DNA binding constants are assumed to act very differently on the systems dynamics, toggle switches with high protein amounts are less prone to asymmetries and can maintain their function more robustly.

4.2.3 Initial conditions

In a similar fashion we assess the asymmetry in initial conditions introduced through to transition from a coexpressed to a switching state as described in section 3.2.2.

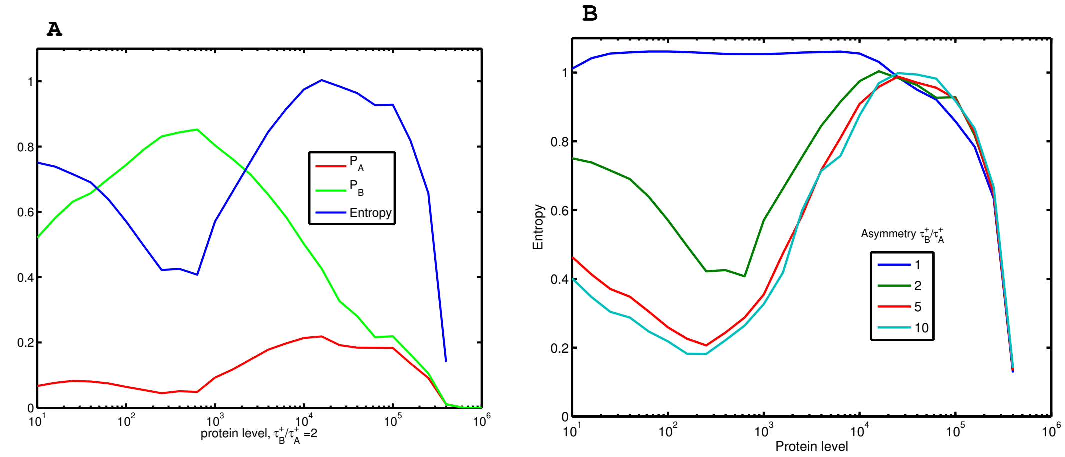

Figure 4.8: Entropy of the two-stage switch. A: Mean protein level μ plotted against P_A, P_B and the entropy at a two-fold binding asymmetry: $\tau^+_\mathrm{B}/\tau^+_\mathrm{A} = 2$. and various mRNA and protein levels. B: Mean protein level μ plotted against the entropy for various asymmetries $\tau^+_\mathrm{B}/\tau^+_\mathrm{A}$ The entropy shows a maximum at protein levels in the range of 10^4 to 10^5, which is conserved across all degrees of asymmetry. The switch is maximally robust in at these protein levels.

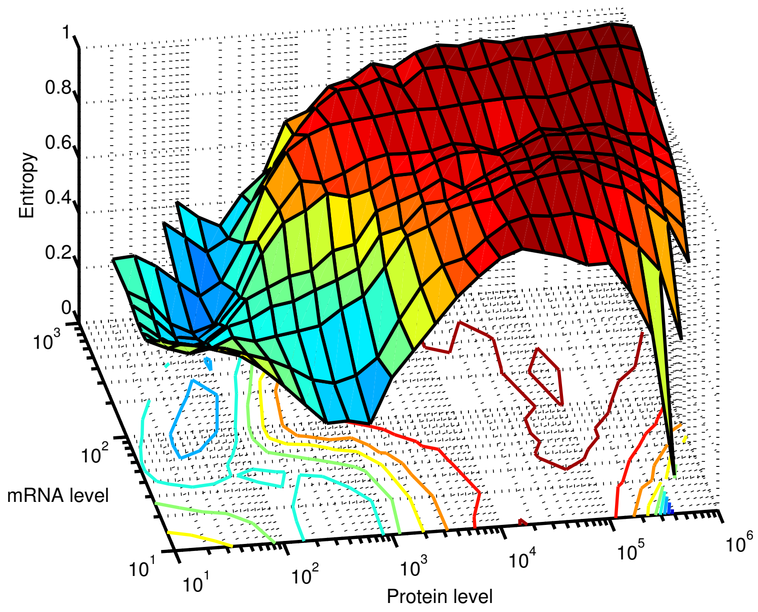

Figure 4.9: The entropy of the system is plotted against various protein levels μ and mRNA levels at a two-fold binding asymmetry ($\tau^+_\mathrm{B}/\tau^+_\mathrm{A} = 2$). Remaining parameters are set according to Table 2.1. The entropy maximum is conserved across mRNA levels.

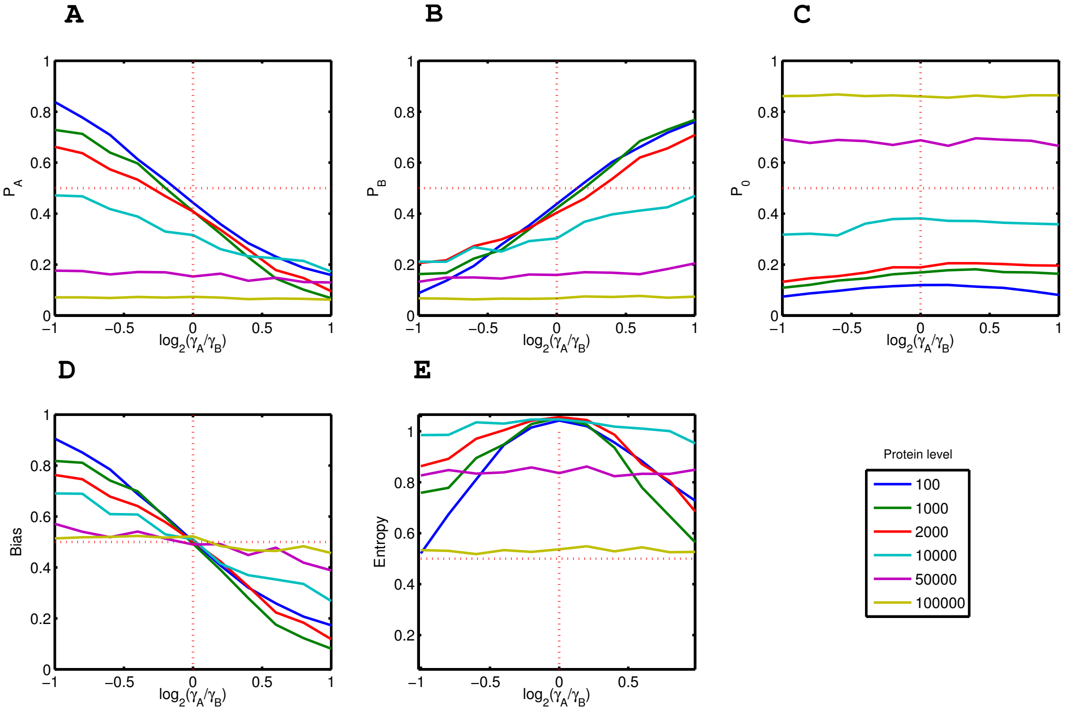

Figure 4.10: Robustness of the two-stage model with respect to mRNA degradation constants γ (α, β, τ^+ and τ^- were chosen from Table 2.1). This leads to mRNA levels in the range of 5 to 20. Estimates of P_A, P_B, P_0, the bias and the entropy are plotted against the mRNA degradation asymmetry for six different protein mean values μ. Systems with $\gamma_\mathrm{A} < \gamma_\mathrm{B}$ ($\log_2(\gamma_\mathrm{A}/\gamma_\mathrm{B}) < 0$) lead to preference of S_A, since lower mRNA degradation γ_A leads to an overall increased protein level of player A compared to player B: $\mu_\mathrm{A} > \mu_\mathrm{B}$, giving strong advantage to A. This advantage is less pronounced for systems with higher mean protein level μ.

Maximal asymmetries are estimated using the upper and lower bounds of the two-stage gene expression distribution (4.3) using a normal approximation:

$$\phi_x^+ = \mu_x + \sigma_x \cdot Z_{\alpha'}$$
$$\phi_x^- = \mu_x - \sigma_x \cdot Z_{\alpha'} \ .$$

$Z_{\alpha'}$ denotes the α'-quantil of the standard normal distribution. For example using $\alpha' = 0.9995$ assures that ϕ_x^+ and ϕ_x^- enclose 99.9% of the distribution's probability mass. Protein initial conditions pairs were chosen within these bounds:

$$\begin{aligned} p_{\mathrm{A}0} &= \mu_\mathrm{A} + i \cdot (\sigma_\mathrm{A} \cdot Z_{\alpha'}) \\ p_{\mathrm{B}0} &= \mu_\mathrm{B} - i \cdot (\sigma_\mathrm{B} \cdot Z_{\alpha'}) \ , \end{aligned} \tag{4.7}$$

$i \in [0, 1]$, where $i = 0$ gives perfectly symmetric initial conditions, $i = 1$ leads to maximal asymmetric initial conditions. mRNA initial conditions were adjusted accordingly:

$$m_{\mathrm{A}0} = p_{\mathrm{A}0} \frac{\delta_\mathrm{A}}{\beta_\mathrm{A}}$$
$$m_{\mathrm{B}0} = p_{\mathrm{B}0} \frac{\delta_\mathrm{B}}{\beta_\mathrm{B}} \ .$$

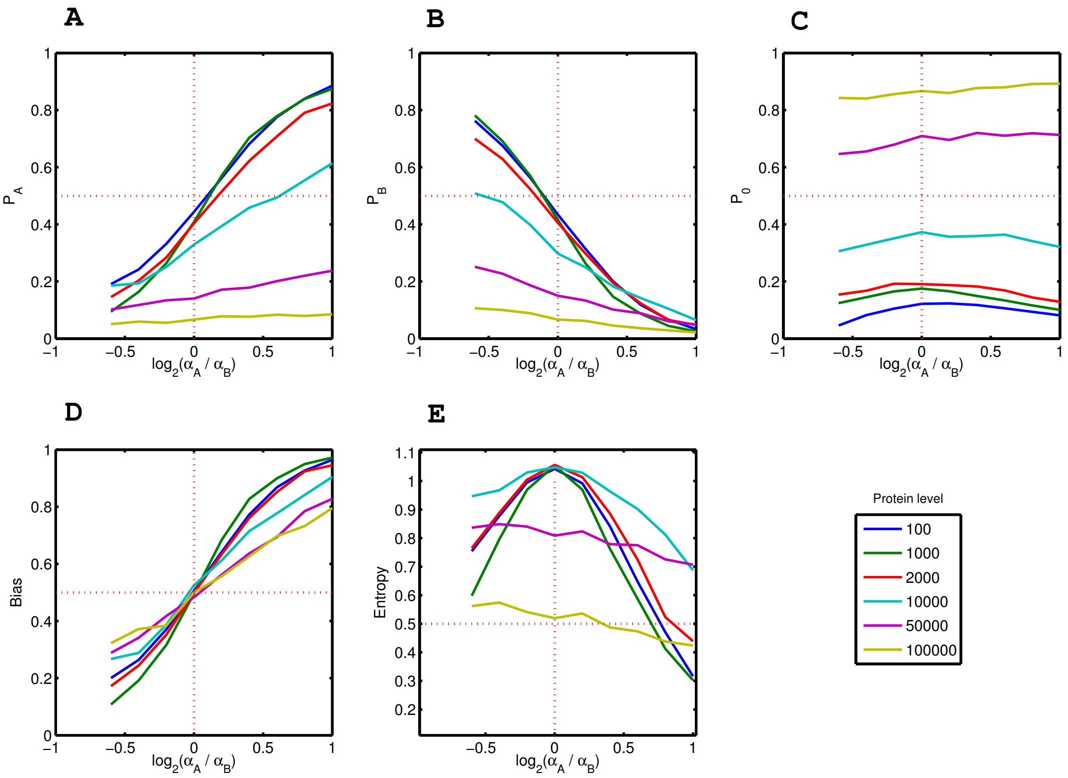

Figure 4.11: Robustness of the two-stage model with respect to mRNA synthesis constants α. Estimates of P_{A}, P_{B}, P_0, the bias and the entropy are plotted against the transcription asymmetry for six different protein mean values μ. Systems with $\alpha_{\mathrm{A}} > \alpha_{\mathrm{B}}$ ($\log_2(\gamma_{\mathrm{A}}/\gamma_{\mathrm{B}}) > 0$) lead to preference of S_{A}, since higher mRNA synthesis α_{A} leads to an overall increased protein level of player A compared to player B: $\mu_{\mathrm{A}} > \mu_{\mathrm{B}}$, giving strong advantage to P. This advantage is less pronounced for systems with higher mean protein level μ. Estimates for $\log_2(\gamma_{\mathrm{A}}/\gamma_{\mathrm{B}}) < -0.5$ are missing, since these parameters decrease the mean protein level μ to ranges where S_{A} and S_0 cannot be discriminated any more.

We perform a analysis of the bias with respect to the first decision in the system (Fig. 4.12) analogous to the one-stage system (see section 3.2.2). The bias regarding the first regime choice is almost independent of the asymmetry in initial conditions for the same reasons as in the one-stage model: The system comes into a deadlock phase immediately, where mRNA and proteins are degraded almost completely and the former advantage of one player is lost. We only see slight deviations from an unbiased decision at strong asymmetries. Strong initial asymmetry might cause that one player is already completely degraded, while the other is still present in significant amounts, driving the system into the regime of this player. However these influences are quite small compared to the asymmetries of binding parameters and mRNA degradation.

Systems with high protein number μ are less influenced than systems with small protein numbers. This stems from the fact that in systems with lower protein levels μ the fold difference $\frac{\phi^+}{\phi^-}$ between opposing edges of the distribution is larger than in systems with high protein level. This fold difference is plotted in Fig. 4.13, revealing that for the 100-protein system, choosing initial conditions from the edges of the distribution results in a 3.5-fold difference, whereas for a 10000-protein system it is ≈ 1.2. Therefore the asymmetric conditions are effectively stronger in low copy number systems, leading to a stronger bias.

Additionally, the protein distribution of the two-stage switch is much broader than the distribution of the one-stage model (due to the noise from the mRNA expression). Therefore the fold difference of initial conditions is much larger in the two-stage system, leading to noticeable bias.

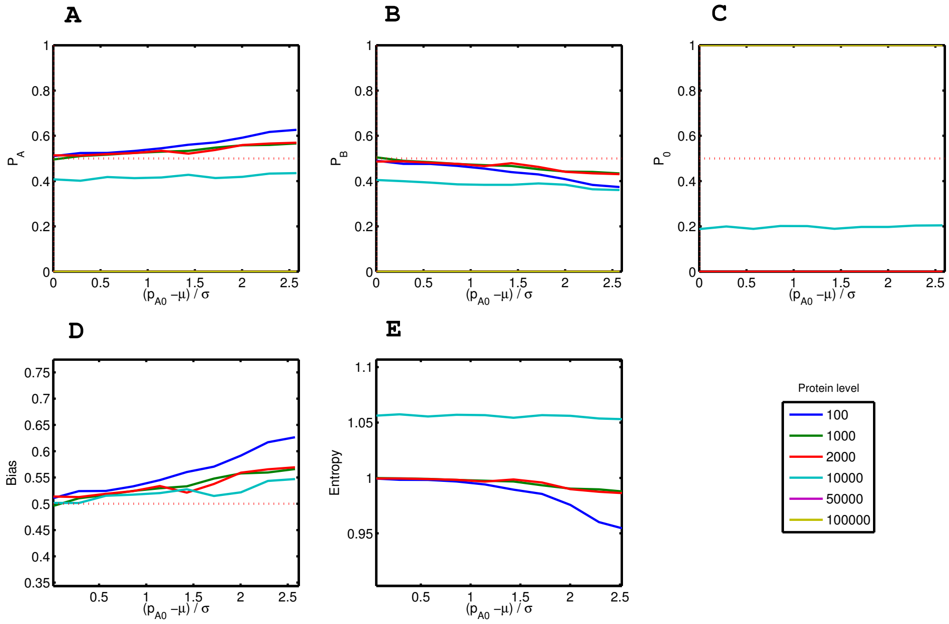

Figure 4.12: Robustness of the two-stage model with respect to switching bias and initial conditions. The initial conditions of the systems were chosen according to equations (4.7). On the x-axis the distance of the initial number $\text{Protein}_\text{A}/\text{Protein}_\text{B}$ from the distribution mean is given in standard deviations. For systems with distance 0, initial conditions for both proteins were chosen from the center of the distribution, resulting in equal amounts. System with large distance have initial conditions chosen from the opposing edges of the distribution.

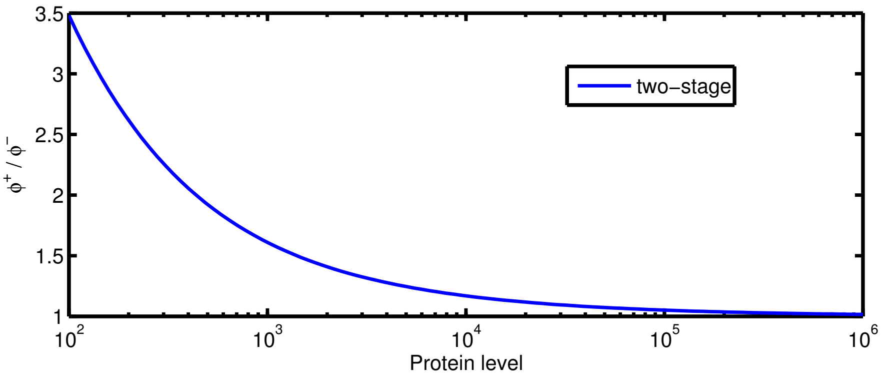

Figure 4.13: Fold difference ϕ^+/ϕ^- between opposing quantils of the protein-distribution plotted against the mean protein level μ of the system, showing that the fold difference decreases rapidly for larger protein levels μ.

Chapter 5

Discussion

In the previous chapters we have analyzed the dynamics of the probabilistic one- and two-stage toggle switch. In this chapter we discuss the results.

5.1 Comparison of models

We start by a general comparison of deterministic and probabilistic models to emphasize the importance of stochastic modeling of regulatory motifs.

5.1.1 Deterministic and probabilistic models

In this work, we presented a probabilistic description of the toggle switch, based on the fact that the system involves low molecule numbers, introducing an inherent stochastic component into the system. To evaluate features like the switching time or the bias, we simulated many timecourses of probabilistic model, extract the feature of interest and in the end take the average value of the feature. One could argue that this averaging in fact reduces to simply considering the deterministic case.

Despite the fact that the ODE solution cannot capture the bistability of the system (Fig. 3.3 and 5.1 B), we want to emphasize that the solution of the ODE does not necessarily – that is, only in case of linear propensities – resemble the mean trajectories of species abundances obtained by many stochastic simulation runs. Fig. 5.1 shows this for the one-stage toggle switch. Two discrepancies occur: Due to the uncertainty of the first regime choice in the probabilistic model, the mean peaks early at ≈ 20. This peak is not captured by the ODE solution (Fig. 5.1 A).

Even more, the steady state of the ODE solution differs from the mean of the stochastic simulations: In the stochastic simulation the regimes are changed frequently after the initial peak. The intermediate S_0 also gains probability. The mean drops and stabilizes at a value of ≈ 14, representing a balance between S_{A}, S_{B} and S_0. The ODE solution shows much simpler dynamics (Fig. 5.1 A), starting from the origin and leading rapidly into a steady state at around 3, which corresponds to S_0 (the deadlock regime).

However, if we use the median instead of the mean as a summary statistic on the stochastic simulations, we see that the ODE solutions resembles the median of the stochastic simulation trajectories.

For small molecular abundances stochastic differential equations also fail (data not shown), since species numbers are quickly driven below 0. Also, mass conservation is not guaranteed ad hoc and needs to be incorporated into the SDE, e.g. by correlated noise for players that are interconnected via mass conservation.

5.1.2 Existing models

Loinger et al. [33] studied the one-stage toggle switch and several modifications of the motif comparing the deterministic with the probabilistic approach. The modifications include the exclusive switch (both players share one promoter so either A or B can be repressed but not both at once), a switch with bound repressor degradation (inhibitors can be degraded even if bound to DNA), a switch where both proteins form an inactive heterodimer and a switch with cooperative binding of the repressors. The main focus of this work was on the bistability of the systems. In the deterministic case bistability was defined as the existence of two positive steady states. Loinger et al. found that both the normal toggle switch as well as the exclusive toggle switch do not exhibit bistability if modeled deterministically. The switches including bound repressor degradation, heterodimers or cooperative binding are bistable in a certain parameter range (characterized by strong repression). In the probabilistic case the regimes S_{A}, S_{B} and S_0 of the switch were defined by a static boundary $\phi = 2$ (see section 3.1.3 for our definition of the regime boundaries). The system was defined as bistable if $P_0 < 0.01$. The probability distribution of systems that fulfill this criterion exhibits two distinct peaks where either A or B dominates but does not have a peak at the deadlock regime. Loinger et al. used numerical integration and Monte Carlo methods to approximate the solution of the CME and found that the normal toggle switch does not fulfill the bistability criterion due to a large probability of the deadlock regime – high P_0 is also observed in our analysis (see e.g. Fig. 3.10 C). The exclusive switch as well as switches with bound repressor degradation, heterodimer formation or cooperative binding are bistable in a limited range of parameters. Note that here the idea of a functioning switch is quite different from ours. We allow for $P_0 > 0.01$ and assess the functionality of the system as switch by requiring long switching and moderate priming times. Loinger et al. also gave an analytical expression of the switching time for the exclusive switch. As we did not consider this modification of the switch a direct comparison cannot be made. However, we found that the switching time is actually geometrically distributed. It is likely that the switching time of the exclusive switch also obeys a geometric distribution, as the switching process is based on the same principles (unbinding of the repressor).

Schultz et al. [54] solved the CME of a one-stage toggle switch (with and without dimerisation) analytically using linear noise and fast transition approximations. They introduced three characteristic parameters, namely repression strength, number of free proteins and an adiabaticity parameter, relating the timescale of gene expression to the timescale of protein-DNA interaction. They investigated how the protein distribution depends on these parameters. Different parameters lead to a different number of peaks in the distribution. For example in the range of parameters used in our work (large adiabaticity, moderately strong repression) the distribution exhibits three peaks, which is in agreement with our findings. Similar to Loinger et al. [33] the criterion for functioning switches was defined as $S_{\text{A}} + S_{\text{B}} \approx 1$ and it was shown that strong repression is necessary to obtain functioning switches. However strong repression leads to high probability for the deadlock regime. Schultz et al. suggested that introducing cooperative binding significantly decreases the deadlock probability and further enhances the switching property.

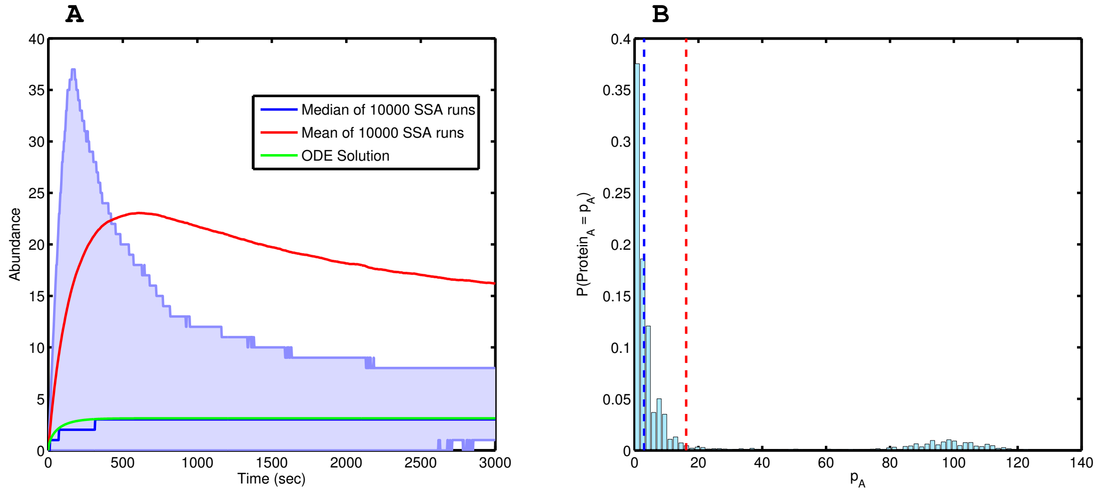

Figure 5.1: Comparison of the ODE solution and simulated trajectories for Protein$_\mathrm{A}$ in the one-stage toggle switch. A: The mean (red) and median (blue) trajectory of Protein$_\mathrm{A}$ obtained by 10000 stochastic simulation and the trajectory of the corresponding ODE (green) are shown. 25% and 75% quantils of the simulations are indicated by shaded area. A clear discrepancy between the ODE solution and the average of the stochastic simulations in terms of transient dynamics and steady state is visible. However the median of the trajectories is very close to the ODE solution. B: Distribution of the amount of Protein$_\mathrm{A}$ at $t = 3000$ sec. The mean (red dashed line) and median (blue dashed line) of the distribution are also indicated.

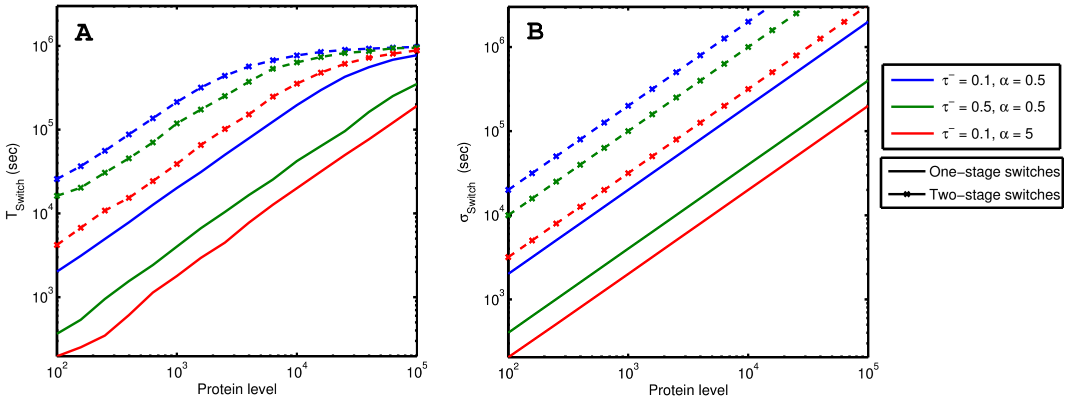

Figure 5.2: Comparison of mean switching time between model classes. Parameters are chosen from Table 2.1. One-stage switch: solid lines, two-stage switch: dashed lines. A: Simulated estimates. B: Analytical estimates. Two-stage models have an overall higher switching time than one-stage models.

Sasai and Wolynes [53] formulated the toggle switch as a quantum many-body problem, solved the CME of the toggle switch (with dimerizing repressors) using the Hartree approximation and determined the parameter range (in terms of repression strength, number of free proteins and an adiabaticity parameter) where the system exhibits two stable states. Sasai and Wolynes found that the system exhibits two stable states if the adiabaticity parameter and the number of free proteins is large, which reflects our finding that large protein numbers reduce switching between S_{A} and S_{B}.

However, to our knowledge no publications exist that investigate the role of mRNA and the influence of asymmetric parameters.

5.1.3 Transition times

In chapters 3 and 4 we defined the transition times as the average times the system needs to change its regime. The switching time (time for a transition from S_{A} and S_{B} to S_0) quantifies how long the toggle switch can memorize its decision. The priming time (time for a transition from S_0 to S_{A} or S_{B}) quantifies how long the toggle switch needs to make a decision. Before, we have shown the results on the switching times of one- and two-stage switches independently. Here, we compare the transition times between equivalent one- and two-stage switches. We call a one-stage switch with mean protein number $\mu^{(1)}(=\mu_{\mathrm{A}}^{(1)}=\mu_{\mathrm{B}}^{(1)})$ equivalent to a two-stage switch with mean protein number $\mu^{(2)}$ if $\mu^{(1)}=\mu^{(2)}$ and both models share the same binding/dissociation constants.

We first compare the mean switching times T_{Switch} of the different models. We find that for different sets of parameters the one-stage switch has smaller switching times than two-stage switches (Fig. 5.2). This means that one-stage switches loose the information on the previous decision earlier and change their regime sooner. The reason lies within the condensed transcription/translation reaction of the one-stage switch. It has a much higher rate compared to the transcription reaction of the two-stage switch to fulfill the equivalence condition $\mu^{(1)}=\mu^{(2)}$. For example a two-stage system with $\alpha=\beta=0.05,\ \gamma=\delta=0.005$ results in a mean of 10 mRNAs and 100 proteins. An equivalent one-stage system has $\alpha=0.5, \delta=0.005$. The synthesis rate α must be 10 fold higher compared to the two-stage system as it must account for the 10 mRNA molecules present in the two-stage system, which increase the effective rate of translation. This leads to higher probability of synthesis during the unbound phase in the one-stage model, ultimately resulting in shorter switching times. Contrary, in the two stage model the probability of mRNA synthesis is smaller, but once an mRNA has been synthesized the binding probability is much higher as not only one but several proteins are translated from this mRNA leading to increased binding propensity. However, the binding probability is already very close to 1 (see equation (4.1.4)), resulting in only light influence of the increased binding propensity on the binding probability.

Additionally we see that, even if the synthesis rates α are equal (solid blue and dashed red line in Fig. 5.2), the two-stage switches have longer switching times. Due to the additionally needed translation process in the two-stage systems, it can happen that even though an mRNA was synthesized, it is degraded before translation and no switching occurs. Comparing the priming times in Fig. 5.3 reveals the opposite trend. Two-stage systems spend less time in the primed state (S_0) and are driven faster into S_{A} or S_{B} than their corresponding one-stage systems. This is explained by the fact that two-stage systems enter the deadlock situation less often. Being in a state where both promoters are unbound, for the two-stage system it is very unlikely to run into a deadlock: Both genes must almost simultaneously produce mRNA

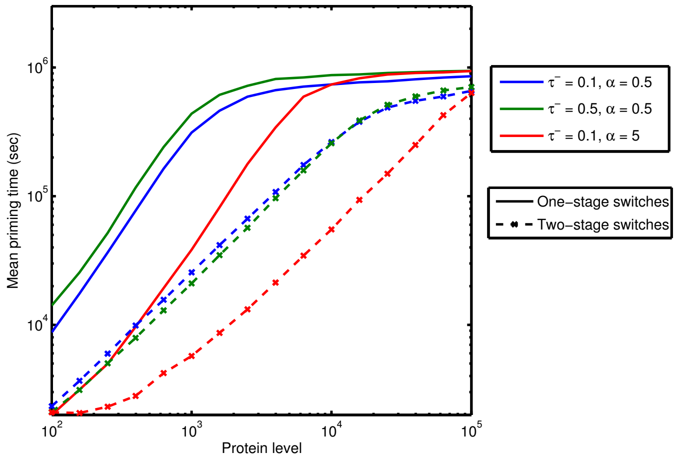

Figure 5.3: Comparison of mean priming time between model classes. Parameters are selected from Table 2.1. Two-stage switches (dashed lines) show decreased priming time compared to one-stage switches (solid line). One-stage models have high probability of entering the deadlock regime, since only the synthesis reaction for both genes has to occur simultaneously. Two-stage models have low probability of entering the deadlock regime due to the two required steps of gene expression.

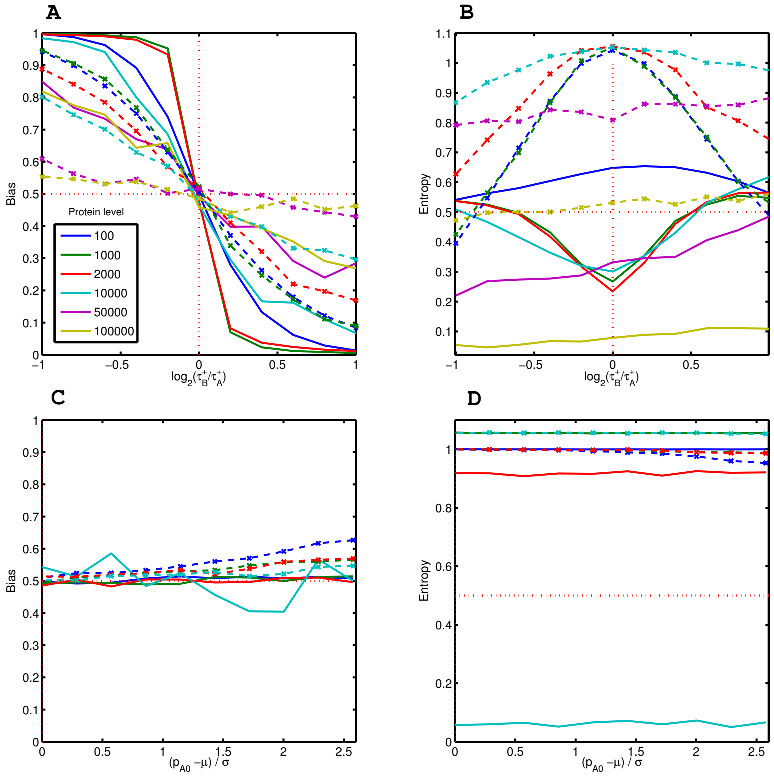

Figure 5.4: Comparison of the switching bias between model classes. One-stage switch: solid lines, two-stage switch: dashed lines. A: Bias for asymmetric binding constants τ^+. B: Entropy for asymmetric binding constants τ^+. C: Bias for asymmetric initial conditions. D: Entropy for asymmetrical initial conditions. Overall, two-stage switches are more robust to asymmetries in binding parameters, but are less robust to asymmetries in initial conditions.

before their promoters are inhibited and both mRNAs must be translated into proteins. In one-stage switches only the simultaneous protein synthesis is required, being more likely.

5.1.4 Bias/Robustness

In chapters 3 and 4 we estimated the switching bias for the one- and two-stage systems introducing different kinds of asymmetries into the system. It turned out that systems with higher average protein levels are less biased regarding the switching behavior. Now we compare the equivalent one- and two-stage switches with respect to the switching bias. Fig. 5.4 A shows that one-stage systems are more biased than the corresponding two-stage systems for asymmetries in binding parameters. The two-stage systems are more sensitive for asymmetries in initial conditions than one-stage systems (Fig. 5.4 C). Considering the entropy, we clearly see a higher entropy of the two-stage systems over the whole range of asymmetries in binding parameters (Fig. 5.4 B), whereas little difference is observed in the entropies regarding initial condition asymmetries (Fig. 5.4 D).

5.1.5 Autoactivation

In the previous chapters we built up and examined models of a regulatory motif consisting of two genes mutually repressing each other. A modification of this motif, where additionally both genes can activate their own expression, occurs frequently in regulatory networks of differentiating cells (for example in the PU.1/GATA-1 mutual inhibition in myloid differentiation [43]). In this regulatory motif, both players can bind to their own promoters, stimulating their own expression and bind to the antagonistic promoters, repressing the other gene.

In this section we evaluate how the additional self-activation acts on the system's dynamics. We assume that a gene can either be bound by the activator or the repressor, but not both at the same time. This corresponds to a close distance between both binding sites on the DNA, leading to steric hindrance. Allowing for simultaneous binding of repressor and inhibitor is possible in general, but transcriptional potency of this double bound promoter is unclear. Including auto-activation into our model is straight forward: We introduce two additional species, representing activator-bound promoter (one for each player), which result from an reaction between activator protein and an unbound promoter. The activator can dissociate, restoring the unbound promoter. If neither repressor nor activator is bound to the promoter, the gene will be transcribed with basal rate κ. Repressors bound to a promoter block transcription completely. Promoters with bound activator lead to full transcription of the gene with rate $\alpha \gg \kappa$. A scheme of the two-stage model extended by self-activation is shown in Fig. 5.5 and the biochemical reactions describing this system are listed in Table 5.1.

The selfactivation has a strong influence on the switching time. Switches with autoactivation show strongly increased switching times compared to the non-autoregulatory switches (Fig. 5.6). Therefore, switches with autoactivation can memorize their decision much longer. On the one hand, even if the repressed promoter is unbound for a short time, due to the low basal transcription rate, synthesis of an mRNA is very unlikely. On the other hand, the dominating species will almost all the time occupy its own promoter, prohibiting a eventually synthesized repressor from binding. Even if a repressor was able to bind the promoter it will quickly be displaced by an activator again. Both facts lead to strongly increased switching time as depicted in Fig. 5.6. Even for high degradation rates/low protein numbers the mean switching time is close to the upper bound of the simulated time. It can be expected that the

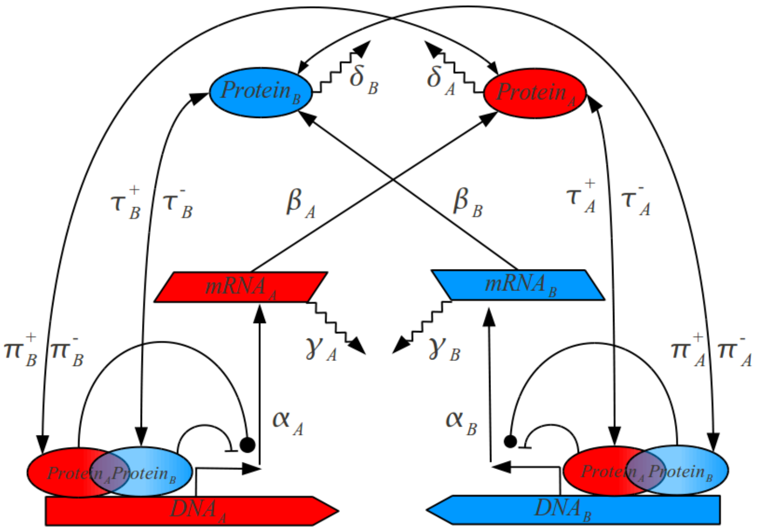

Figure 5.5: Scheme of the two-stage toggle switch with autoactivation. Species associated with player A are shown in red, species associated with player B are shown in blue. Reactions/Interactions are indicated as arrows, jagged arrows indicate degradation reactions. The two-stage model of Fig 4.1 is extended by the binding (unbinding) of the proteins to their own promoter with rate π^+ (π^-). Autoactivation and inhibition are mutually exclusive – the promoter is either bound by the activator protein or by the repressor protein. Depending on the promoter state, the transcription rate is basal (κ), full (α) or completely inhibited.

$$
\begin{aligned}
\mathrm{DNA_B} &\xrightarrow{\kappa_\mathrm{B}} \mathrm{DNA_B} + \mathrm{mRNA_B} & \qquad \mathrm{DNA_A} &\xrightarrow{\kappa_\mathrm{A}} \mathrm{DNA_A} + \mathrm{mRNA_A} \\
\mathrm{DNA_B^{act}} &\xrightarrow{\alpha_\mathrm{B}} \mathrm{DNA_B^{act}} + \mathrm{mRNA_B} & \qquad \mathrm{DNA_A^{act}} &\xrightarrow{\alpha_\mathrm{A}} \mathrm{DNA_A^{act}} + \mathrm{mRNA_A} \\
\mathrm{mRNA_B} &\xrightarrow{\gamma_\mathrm{B}} \emptyset & \qquad \mathrm{mRNA_A} &\xrightarrow{\gamma_\mathrm{A}} \emptyset \\
\mathrm{mRNA_B} &\xrightarrow{\beta_\mathrm{B}} \mathrm{mRNA_B} + \mathrm{Protein_B} & \qquad \mathrm{mRNA_A} &\xrightarrow{\beta_\mathrm{A}} \mathrm{mRNA_A} + \mathrm{G} \\
\mathrm{Protein_B} &\xrightarrow{\delta_\mathrm{B}} \emptyset & \qquad \mathrm{Protein_A} &\xrightarrow{\delta_\mathrm{A}} \emptyset \\
\mathrm{Protein_B} + \mathrm{DNA_A} &\xrightarrow{\tau^+_\mathrm{B}} \mathrm{DNA_A^{inh}} & \qquad \mathrm{Protein_A} + \mathrm{DNA_B} &\xrightarrow{\tau^+_\mathrm{A}} \mathrm{DNA_B^{inh}} \\
\mathrm{DNA_A^{inh}} &\xrightarrow{\tau^-_\mathrm{B}} \mathrm{Protein_B} + \mathrm{DNA_A} & \qquad \mathrm{DNA_B^{inh}} &\xrightarrow{\tau^-_\mathrm{A}} \mathrm{P} + \mathrm{DNA_B} \\
\mathrm{Protein_B} + \mathrm{DNA_B} &\xrightarrow{\pi^+_\mathrm{B}} \mathrm{DNA_B^{act}} & \qquad \mathrm{Protein_A} + \mathrm{DNA_A} &\xrightarrow{\pi^+_\mathrm{A}} \mathrm{DNA_A^{act}} \\
\mathrm{DNA_B^{act}} &\xrightarrow{\pi^-_\mathrm{B}} \mathrm{Protein_B} + \mathrm{DNA_B} & \qquad \mathrm{DNA_A^{act}} &\xrightarrow{\pi^-_\mathrm{A}} \mathrm{Protein_A} + \mathrm{DNA_A}
\end{aligned}
$$

Table 5.1: List of reactions for the two-stage toggle switch including self-activation. The former two-stage model is extended by three reactions for each species, representing the protein binding its own promoter, the according dissociation and the full transcription of the activated promoter. The self-bound promoter DNA$^{\mathrm{act}}$ has increased transcriptional rate α, the unbound promoter has basal transcription rate κ.

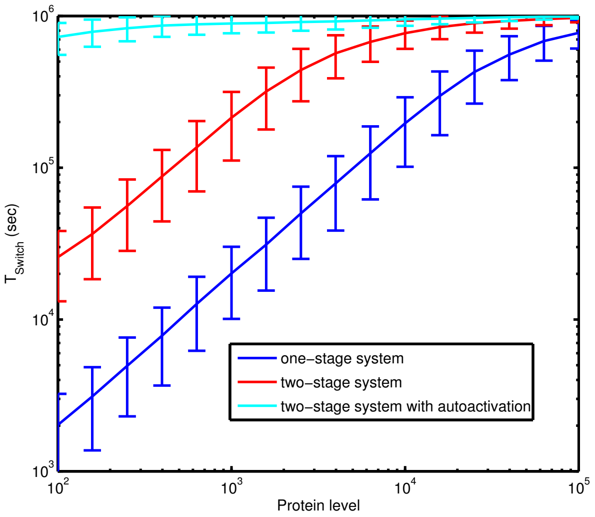

Figure 5.6: Simulated mean switching times of the autoregulatory two-stage system compared to equivalent systems without autoactivation. Errorbars indicate standard deviations σ_{Switch} of the estimates. Switches with autoactivation have strongly increased switching time. The dominating protein protects its own promoter from inhibition by self-binding. Because of the low basal transcription rate, successful synthesis of the repressed protein is very unlikely, also increasing the switching time. Due to the maximal simulation time of 10^6 seconds, estimates close to this boundary are biased and assumed to be even higher.

true mean switching time is much higher than the simulated estimates (due to the maximal simulation time of 10^6 sec). Hence, self-activation can stabilize the system's regime choice by strongly increasing the time until the systems forgets a previously made choice. Moreover autoactivation increases the system's robustness against parameter asymmetries as shown in Fig. 5.7. Whereas we observed strong bias in the non-autoregulatory systems by introducing asymmetric binding parameters, here we see only a weak bias of the systems (Fig. 5.7 A). Two-state systems with $\mu = 10000$ for example showed considerable bias in Fig. 4.7 C, whereas here its bias is almost constantly around 0.5. Also the entropy is quite constant for systems except for $\mu = 100$ (Fig. 5.7 B). These systems are still influenced by the asymmetries.

Considering asymmetric initial conditions shows the inverse trend: Here autoregulatory systems exhibit a stronger bias than non-autoregulatory systems (Fig. 5.7 C,D). This stems from the fact that, through autoactivation the system can escape the deadlock situation and the complete degradation of both players. Through binding its own promoter one of the two species can rescue themselves from degradation, moving the system into that player's regime. This active intervention depends on the amounts of proteins present, since the propensity of the binding reactions is linear in the number of proteins. If e.g. there are more proteins of P than proteins of G, the probability is higher that P will bind its own promoter and drive the system into S_{A}. This explains the increased bias for autoregulatory systems.

5.2 Biological interpretation

In this work, we studied a stochastic model of the toggle switch and analyzed the system's features and dynamics. What do these results imply in the context of biology, especially when

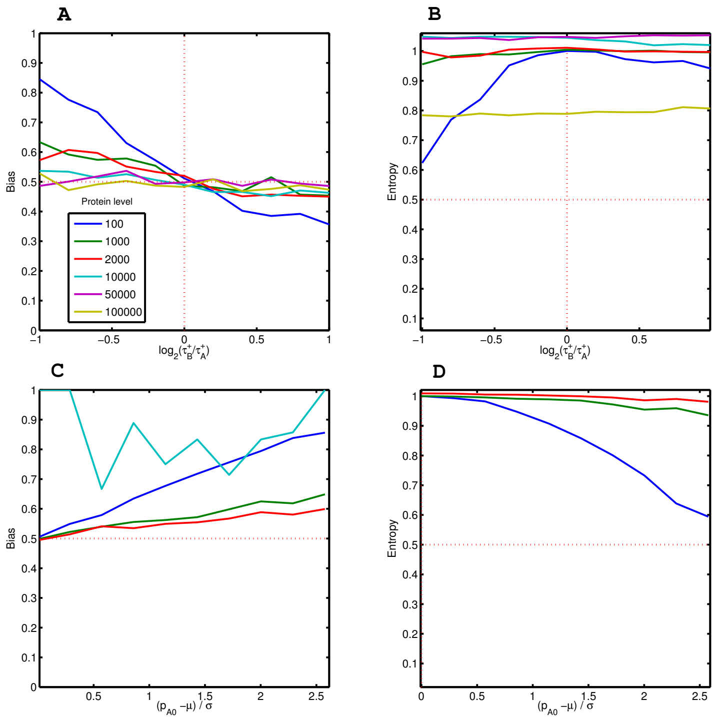

Figure 5.7: Robustness of the autoregulatory two-stage switch. A,B: Bias and entropy with respect to binding constants τ^+. Autoregulatory switches are robust to asymmetries in binding constants, resulting in a bias close to 0.5 and entropies close to 1. C,D: Bias and entropy with respect to initial conditions. Autoregulatory systems are less robust to asymmetries in initial conditions. Higher protein levels μ increase the robustness. Strong fluctuations of the 10000-protein switch in C results from the high probability of the system in S_0, similar to Fig. 3.11.

we consider cell lineage differentiation? In this section we give a biological interpretation of the main results.

5.2.1 Transition times

We found that the switching time – the system's waiting time in the regime dominated by P (S_A) or G (S_B) – strongly depends on the average protein level μ during theses regimes. The more proteins are present, the longer the system will be locked in S_A or S_B, independent of the chosen model. The toggle switch is thought to act as a memory unit in cells. If a differentiating cell uses the toggle switch for lineage decision (S_A corresponds to one lineage, S_B to the other), the toggle switch has to maintain its regime for a long time, until the differentiating cell has completely changed into the new lineage. Therefore the protein levels involved in the toggle switch have to be high: For example, if we require that the mean switching time should be 5 days ($4 \cdot 10^5$ seconds, chosen arbitrarily), from Fig. 4.5 A we find, depending on parameters, a mean protein level of 10^3 to 10^4.

However, we showed in sections 3.1.5 and 4.1.4 that the switching time is geometrically distributed. Recall that the probability density function of the geometric distribution is monotonically decreasing. Therefore, short switching times are always more probable than longer switching times, regardless of the mean switching time, which is not the most probable value of the distribution. Considering the toggle switch in a differentiating cell, this aspect of the switching time is not favorable, since even if the mean switching time is high, a large fraction of the differentiating cell population will quickly loose the previous lineage decision.

Instead, one should require that, e.g. 95% of the population must have a switching time longer than five days. This corresponds to the claim that $1 - F_\lambda(5 \text{ days}) > 0.95$, where F_λ is the cumulative distribution function of the geometric distribution with parameter λ, which depends on the parameters of the system (see equation (3.9)). To fulfill this, the system must have a much higher protein level, strongly decreasing λ (see equation (4.5) and (4.6)). Now the geometric distribution flattens, getting more similar to a uniform distribution. The protein levels necessary to establish that 95% of the population have a switching time longer than 5 days is thereby estimated to be between 10^4 to 10^5, which is one order of magnitude larger than the value obtained by using the mean switching time. To reduce this number, autoactivation can be used to further stabilize the regimes (shown in section 5.1.5), leading to a protein number of 10^3 to 10^4.

Considering the priming time – the system's waiting time in S_0 – we find also a strong dependence on the protein level/degradation rates. Interpreting this quantity is more difficult than for the switching time and depends on the idea of how cells are instructed to commitment.

1. If we assume no external signal that forces the cell to commit (the signal would only tell the cell to start commitment, but not into which direction), the priming time could be seen as the time during which the cell can proliferate. Due to stochastic fluctuations the cell is driven out of this proliferation state and into S_A or S_B, committing itself. The priming time of the system would control the ratio between cells that are proliferating (not committing during one cell cycle) and cells that commit to one lineage.

2. Assuming an external signal that instructs the cell to commit (but not to what lineage; see section 4.2.3), the priming time does not have a biological function. Here, the priming time is simply the phase where the system is in the deadlock regime. Then, the system eventually escapes the deadlock, committing towards either S_A or S_B. The priming time becomes a limiting factor here, since it can become so large that cells will not commit during lifetime (consider for example the 50000 protein system in Fig. 4.12, which has $P_0 = 1$). Here, a balance between switching time, which should be as high as possible, and priming time, which should be as small as possible, must be found. Changing the molecular interactions in our model, for example the inclusion of heterodimer formation, could reduce the priming time [33], granting long switching through high protein levels, but reducing the otherwise long priming time.

Comparing the switching and priming times between one- and two-stage models shows that two-stage systems have longer switching times and shorter priming times than corresponding one-stage models. This reveals an interesting effect of the mRNA stage in a cell. This partitioning of gene expression into two separate slow processes (instead of one fast synthesis process) helps to stabilize the switch, enabling a differentiation cell to maintain its lineage choice longer and furthermore to commit faster towards one lineage.

5.2.2 Robustness

The second major result of this work is on the robustness of the system to parameter asymmetries. We evaluated the influence of asymmetries on the system's switching bias, that quantifies if the system's shows preference for one committed regime over the other. We found that toggle switches with high protein levels μ are less influenced by the asymmetries resulting in a smaller switching bias. These systems robustly maintain their function as an unbiased decision maker. This results from the fact that the first decision in the switch is always unbiased. Toggle switches with high protein levels have longer switching times and therefore can preserve this unbiased decision.

Considering the toggle switch in differentiating cells, an unbiased switching is favorable. Otherwise, the cell will not be able to commit to both lineages with equal probability. The toggle switch will decide unbiased if the system is completely symmetric, which is intuitive. However, parameters that are involved in DNA-protein interactions are likely to be asymmetric, due to the variety of protein-DNA interaction mechanisms, consisting of many different combinations of DNA-sequences and binding domains. Additionally we showed that the initial conditions can be asymmetric due to a transition from coexpression to mutual inhibition (see sections 3.2.2 and 4.2.3). In order to function robustly and to compensate for this inherent asymmetries a differentiating cell can utilize high protein levels that will keep the decisions unbiased. Considering Fig. 4.9 A, we find the number of proteins leading to an maximal robust system between $10^4 < \mu < 10^5$ for a mean mRNA level of 10.

Both the analysis of transitions times and the robustness show that protein levels in the range of 10^4 to 10^5 (corresponding protein half lives: 3.8 hours to 38 hours) are advantageous in a cellular context. This surprisingly high number is in good agreement with measurements of PU.1 abundance in differentiating common myeloid progenitors (experiments done by Timm Schroeder).

Last, we point out how the bias of the switch can be modulated by the cell. An unbiased decision might not be favorable for an organism in extreme situations, e.g. physical strain. Here, the organism requires special cell types whereas others are not needed. Considering e.g. endurance sports, over a longer time the organism will increase the amount of red blood cells to account for the increased need for oxygen supply. However, megakaryocyte/erythroid progenitors give rise to both red blood cells and megakaryocytes [44]. This lineage decision is thought to be maintained by a toggle switch including the genes EKLF and Fli. Given an unbiased switch, increasing the amount of red blood cells will simultaneously increase the amount of megakaryocytes. However, in our model of the toggle switch this will create an excess of megakaryocyte, which are not needed to cope with the need for increased oxygen supply. There are two theories on how the control the ratio between different progeny [11, 48]:

1. Permissive control: Cells differentiate unbiasedly and the ratio between different progeny is adjusted by modulating the survival and proliferation rates of progeny. The organism would increase the survival rate for red blood cells and induce apoptosis in megakaryocytes.

2. Instructive control: The bias the toggle switch in the progenitor itself is changed. Progenitors will more likely give rise to red blood cells then to megakaryocytes.

   However, this system is quite robust and will hardly be influenced if e.g. binding rates are changed towards one lineage. To tilt the switch into one direction, our model suggests that an asymmetry in the mRNA synthesis α will have the strongest influence on the switching bias (see section 4.2.2). Such asymmetry can be reversibly introduced by modulating e.g the rate of mRNA export from the nucleus. Thereby the overall time until an mRNA is ready for translation is changed. This mRNA export from the nucleus is highly specific and mediated by the nuclear pore complex [1], which recognizes signals (e.g bound proteins) on the mRNA and regulates the export to the cytoplasm through the nuclear pores.

5.3 Outlook

During this work we considered a purely theoretical toggle switch model to get an impression of the systems dynamics and features and to relate these results to reality. However, until now we have not considered a major feature of living cells, namely cell growth and division. As toggle switches are thought to control lineage commitment in differentiating cells where everything evolves around growth and division, including both is necessary to get a better understanding of the involved mechanisms. Cell growth is expected to influence the system's dynamics, because increasing cellular volume will increase dilution, reducing the effective reaction constants. On the other side, during the cell cycle DNA will be replicated and at some time two copies of both genes are available, potentially increasing the transcriptional rates, as two templates are present. Cell division will also disturb the state of the cell by redistributing all present mRNA and protein to two daughter cells, restarting the system with altered initial conditions. Even the promoter configuration will be reset, since DNA-bound proteins are removed during replication. Including cell growth and cell cycle creates additional noise in the system – e.g. due to asymmetric cell division – and helps to get a more realistic idea of the cellular dynamics in differentiation.

To ultimately compare the results found during this work with measured protein levels, one could try to further modify the model to better resemble experimental findings. If we consider the PU.1/GATA-1 toggle switch, molecular interactions are far more complex than in our model, e.g. involving the formation of a heterodimer between antagonistic proteins. According to Loinger et al. [33] an interaction between the antagonists leads to different dynamics in the one-stage system, e.g. the deadlock regime (S_0) vanishes. Also the mutual inhibition between PU.1 and GATA-1 is not realized by simple binding of the repressor proteins to the promoter, but involves an additional cofactor pRB [47], introducing a regulatory asymmetry into the system, as pRB is only needed to inhibit GATA-1. The effects of these molecular details have been assessed by Krumsiek [29] using a deterministic model. In our model we do not allow for degradation of protein that are bound to DNA, assuming sterical hindrance of the proteasome – a large complex that actively degrades proteins – in close distance to DNA. As suggested by Loinger et al. [33] degradation of DNA bound proteins can lead to different dynamics. As the above assumption is neither supported nor disproved by experiments, an extension of our model, that allows for bound-protein degradation might be insightful. Overall, including more available molecular details might reveal further aspects of the system's dynamics – at the cost of increasing model complexity.

Finally, extensions of the applied methods are promising. The use of a more sophisticated simulation algorithm allows for analysis of a greater space of parameters. Extensions of the hybrid algorithm introduced by Haseltine and Rawlings [19], which incorporates the dynamic assignment of fast and slow reactions, would allow its application to the toggle switch model. The original algorithm could not be applied to the toggle switch, since there is a ongoing change in the fast and slow reaction subsets (depending in which regime the system is). A static partition into slow and fast subsets is not possible. A simulation algorithm that is capable of cell growth and cell cycle is also available [35]. This is an extension of the original stochastic simulation algorithm [17] which includes time dependent propensities that account for the exponentially growing cell volume. The cell cycle time is fixed and cell division is modeled by equally dividing molecular species once cell cycle time has been reached. Additionally Rathinam et al. [46] recently proposed a method for the computation of parameter sensitivity in stochastic systems, which could be applied to further investigate the influence of parameter asymmetries on the dynamics of the toggle switch. Recent advances in numerical solutions of the CME [66] or advanced finite state projection algorithms [41, 45] could be applied to the toggle switch, allowing for analysis of the complete probability distribution. Furthermore, an analytical solution of the CME using the methods of Sasai and Wolynes [53] and Walczak et al. [61] would improve the feasibility of the system description. Ultimately, model selection and inference of system parameters from experimentally observed realizations of the system using single cell data will lead to a comprehensive understanding of the cellular regulatory systems.

Appendix A

ODE steady state solutions

A.1 Deterministic one-stage toggle switch

Using symmetric parameters $\alpha = \alpha_\mathrm{A} = \alpha_\mathrm{B}$, $\delta = \delta_\mathrm{A} = \delta_\mathrm{B}$, $\tau^+ = \tau^+_\mathrm{A} = \tau^+_\mathrm{B}$ and $\tau^- = \tau^-_\mathrm{A} = \tau^-_\mathrm{B}$ results in two following steady state solutions of equations (3.2):

$$p_\mathrm{A}{}^{(1)} = p_\mathrm{B}{}^{(1)} = -\frac{\tau^- - \frac{\sqrt{\tau^-}\sqrt{4\alpha\tau^+ + \delta\tau^-}}{\sqrt{\delta}}}{2\tau^+} \tag{A.1}$$

$$d_\mathrm{A}{}^{(1)} = d_\mathrm{B}{}^{(1)} = \frac{2}{1 + \frac{\sqrt{4\alpha\tau^+ + \delta\tau^-}}{\sqrt{\delta\tau^-}}} \tag{A.2}$$

$$p_\mathrm{A}{}^{(2)} = p_\mathrm{B}{}^{(2)} = -\frac{\tau^- + \frac{\sqrt{\tau^-}\sqrt{4\alpha\tau^+ + \delta\tau^-}}{\sqrt{\delta}}}{2\tau^+} \tag{A.3}$$

$$d_\mathrm{A}{}^{(2)} = d_\mathrm{B}{}^{(2)} = \frac{2}{1 - \frac{\sqrt{4\alpha\tau^+ + \delta\tau^-}}{\sqrt{\delta\tau^-}}} \tag{A.4}$$

The first solutions is positive, the second is negative (given all parameters are positive).

A.2 Deterministic two-stage toggle switch

Using symmetric parameters $\alpha = \alpha_\mathrm{A} = \alpha_\mathrm{B}$, $\beta = \beta_\mathrm{A} = \beta_\mathrm{B}$, $\gamma = \gamma_\mathrm{A} = \gamma_\mathrm{B}$, $\delta = \delta_\mathrm{A} = \delta_\mathrm{B}$, $\tau^+ = \tau^+_\mathrm{A} = \tau^+_\mathrm{B}$ and $\tau^- = \tau^-_\mathrm{A} = \tau^-_\mathrm{B}$ results in two following steady state solutions of equations (4.1):

$$m_\mathrm{A}{}^{(1)} = m_\mathrm{B}{}^{(1)} = -\frac{\delta\tau^- - \frac{\sqrt{\delta\tau^-}\sqrt{4\alpha\beta\tau^+ + \gamma\delta\tau^-}}{\sqrt{\gamma}}}{2\alpha\tau^+} \tag{A.5}$$

$$p_\mathrm{A}{}^{(1)} = p_\mathrm{B}{}^{(1)} = -\frac{\tau^- - \frac{\sqrt{\tau^-}\sqrt{4\alpha\beta\tau^+ + \gamma\delta\tau^-}}{\sqrt{\gamma\delta}}}{2\tau^+} \tag{A.6}$$

$$d_\mathrm{A}{}^{(1)} = d_\mathrm{B}{}^{(1)} = \frac{2}{1 + \frac{\sqrt{4\alpha\beta\tau^+ + \gamma\delta\tau^-}}{\sqrt{\gamma\delta\tau^-}}} \tag{A.7}$$

$$m_{\mathrm{A}}{}^{(2)} = m_{\mathrm{B}}{}^{(2)} = -\frac{\delta\tau^{-} + \frac{\sqrt{\delta\tau^{-}}\sqrt{4\alpha\beta\tau^{+}+\gamma\delta\tau^{-}}}{\sqrt{\gamma}}}{2\alpha\tau^{+}} \tag{A.8}$$

$$p_{\mathrm{A}}{}^{(2)} = p_{\mathrm{B}}{}^{(2)} = -\frac{\tau^{-} + \frac{\sqrt{\tau^{-}}\sqrt{4\alpha\beta\tau^{+}+\gamma\delta\tau^{-}}}{\sqrt{\gamma\delta}}}{2\tau^{+}} \tag{A.9}$$

$$d_{\mathrm{A}}{}^{(2)} = d_{\mathrm{B}}{}^{(2)} = \frac{2}{1 - \frac{\sqrt{4\alpha\beta\tau^{+}+\gamma\delta\tau^{-}}}{\sqrt{\gamma\delta\tau^{-}}}} \tag{A.10}$$

The first solutions is positive, the second is negative (given all parameters are positive).

Appendix B

Accuracy of Gillespie's algorithm

In order to choose a reasonable number of simulation runs when estimating the regime probabilities P_{A}, P_{B} and P_0 in sections 3.2 and 4.2, we calculate these estimates for different numbers of simulation runs. We perform this analysis only for the symmetric system and assume that the results in terms of accuracy are also valid for asymmetric systems. Fig. B.1 and Fig. B.2 show that the accuracy of the estimates increase with increasing number of simulation runs, as expected. However, using more than 1000 simulations does not improve accuracy much, the estimates of the probabilities do not change much beyond this value.

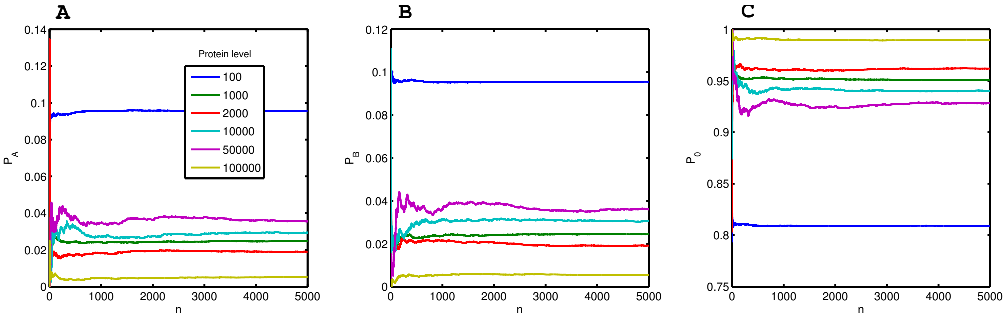

Figure B.1: Estimates of P_{A} (A), P_{B} (B) and P_0 (C) for different numbers n of simulation runs in the one-stage model. For low n estimates fluctuate heavily, but stabilize as more simulations are performed.

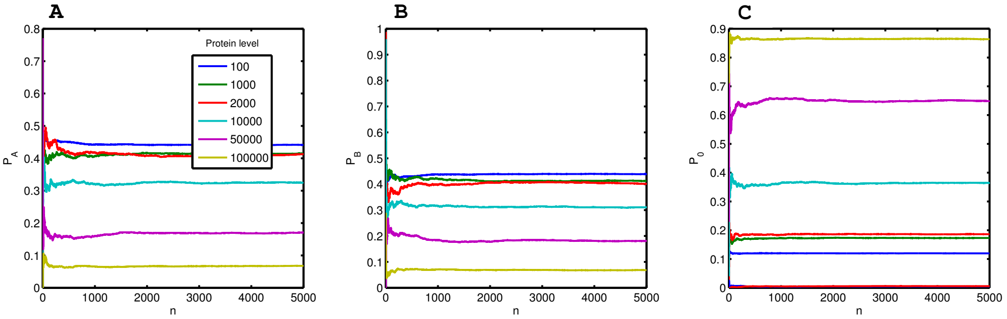

Figure B.2: Estimates of P_{A} (A), P_{B} (B) and P_0 (C) for different numbers n of simulation runs in the two-stage model. Low n leads to strong fluctuations of the estimates, but the variation decreases as more simulations are performed.

Bibliography

[1] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., , and Watson, J.D. *Molecular Biology of the Cell, 4th edition.* Taylor and Francis, 2002.

[2] Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits.* Chapman and Hall/CRC, 2006.

[3] Alon, U. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461, 2007.

[4] Barabasi and Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] Barzel, B. and Biham, O. Calculation of switching times in the genetic toggle switch and other bistable systems. *Phys Rev E Stat Nonlin Soft Matter Phys*, 78(4 Pt 1):041919, 2008.

[6] Cao, Y., Gillespie, D.T., and Petzold, L.R. The slow-scale stochastic simulation algorithm. *J Chem Phys*, 122(1):14116, 2005.

[7] Cao, Y., Gillespie, D.T., and Petzold, L.R. Efficient step size selection for the tau-leaping simulation method. *J Chem Phys*, 124(4):044109, 2006.

[8] Chickarmane, V., Enver, T., and Peterson, C. Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Comput Biol*, 5(1):e1000268, 2009.

[9] Dahlberg, M.E. and Benkovic, S.J. Kinetic mechanism of dna polymerase i (klenow fragment): identification of a second conformational change and evaluation of the internal equilibrium constant. *Biochemistry*, 30(20):4835–4843, 1991.

[10] Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[11] Enver, T., Heyworth, C.M., and Dexter, T.M. Do stem cells play dice? *Blood*, 92(2):348–51; discussion 352, 1998.

[12] Gardner, T.S., Cantor, C.R., and Collins, J.J. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[13] Gibson, M.A. and Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000. ISSN 1089-5639.

[14] Gillespie, D.T. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992. ISSN 03784371.

[15] Gillespie, D.T. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000. ISSN 00219606.

[16] Gillespie, D.T. Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys*, 115:1716–1733, 2001.

[17] Gillespie, D.T. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58:35–55, 2007.

[18] Glass, L. and Kauffman, S.A. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*, 39(1):103–129, 1973.

[19] Haseltine, E. and Rawlings, J. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of Chemical Physics*, 117(15):6959–6969, 2002.

[20] Haseltine, E. and Rawlings, J. On the origins of approximations for stochastic chemical kinetics. *J Chem Phys*, 123(16):164115, 2005.

[21] Huang, S., Guo, Y.P., May, G., and Enver, T. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev Biol*, 305(2):695–713, 2007.

[22] Iyer-Biswas, S., Hayot, F., and Jayaprakash, C. Stochasticity of gene products from transcriptional pulsing. *Phys Rev E Stat Nonlin Soft Matter Phys*, 79(3 Pt 1):031911, 2009.

[23] Kaern, M., Elston, T.C., Blake, W.J., and Collins, J.J. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–464, 2005.

[24] Kauffman, S.A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*, 22(3):437–467, 1969.

[25] Kennell, D. and Riezman, H. Transcription and translation initiation frequencies of the escherichia coli lac operon. *J Mol Biol*, 114(1):1–21, 1977.

[26] Kepler, T.B. and Elston, T.C. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys J*, 81(6):3116–3136, 2001.

[27] Kitano, H. Biological robustness. *Nat Rev Genet*, 5(11):826–837, 2004.

[28] Kloeden, P. and Platen, E. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

[29] Krumsiek, J. *Computational modeling of regulatory networks in hematopoietic differentiation*. Master's thesis, Technische Universitaet Muenchen, 2009.

[30] Laiosa, C.V., Stadtfeld, M., and Graf, T. Determinants of lymphoid-myeloid lineage diversification. *Annu Rev Immunol*, 24:705–738, 2006.

[31] Li, H., Cao, Y., Petzold, L.R., and Gillespie, D.T. Algorithms and software for stochastic simulation of biochemical reacting systems. *Biotechnol Prog*, 24(1):56–61, 2008.

[32] Little, J.W., Shepley, D.P., and Wert, D.W. Robustness of a gene regulatory circuit. *EMBO J*, 18(15):4299–4307, 1999.

[33] Loinger, A., Lipshtat, A., Balaban, N.Q., and Biham, O. Stochastic simulations of genetic switch systems. *Phys Rev E Stat Nonlin Soft Matter Phys*, 75(2 Pt 1):021904, 2007.

[34] Lok, L. and Brent, R. Automatic generation of cellular reaction networks with moleculizer 1.0. *Nat Biotechnol*, 23(1):131–136, 2005.

[35] Lu, T., Volfson, D., Tsimring, L., and Hasty, J. Cellular growth and division in the gillespie algorithm. *Syst Biol (Stevenage)*, 1(1):121–128, 2004.

[36] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[37] Mitchell, K.J. The genetics of brain wiring: from molecule to mind. *PLoS Biol*, 5(4):e113, 2007.

[38] Mueller Herold, U. General mass-action kinetics. positiveness of concentrations as structural property of horn's equation. *Chemical Physics Letters*, 33(3):467–470, 1975. ISSN 0009-2614.

[39] Munsky, B. and Khammash, M. Transient analysis of stochastic switches and trajectories with applications to gene regulatory networks. *IET Syst Biol*, 2(5):323–333, 2008.

[40] Munsky, B. and Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys*, 124(4):044104, 2006.

[41] Munsky, B.E. *The Finite State Projection Approach for the Solution of the Master Equation and its Applications to Stochastic Gene Regulatory Networks.* Ph.D. thesis, University of California, 2008.

[42] Newman, M.E.J. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, 2006.

[43] Okuno, Y., Huang, G., Rosenbauer, F., Evans, E.K., Radomska, H.S., Iwasaki, H., Akashi, K., MoreauGachelin, F., Li, Y., Zhang, P., Goettgens, B., and Tenen, D.G. Potential autoregulation of transcription factor pu.1 by an upstream regulatory element. *Mol Cell Biol*, 25(7):2832–2845, 2005.

[44] Orkin, S.H. and Zon, L.I. Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644, 2008.

[45] Peles, S., Munsky, B., and Khammash, M. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *J Chem Phys*, 125(20):204104, 2006.

[46] Rathinam, M., Sheppard, P.W., and Khammash, M. Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks. *J Chem Phys*, 132(3):034103, 2010.

[47] Rekhtman, N., Choe, K.S., Matushansky, I., Murray, S., Stopka, T., and Skoultchi, A.I. Pu.1 and prb interact and cooperate to repress gata-1 and block erythroid differentiation. *Mol Cell Biol*, 23(21):7460–7474, 2003.

[48] Rieger, M.A., Hoppe, P.S., Smejkal, B.M., Eitelhuber, A.C., and Schroeder, T. Hematopoietic cytokines can instruct lineage choice. *Science*, 325(5937):217–218, 2009.

[49] Roeder, I. and Glauche, I. Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors gata-1 and pu.1. *J Theor Biol*, 241(4):852–865, 2006.

[50] Roussel, M.R. and Zhu, R. Stochastic kinetics description of a simple transcription model. *Bull Math Biol*, 68(7):1681–1713, 2006.

[51] Roussel, M.R. and Zhu, R. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys Biol*, 3(4):274–284, 2006.

[52] Samad, H., Khammash, M., Petzold, L., and Gillespie, D. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15(15):691–711, 2005.

[53] Sasai, M. and Wolynes, P.G. Stochastic gene expression as a many-body problem. *Proc Natl Acad Sci U S A*, 100(5):2374–2379, 2003.

[54] Schultz, D., Onuchic, J.N., and Wolynes, P.G. Understanding stochastic simulations of the smallest genetic networks. *J Chem Phys*, 126(24):245102, 2007.

[55] Shahrezaei, V. and Swain, P.S. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci U S A*, 105(45):17256–17261, 2008.

[56] Singh, K., Srivastava, A., Patel, S.S., and Modak, M.J. Participation of the fingers subdomain of escherichia coli dna polymerase i in the strand displacement synthesis of dna. *J Biol Chem*, 282(14):10594–10604, 2007.

[57] Swain, P.S., Elowitz, M.B., and Siggia, E.D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A*, 99(20):12795–12800, 2002.

[58] Thattai, M. and van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–8619, 2001.

[59] Varma, A., Morbidelli, M., and Wu, H. *Parametric sensitivity in chemical systems.* Cambridge University Press, 1999.

[60] Waddington, C.H. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology.* George Allen & Unwin, 1957.

[61] Walczak, A., Mugler, A., and WIggins, C. Analytic methods for modeling stochastic regulatory networks. pages –, 2010.

[62] Warren, L., Bryder, D., Weissman, I.L., and Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital rt-pcr. *Proc Natl Acad Sci U S A*, 103(47):17807–17812, 2006.

[63] Watts, D.J. and Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[64] Wittmann, D.M., Bloechl, F., Truembach, D., Wurst, W., Prakash, N., and Theis, F.J. Spatial analysis of expression patterns predicts genetic interactions at the mid-hindbrain boundary. *PLoS Comput Biol*, 5(11):e1000569, 2009.

[65] Wittmann, D.M., Krumsiek, J., Rodriguez, J.S., Lauffenburger, D.A., Klamt, S., and Theis, F.J. Transforming boolean models to continuous models: methodology and application to t-cell receptor signaling. *BMC Syst Biol*, 3:98, 2009.

[66] Wolf, V., Goel, R., Mateescu, M., and Henzinger, T.A. Solving the chemical master equation using sliding windows. *BMC Syst Biol*, 4(1):42, 2010.