Stress responses in *Escherichia coli* and HIV as model systems of adaptation to the environment

A modeling approach based on stochastic dynamics

Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Kaveh Pouran Yousef, M.Sc. Bioinformatik

Berlin, Juli 2013

Betreuer: Prof. Dr. Christof Schütte
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Arnimalle 6
14195 Berlin

Gutachter: Prof. Dr. Christof Schütte

Prof. Dr. Niko Beerenwinkel (ETH Zürich)

Tag der Disputation: 25. November 2013

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Optimal, robust and fine-tuned regulation in biological systems does not necessarily go along with deterministic behavior. While from an engineering point of view noise is usually regarded as an unwanted side effect, living systems exhibit a very versatile relationship to stochastic fluctuations associated with the transmission of signals. On the one hand, erroneous transduction and interpretation of noisy environmental stimuli is known to have a strong potential to disrupt the cellular function and cause diseases. On the other hand, stochasticity is necessary for generating heterogeneity of cellular functions and fates, vital for a survival in changing environments. A dissection of mechanisms which enable a trade-off between these two roles has become an essential objective for a holistic dynamical understanding of cellular processes.

Today's life scientists have access to a vast systematic repository of data describing biological structure, organization and interaction. Almost seventy years after Schrödinger's physical focus on living matter in "What is life?", sixty years after the publication of the structure of the DNA and over a decade after the complete sequence of the human genome became available, it is clear that no new physical laws have to be formulated to describe the mechanisms governing life. Nevertheless many concepts related to biological self-regulation and organization or, to go a step further, the abstract concept of intelligence still retain a touch of magical fascination. On the most basic level, at the frontier between living and non-living matter, viruses exhibit a certain, often harmful form of intelligence by being able to rapidly adapt to various environments and hostile drugs. Physically, this adaptation can be considered as a stochastic system in a potential landscape. It resides in a genetic equilibrium state and possibly sometimes exhibits a transition, induced by mutational noise, to another distant equilibrium. A variation in the environment, say by drug addition, gives rise to a novel potential landscape, and begins the viral search for the optimum anew.

In the most intensively studied prokaryotic species, *Escherichia coli*, genomic changes due to mutation and horizontal gene transfer also play an important role in adapting to various environments. However the bacteria possess a whole arsenal of mechanisms for generating phenotypic responses to stress conditions which makes their genomes more robust with respect to selection pressure. The various signaling cascades incorporating ubiquitous two-component modules and second messenger systems exhibit a high degree of orderliness. In recent years, however, an increasing number of studies addressed the role of stochastic fluctuations in bacterial gene

regulation and signal transduction [73]. It has been shown that the bacterial cell constantly deals with *intrinsic noise* originating from small sizes of populations of crucial molecules, such as the mRNA. Furthermore, fluctuations in global biochemical parameters, such as the reaction rates, give rise to what has been coined as the *extrinsic noise*. Most systems set upper tolerance limits to such fluctuations, for instance by using feedback loops or regulatory checkpoints. In certain cases, however, stochastic fluctuations are exploited in order to generate heterogeneity of phenotypic or genotypic responses to external stimuli. These properties are encountered throughout the domain of prokaryotes. For instance, it is has been shown that upon environmental stress, the Gram-positive *Bacillus subtilis* stochastically switches from its normal vegetative state to the state of competence, where it has an increased ability to take up extracellular DNA [18].

From a mathematical perspective of system design there is no contradiction between tight regulation and optimal adaptation on the one hand and the stochasticity of the dynamics on the other hand. Thus the tight control can be imposed over the deterministic properties of the underlying probability distributions of the molecules involved in the regulation. For instance, in the case of the viral evolution, the potential energy $V(x)$, associated with the probability for each genotype x, might vary with external conditions. A change in the potential induces a new genotypic distribution corresponding to an optimal adaption of the population to the new environment. The potential can also be subject to control e.g. by a bacterial signaling system which regulates the probability distribution for expressing a particular phenotype as a reaction to stress conditions.

In fig. 1.1 the probabilistic adaptation to environment is exemplified using two different double-well potentials. Initially, the potential energy $V(x)$ has two equal minima which correspond to two distinct phenotypic or genotypic traits (figure 1.1, left). First, the system stochastically fluctuates in the right well and after some time, eventually jumps over the energy barrier to the left well. In this situation the trait of interest is characterized by a bistable expression. In the second situation, a change in environmental conditions and possible corresponding stress response gives rise to a new asymmetric potential function (figure 1.1, right). This time the stochastic fluctuations are not large enough to let the system jump over the energy barrier to the left. The resulting system exhibits a monostable probability distribution, tightly concentrated around a single trait.

The focus of this thesis is on stochastic modeling of adaptation processes and stress responses. Although established mathematical methodology is used, the presented applications give rise to novel tools and insights within the study field of Systems Biology. The thesis is based on two main projects. In the first one an analysis of a signaling system of the bacterium *Escherichia Coli* is conducted. This system is part of the general stress response, induced as a reaction to a set of stressful environmental conditions (chapter 3). In the second project an evolutionary adap-

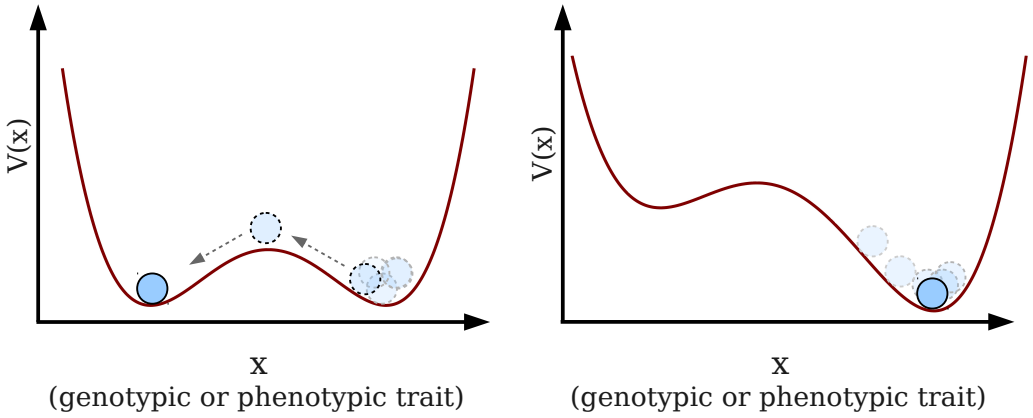


Figure 1.1: **Potential energy resulting from environmental conditions as a determinant of a genotypic or phenotypic distribution.**
Realizations of a stochastic system sketched as a ball in a double-well potential subject to random fluctuations. A variation of the potential function leads to a change of the probabilities to encounter the system in either of the two equilibria. **Left**: An exemplary symmetric potential enables noise-induced transitions from the left well to the right one. **Right**: In an asymmetric potential the energy barrier is to large to enable a transition.

tation process of the human immunodeficiency virus (HIV) in the presence of drug application is studied (chapter 4). The similarity of the two projects arises from the fact that both systems give rise to state-discrete stochastic dynamics where the time-evolution of the probability distribution can be studied using the same mathematical framework. However the addressed questions significantly differ from each other, as described in the following.

Dynamics of bacterial signaling induced by the general stress response

The protein RpoS is the master regulator of the general stress response controlling 10% of the whole genome of *Escherichia coli*. A crucial consequence of RpoS-induction is the reduction of bacterial metabolism and expression of a protein network which ultimately gives rise to the synthesis of biofilm, a protecting substance promoting aggregation in bacterial colonies. Key to dissecting the underlying regulatory mechanisms is an understanding of the stochastic dynamics induced by the second messenger signaling molecule cyclic di-GMP (c-di-GMP) [35]. Based on well-known qualitative principles of c-di-GMP regulation, a Markov jump process model is derived in chapter 3. The stationary properties of the underlying probability distribution, noise-reduction strategies of the system and first passage time statistics are analyzed. These results enable to study the dynamics of the biofilm expression system, which incorporates c-di-GMP signaling. Using a bifurcation analysis, it is

shown that the analyzed system exhibits bistable dynamics within a realistic range of parameters. Finally, based on expression measurements, evidence is found that this system is responsible for the phenotypic heterogeneity of biofilm formation in *E. Coli* populations. The results of chapter 3 yield, to our knowledge, the first model of the stochastic dynamics of the stress-induced biofilm formation network. They deliver optimality arguments for the particular architecture of the qualitative network and explain the experimental measurements of biofilm expression data. It is left for following studies to extend these results for a more detailed analysis of transition dynamics of the bistable biofilm-synthesis system and a more global dynamical understanding of the RpoS-controlled network.

Studying evolution of the HIV genome using a stochastic model of viral growth

In the presence of drugs inhibiting the reverse transcription, susceptible HIV-strains ultimately become resistant by exhibiting genetic mutations which reduce the effect of drug action. This evolutionary escape mechanism, as a response to environmental stress, can be considered as a stochastic system on a potential landscape. Due to a large number of amino acid residues associated with the structure and function of the reverse transcriptase enzyme (RT), the corresponding potential landscape of drug escape is complex and high-dimensional. In chapter 4 certain regions of this landscape are inferred from experimental data of viral growth and the associated drug-induced mutations. As in the preceding chapter, a Markov jump process is used as a model for the viral population growth given a certain genetic background and drug regimen. As a result, the effect of fitness loss and resistance gain associated with an extensive set of mutations of the RT-enzyme is estimated using stochastic viral growth dynamics. Furthermore the presented model explains the observed direction of the genetic evolution of HIV in the presence of drugs and after their removal. The results of chapter 4 enhance the so far available knowledge about resistance development mechanisms of HIV and contribute to predicting the direction of its genetic evolution. The results of this chapter have been published in [60].

Acknowledgements

My gratitude is firstly owed to Professor Dr. Christof Schütte for the valuable advice and constant support. He gave me the chance to do research in the Biocomputing group where I was surrounded by so many brilliant people and where the solution to each scientific and non-scientific problem often was just a cup of coffee away. My special thanks are owed to Dr. Max von Kleist. I was lucky to have him as a mentor who shared his scientific expertise and gave me confidence in this project. Without him this thesis would be a completely different one. I'm indebted to Professor Dr. Regine Hengge. She introduced me to the fascinating world of procaryotes and her mentoring and expert advice were a very important help and motivation

throughout the doctoral studies. I also have to thank my friends and the people from the Biocomputing group who strongly contributed to making the studies a pleasant time and helped me to keep my smile even in difficult periods. I would like to thank my parents for their constant love and trust. My deepest gratitude goes to Fenna for being there and for the infinite patience. Without her this would not have been even possible.

Chapter 2

Stochastic processes in biological applications

2.1 Introduction

Stochastic modeling is becoming an increasingly important tool in computational simulation of biological processes. Applications range from protein folding, system models of cellular networks and evolutionary dynamics of the genome. In many of such applications system averages do not rigorously represent the qualitative picture behind the real-life system and sometimes even distort it. A classical example is given by multistable systems, describing conformational dynamics of proteins or bistable phenotypes of cells in a culture. In such systems stochastic modeling often reveals noise-induced transitions between various stable system states, which might have important biological implications. However, this switching behavior can not be observed if the stochastic fluctuations are neglected, leading to a completely different qualitative dynamics.

A further example showing the role of stochastic modeling arises from situations where extinction has a significant qualitative impact. Consider a population of pathogenic cells (e.g. bacteria) in the human organism of size X, subject to a medical treatment by drug application. Usually, the success of a medical treatment is determined by the time that it takes to eradicate the pathogen from the patients organism. The simplest model for such a system consists of a population growth e.g. by cell division of the pathogen and the population decay due to cell death. Obviously these two processes do not take place in a continuous manner but have rather a random nature. The number of random events of each kind within a certain time interval might be a function of various parameters such as current number of pathogens, drug concentration, body temperature etc. In the case that these are linear functions of the population size with constant parameters a and b, the expected level of pathogens can be modeled by an ordinary differential equation (ODE):

$$\frac{dx(t)}{dt} = ax(t) - bx(t), \tag{2.1}$$

where $x(t)$ represents the expected number of cells (which can be considered as concentration rather than cell counts) at time t. Denoting by x_0 the initial population

size, the solution at time t is obviously given by an exponential function

$$x(t) = x_0 \cdot \exp[(a - b)t].$$

In a situation where, for instance due to drug application, the ratio $(a - b)$ is negative, it is of interest to determine the time point at which the population goes extinct. However, by analyzing the exponential solution of the above ODE, we find that the expected number of cells approaches zero but never completely vanishes. In contrast, a stochastic simulation, discussed further below in this chapter, allows to exactly analyze the statistics of extinction times. In figure 2.1 two stochastic realizations of the population dynamics model are depicted along with their mean level, modeled by the ODE (2.1). The two stochastic realizations (blue trajectories)

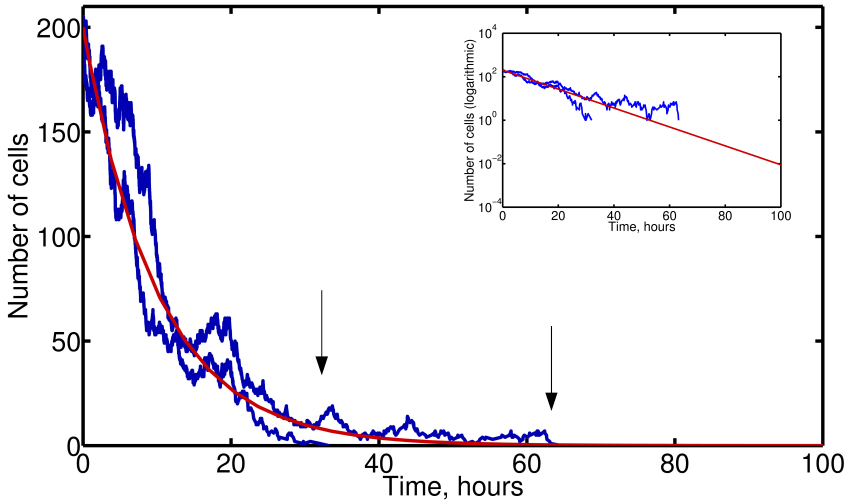


Figure 2.1: **Deterministic dynamics does not reflect the extinction times.** Shown are two stochastic simulations (blue) and the corresponding deterministic simulation of the mean level (red) of a pathogenic population decay e.g. under drug application. The mean level, given by an exponential function, continuously decreases but never hits 0 (as shown in the inlay figure, where a logarithmic y-axis is used). In contrast, two individual stochastic realizations indicate different extinction times (marked by arrows), giving rise to the probability distribution of extinction.

indicate extinction around 25 and 59 hours after the start of drug application, respectively. These might represent the individual treatment outcomes for two different patients. In contrast, the mean level (red) modeled by the ODE (2.1) exhibits a continuous convergence to zero without indicating the statistics of extinction times.

This chapter is dedicated to stochastic processes and related topics playing important roles in computational biology. In particular, the methodological framework presented here is tightly connected to biological questions addressed. In chapter 3 an analysis of the stress-induced dynamics of curli formation is conducted, an extracellular polymeric substance leading to aggregated bacterial biofilm colonies.

We ask how discrete on/off expression of curli on the single cell level is related to the continuous average expression measured in the cell population. A stochastic model is derived, describing the molecular biological network which is suggested here to be the source for the switching behavior between the curli-on and curli-off states. The theoretical framework of Markov jump processes (MJPs), derived in this chapter, enables an analysis of dynamical properties induced by the signaling network and quantify the impact of the discrete nature of biomolecular kinetics and the resulting stochastic fluctuations. In chapter 4 the probability density of first passage times is analyzed in order to model the population growth of HIV underlying a simultaneous genetic evolution as a stress response to drug application. The theory of MJPs allows to mechanistically derive and explain the variability of the viral growth dynamics. This gives an insight into the principles of the high genetic flexibility of the virus with respect to the stress conditions imposed by different medical treatments.

In parallel to Markov jump processes and the Chemical Master Equation, as the theoretical cornerstones, different affiliated topics will be discussed. In most of the applications in the following chapters the connection between the analyzed stochastic process and a related deterministic equation plays an important role. In chapter 3 the analysis of an approximating ODE at its fixed points will enable to find the stationary states of the underlying stochastic system. Here we present the framework which allows to conduct this analysis and we will discuss the limits of the approximation by considering the moments equations associated with the probability density of MJPs. Since stochastic differential equations (SDEs) can be considered as a link between the world of state-discrete Markov processes and the purely continuous ODEs, some space will be dedicated to shed light on the relationship between the three different frameworks. Notably, the Linear Noise approximation of the Chemical Master Equation will be discussed in this chapter and applied to interaction dynamics of HIV and host cells in chapter 4. This enables a considerable acceleration of sampling by using SDEs for parameter estimation and fitting of the first passage time moments.

Finally, a framework for model inference in biochemical reaction systems will be presented in the end of this chapter. To our knowledge it is the first time that the methodology for estimating infinitesimal generator matrices is adopted for inferring biochemical reaction rate constants. It will be shown that despite prohibitively large state spaces, the problem of estimating the large set of infinitesimal jump rates can be reduced to the inference of a few reaction rate constants. The advantages of this approach and possible solutions to the problem of discrete process observations will be discussed.

2.2 Markov Processes

Stochastic processes and the Markov property

A central object in the analysis of dynamical systems with time evolution subject to random fluctuations is the notion of a stochastic process. Its definition is derived using the triple $(\Omega, \mathcal{A}, \mathbb{P})$, the so called *probability space*, where Ω is a sample space, $\mathcal{A}$ is the corresponding σ-algebra i. e. the set of all possible subsets on the sample space and $\mathbb{P} : \mathcal{A} \mapsto [0, 1]$ is a probability measure.

Definition 1. *A collection of random variables* $X = \{X(t) : \Omega \mapsto \mathrm{S}\}$ *is called a* ***stochastic process*** *with time index* $t \in \mathbb{R}^+$ *and a space of all possible simple outcomes* S. *For a given* $\omega \in \Omega$ *the set* $\{X(t, \omega)\}$ *is called a* ***realization*** *of the stochastic process* X.

A stochastic process $\{X(t)\}$ thus defines a mapping from the sample space to the space of all possible simple outcomes S, also referred to as *state space*. The probability measure defined on the corresponding σ-algebra ensures that the elements of $\{X(t)\}$ are random variables. The dependence on a non-negative real variable t in definition 1 implies that a stochastic process can be regarded as a sequence of random variables evolving in time. The probability measure $\mathbb{P}$ completely characterizes this evolution, since for a given finite sequence of successive time points $t_1, t_2, t_3, .., t_n$ a realization of a stochastic process obeys the joint probability

$$\begin{aligned} & \mathbb{P}[X(t_1) = x_1, X(t_2) = x_2, .., X(t_n) = x_n] \\ = \; & \mathbb{P}[X(t_1) = x_1] \times \mathbb{P}[X(t_2) = x_2 | X(t_1) = x_1] \\ \times \; & \mathbb{P}[X(t_3) = x_3 | X(t_1) = x_1, X(t_2) = x_2] \times ... \\ \times \; & \mathbb{P}[X(t_n) = x_n | X(t_1) = x_1, X(t_2) = x_2, ..., X(t_{n-1}) = x_{n-1}]. \end{aligned} \tag{2.2}$$

For instance if for each realization the joint probability is given by a multivariate Gaussian distribution

$$\begin{aligned} \mathbb{P}[X(t)] &= \mathbb{P}[X(t_1) = x_1, X(t_2) = x_2, .., X(t_n) = x_n] \\ &= \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (X(t) - \mu)^T \Sigma^{-1} (X(t) - \mu) \right], \end{aligned} \tag{2.3}$$

then $\{X(t)\}$ is called a *Gaussian process*. In this equation μ and Σ are the corresponding mean vector and the covariance matrix, respectively. $|\Sigma|$ denotes the matrix determinant. For instance if at time $t_1 = 0$ it holds for every realization that $X(t_1) = 0$ and $X(t_k) - X(t_{k-1}) \sim \mathbb{N}(0, t_k - t_{k-1})$ for $t_{k-1} < t_k$, then it can be shown [23] that the corresponding process is *Gaussian*. It is called the *Wiener Process* and it models the phenomenon of Brownian motion, that will be referred to as dB_t.

The characterization of realizations of a stochastic process in terms of the joint probability of the corresponding random variables is usually rather intractable.

However, for a certain class of stochastic processes this joint probability can be reformulated in simpler form using the conditional probability.

Definition 2. *If for a stochastic process $X(t)$ it holds true that*

$$\mathbb{P}[X(t_n)|X(t_1), X(t_2), ..., X(t_{n-1})] = \mathbb{P}[X(t_n)|X(t_{n-1})] \tag{2.4}$$

*for all $t \in \mathbb{R}^+$ then $X(t)$ is a **Markov process**. In addition, given that the one-step conditional probability on the right hand side does not depend on individual time points but their difference $t_n - t_{n-1}$ then the Markov process is called **homogeneous**.*

A homogeneous Markov process is uniquely characterized by
(a) the function $p : \mathbb{R}^+ \times \mathrm{S} \times \mathrm{S} \mapsto [0, 1]$ with

$$p(t, x, y) := \mathbb{P}[X(t) = y|X(0) = x],$$

*called **stochastic transition function** with the properties $p(t, x, y) \geq 0$ for all $x, y \in \mathrm{S}$ and $\sum_{y\in\mathrm{S}} p(t, x, y) = 1$.*
(b) and its initial probability distribution is

$$P_0(x) := \mathbb{P}[X(0) = x].$$

Obviously, the *Markov property* defines a subclass of stochastic processes for which the joint probability of successive realizations can be simplified to the product of one-step conditional probabilities

$$\begin{aligned}
&\mathbb{P}[X(t_1) = x_1, X(t_2) = x_2, .., X(t_n) = x_n] = \\
&\mathbb{P}[X(t_n) = x_n|X(t_{n-1}) = x_{n-1}] \times \\
&\mathbb{P}[X(t_{n-1}) = x_{n-1}|X(t_{n-2}) = x_{n-2}] \times ... \times \mathbb{P}[X(t_2) = x_2|X(t_1) = x_1],
\end{aligned}$$

which is analytically by far better tractable than the general case in eq. (2.2). Furthermore, for a homogeneous Markov process with a constant time increment $t_k - t_{k-1} = s$ for all $k \in \mathbb{N}$ this further simplifies to

$$\begin{aligned}
&\mathbb{P}[X(t_1) = x_1, X(t_2) = x_2, .., X(t_n) = x_n] = \\
&P_0(x_1) \prod_{i=1}^{n-1} p(s, x_{i+1}, x_i).
\end{aligned} \tag{2.5}$$

Notably, the Wiener process fulfills the Markov property and it is homogeneous since for every given realization $X(t), t \in \{t_1, ..., t_n\}$

$$X(t_k) = X(t_{k-1}) + \triangle X$$

with $\triangle X \sim \mathbb{N}(0, t_k - t_{k-1})$. This shows that the value at the current time point only depends on its difference to the preceding one.

Note that the time variable can also be chosen to be discrete. In that case a

realization of the Markov process is defined only by integer indices $X = \{X_k\}_{k\in\mathbb{N}}$. It is then only of interest how many discrete steps it takes to go from state X_i to state X_j. The transition probability density becomes time-independent, changing probability of a realization (2.5) to

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, .., X_n = x_n] = P_0(x_1) \prod_{i=1}^{n-1} p(x_{i+1}, x_i). \tag{2.6}$$

In contrast to the Wiener process, these processes are discrete in time and in space and are referred to as **Markov chains**. As an example, consider the state space of genetic mutations associated with the amino acid Valine at position 106 in the reverse transcriptase enzyme of HIV (figure 2.2). As shown in chapter 4, this mutation confers the virus resistance with respect to the non-nucleoside reverse-transcriptase inhibitor NVP. The reference wild type viral strain is known to have the amino acid Valine at this position which can be coded by four different base triples (codons) within the red box in fig. 2.2. Single-base mutation events can lead to transitions between different states. Three of the possible base-triples coding for Valine have the same transition characteristics since they have a distance of one mutation to codons of Alanine and Isoleucine (the upper three codons in the red box in fig. 2.2). Thus these three codons are subsumed as a single state 1. Since transitions between these three codons are possible, the state 1 has a self-transition arrow. Furthermore, the three codons can mutate to the fourth codon of Valine (GUG), which constitutes a separate state 2 due to different transition properties. Similarly the states 3 and 4 refer to the amino acid Alanine, the state 5 refers to the amino acid Isoleucine and the state 6 belongs to Methionine. By denoting the

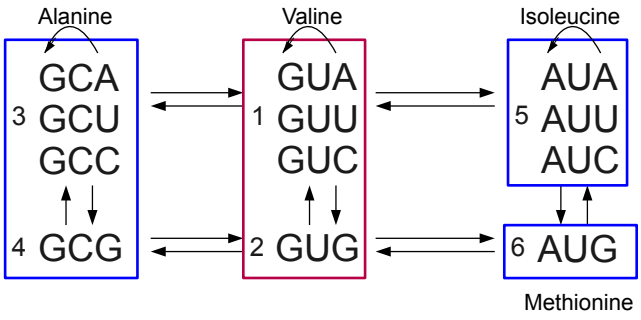


Figure 2.2: **Transition graph induced by mutations of the amino acid at position 106 of the reverse transcriptase enzyme of HIV.**
Transitions between different states are induced by single-base mutation events. The impact of the associated mutations on the resistance of HIV towards treatment is analyzed in chapter 4.

probability of transition between state x and y as p_{xy}, a *transition matrix* $\mathbf{P}$ can be set up that describes how probable it is to jump from one state to another. The

transition graph in figure 2.2 gives rise to the following transition matrix:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & p_{13} & 0 & p_{15} & 0 \\ p_{21} & 0 & 0 & p_{24} & 0 & p_{26} \\ p_{31} & 0 & p_{33} & p_{34} & 0 & 0 \\ 0 & p_{42} & p_{43} & 0 & 0 & 0 \\ p_{51} & 0 & 0 & 0 & p_{55} & p_{56} \\ 0 & p_{62} & 0 & 0 & p_{65} & 0 \end{pmatrix}.$$

Since probabilities of realizations of this process are only defined in terms of the mutational steps involved and do not consider how much time it takes to mutate, the described process induces a time-discrete Markov chain. For instance, assigning all initial probability mass to state 1 (Valine) by

$$P_0(1) = 1,$$

the probability of the realization of the mutational sequence GUA-GUU-AUU-AUG (Val-Val-Ile-Met) is given by

$$\mathbb{P}[X(t_1) = 1,\ X(t_2) = 1,\ X(t_3) = 5,\ X(t_4) = 6] = P_0(1) \cdot p_{11} \cdot p_{15} \cdot p_{56}.$$

2.2.1 Markov jump processes

So far, none of the definitions makes any assumption about the continuity of the state space S. However, in the present work we are mainly interested in Markov processes with a countable state-space S and continuous time t which can be regarded as a mix between the time and space continuous Wiener process and the fully discrete Markov chain. Also, we have not yet discussed how realizations of Markov processes can be generated (although rather simple in the case of the RT-mutational Markov Chain, it requires an additional theory for generating sample paths of the Wiener process).

The discreteness of S requires an additional assumption about the process. That is, the stochastic transition function needs to fulfill the following condition

$$p(0, y, x) = \delta_{x,y},$$

for all $x, y \in \mathrm{S}$, where $\delta_{x,y}$ denotes the Dirac delta function with $\delta_{x,y} = 1$ if $x = y$, and $\delta_{x,y} = 0$, otherwise. This property ensures that no transitions are possible in zero time. In addition, the continuity of the transition function is assumed

$$\lim_{t \to 0^+} p(t, y, x) = \delta_{x,y},$$

which makes sure that Markov jump processes are *right-continuous*. Furthermore using the Markov property it can be shown that the stochastic transition function of a homogeneous Markov process fulfills the **Chapman-Kolmogorov equation**:

$$p(s + t, x, y) = \sum_{z \in \mathrm{S}} p(s, x, z) p(t, z, y). \tag{2.7}$$

Equation (2.7) states that the probability of going from some state x to a state y results from the sum of all possibilities to go from x to one of the possible intermediate states and then to go from the intermediate state to the end state y. This equation is a direct consequence of the Markov property. The sum over all possible intermediate states can be interpreted here as a set-theoretical union over the alternative paths or it can be regarded as a logical `OR`. The transition probabilities can also be expressed in a matrix notation. As a result, the Chapman-Kolmogorov property reads

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s).$$

where

$$\mathbf{P}(t)_{x,y} = p(t,x,y).$$

is called the *transition matrix* (or transition *semi-group*), as in the case of Markov chains. Importantly, due to the probability normalization condition the transition probabilities associated with a particular state must sum to 1. Owing to this property, the transition matrix qualifies as a *stochastic matrix*, having a row sum of one and non-negative entries.

As shown in equation (2.5), the stochastic transition function takes some initial probability distribution P_0 and assigns the probability $P(t)$ to each state of the state space at time t defined as a vector $P(t) := \mathbb{P}[X = x, t]$. The propagation of the probability in time is then expressed by

$$P(t)^T = P_0^T \mathbf{P}(\mathbf{t}). \tag{2.8}$$

Furthermore, if a given probability distribution P_0 fulfills the following equality

$$P_0^T = P_0^T \mathbf{P}(\mathbf{t}), \tag{2.9}$$

then P_0^T is referred to as the **stationary probability distribution** of the Markov jump process. This distribution is time-invariant i.e. a system started in P_0 stays there forever.

So far we have defined a Markov process on a countable state-space S. By requiring that it is right-continuous we made sure that the transitions between the discrete states are well defined. In the following we will study the properties of the jump-like transitions between the states of the process and refer to this class of processes as *Markov jump processes*.

The infinitesimal generator and the Master Equation

The right-continuity of Markov jump processes is preliminary for deriving expressions for their time dynamics. So far it is clear that transitions between discrete states must take place but since these are random jumps, how can their statistics be quantified? The basic idea behind the following framework is the analysis of

infinitesimal time intervals which are so small that only one jump at a time can occur. The following limit sets up a connection between the *propensity* for a particular jump to take place, out of all possible jumps, and the respective transition probabilities.

Proposition 1. *Given a Markov jump process with a semigroup P(t) then the limit*

$$\mathbf{A} = \lim_{t \to 0^+} \frac{\mathbf{P}(t) - \mathrm{Id}}{t} \tag{2.10}$$

exists and defines the infinitesimal generator $\mathbf{A} = (a(x,y))_{xy \in S}$ *with* $-\infty \leq a(x,x) \leq 0 \leq a(x,y) \leq \infty$.

A proof can be found in [9]. Note that the above limit is defined entrywise, from which follows the existence of two scalar limits:

$$a(x,x) = \lim_{t \to 0^+} \frac{p(t,x,x) - 1}{t},$$

and

$$a(x,y) = \lim_{t \to 0^+} \frac{p(t,x,y)}{t}.$$

Since for the transition matrix it holds $\sum_{\substack{y \in \mathrm{S} \\ y \neq x}} \mathbf{P}(t)_{xy} + \mathbf{P}(t)_{xx} = 1$ it immediately follows from the limit (2.10) that

$$a(x,x) = -\sum_{\substack{y \in \mathrm{S} \\ y \neq x}} a(x,y). \tag{2.11}$$

The infinitesimal generator, defined in this way, is a matrix containing transition rates between the states in S. While the transition probability matrix is characterized by a row-sum of 1, the infinitesimal generator has a row-sum of 0. Notably, this matrix form assumes a finiteness of the state space. Markov jump processes, which are not constrained to a finite S, can not be described in this way. Alternatively, appropriate boundary conditions ("exit-states") must be implemented which reduce the loss of probability mass and the resulting error (see for instance [53]). The diagonal entries have furthermore an implication on the time dynamics by determining how long the process stays in the respective state.

Theorem 1. *Given a Markov jump process* $X(t)$ *on a state space* S *then the time*

$$\tau(t) = \inf\{s \geq 0 : X(t+s) \neq X(t)\}$$

is called the ***residual life time****. The tail distribution of* $\tau(t)$ *is given by*

$$\mathbb{P}[\tau(t) > s | X(t) = x] = \exp(a(x,x)s) \tag{2.12}$$

and the probability of going from state x *to state* y *upon jumping from* x*, with* $x, y \in \mathrm{S}$ *is*

$$\mathbb{P}[X(t + \tau(t)) = y | X(t) = x] = \frac{a(x,y)}{|a(x,x)|}. \tag{2.13}$$

The above result for the residual life time can be shown by considering some infinitesimal time interval ds, cf. [25]. Due to the definition of the infinitesimal generator, the probability not to jump away from a state x in this time interval is

$$\mathbb{P}[X(t+s+ds)=x|X(t)=x] = \left(1-\sum_{y\in S} a(x,y)\cdot ds.\right)\cdot \mathbb{P}[X(t+s)=x|X(t)=x].$$

The above equation can also be written as

$$\begin{aligned} & \frac{1}{ds}\cdot\Big(\mathbb{P}[X(t+s+ds)=x|X(t)=x]-\mathbb{P}[X(t+s)=x|X(t)=x]\Big) \\ = \;& a(x,x)\cdot\mathbb{P}[X(t+s)=x|X(t)=x], \end{aligned} \tag{2.14}$$

where the sum over the jump rates was replaced by the diagonal entry of the infinitesimal generator (eq. 2.11). This yields a linear differential equation for the probability of *not jumping* away from the state x in the interval $[t,t+s]$:

$$\frac{d\mathbb{P}[X(t+s)=x|X(t)=x]}{ds} = a(x,x)\cdot\mathbb{P}[X(t+s)=x|X(t)=x].$$

The solution of this ODE is an exponential function given by

$$\mathbb{P}[X(t+s)=x|X(t)=x] = P_0(x)\cdot\exp(a(x,x)s) = \exp(a(x,x)s),$$

where the deterministic initial condition results in $P_0(x)=1$. The residual life time is thus an exponentially decreasing function since $-\infty \leq a(x,x) \leq 0$. The time when the process jumps away from a state x determines the time interval that the process has stayed in this state, which is equivalent to the definition of the residual life time. This proves the first statement of the theorem. The second statement follows from the entrywise limit in equation (2.10).

Note that since $a(x,x)\leq 0$, the residual life time is a monotonically decreasing function of s, implying that the probability of a jump increases with an increasing time that the process has spent in a particular state. Importantly, the two random variables representing the residual life time and the transition probability to the next state are independent of each other. A proof can be found in [9]. This leads to a representation of a MJP as a process integrating a random time sequence $\{T_k\}_{k\in\mathbb{N}}$ with $T_{k+1}=T_k+\tau(T_k)$ with a Markov chain which is characterized by the transition probabilities $p(T_k,x,y)$. The latter is referred to as an **embedded Markov chain** and its properties are discussed in more detail in [9].

Theorem 1 implies that diagonal entries of the infinitesimal generator determine the expected residence time in each state as the mean of the exponential distribution $\mathbb{E}(\tau(t)) = 1/|a(x,x)|$ and assign an individual probability of going to each state of the state space upon jumping. These results immediately yield an algorithm for generating realizations of a Markov jump process. It is based on an alternate

sampling of the residual life time $\tau(t)$ in the current state x from the exponential distribution $\tau(t) \sim \exp(a(x,x)t)$ and drawing the next state y by sampling from the probability (2.13). This sampling approach is referred to as the Stochastic Simulation Algorithm (SSA) or Gillespie method [24].

It was shown in a previous section by equation (2.8) that the time-dependent transition matrix propagates probability distributions of the Markov jump process in time. In many situations one does not have the transition matrix for each time point of interest. The SSA-algorithm for generating realizations of the MJP yields an alternative approach through sampling a sufficiently large amount of trajectories. Although this method is vital in the analysis of Markov jump processes, it can be computationally very demanding. The run time increases firstly with an increasing number of different possible transition steps and with increasing transition rates leading to a high amount of transitions which need to be sampled. This problem gets worse when very different time scales are involved leading to prohibitively large simulation times for sampling rare transition events.

The third possible approach for propagating probability distributions is based on a system of differential equations for the probability density for each state and the infinitesimal generator as its rate matrix. To this end the following results must be first stated:

Theorem 2. *Given a Markov process with a transition semigroup* $\mathbf{P}(t), t \geq 0$ *and an infinitesimal generator* $\mathbf{A}$*. Then the limit*

$$\frac{d}{dt}p(t,x,y) = \lim_{h\to 0} \frac{p(t,x+h) - p(t,x,y)}{h}$$

for all $x, y \in S$ *exists and is given by the* ***Kolmogorov backward equation***

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{A}\mathbf{P}(t).$$

If in addition it holds true that

$$\sum_{y\in S} -p(t,x,y)a(y,y) < \infty$$

for all $t \geq 0$ *and* $x \in S$ *then the above limit is equivalently given by the* ***Kolmogorov forward equation***

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{A}.$$

A proof for both equations follows immediately from the time-limit (2.10) of the infinitesimal generator matrix.

The Kolmogorov backward equation can be combined with equation (2.8) for probability transport using the transition matrix to yield

$$\begin{aligned} P_0^T \frac{d}{dt}\mathbf{P}(t) &= P_0^T \mathbf{P}(t)\mathbf{A} \leftrightarrow \\ \frac{d}{dt}\left(P_0^T \mathbf{P}(t)\right) &= P_0^T \mathbf{P}(t)\mathbf{A}, \end{aligned}$$

which yields the **Master Equation**, propagating the probability densities function (PDF) in time:

$$\begin{aligned} \frac{d}{dt}P(t)^T &= P(t)^T \mathbf{A} \leftrightarrow \\ \frac{d}{dt}P(t) &= \mathbf{A}^T P(t), \end{aligned}$$

with the general solution based on the matrix exponential

$$P(t)^T = P_0^T \exp(\mathbf{A}^T t).$$

This representation of the probability transport is only constrained to MJPs with a finite state space since the matrix form of the infinitesimal limit in theorem 2 is not defined in the case of an infinite state space. However the Master Equation can also be expressed elementwise for each state $x \in \mathrm{S}$

$$\frac{d}{dt}P(x,t) = \sum_{\substack{y \in \mathrm{S} \\ y \neq x}} P(y,t)a(y,x) - P(x,t)a(x,y), \tag{2.15}$$

where $P(x,t) := \mathbb{P}(X = x, t)$ is a column vector assigning probabilities to each state in S. It is a significant virtue of the Master Equation to enable expressing the probability transport also for processes with an infinite state space. However it is still not guaranteed that this equation can be solved.

2.2.2 Stochastic reaction kinetics

A rigorous approach to modeling (bio-)chemical reaction systems is based on a discrete formulation of molecular species numbers reacting with each other through a set of reaction channels. If the reactions are viewed as jumps of the system on a discrete state space, given by possible permutations of molecular numbers, then such a system can be described by a Markov jump process. To this end, following assumptions must be fulfilled: the system must be well-stirred i.e. the probability of finding any molecule within a subvolume δV is uniformly distributed according to $\delta V/V$, where V denotes the total volume. Furthermore the system must be at thermal equilibrium, i.e. the velocity of each molecule is a random variable determined by the Maxwell-Boltzmann distribution [26].

Let a system fulfilling above assumptions and consisting of n distinct chemical species $\mathbf{x} = \{X_1, X_2, ..., X_n\}$ be given which can react with each other, amounting to m possible biochemical reactions. According to chemical reaction kinetics, each reaction can be considered as a conversion of certain amount of educts to a certain amount of products. Assuming that reaction $j \in 1..m$ consumes $\mathbf{N}_j = \{N_{1j}, N_{2j}, ..., N_{nj}\}^T$ molecules and produces $\mathbf{M}_j = \{M_{1j}, M_{2j}, ..., M_{nj}\}^T$ molecules of species $x_1, x_2, ..., x_n$, the chemical reaction j can be described by

$$N_{1j} + N_{2j}, ..., N_{nj} \xrightarrow{c_j} M_{1j} + M_{2j}, ..., M_{nj}, \tag{2.16}$$

where c_j denotes the corresponding reaction rate constant with units events per time unit (e.g. s^{-1}).

According to the law of mass action [32], the probability that a reaction occurs is proportional to the length of the time interval and the number of the possible ways that the substrate molecules can interact with each other. In line with this law a rigorous derivation in the context of stochastic dynamics was given by Gillespie [26]. It states that the probability that a reaction j occurs in the time interval $[t, t + dt]$ is given by $w_j(\mathbf{x})dt + o(dt)$ and that the probability that more than one reaction occurs is proportional to $o(dt)$. Thus in an infinitesimally small time interval at most one reaction can take place. The function $w_j(\mathbf{X})$ is referred to as the **propensity** for the occurrence of j-th reaction and is given by

$$w_j(\mathbf{X}) = \begin{cases} c_j \prod_{i=1}^{n} \frac{X_i!}{N_{ij}!(X_i - N_{ij})!}, & \text{if } X_i \geq N_{ij} \text{ for all } i = 1, .., n, \\ 0, & \text{otherwise.} \end{cases} \tag{2.17}$$

This equation results as a product of the reaction rate c_j and the number of all possible ways that the educt molecules can react with each other.

A comparison of above equation with eq. (2.10) suggests an interpretation of the propensity function as an infinitesimal jump rate. Assume that the reaction j makes the Markov process jump from state q to state r, which correspond to the species vectors $\mathbf{x} = \{X_1 = x_1, X_2 = x_2, ..., X_n = x_n\}$ and $\mathbf{y} = \{X_1 = y_1, X_2 = y_2, ..., X_n = y_n\}$, respectively. If the state space S is finite, then the propensity function $w_j(\mathbf{X})$ gives rise to the entries (q, r) of the corresponding infinitesimal generator matrix by

$$a(q, r) = w_j(\mathbf{x}), \tag{2.18}$$

and accordingly the diagonal entries are given by

$$a(q, q) = - \sum_{r,q \in S, r \neq q,} a(q, r).$$

Let $\mathbf{v}_j = \mathbf{M}_j - \mathbf{N}_j$ be the net change of molecular species caused by reaction j. Then the Master Equation (2.15) can be reformulated as

$$\frac{d}{dt} P(\mathbf{x}, t) = \sum_{j=1}^{m} P(\mathbf{x} - \mathbf{v}_j, t) w_j(\mathbf{x} - \mathbf{v}_j) - P(\mathbf{x}, t) w_j(\mathbf{x}). \tag{2.19}$$

Equation (2.19) is referred to as the **the Chemical Master Equation** (CME) [26] and it describes the evolution of the PDF of a Markov jump process whose infinitesimal jump rates result from the reaction propensities $w_j(\mathbf{X})$.

2.3 Moment dynamics and macroscopic limit equations

2.3.1 Time evolution of moments in one dimension

Often the exact solution of the Master Equation is not available and sampling using the SSA algorithm [24] might be computationally too expensive. If one is not interested in the entire probability density but rather in certain properties, such as its mean and variance, the differential equations for the statistical moments are particularly helpful. They yield significant information about the time evolution of the properties of the probability density with lower computational costs. Thus, the solution of the CME, which is an ODE system containing an evolution equation of the probability distribution in each state, is reduced to the solution of an ODE for the PDF moments for each species of the system.

The derivation of an equation for the k-th moment of the sought probability density is based on its corresponding Master Equation. For brevity, we derive it in the following for a one-dimensional Markov jump process, cf. [25]. The k-th moment $\langle X^k \rangle$ of the probability density[1] $P(x,t)$ with $x \in \mathrm{S}$ is given by $\sum_{x\in \mathrm{S}} x^k P(x,t)$, where $P(x,t) := \mathbb{P}(X = x, t)$. Accordingly, by taking the time derivative and subsequently using the corresponding Master Equation (2.19) for a single chemical species X, one obtains

$$\begin{aligned}\frac{d}{dt}\langle X^k\rangle &= \frac{d}{dt}\sum_{x\in \mathrm{S}} x^k P(x,t) = \sum_{x\in \mathrm{S}} x^k \frac{\partial}{\partial t} P(x,t) \\ &= \sum_{x\in \mathrm{S}}\sum_{j=1}^{m} x^k P(x-v_j,t) w_j(x-v_j) - \sum_{x\in \mathrm{S}}\sum_{j=1}^{m} x^k P(x,t) w_j(x).\end{aligned}$$

In the last equation the summation index in the first infinite sum can be changed from x to $x - v_j$:

$$\begin{aligned}\frac{d}{dt}\langle X^k\rangle &= \sum_{x\in \mathrm{S}}\sum_{j=1}^{m} (x+v_j)^k P(x,t) w_j(x) - \sum_{x\in \mathrm{S}}\sum_{j=1}^{m} x^k P(x,t) w_j(x) \\ &= \sum_{x\in \mathrm{S}}\sum_{j=1}^{m} \left[(x+v_j)^k - x^k\right] P(x,t) w_j(x)\end{aligned}$$

[1] The notation $\langle X \rangle$ will from now on refer to the expected value of the random variable X.

Now the term $(x + v_j)^k$ can be expanded using the binomial coefficient

$$\begin{aligned}\frac{d}{dt}\langle X^k\rangle &= \sum_{x\in \mathrm{S}}\sum_{j=1}^{m}\left[\sum_{i=1}^{k}\binom{k}{i}v_j^i x^{k-i}\right]P(x,t)w_j(x)\\ &= \sum_{i=1}^{k}\binom{k}{i}\sum_{x\in \mathrm{S}}x^{k-i}\sum_{j=1}^{m}v_j^i w_j(x)P(x,t)\end{aligned}$$

Using the definition of expectation, the above equation finally yields

$$\frac{d}{dt}\langle X^k\rangle = \sum_{i=1}^{k}\binom{k}{i}\left\langle X^{k-i}\sum_{j=1}^{m}v_j^i w_j(X)\right\rangle. \tag{2.20}$$

Eq. (2.20) is an ODE for the k-th moment of the Chemical Master Equation (2.19) in a one-species system. It depends on the first $k-1$ moments multiplied by the polynomial, which results from the product of propensities and stoichiometric state changes. If the propensity $w_j(X)$ is at most a linear function of x then the highest order of x in the equation is k. Thus a k-dimensional system of ODEs has to be solved simultaneously in order to obtain the closed solution for the first k moments. However, if $w_j(X)$ has an order higher than one (e.g. quadratic in x) then the moment $k+1$ appears and an additional equation for its time evolution has to be solved. This equation, in turn, includes the moment $k+2$ and so forth. Thus a closed solution of the system is not possible if the propensity $w_j(X)$ of any reaction j is not constant or linear.

Example: mono- and bimolecular reactions

As an example, a comparison of two reaction systems with a single species can be drawn. The first system is given by a synthesis reaction

$$\mathrm{X} \xrightarrow{c_1} 2\mathrm{X},$$

which is referred to as a *monomolecular reaction*, since it consumes one molecule of the species X. The second system is described by a degradation reaction

$$2\mathrm{X} \xrightarrow{c_2} \mathrm{X},$$

which is a *bimolecular reaction*. According to equation (2.17), the propensity functions of the two systems result in $w_1(X) = c_1 X$ and $w_2(X) = \frac{1}{2}c_2 X(X-1)$. If one is interested in the time-evolution of the mean of the probability distribution of the two systems, eq. (2.20) for $k = 1$ can be derived, resulting in

$$\frac{d}{dt}\langle X\rangle = \left\langle\sum_{j=1}^{m}v_j w_j(X)\right\rangle. \tag{2.21}$$

For the first reaction system one obtains

$$\frac{d\langle X\rangle}{dt} = c_1\langle X\rangle,$$

where the linearity of the propensity function yields an ODE with a closed solution given by $\langle X(t)\rangle = \langle X_0\rangle \exp(c_1 \cdot t)$.

The time evolution of the mean in the second reaction system results in

$$\begin{aligned} \frac{d\langle X\rangle}{dt} &= \frac{1}{2}c_2\langle X(X-1)\rangle, \\ &= \frac{1}{2}c_2\left(\langle X^2\rangle - \langle X\rangle\right). \end{aligned}$$

Since in general $\langle Y^2\rangle \neq \langle Y\rangle\cdot\langle Y\rangle$, the equation for the second moment must be solved simultaneously in order to solve the equation for the mean. As outlined above, the second moment equation will however include the third moment, giving rise to an infinite series of higher order moments. Thus in contrast to the monomolecular reaction system, a closed-form solution is not available for a bimolecular system and the evolution of the mean can not be computed exactly. In this case, for instance, an approximation can be found by truncating higher order moments, as described in [21].

2.3.2 Time evolution of moments in multiple dimensions

The above equations for the PDF moments of a single-species reaction system can also be generalized to systems with arbitrary number of species. As for instance stated in [21], the multidimensional first moment, describing the evolution of the mean of the PDF is given as the solution of the ODE

$$\frac{d\langle \mathbf{X}\rangle}{dt} = \mathbf{S}\cdot\langle \mathbf{w}(\mathbf{X})\rangle, \tag{2.22}$$

where

$$\mathbf{S}_{ij} = M_{ij} - N_{ij},$$

is the *stoichiometric matrix* containing the jump sizes induced by individual reactions and the propensity vector is given by $\mathbf{w}(\mathbf{X}) = [w_1(\mathbf{X}), w_2(\mathbf{X}), ..., w_m(\mathbf{X})]$. Equation (2.22) is a generalization of eq. (2.21) for an arbitrary number of chemical species.

Accordingly, the time evolution of the multidimensional second moment, $\mathbf{\Sigma} = \mathbf{X}\cdot\mathbf{X}^T$, is given by

$$\frac{d\mathbf{\Sigma}}{dt} = \langle \mathbf{S}\cdot\mathbf{w}(\mathbf{X})\cdot\mathbf{X}^T\rangle + \langle \mathbf{X}\cdot[\mathbf{S}\cdot\mathbf{w}(\mathbf{X})]^T\rangle + \langle \mathbf{D}(\mathbf{X})\rangle, \tag{2.23}$$

where the positive-definite matrix $\mathbf{D}\in\mathbb{R}^{n\times n}$ is referred to as the *diffusion matrix*, and it is defined as

$$\mathbf{D}(\mathbf{X}) = \mathbf{S}\cdot\mathrm{diag}(\mathbf{w}(\mathbf{X}))\cdot\mathbf{S}^T.$$

The matrix of second moments $\mathbf{\Sigma}$ corresponds to a non-centralized covariance. Thus its diagonal entries yield $\mathbf{\Sigma}_{ii} - (\langle X_i\rangle)^2$ the variance in the i-th component (i.e. chemical species) of the probability distribution of the Markov jump process. Accordingly,

$\mathbf{\Sigma}_{ij} - \langle X_i \rangle \langle X_j \rangle$ is the covariance between the i-th and the j-th component, where $i, j \in [1, ..., n]$. Analogously to a single-species system, the ODEs (2.22) and (2.23) can be solved if all reaction propensities are of zero-th or first order in $\mathbf{X}$, cf. [21, 75].

2.3.3 Macroscopic reaction rate equations and the relation to the stochastic chemical kinetics

A classical approach to modeling reaction systems is based on a deterministic description with continuous variables rather than integral ones. This is achieved by introducing a scaling parameter Ω which is proportional to the system size e.g. the volume. The macroscopic variable $\bar{\mathbf{X}} = \{\bar{X}_1, \bar{X}_2, ..., \bar{X}_n\}$ and the stochastic variable $\mathbf{X}$ are then related to each other through

$$\bar{\mathbf{X}} = \frac{\mathbf{X}}{\Omega}. \tag{2.24}$$

If the scaling parameter is assumed to be the volume of the system then the units of the macroscopic variable are in molecules or mole per unit volume, i.e. concentration (e.g. $mol \cdot L^{-1}$).

In macroscopic models of chemical kinetics the reaction rate function $a_j(\bar{\mathbf{X}})$ is derived from the law of mass action [32] by assuming proportionality of the size of the reaction rates to the concentration of educt species $\bar{X}_i$ and the reaction stoichiometries. The rate of j-th reaction is thus given by

$$a_j(\bar{\mathbf{X}}) = k_j \prod_{i=1}^{n} \bar{X}_i^{N_{ij}}, \tag{2.25}$$

where k_j denotes the basic reaction rate constant with its units given in events per unit of time and volume (e.g. $s^{-1}L^{-1}$). Note that in contrast to the stochastic propensity function (2.17), the macroscopic reaction rate does not consider the detailed configurations of reacting molecular species. Thus e.g. for a bimolecular reaction ($N_{ij} = 2$) the stochastic propensity is given by $w(X) = c\frac{1}{2}X(X-1)$ while the macroscopic rate is $a(\bar{X}) = k\bar{X}^2$, where c and k denote the stochastic and macroscopic reaction constants, respectively.

As a second consequence of the law of mass action, the rate of change of the concentration of the species $\bar{X}_i$ by reaction j is proportional to the molecular difference caused by this reaction, i.e. $\mathbf{S}_{ij} = M_{ij} - N_{ij}$. Combining this with the reaction rate (2.25) yields an ODE for the time evolution of the species vector $\bar{\mathbf{X}}$

$$\frac{d\bar{\mathbf{X}}}{dt} = \mathbf{S} \cdot \mathbf{a}(\bar{\mathbf{X}}) = f(\bar{\mathbf{X}}), \tag{2.26}$$

where $\mathbf{a}(\bar{\mathbf{X}}) = [a_1(\bar{\mathbf{X}}), a_2(\bar{\mathbf{X}}), ..., a_n(\bar{\mathbf{X}})]^T$ is the vector-valued reaction rate. The function $f(\bar{\mathbf{X}})$ can be considered as a vector field describing the gradient of $\bar{\mathbf{X}}$ at time t and its integral yields the evolution of its concentration in time, $\bar{\mathbf{X}}(t) =$

$\int_{t_0}^{t} f(\bar{\mathbf{X}}(\hat{t}))d\hat{t}$, given an initial value $\bar{\mathbf{X}}(t_0) = \bar{\mathbf{X}}_0$.

Although the analytical solution of the macroscopic equation (2.26) can only be found in a limited number of situations (e.g. $f(\bar{\mathbf{X}})$ linear), in general a numerical solution can be obtained by standard methods [19]. Notably, it is by far less costly than the solution of the CME since the dimensionality of the macroscopic equation is determined by the number of chemical species, while the CME is an ODE system having the size of the whole state space S, which, if finite, can be prohibitively large and numerically intractable.

The difference between the stochastic propensity function and the deterministic reaction rate gives a first intuitive insight into the accuracy of the two approaches in dependence of the system size. While in a "small" system with a few molecules the exact configurations of possible molecular collisions may have a significant impact on the dynamics, in a "large" system these specifics become negligible and can be approximated by a proportionality factor. Also, if the number of molecules of species i is large, the relative jumps $\mathbf{S}_{ij}$ in molecular numbers caused by reaction j become negligible. However if this condition does not hold, the discretely occurring stochastic reaction firings dominate the dynamics.

In order to formalize the above statements and rigorously assess the role of the system size, the stochastic propensity (2.17) can be formulated in terms of the concentration vector $\bar{\mathbf{X}}$ scaled by Ω. For clarity, the original propensity function can be first restated as

$$\begin{aligned} w_j(\mathbf{X}) &= c_j \prod_{i=1}^{n} \frac{X_i!}{N_{ij}!(X_i - N_{ij})!} \\ &= c_j \frac{1}{\prod_{i=1}^{N} N_{ij}!} \prod_{i=1}^{n} \prod_{h=0}^{N_{ij}-1} (X_i - h), \end{aligned}$$

if $X_i \geq N_{ij}$ for all $i \in [1, 2, .., n]$. The corresponding Ω-scaled propensity function now reads (cf. [48]):

$$w_j(\bar{\mathbf{X}}) = \frac{w_j(\bar{\mathbf{X}}\Omega)}{\Omega} = \frac{c_j \Omega^{|N_j|-1}}{\prod_{i=1}^{N}(N_{ij}!)} \prod_{i=1}^{n} \prod_{h=0}^{N_{ij}-1} (\bar{X}_i - \frac{h}{\Omega}), \tag{2.27}$$

if $\bar{X}_i \geq \frac{N_{ij}}{\Omega}$ for all $i = 1, ..., n$ and $|N_j| = \sum_i^n N_{ij}$. It describes the reaction propensity in terms of molecular concentrations rather than discrete numbers. The correspondence of the Ω-scaled propensity function and the macroscopic reaction rate becomes obvious if the following relation between the basic reaction constants is invoked:

$$k_j = \frac{c_j \ \Omega^{|N_j|-1}}{\prod_{i=1}^{N}(N_{ij}!)}. \tag{2.28}$$

Using this relation, eq. (2.27) can be restated in terms of the basic macroscopic reaction constant instead of the stochastic one:

$$w_j(\bar{\mathbf{X}}) = \frac{w_j(\bar{\mathbf{X}}\Omega)}{\Omega} = k_j \prod_{i=1}^{n} \prod_{h=0}^{N_{ij}-1} (\bar{X}_i - \frac{h}{\Omega}). \tag{2.29}$$

The last two equations show that if each reaction is of order one i.e. ($|N_j| = 1$) for all $j = [1, ..., m]$, then the stochastic and the deterministic rates are equivalent and the solution $\bar{\mathbf{X}}$ of the macroscopic equation (2.26) corresponds to the mean $\langle \mathbf{X} \rangle$ of the Markov jump process PDF, given by eq. (2.22) since

$$\begin{aligned} \frac{d\langle \mathbf{X} \rangle}{dt} &= \mathbf{S}\langle \mathbf{w}(\mathbf{X}) \rangle = \mathbf{S}\langle \mathbf{N}^{\mathrm{T}}\mathbf{k}\mathbf{X} \rangle = \mathbf{S}\mathbf{N}^{\mathrm{T}}\mathbf{k}\langle \mathbf{X} \rangle \\ &= \mathbf{S} \cdot \mathbf{a}(\langle \mathbf{X} \rangle), \end{aligned} \tag{2.30}$$

where $\mathbf{N}_{ij}$ denotes the number of molecules of species i consumed by reaction j, cf. eq. (2.25). If some reactions are of order zero, i.e. $|N_j| = 0, j \in [1, .., m]$, then this equivalence also holds, however the reaction rates have to be converted using the factor Ω, cf. eq. (2.28). In these two special cases, due to the order ≤ 1 of all reactions, $\mathbf{N}$ is a sparse matrix where each column contains at most a 1. Moreover, the vector $\mathbf{k} = [k_1, k_2, ..., k_m]^{\mathrm{T}}$ contains the basic macroscopic reaction constants.

Equation (2.30) implies that if under described conditions the initial values of the two equations (2.22) and (2.26) are chosen to be equal, then their solutions are equivalent. At equilibrium states this also holds for reactions of order higher than 1, which can be shown by linearizing the propensity function at the macroscopic fixed points. Furthermore, from eq. (2.28) the conversion between macroscopic and stochastic reaction constants can be deduced:

1. $c_j = \Omega k_j$ if $|N_j| = 0$ (0-th order),
2. $c_j = k_j$ if $|N_j| = 1$ (1-st order),
3. $c_j = \frac{2k_j}{\Omega}$ if $|N_j| = 2$ (2-nd order).

2.3.4 The large volume limit of the CME and the Linear Noise Approximation

With the exceptions discussed above, in general, the solution of the macroscopic equation is not equivalent to the mean of the stochastic system. However, as shown in eq. (2.29), an increasing Ω lets the scaled jumps N_{ij}/Ω vanish, leading to a deviation of $O(\Omega^{-1})$ between the macroscopic rate and the stochastic propensity, cf. [48]. In the thermodynamic limit, i.e. $\Omega \to \infty$ and $X \to \infty$, one expects the concentration X/Ω to remain constant while the relative jumps N_{ij}/Ω become negligible. In fact, it was shown by T. Kurtz that within this limit the mean solution of the Chemical Master Equation approaches the solution of the macroscopic equation [43].

Various methods for approximating the solution of the CME resulted from considering its limit behaviour. These include a decomposition of the reaction system into stochastic and macroscopic components [49] or an approximation of the state-discrete process described by the CME using a state-continuous process, given by the Chemical Langevin equation (CLE) [27]. The latter is a powerful method reducing the dimensionality of the CME while keeping its stochastic nature. However the CLE does not describe how its approximation accuracy depends on system size. Van Kampen introduced a rigorous method for approximating the CME by adding perturbation terms to the corresponding macroscopic solution [77]. To its lowest order this expansion yields a linear Fokker Planck equation (FPE) with a Gaussian solution. Its sample paths are given by a stochastic differential equation, similar to the CLE. This method is referred to as the *Linear Noise Approximation*. In the following an outline of this method is given and an application to HIV dynamics is discussed.

The Linear Noise Approximation is based on the idea of replacing the space-discrete stochastic variable $\mathbf{X}$ by the corresponding macroscopic variable $\bar{\mathbf{X}}$, (see eq. 2.24) and a new space-continuous stochastic variable $\boldsymbol{\xi}$, describing the fluctuations around $\bar{\mathbf{X}}$. This change of variables is conducted by a linear ansatz where the fluctuations are scaled as a square root of system size

$$\mathbf{X} = \Omega\bar{\mathbf{X}} + \Omega^{1/2}\boldsymbol{\xi}, \tag{2.31}$$

where $\boldsymbol{\xi} = [\xi_1, \xi_2, ..., \xi_n]$ is a vector of stochastic fluctuations in the dimension of each chemical species. Its probability law is described by a new probability distribution $\Pi(\boldsymbol{\xi}, t)$ instead of $P(\mathbf{X}, t)$. The scaling $\Omega^{1/2}$ of the stochastic variable can be justified in this ansatz by observing that in the thermodynamic limit the difference between the macroscopic equation and Chemical Langevin equation (describing a state-continuous stochastic process) is proportional to $\Omega^{1/2}$ [28].

Due to the change of the probability function $P(\mathbf{X}, t) \to \Pi(\boldsymbol{\xi}, t)$, the time derivative of the original probability distribution in terms of the new one is given as

$$\frac{\partial P(\mathbf{X}, t)}{\partial t} = \frac{\partial \Pi(\boldsymbol{\xi}, t)}{\partial t} - \sum_{i=1}^{n} \Omega^{1/2} \frac{\partial \Pi(\boldsymbol{\xi}, t)}{\partial \xi_i} \frac{\partial \bar{X}_i}{\partial t}, \tag{2.32}$$

where the chain rule is used to obtain the partial derivatives of $\Pi(\boldsymbol{\xi}, t)$. Furthermore, it is used that the discrete stochastic variable $\mathbf{X}$ is fixed w.r.t. time, leading to the following relation (see eq. (2.31)):

$$\begin{aligned} 0 \quad &= \quad \frac{\partial X_i}{\partial t} = \Omega \frac{\partial \bar{X}_i}{\partial t} + \Omega^{1/2} \frac{\partial \xi_i}{\partial t} \\ &\implies \frac{\partial \xi_i}{\partial t} = -\Omega^{1/2} \frac{\bar{X}_i}{\partial t}. \end{aligned}$$

Equation (2.32) yields the left hand side of the evolution equation for the probability distribution $\Pi(\boldsymbol{\xi}, t)$. In order to derive its right-hand side, the Chemical Master Equation (2.19) can be restated as

$$\frac{\partial}{\partial t}P(\mathbf{X},t) = \Omega\sum_{j=1}^{m}\left[\left(\prod_{i=1}^{n}\mathbf{E}_i^{-S_{ij}}\right)-1\right]w_j\left(\frac{\mathbf{X}}{\Omega}\right)P(\mathbf{X},t),$$

where the step operator notation is used in order to express the molecular jumps

$$\mathbf{E}_i^h f(\mathbf{X}) = f(...,X_i+h,...),$$

and the scaled propensity function is incorporated, which leads to the pre-factor Ω, cf. eq. (2.29). In order to assess the dependence of the step operator on the system size Ω, it can be expanded in the limit $\Omega\to\infty$, around the point $h/\sqrt{\Omega}=0$. By using eq. (2.31), this yields

$$\begin{aligned}\mathbf{E}_i^h f(\mathbf{X}) &= f(...,X_i+h,...) = f(...,\Omega\bar{X}_i+\sqrt{\Omega}(\xi_i+\frac{h}{\sqrt{\Omega}}),...)\\ &\approx f(\mathbf{X})+\frac{h}{\sqrt{\Omega}}\frac{\partial}{\partial\xi_i}+\frac{h^2}{2\Omega}\frac{\partial^2}{\partial\xi_i^2}+\dots\ .\end{aligned}$$

Similarly, the Ω-scaled propensity function can be expanded at the limit $\Omega\to\infty$ and $X\to\infty$, i.e. X/Ω constant. As shown by eq. (2.29), to the lowest order the propensity function is equivalent to the macroscopic reaction rate $a(\bar{\mathbf{X}})$ as

$$w_j(\bar{\mathbf{X}}) = a_j(\bar{\mathbf{X}})+\frac{1}{\sqrt{\Omega}}\sum_{i=1}^{n}\frac{\partial a_j(\bar{\mathbf{X}})}{\partial\bar{X}_i}\xi_i+\frac{1}{\sqrt{2\Omega}}\sum_{i=1}^{n}\frac{\partial^2 a_j(\bar{\mathbf{X}})}{\partial\bar{X}_i^2}\xi_i+\dots\ .$$

Using eq. (2.32) as the left-hand side and the Master Equation with expanded step operator and propensities as the right-hand side, yields an evolution equation for the probability of the new stochastic variable $\boldsymbol{\xi}$

$$\begin{aligned}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial t}-\sum_{i=1}^{n}\Omega^{1/2}\frac{\partial\Pi(\boldsymbol{\xi},t)}{\partial\xi_i}\frac{\partial\bar{X}_i}{\partial t} &=\\ \Omega\sum_{j=1}^{m}\left[\left(1+\frac{h}{\sqrt{\Omega}}\frac{\partial}{\partial\xi_i}+\frac{h^2}{2\Omega}\frac{\partial^2}{\partial\xi_i^2}...\right)-1\right]w_j(\bar{\mathbf{X}})\Pi(\boldsymbol{\xi},t),&\end{aligned}\tag{2.33}$$

where the expanded form of $w_j(\bar{\mathbf{X}})$ is omitted for brevity of notation. Note also that in the Chemical Master Equation the probability distribution of the original stochastic variable $P(\mathbf{X},t)$ is replaced by $\Pi(\boldsymbol{\xi},t)$

Equation (2.33) describes the behaviour of the state-continuous stochastic variable $\boldsymbol{\xi}$ whose probability distribution approximates the probability of the original Markov jump process with increasing number of Ω-terms included before truncation. Truncating this equation at order $\Omega^{1/2}$, yields

$$\frac{d\bar{X}_i}{dt}\frac{\partial\Pi}{\partial\xi_i} = [\mathbf{S}\cdot\mathbf{w}(\bar{\mathbf{X}})]_i\frac{\partial\Pi}{\partial\xi_i},$$

which is satisfied if the original macroscopic equation (2.26) holds

$$\frac{d\bar{X}_i}{dt} = [\mathbf{S} \cdot \mathbf{w}(\bar{\mathbf{X}})]_i \equiv f_i(\bar{x}).$$

To the next order Ω^0 the expansion yields

$$\frac{\partial \Pi}{\partial t} = -\sum_{i,j} \Gamma_{ij} \partial_i (\xi_j \Pi) + \frac{1}{2} \sum_{i,j} D_{ij} \partial_{ij} \Pi, \tag{2.34}$$

where $\Gamma_{ij}(t) = \frac{\partial f_i}{\partial \bar{X}_j}$ is the Jacobi matrix of the macroscopic function $f(\bar{\mathbf{X}})$ and $\mathbf{D} = \mathbf{S} \cdot \mathrm{diag}[\mathbf{w}(\bar{\mathbf{X}})] \cdot \mathbf{S}^T$ corresponds to the diffusion matrix in equation (2.23). This is a linear (in $\boldsymbol{\xi}$) partial differential equation (PDE) describing the evolution of the probability distribution of fluctuations around the macroscopic trajectory and it is referred to as *Fokker-Planck equation* (FPE). This order of truncation thus yields an n-dimensional PDE, where n is the number of chemical species, approximating the CME, given by a large system of ODEs having the size of the state space S.

Computation of the first two moments of the FPE can be conducted by multiplying eq. (2.34) by $\boldsymbol{\xi}$ and $\boldsymbol{\xi} \cdot \boldsymbol{\xi}^T$. This yields for the mean $\langle \boldsymbol{\xi} \rangle$ and non-centralized covariance $\boldsymbol{\Sigma}$ of stochastic fluctuations, cf. [75]:

$$\frac{d\langle \boldsymbol{\xi} \rangle}{dt} = \boldsymbol{\Gamma} \langle \boldsymbol{\xi} \rangle,$$
$$\frac{d\boldsymbol{\Sigma}}{dt} = \boldsymbol{\Gamma} \cdot \boldsymbol{\Sigma} + \boldsymbol{\Sigma} \cdot \boldsymbol{\Gamma}^T + \boldsymbol{D}.$$

The first equation suggests that if the initial condition of $\boldsymbol{\xi}$ is zero, then the mean $\langle \boldsymbol{\xi} \rangle$ will remain zero forever. This implies that the solution of the stochastic process is concentrated around the macroscopic trajectory, due to ansatz (2.31). The second equation is equivalent to the eq. (2.23) for the second moment of the original Markov jump process if all propensities are linear in X. In a general non-linear case $\boldsymbol{\Sigma}$ can be computed at fixed points by linearization of $f(\bar{\mathbf{X}})$, cf. [75].

Notably, the Fokker-Planck equation (2.34) is equivalent to a stochastic differential equation describing the time-evolution of sample paths of the underlying continuous-time process [54]

$$d\mathbf{X} = f(\mathbf{X})dt + \mathbf{D}^{\frac{1}{2}} d\mathbf{B}_t, \tag{2.35}$$

where $d\mathbf{B}_t$ denotes multidimensional Brownian motion and $\mathbf{D}^{\frac{1}{2}}$ is a matrix satisfying

$$\mathbf{D}^{\frac{1}{2}} \cdot \left[\mathbf{D}^{\frac{1}{2}}\right]^T = \mathbf{D}.$$

A numerical solution of this equation can be used in order to obtain sample paths from the probability distribution $\Pi(\boldsymbol{\xi}, t)$, which is usually significantly more efficient than sampling from the Master Equation of the original process [42].

Example: Linear Noise Approximation of the CME for HIV-host dynamics

Consider a system of HIV-particles V^* which infect immune cells, thus generating infected cells T^*. In turn, infected cells give rise to new viruses. Assuming constant linear death propensities for both species the described scenario gives rise to the following reaction system:

$$\begin{aligned} \mathrm{V}^* &\xrightarrow{c_1} \varnothing, \\ \mathrm{V}^* &\xrightarrow{c_2} \mathrm{T}^*, \\ \mathrm{T}^* &\xrightarrow{c_3} \varnothing, \\ \mathrm{T}^* &\xrightarrow{c_4} \mathrm{T}^* + \mathrm{V}^*. \end{aligned} \tag{2.36}$$

This set of reactions corresponds to the following propensity vector

$$\mathbf{w}(\mathbf{X}) = \{c_1 \cdot \mathrm{V}^*,\ c_2 \cdot \mathrm{V}^*, c_3 \cdot \mathrm{T}^*, c_4 \cdot \mathrm{T}^*\}^T,$$

where $\mathbf{X} = \{X_1, X_2\}^T := \{\mathrm{T}^*, \mathrm{V}^*\}$. The corresponding CME results in, cf subsection 2.2.2:

$$\begin{aligned} \frac{\partial P(x_1, x_2, t)}{\partial t} &= c_1(x_2+1)\ P(x_1, x_2+1, t) \\ &+ c_2(x_2+1)\ P(x_1-1, x_2+1, t) \\ &+ c_3(x_1+1)\ P(x_1+1, x_2, t) \\ &+ c_4 x_1\ P(x_1, x_2-1, t) \\ &- ([c_1+c_2]x_2 + [c_3+c_4]x_1)\ P(x_1, x_2, t), \end{aligned} \tag{2.37}$$

where it is defined $P(x_1, x_2, t) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \text{time} = t)$. The macroscopic dynamics of the corresponding continuous variable $\bar{\mathbf{X}}$ is described by the ODE

$$\frac{d\bar{\mathbf{X}}}{dt} = \mathbf{S} \cdot \mathbf{a}(\bar{\mathbf{X}}) = f(\bar{\mathbf{X}}). \tag{2.38}$$

where the macroscopic function is given by

$$f(\bar{\mathbf{X}}) = \begin{pmatrix} c_2 \cdot \dfrac{X_2}{\Omega} - c_3 \dfrac{X_1}{\Omega} \\ -(c_1+c_2)\dfrac{X_2}{\Omega} - c_4 \dfrac{X_1}{\Omega} \end{pmatrix},$$

and the stoichiometric matrix $\mathbf{S}$ is

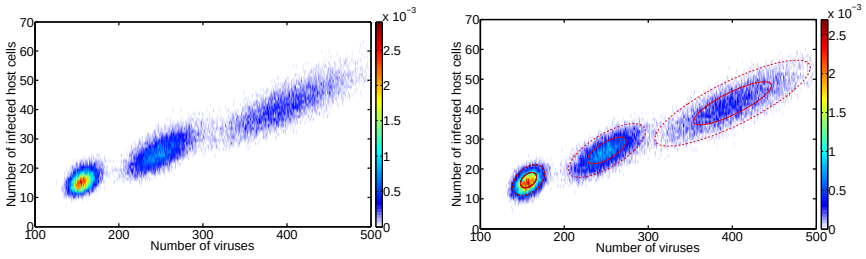
$$\mathbf{S} = \begin{pmatrix} 0 & 1 & -1 & 0 \\ -1 & -1 & 0 & 1 \end{pmatrix}.$$

Accordingly, the Jacobi and the diffusion matrix in the Fokker-Planck equation (2.34), are given as

$$\Gamma = \begin{pmatrix} -c_3 & c_2 \\ c_4 & -(c_1+c_2) \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} c_2\left(\frac{X_2}{\Omega}\right) + c_3\left(\frac{X_1}{\Omega}\right) & -c_2\left(\frac{X_2}{\Omega}\right) \\ -c_2\left(\frac{X_2}{\Omega}\right) & (c_1+c_2)\frac{X_2}{\Omega} + c_4\frac{X_1}{\Omega} \end{pmatrix}.$$

In the following we compared the dynamics of the Markov jump process induced by the HIV-host interaction model with the approximating continuous stochastic process. To this end we have sampled the corresponding probability distributions at three discrete time points $t \in \{0.5, 1, 1.5\}$ (days). The parameter values were chosen to be $c_1 = 0.01, c_2 = 0.1, c_3 = 0.01, c_4 = 10$, in units $(\text{days})^{-1}$. The result is shown in figure 2.3. In order to obtain sample paths of the Markov jump process, we have generated 10^4 trajectories using Gillespie's SSA-algorithm. It is shown in fig. 2.3 a, where the three probability clouds are from left to right the solutions of the CME at the three discrete time points 0.5, 1 and 1.5, respectively.



(a) (b)

Figure 2.3: **Time evolution of the probability distribution of the HIV-host-system sampled at three discrete time points.**
The probability distribution was sampled at discrete time points $t = 0.5$, $t = 1$ and $t = 1.5$ (days) corresponding to the three distribution clouds, from left to right. **(a)** Gillespie's SSA-algorithm was used to sample the evolution of the probability described by the Master Equation (2.37). 10^4 sample paths were generated for each of the three time points. **(b)** The SDE (2.35) was solved numerically using the Euler-Maruyama integration method to yield 10^4 sample trajectories. In addition, two ellipses are plotted, indicating the 1- and 2-standard deviation area of the multivariate Gaussian distribution, as the analytical solution of the Fokker-Planck equation (2.34).

For sampling the probability distribution of the continuous-state stochastic process, we generated 10^4 sample paths of SDE (2.35) by using the Euler-Maruyama integration method [42], as shown in fig. 2.3 b. In addition, at each of the three time points two ellipses are plotted which indicate the areas of 1- and 2-standard deviation of the exact Gaussian solution of the Fokker-Planck equation (2.34). This experiment visualizes the equivalence of the first two moments of the discrete and continuous stochastic process. As discussed above the good approximation can be accounted to the linearity of the reaction propensities of the system (2.36).

In this section a theoretical framework was discussed which approximates the Chemical Master Equation, as a K-dimensional integro-differential equation by a Fokker-Planck equation, which is an n-dimensional linear partial differential equation. K is the size of the discrete state space S and n is the number of system variables (e.g. chemical species) which is in most biochemical applications smaller than K by several orders of magnitude. In comparison to the SSA algorithm, simulation of reaction kinetics can be significantly accelerated by diffusion approximation, described by the Fokker-Planck equation and sampling using Stochastic Differential equations of type (2.35). The accuracy of the LNA is limited by the ansatz (2.31) where the size of fluctuations must be small relative to the macroscopic solution and it increases as the thermodynamic limit is approached. More results on the limits of the approximation accuracy in various applications can be found in in [22].

2.4 Statistics of first passage times

Often one is not only interested in the probability distribution of a stochastic process on the state space but also in the dynamical properties at its boundaries. Naturally, the property of interest might be itself a random variable induced by individual realizations of the process. This might be the first time at which the process leaves a certain boundary or with regard to figure (2.3), the time that the system reaches a certain state, e.g. 400 viruses, for the first time. In this section we discuss how the statistics of time-related random variables can be computed by focusing on *first passage times* statistics of a Markov jump process.

Consider a Markov jump process $\{X(t)\}_{t\in\mathbb{R}^+}$ with an initial probability distribution $P_0(X,t)$. We are interested in the time at which the process reaches a certain boundary $X = x_{\max}$ for the first time. This boundary can be defined with respect to each of the n dimensions of the process or only their subset. For instance for the model (2.36) of a two-dimensional system of viruses V^* and host cells T^* one might be interested in the time that it takes the viruses to reach a certain population size $\mathrm{V}^*_{\max}$, independent of the dynamics of the host cells to which the virus dynamics is coupled. This problem is explained in figure 2.4 in terms of two sets of states n_0 and n_1. The first passage time is thus defined as the time that the system, started in a state $X_0 = \{\mathrm{T}^*_0, \mathrm{V}^*_0\} \in n_0$, spends in the set n_0 before it reaches the set of states n_1. An analysis of the first passage time statistics can be conducted using results from [25]. The starting point of this analysis is the probability for the system to be in the set of states n_0 and not having reached n_1 yet, that we denote by $G(n_0, n_1, t)$. Obviously, this quantity depends on the probability density $P(\mathrm{T}^*, \mathrm{V}^*, t)$ of the Markov jump process and the idea is to sum this probability in the region n_0. However, since the state space S is theoretically unlimited, an artificial boundary $\mathrm{V}^*_{\max}$ must be imposed. In order to detect the unique first time point of crossing this level of virions, a boundary condition must be defined on $\mathrm{V}^*_{\max}$, which prevents the process from jumping back to n_0, once it is in n_1. This can be implemented by setting the

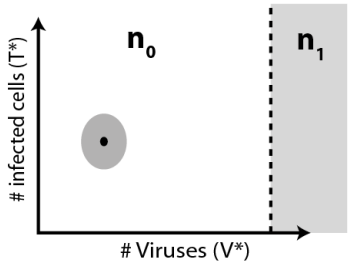


Figure 2.4: **Interpretation of the passage time problem as a state space, separated into two sets.**
The probability distribution of the system, sketched as the standard deviation area (grey oval) around the mean (black point). The statistics of the first time when this probability distribution crosses the dashed boundary gives rise to the first passage time distribution.

propensity of degradation of virions $w_-(\mathrm{V}^*)$ in reaction system (2.36) to zero if the process is n_1:

$$w_-(\mathrm{V}^*) = \begin{cases} c_1\mathrm{V}^*, & \text{if } \mathrm{V}^* < \mathrm{V}^*_{\max}, \\ 0, & \text{else.} \end{cases}$$

The boundary condition implies that n_1 is absorbing i.e. once the system reaches the set of states n_1, it stays there forever. By changing this propensity, we do not affect the statistics of the first passage time, since we are not interested in the dynamics of the system after it reaches n_1. Now the probability of the system to be in set n_0 at time t is the cumulative probability density of the stochastic process over n_0:

$$G(n_0, n_1, t) = \sum_{\mathrm{T}^*=\mathrm{T}_0}^{\infty} \sum_{\mathrm{V}^*=\mathrm{V}_0}^{\mathrm{V}^*_{\max}} P(\mathrm{T}^*, \mathrm{V}^*, t),$$

In combination with the above boundary condition $G(n_0, n_1, t)$ becomes the probability for the system *not to have reached* the set n_1 by time t yet and $1-G(n_0, n_1, t)$ is the probability to reach n_1 for the first time. Let us denote by $T(n_0 \to n_1)$ the random variable for the first passage time. Then its cumulative density function reads

$$\begin{aligned} & 1 - G(n_0, n_1, t) = \mathbb{P}\left[T(n_0 \to n_1) \le t\right] \\ \equiv \quad & \int_0^t \mathbb{P}\left[T(n_0 \to n_1) = \hat{t}\right] d\hat{t}, \end{aligned}$$

Differentiating this relation finally yields the PDF of the first passage time, cf. [25]

$$\mathbb{P}\left[T(n_0 \to n_1) = t\right] = -\frac{\partial}{\partial t} G(n_0, n_1, t). \tag{2.39}$$

This result can also be used to compute the k-th moment of the probability density as

$$T_k(n_0 \to n_1) = \int_0^\infty t^k \left[-\frac{\partial}{\partial t} G(n_0, n_1, t)\right] dt.$$

An integration of this equation by parts yields a recursive equation the k-th moment of the first passage time based on $(k-1)$-th moment:

$$T_k(n_0 \to n_1) = k \cdot \int_0^\infty t^{k-1} G(n_0, n_1, t) dt \qquad (k \geq 1). \tag{2.40}$$

By noting that the zero-th moment is always 1, the successive moments can be computed if the time evolution of the cumulative probability $G(n_0, n_1, t)$ is known.

In some special cases the solution of (2.40) can be found exactly. Consider the following two chemical reactions giving rise to a one-dimensional birth-death process

$$\xrightarrow{b(X)} \mathrm{X} \xrightarrow{d(X)} \varnothing.$$

In this one-dimensional case the boundary between some set of states n_0 and n_1 becomes a single state. Thus one can ask how long it takes the process to start in state x_0 and reach the state x_1 for the first time, where $x_0 < x_1$. It can be shown [25] that the k-th moment of the first passage time distribution is given by

$$T_k(x_0 \to x_1) = \begin{cases} D_k(x_1 - 1, x_1), & \text{if } x_0 = x_1 - 1, \\ D_k(x_0, x_1) + T_k(x_0 + 1 \to x_1), & \text{if } 0 \leq x_0 \leq x_1 - 2. \end{cases} \tag{2.41}$$

where the quantity D_k is defined in terms of the next lower order moment

$$D_k(x_0, x_1) = \begin{cases} \dfrac{k \cdot T_{k-1}(0 \to x_1)}{b(0)}, & \text{if } x_0 = 0, \\ \dfrac{k \cdot T_{k-1}(x_0 \to x_1)}{b(x_0)} + \dfrac{d(x_0)}{b(x_0)} \cdot D_k(x_0 - 1, x_1), & \text{if } 1 \leq x_0 \leq x_1 - 1. \end{cases}$$

In a general case with multiple chemical species the evaluation of the integral (2.40) is not possible and approximations need to be found. In the outlook of chapter 4 it is suggested to derive a Linear Noise Approximation of a reaction system with two species and to sample the first passage times by numerically solving the corresponding stochastic differential equations.

2.5 Estimation of reaction rate constants

In many applications model parameters are not known and need to be estimated from experimental measurements. The general approach to this problem is based on finding a set of parameters Θ such that a normed distance of the observed data to the model is minimized,

$$\arg\min_{\Theta} ||X(\Theta) - Y||_2, \tag{2.42}$$

where $X(\Theta)$ and Y denote the model and observation, respectively and $||.||_2$ is a 2-norm. This problem can also be considered in a probabilistic sense by searching

for a set of parameters such that the probability of observations Y w.r.t. to the model X is maximized. This approach is referred to as the *maximum likelihood method*. If Y is an observation of a sample path of a Markov jump process then the goal is a maximization of the joint probability of the observed jumps with respect to model parameters. Obviously, in situations where the stochastic process is not fully observed, inference becomes aggravated. Firstly, this is due to a naturally given reduced information content of the data leading to a larger estimation bias. Secondly, for Markov jump processes an additional challenge of estimating the unobserved path arises. In this section we discuss the problems associated with this task in the context of biochemical kinetics and along with possible methods for solving them.

2.5.1 ML-estimation for fully observed processes

Given a Markov jump process $\{X(t)\}_{t\in\mathbb{R}^+}$ on a state space S, describing the dynamics of a system of m coupled biochemical reactions, where reaction j is of general type

$$N_{1j} + N_{2j}, ..., N_{nj} \xrightarrow{c_j} M_{1j} + M_{2j}, ..., M_{nj},$$

for all $j \in \{1, .., m\}$. Accordingly, denote by $\boldsymbol{\nu}_j = \{\nu_{1j}, \nu_{2j}, ..., \nu_{nj}\}$ the state change vector defined as $\nu_{ij} = M_{ij} - N_{ij}$ and let each state $k \in S$ be uniquely characterized by the vector of chemical species numbers $\mathbf{x}_k := \{X_1 = x_1, X_2 = x_2, X_3 = x_3, ..., X_n = x_n\}$. If S consists of a finite set of d states, then the probability density at time t is given by

$$P(X, t) = P_0^T \exp(t\mathbf{A}^T),$$

where P_0^T is the initial probability. Each non-diagonal entry $\mathbf{A}_{kl}$ of the infinitesimal generator matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is zero, unless a reaction j exists with $\mathbf{x}_l = \mathbf{x}_k + \boldsymbol{\nu}_j$ yielding, cf. eq. (2.17)

$$\mathbf{A}_{kl} = c_j \prod_{i=1}^{n} \binom{x_i}{N_{ij}}. \tag{2.43}$$

Furthermore, the diagonal entries are

$$\mathbf{A}_{kk}(x) = -\sum_{\substack{k,q\in\mathbf{S}\\ k\neq q}} \mathbf{A}_{kq}.$$

Let an observed sample path $X(t)$ of the Markov jump process be given by $X := \{X(t_1) = \mathbf{x}_1,\ X(t_2) = \mathbf{x}_2,\ X(t_3) = \mathbf{x}_3...\}$. Then the joint probability of the sample path is obtained from the product

$$\mathbb{P}(X) = P(\mathbf{x}_1, t_1) \cdot P(\mathbf{x}_2, t_2) \cdot P(\mathbf{x}_3, t_3) \cdot ..., \tag{2.44}$$

where we defined $P(\mathbf{x}_i, t_i) := \mathbb{P}(X(t_i) = \mathbf{x}_i)$. Note that in most applications the individual state probabilities in equation (2.44) are not available, since in a situation

where the state space becomes prohibitively large, the computation of the sample path probability using the matrix exponential of the infinitesimal generator becomes infeasible. However, since all jumps along the sample path are observed, this probability can be expressed as a product of conditional probabilities which in turn can be reformulated using the infinitesimal jump rates i.e. reaction propensities. This fact makes infeasible matrix computations unnecessary, yielding

$$\begin{aligned}\mathbb{P}(X) &= P(\mathbf{x}_1, t_1) \cdot P(\mathbf{x}_2, t_2) \cdot P(\mathbf{x}_3, t_3) \cdot ... \\ &= P(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1) \cdot P(\mathbf{x}_3, t_3 | \mathbf{x}_2, t_2) \cdot ... \\ &= P(\mathbf{x}_1, t_1) \cdot \mathbf{A}_{12} \cdot \exp(\mathbf{A}_{11}[t_2 - t_1]) \cdot \mathbf{A}_{23} \cdot \exp(\mathbf{A}_{22}[t_3 - t_2]) \cdot ... \ ,\end{aligned}$$

where in the second equation the Markov property is used. The last equation gives rise to the *likelihood function* of the parameters of the model, i.e. the infinitesimal generator, given a sample path observation:

$$\mathfrak{L}(\mathbf{A}) := \mathbb{P}(Y|\mathbf{A}) = \prod_{i=1}^{d} \prod_{j \neq i} \mathbf{A}_{ij}^{H_{ij}} \exp(\mathbf{A}_{ii} R_i), \tag{2.45}$$

where the entries H_{ij} of the matrix H are given by the observed number of transitions from state i to state j. R_i is the total amount of time that the process spends in a state i

$$R_i = \int_0^T \chi_{\{i\}}(X(s)) \ \mathrm{ds},$$

where the characteristic function is $\chi(X(s)) = 1$ if $X(s) = i$ and zero otherwise. Obviously, the likelihood function is a conditional probability of the sample path given a fixed set of model parameters. This yields a framework which enables to conduct an optimization over the unknown parameter space merely based on path probabilities. In contrast to *Bayesian estimation*, likelihood-based inference does not yield the whole (posterior) distribution of the parameters but a single parameter value, also known as the *maximum likelihood* (ML) estimate. Despite of its lower information content, the ML-approach can be computationally more efficient, compared to the usually sampling-based Bayesian approach. In particular, this is the case if the optimization can be conducted analytically. In order to obtain an analytical optimizer for the likelihood function (2.45) its log must be taken first, for a better numerical stability:

$$\log \mathfrak{L}(\mathbf{A}) = \sum_{i=1}^{d} \sum_{j \neq i} \log(a_{ij}) H_{ij} - a_{ij} R_i.$$

Note that both sums in the above equation are over the entire state space S, which can be prohibitively large. Due to a relatively low number of reactions (compared to the size of the state space) in biochemical kinetic systems, the connectivity of states is usually lower than in other applications. As a result, the infinitesimal generator matrix has a very sparse structure. This property can be exploited in the

log-likelihood function by rewriting the second sum in the above equation in terms of the set of m chemical reactions instead of the set of d states:

$$\begin{aligned}\log \mathfrak{L}(\mathbf{A}) &= \sum_{i=1}^{d} \sum_{\substack{q=1 \\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{m} \log\left(a_{ij}\right) H_{ij} - a_{ij} R_i \\ &= \sum_{i=1}^{d} \sum_{\substack{q=1 \\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{m} \log\left[c_q \cdot \prod_{k=1}^{n} \binom{x_k^i}{N_{kq}}\right] H_{ij} - c_q \cdot \prod_{k=1}^{n} \binom{x_k^i}{N_{kq}} R_i,\end{aligned}$$

where in the second equality the definition of reaction propensities (2.43) is used. Note that x_k^i denotes the number of molecules of the k-th species in the state $i \in S$ of the corresponding Markov jump process. Notably the optimization of the reformulated likelihood function is not any more conducted w.r.t the infinitesimal generator. The new likelihood merely depends on a set of reaction rate parameters $\Theta = \{c_1, c_2, ..., c_m\}$. The two functions are equivalent since the parameter set Θ gives rise to the infinitesimal generator as described in equation (2.43). However, the latter approach is by far more efficient, since, as noted above, it usually holds that $m << d$. Differentiation of the new log-likelihood equation w.r.t. parameter c_q yields

$$\frac{\partial \log \mathfrak{L}(\Theta)}{\partial c_q} = \sum_{\substack{i=1 \\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{d} \left[\frac{H_{ij}}{c_q} - \prod_{k=1}^{n} \binom{x_k^i}{N_{kq}} R_i\right].$$

Finally, the zero of the likelihood derivative leads to the maximum-likelihood estimator $\widehat{c}_q$ for the rate constant of the q-th chemical reaction:

$$\widehat{c}_q = \sum_{\substack{i=1 \\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{d} H_{ij} \cdot \left(\sum_{i=1}^{d} \prod_{k=1}^{n} \binom{x_k^i}{N_{kq}} R_i\right)^{-1}. \tag{2.46}$$

Example: prey-predator dynamics

As an example consider a Markov jump process induced by a reaction system of preys (X_1) and predators (X_2). The prey population grows using natural resources, described by a constant parameter a. The predator population consumes prey species in order to grow, and dies with a linear propensity. By neglecting the natural extinction of preys, the described system is modeled by three reactions

$$\begin{aligned} R_1: &\quad \mathrm{a} + \mathrm{X}_1 \xrightarrow{c_1} 2\,\mathrm{X}_1, \\ R_2: &\quad \mathrm{X}_1 + \mathrm{X}_2 \xrightarrow{c_2} 2\,\mathrm{X}_2, \\ R_3: &\quad \mathrm{X}_2 \xrightarrow{c_3} \emptyset. \end{aligned}$$

By denoting the corresponding probability of species numbers by $P(x_1, x_2, t) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \text{time} = t)$, the Chemical Master Equation of this system reads

$$\begin{aligned}\frac{\partial P(x_1,x_2,t)}{\partial t} &= ac_1(x_1-1)P(x_1-1,x_2,t)\\ &+ c_2(x_1+1)(x_2-1)P(x_1+1,x_2-1,t)\\ &+ c_3(x_2+1)(x_2-1)P(x_1,x_2+1,t)\\ &- (ac_1x_1 + c_2x_1x_2 + c_3x_2)P(x_1,x_2,t).\end{aligned}$$

Figure 2.5 depicts a sample trajectory of the system obtained using Gillespie's simulation algorithm with the following parametrization: $c_1 = 2, c_2 = 0.1, c_3 = 1, a = 1$. All initial probability mass was concentrated on one state $P(100, 100, t) = 1$, which starts the trajectory in a region of the phase-space where the system exhibits oscillatory dynamics. Using equation (2.46) the maximum likelihood estimators can be

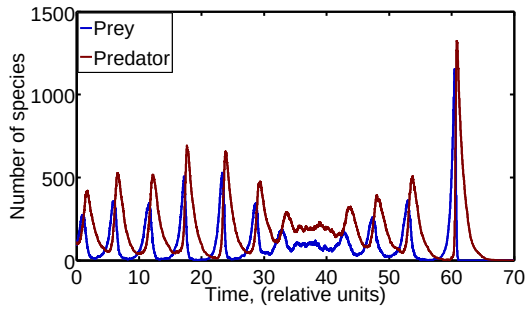


Figure 2.5: **Trajectory of the Markov jump process induced by the stochastic prey-predator system.**
The sample paths were generated using the Stochastic Simulation algorithm [24] and the initial state $X_0 = \{100, 100\}$. In this part of the phase space the system exhibits stochastic oscillations.

derived for all three reaction rate parameters:

$$\widehat{c}_1 = \sum_{\substack{i=1\\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{d} H_{ij} \cdot \left(\sum_{i=1}^{d}\prod_{k=1}^{n} a x_1^i R_i\right)^{-1}, \tag{2.47}$$

$$\widehat{c}_2 = \sum_{\substack{i=1\\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{d} H_{ij} \cdot \left(\sum_{i=1}^{d}\prod_{k=1}^{n} x_1^i x_2^i R_i\right)^{-1}, \tag{2.48}$$

$$\widehat{c}_3 = \sum_{\substack{i=1\\ \mathbf{x}_j=\mathbf{x}_i+\nu_q}}^{d} H_{ij} \cdot \left(\sum_{i=1}^{d}\prod_{k=1}^{n} x_2^i R_i\right)^{-1}, \tag{2.49}$$

where by x_j^i we denote the population size of species X_j in state $i \in S$. For parameter estimation 100 trajectories were sampled with each having a fixed length of 500 jumps. The estimation results are depicted in figure 2.6.

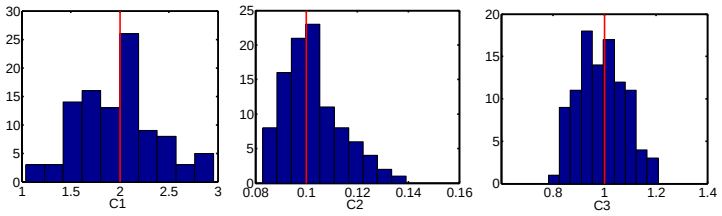


Figure 2.6: **Statistics of ML-estimates for the reaction rates of the stochastic prey-predator model.**
Estimation was conducted for 100 trajectories, each having a length of $N = 500$ jumps. The red vertical lines mark the original model parameters: $c_1 = 2,\ c_2 = 0.1,\ c_3 = 1$.

2.5.2 Discussion of ML-estimation for discretely observed processes

For deriving the maximum likelihood estimators in the previous section full observation of sample paths were assumed. This enabled to compute the path probability (2.44) without an explicit computation of the probability density of the Markov jump process on the state space. To this end successive state probabilities were expressed in terms of conditional probabilities of jumps between states. The latter only requires the knowledge of infinitesimal jump rates, given by the propensities of the chemical reaction system. In practice experimental observations are often available only at selected discrete time points. In this case the path probability (2.44) can not be expressed merely using chemical reaction rates since the jumps of the system are not completely observed. If the state space S is sufficiently small, then the probability density $P(X, t)$ can be computed using the matrix exponential yielding the discrete likelihood function

$$\mathfrak{L}(\mathbf{A}) = \prod_{i=1}^{J} P(X = \mathbf{x}_i, t_i),$$

where J is the total number of observations. In contrast to the continuous likelihood function (2.45), the discrete likelihood can not be optimized analytically and numerical optimization methods need to be applied. In order to avoid a direct solution of the CME, the probability density can be sampled using the SSA-algorithm method, cf. [76]. However, this method suffers from a large computational overhead which further increases if rare state transitions need to be sampled.

As an alternative to the discrete likelihood function, the dynamics of the unobserved path segments can be estimated along with the model parameters. The main idea of this method is based on computing the expectation of path statistics given the discretely observed process $\mathbb{E}[R_i|X]$ and $\mathbb{E}[H_{ij}|X]$. These expectations are computed by summing over the expected time spent in each state and the expected number of transitions during each unobserved time period τ_s. The assumption of

time-homogeneity yields [8, 51]:

$$\begin{aligned}\mathbb{E}[R_i(T)|X] &= \sum_{s=1}^{r}\sum_{k,l=1}^{d} c_{kl}(\tau_s)\ \mathbb{E}[R_i(\tau_s)|X(\tau_s)=l, X(0)=k],\\ \mathbb{E}[H_{ij}(T)|X] &= \sum_{s=1}^{r}\sum_{k,l=1}^{d} c_{kl}(\tau_s)\ \mathbb{E}[H_{ij}(\tau_s)|X(\tau_s)=l, X(0)=k],\end{aligned}$$

where $R_i(T)$ and $H_{ij}(T)$ denote the path statistics in the time interval $[0,T]$, where T is the total observation time. By c_{kl} we denote the observed number of transitions between states $k,l \in S$. The quantities $\mathbb{E}[R_i(\tau_s)|X(\tau_s)=l, X(0)=k]$ and $\mathbb{E}[H_{ij}(\tau_s)|X(\tau_s)=l, X(0)=k]$ can be computed using an eigenvalue decomposition of the infinitesimal generator [51]. If this decomposition is numerically feasible (due to a small state space), the expected path statistics and the parameters can be simultaneously estimated using the expectation maximization algorithm (EM) [51] .

The size of the state-space puts a restriction on the above estimation method for discretely observed processes. For instance, if a biochemical system consists of three species with a maximum number of 100 molecules, then the infinitesimal generator is of size $\mathbf{A} \in \mathbb{R}^{10^6\times 10^6}$. This makes a repeated eigenvalue decomposition of the matrix $\mathbf{A}$ numerically infeasible. A possible method for obtaining the path statistics during unobserved time intervals τ_s is given by end-point conditioned sampling [38]. This method is based on rejection sampling of SSA-sample paths. Thus, in order to compute $\mathbb{E}[R_i(\tau_s)|X(\tau_s)=l, X(0)=k]$ a path is sampled starting in state k at time $t=0$ and only accepted if it hits state l at time $t=\tau_s$. However, besides introducing an estimation error proportional to the length of unobserved path segments, this method becomes inefficient if the number of species in the system is large cf. [14]. As an alternative, estimation of stochastic model parameters for incompletely observed processes can be conducted by fitting first passage time moments. This approach will be introduced in chapter 4.

Chapter 3

Dynamics of stress-mediated c-di-GMP regulation in *Escherichia coli*

3.1 Introduction

Bacteria populate almost all environments due to their ability to adopt to extreme conditions. Most of them are a vital part of the various ecosystems. For example, it is estimated that the amount of bacteria in the human organism exceeds the number of native cells by a factor of ten [7]. While most bacteria coexist with the human organism, some pathogenic species cause severe infections. Antibiotics used to protect against these infections are often counteracted by the inherent robustness of bacteria towards changes in external conditions.

The formation of biofilm colonies and the associated curli fibers is a primary mechanism protecting bacterial species against stress-inducing conditions such as antibiotics [33, 72]. In *Escherichia coli*, a Gram-negative species which is a major part of the human intestinal ecosystem, the synthesis of biofilm and curli fibers is under the control of the master regulator protein of the general stress response RpoS (also referred to as sigma^{S}) [4, 5, 17]. In recent years large progress has been made in understanding the details of signaling within the curli expression network, yielding a highly resolved picture of the interactions of individual genes and proteins [57, 71, 80]. The elaborate signal transduction controlled by RpoS reveals properties of a finely tuned system conferring bacterial populations the ability to precisely regulate the time point and the amount of cells which produce curli fibers and of those which do not. Thus, a discrete decision making process has been observed using fluorescence microscopy in *E.coli* [67] and using single cell measurements in a close genetic relative Salmonella [31]. Furthermore, the major signaling molecule of this system, cyclic di-GMP (c-di-GMP), is characterized by relatively low molecular numbers in the cell, suggesting that stochastic fluctuations play a crucial role in regulating the system. The resulting dynamical complexity indicates that a complete understanding of the regulatory mechanisms behind the formation of curli fibers requires an interaction between experimental efforts and theoretical modeling. The latter has the ability to resolve the time dynamics behind the static interactions deciphered by the experimental approaches. Furthermore, certain biologically relevant highly non-linear signal transduction characteristics may only be deciphered

by mathematical analysis [55].

In this work we build a realistic stochastic model describing the signaling properties of c-di-GMP and incorporate these findings into a larger mechanistic picture which explains how bistable decision making between curli-on and curli-off states might be realized in *E. Coli*. This analysis compensates for the currently limited availability of precise *in-vivo* measurements of c-di-GMP dynamics and explains how signal transduction characteristics of c-di-GMP influence the complexity of the curli phenotypes. Prior to presenting the modeling results, in the following some preliminary biological aspects of the system are introduced, comprising the RpoS-controlled signal transduction and the role of c-di-GMP in the expression of curli fimbriae.

General stress response and formation of curli fimbriae

The cornerstone of bacterial gene regulation is the control of different genetic programs by *master regulators* (also called sigma-factors). This importance arises from the requirement for the key transcriptional enzyme RNA-polymerase (RNAP) to bind one of the various master regulators proteins before it can start the transcription process. The resulting complex is referred to as RNAP-holoenzyme and its sigma-subunit is responsible for a specific recognition of binding sites on the DNA. Since distinct functional groups of genes possess different sigma factor binding sites, the type of the sigma factor in the RNAP-holoenzyme determines the genetic program in the cell [34, 68].

In contrast to the vegetative regulator sigma^{70}, which controls the expression of genes needed during normal metabolic life style, the master regulator RpoS controls the genetic program during entry into the stationary growth phase or upon exposure to various stress conditions such as nutritional shortage or extreme temperatures [36, 37, 40]. There are approximately 480 genes controlled by RpoS, which accounts for 10% of the *E. Coli* genome [80]. A subset of these genes and their protein products are responsible for a transition from a unicellular planktonic life style to a multicellular biofilm aggregation and the resulting production of the main building brick of biofilm, the curli fibers (fimbriae). This is achieved by the expression of the protein **CsgD**, the transcription factor for the *csgBAC* operon which codes for the key biochemical components of curli fibers [45, 58, 65, 80]. The production of curli enables the cells to adhere to surfaces and to each other and to create bacterial communities in which single cells are better protected against external stress conditions such an antibiotic treatment. Furthermore, during the shift to this sessile adhesive life style the metabolism and virulence of bacteria is strongly reduced which enables an adaptation to the possibly limited amount of resources [37].

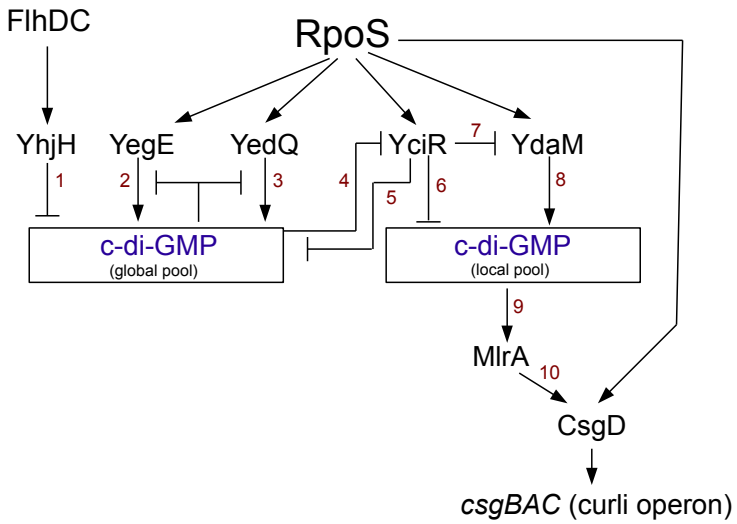


Figure 3.1: **Molecular interactions and antagonism of the RpoS-induced signaling cascade.**
The expression of a set of elaborately interacting DGC- and PDE-enzymes is controlled by the master regulator of the general stress response RpoS. The respective synthesis product and degradation substrate of these enzymes, the second messenger c-di-GMP, is the central signaling molecule within this network. Ultimately, the induction of RpoS leads to the expression of curli fimbriae, a building brick of bacterial biofilm.

The role of c-di-GMP in stress-mediated signal transduction

The expression of the biofilm regulator CsgD is controlled by an interplay in a complex signaling network, as shown in fig. 3.1. In this diagram production and transcriptional activation is subsumed by “positive regulation“ and indicated by arrows with reference numbers in fig. 3.1. Degradation and deactivation is subsumed by ”negative regulation“ and indicated by a perpendicular line fig. 3.1. The key component in the curli expression system is the second messenger molecule bis-(3’-5’)-cyclic-di-guanosine monophosphate, also referred to as **cyclic-di-GMP** (c-di-GMP) which maintains signal transduction by interacting with different effector molecules in the network [35]. The production and degradation of c-di-GMP is catalyzed by two distinct sets of enzymes: diguanylate cyclases (DGCs) and phosphodiesterases (PDEs), respectively [11, 13, 16, 56]. As shown in fig. 3.1, the c-di-GMP-producing DGC-enzymes in the curli expression network are the proteins YegE, YedQ and YdaM and the c-di-GMP-degrading enzymes are YhjH and YciR. The stress-dependent expression of DGCs and PDEs suggests that high or low c-di-GMP levels can code for different signals through different binding affinities of the effector molecules [35].

It is currently assumed that besides a global, freely diffusible pool of c-di-GMP, there are other local, physically separated pools with a functional and temporal restriction of appearance in the network [35, 46]. The different pools possess common

properties concerning the regulation of the c-di-GMP levels. The global c-di-GMP pool is produced by YegE and YedQ (reactions 2 and 3 in fig. 3.1, resp.) and degraded by YhjH and presumably YciR (reactions 1 and 5, resp.). The local c-di-GMP pool is produced by YdaM (reaction 8) and it is presumably also subject to degradation by YciR (reaction 6), as shown in fig. 3.1 [80]. Furthermore, the catalytic synthesis of c-di-GMP is down-regulated by binding of c-di-GMP to a secondary allosteric site (I-site) of the DGC enzyme (YegE and YedQ). This feedback control by **product inhibition** has been suggested to prevent an excessive production and to filter large stochastic fluctuations [15]. In the curli expression network this property is exhibited by YegE and YedQ.

The described regulatory properties of c-di-GMP production and degradation have been observed in various bacterial species [35]. This suggests a recurring and modular character of this system which can be considered as a network motif consisting of simple activation, simple degradation and a negative auto-regulation, shown in fig. 3.2. (see [2] for a review on network motifs). Obviously, the relation between

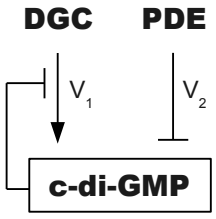


Figure 3.2: **A general c-di-GMP regulation module with product inhibition.**

the rate function of catalytic synthesis V_1 and the rate function of catalytic degradation V_2 determines the cellular levels of c-di-GMP (fig. 3.2). However it is not clear how the expression levels of the enzymes, their catalytic properties and the product inhibition interact with each other in this system to set up a certain cellular level of c-di-GMP.

The role of c-di-GMP signaling in the expression of curli fimbriae becomes evident in fig. 3.1. There is an indirect and elaborate signaling path from RpoS induction to the expression of the *csgBAC* operon. During the entry into the stationary growth phase (i.e. reduced growth rate due to a high cell density) of an *E.Coli* colony all important DGC and PDE enzymes are expressed [80]. Before MlrA, the key activator protein of the transcription factor CsgD, can bind to the DNA and enable the expression of the *csgD* gene (reaction 9 in fig. 3.1) it needs to be activated by binding c-di-GMP (reaction 8). Recent findings suggest that MlrA can only bind c-di-GMP by a complex formation with a YdaM dimer [46]. This c-di-GMP is produced by YdaM (reaction 7) and transmitted to MlrA within the complex and thus it represents an example for a local pool of c-di-GMP generated by molecular sequestration. Paradoxically, RpoS does not only induce the

expression of YdaM and MlrA but also the phosphodiesterase YciR which has been recently shown to prevent the complex formation of YdaM and MlrA [46] (reaction 7). Furthermore, in the same study it has been shown that the inactivating function of YciR is inhibited by binding of the global (freely diffusible) c-di-GMP to YciR (reaction 4). Thus, RpoS not only induces the expression of YciR, the inhibitor of the transcription factor MlrA, but it also induces the inhibitor of the inhibitor, the global c-di-GMP pool produced by the DGC enzymes YegE and YedQ.

The complex signaling system described above, does not obey the principle of parsimony since in an alternative, more efficient architecture MlrA could be directly activated by global c-di-GMP and thus contribute to a saving of resources. With regard to the high genomic variability of bacteria due to horizontal gene transfer and mutation and the resulting selection pressure [64], this suggests that the indirect signaling path is not a result of a random assembly but it must have a functional role which confers a certain evolutionary benefit. A possible role of this signaling system could be the generation of mutual inhibition between global c-di-GMP and the key inhibitor of curli formation YciR, giving rise to a double negative feedback loop (fig. 3.1) and possibly result in a bistable system. This is in line with experimental data, indicating that the expression level of curli in single cells is not simply proportional to the induction level of RpoS but it obeys an all-or-nothing principle suggesting a bistable signaling mechanism [31, 67]. The potential of double negative feedback loops to generate bistable bahaviour [70] and the intrinsic stochasticity of c-di-GMP dynamics suggests that a deeper dynamical understanding of this system is needed in order analyze the potential role of c-di-GMP in bistable curli expression.

3.1.1 Aims, scope and modeling strategy

In the present study we derive a mathematical model of c-di-GMP regulation in the curli expressing network by considering the dynamical properties of the DGC- and PDE-enzymes as parameters of the catalytic rate functions V_1 and V_2 (fig. 3.2). This model enables a basic initial analysis of the dependance of the steady-state levels of c-di-GMP on these parameters. Firstly, this explains how c-di-GMP levels increase during induction of the stress response and RpoS-activation and stabilize at a new, higher steady state. Furthermore the model is used to compare the steady-state of a c-di-GMP module with and without product inhibition and to deduce the role of this feedback regulation in the control of c-di-GMP levels. Secondly, by explicitly considering the inherent stochastic character of c-di-GMP regulation, the stationary probability distribution of the resulting Markov jump process allows to analyze the variability of the dynamics in dependance on the rate functions. This gives an insight into the regulation of signaling noise in the curli synthesis network and the decisive role of product inhibition in noise reduction. Thirdly, due to its central role as the main signal transduction molecule in the curli expression network (fig. 3.1), the velocity of c-di-GMP regulation is a key indicator of the responsiveness of the curli production system. The first passage times of the stochastic model enable an

analysis of the response times of c-di-GMP modules, depending on the expression levels of the DGCs and PDEs.

In the second part of this study the impact of c-di-GMP regulation on the dynamics of curli expression is analysed. The main focus is on the effect of the interaction between the global c-di-GMP and YciR. A system of ordinary differential equations is derived which approximates the stationary states of the stochastic dynamics of this interaction and a bifurcation analysis yields realistic parameter regimes which enable bistable behaviour. Finally, the model is used to explain experimental expression measurements of the curli gene *csgB* and yield new evidence for the bistable signaling mechanism leading to population heterogeneity of curli production in *E.Coli* colonies.

3.2 Results

Cyclic di-GMP is a ubiquitous signaling molecule in the bacterial world and, as described above, it plays a decisive role in the regulation of curli production in *E.Coli*. The number of c-di-GMP molecules are estimated in a range from a few molecules to a few thousands [35], implying that low molecular numbers may be involved in signaling and generate significant stochasticity. In order to account for noise effects in signal transduction, explicit stochastic modeling of c-di-GMP dynamics is required. In the first section of this chapter kinetic properties of c-di-GMP regulation are deduced which result from involved biochemical reactions. Afterwards, the according simple birth-and-death Chemical Master Equation is solved and the corresponding signal transduction properties of this system are analysed. In the last section the simple regulatory model of c-di-GMP dynamics is expanded for an interaction with the potential key inhibitory molecule YciR and the impact of c-di-GMP regulation on curli production is studied.

3.2.1 Enzyme kinetics of DGCs and PDEs

The properties of signal transduction in the curli production network are tightly connected to the dynamics of c-di-GMP regulation. In order to derive the equations for the regulatory dynamics of c-di-GMP levels in the cell, the underlying catalytic principles, as depicted in fig. 3.3, have to be reviewed. Accordingly, the synthesis of a c-di-GMP molecule is catalysed by a diguanylate cyclase (DGC), denoted by E_1 below. This reaction consumes two molecules of guanosine triphosphate (GTP) [12, 66] and involves the formation and dissociation of an enzyme-substrate complex as intermediate reactions. An additional aspect of c-di-GMP synthesis is the product inhibition property which can be described by further intermediate reactions where the inhibitor c-di-GMP either binds to the free enzyme or to the enzyme-substrate complex. This results in the reaction system depicted in fig. 3.4, in which X denotes a molecule of c-di-GMP and S denotes the substrate GTP. Since DGC molecules have a high affinity for dimerization, E_1 denotes a DGC dimer. C-di-GMP (X) has

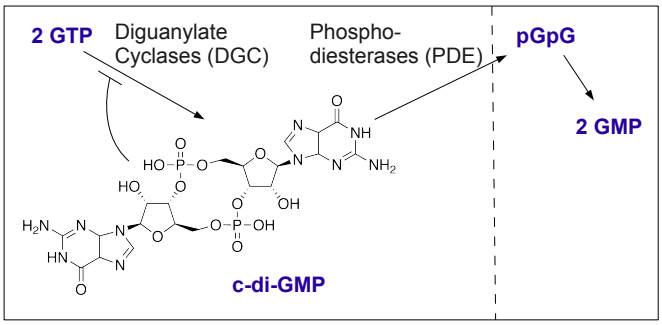


Figure 3.3: **Basic biochemical reactions involved in c-di-GMP production and degradation in bacteria**.
C-di-GMP synthesis is catalyzed by a diguanylate cyclase (DGC) from two molecules of guanosine triphosphate (GTP). The degradation of c-di-GMP is catalyzed by a phospho-diesterase (PDE) by breaking it down to 5'-phosphoguanylyl-(3'-5')-guanosine (pGpG). Finally, pGpG is degraded to two guanosine monophosphate (GMP) molecules.

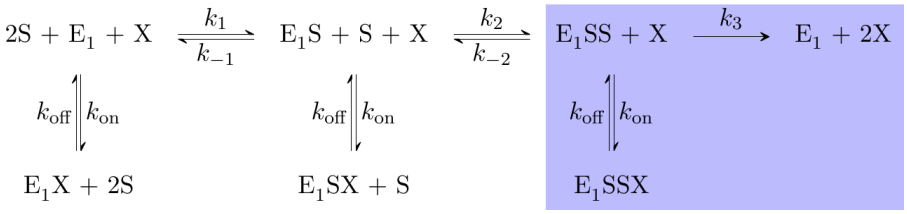


Figure 3.4: **Catalytic reaction system of c-di-GMP synthesis.**
In the scheme, X denotes a molecule of c-di-GMP, S denotes the substrate GTP and E_1 is a DGC enzyme dimer. A complex containing an enzyme molecule and two substrates (E_1SS) is able to react to one c-di-GMP molecule X with a reaction rate k_3. Due to excess availability of GTP (S), the Michaelis-Menten equilibrium can be assumed to be shifted to the last three reactions of the system (blue box).

obviously two roles in the system: firstly it acts as an inhibitor at different stages of the catalysis reaction with association and dissociation rates k_{on} and k_{off} and secondly c-di-GMP is the final product. According to Michaelis-Menten kinetics [10, 39], on the time scale of product synthesis, the association and dissociation of enzyme-substrate complex can be assumed to be at equilibrium. Due to an excess availability of the substrate GTP (S in fig. 3.4) in the cell [35], within this equilibrium, the enzymes E_1 are likely to have bound at least one substrate molecule. Thus, denoting the maximal velocity of product synthesis by $V_{\text{max}\,1} = E_1SS \cdot k_3$, a good approximation to the production rate function of c-di-GMP is

$$V_1 = \frac{V_{\text{max}\,1} S}{(K_{\text{m}} + S) \cdot c}, \tag{3.1}$$

where $c = 1 + X/K_i$ [41] and K_{m} is the corresponding Michaelis-Menten constant with $K_{\text{m}} = (k_{-2} + k_3)/k_2$. The binding affinity of c-di-GMP to the allosteric site of the DGC enzyme is given by $K_i = k_{\text{off}}/k_{\text{on}}$. The excess availability of the substrate

GTP allows to assume that $S >> K_\mathrm{m}$ and $S + K_\mathrm{m} \approx S$. Consequently, eq. (3.1) results in a saturated kinetic rate

$$V_1 = \frac{V_{\max 1}}{1 + X/K_i}, \tag{3.2}$$

which will be assumed in the following as the (approximated) production rate of c-di-GMP. The second key reaction system involved in c-di-GMP regulation is its degradation by PDE enzymes.

$$\mathrm{E_2 + X} \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} \mathrm{E_2X} \xrightarrow{k_5} \mathrm{P}$$

where $\mathrm{E_2}$ denotes a PDE molecule, X denotes a c-di-GMP molecule and P is the product of this reaction, pGpG. Using the equilibrium assumption of the Michaelis-Menten model the reaction rate function of this system results in

$$V_2 = \frac{V_{\max 2} X}{X + K_\mathrm{m}}, \tag{3.3}$$

where X denotes the number of c-di-GMP molecules, $V_{\max 2} = (E_2 + E_2X) \cdot k_5$ and the Michaelis-Menten constant $K_\mathrm{m} = (k_{-4} + k_5)/k_4$.

3.2.2 Signaling properties of c-di-GMP modules

The results of the previous section can be used to study the influence of the molecular properties and expression levels of the DGC and PDE enzymes on the signal transduction by c-di-GMP. To this end the dynamics of a c-di-GMP module can be considered as a one-dimensional stochastic birth-and-death process. The time evolution of its probability distribution is described by the following Chemical Master Equation (CME):

$$\begin{aligned} \frac{\partial \mathrm{P}(x,t)}{\partial t} &= \frac{V_{\max 1}}{1 + (x-1)/K_i} \cdot \mathrm{P}(x-1,t) + \frac{V_{\max 2}\ (x+1)/K_\mathrm{m}}{1\ +\ (x+1)/K_\mathrm{m}} \cdot \mathrm{P}(x+1,t) \\ &- \left[\frac{V_{\max 1}}{1 + x/K_i} + \frac{V_{\max 2}\ x/K_\mathrm{m}}{1\ +\ x/K_\mathrm{m}} \right] \cdot \mathrm{P}(x,t), \end{aligned} \tag{3.4}$$

where $P(x,t) := \mathbb{P}(X = x, \mathrm{time} = t)$ is the probability of x c-di-GMP molecules at time t. The birth and death rates correspond to the synthesis rate V_1 (eq. 3.2) and degradation rate V_2 (eq. 3.3), respectively. By setting $\partial P_s(x,t)/\partial t = 0$ the time-independent stationary solution $P_s(x)$ of this equation can be derived [24] yielding

$$\begin{aligned} P_s(x) &= \left(1 + \sum_{n=1}^{x} \prod_{j=1}^{n} \frac{K_i K_\mathrm{m} + K_i \cdot j}{K_i + j - 1} \cdot \frac{V_{\max 1}}{V_{\max 2}} \right)^{-1} \\ &\cdot \prod_{j=1}^{x} \frac{K_i K_\mathrm{m} + K_i \cdot j}{K_i + j - 1} \cdot \frac{V_{\max 1}}{V_{\max 2}}. \end{aligned} \tag{3.5}$$

This distribution has an important implication for the biological system. At least during the stress-induced induction of the master regulator RpoS, c-di-GMP levels have to be maintained at a steady state to ensure signal transduction. Otherwise, assuming that the system is non-oscillatory, c-di-GMP either would go extinct or it would flood the cell.

Equation (3.5) shows that the ratio $V_{\max 1}/V_{\max 2}$ of the maximal catalytic rates of the DGCs and PDEs determines the stationary probability distribution. According to the corresponding Michaelis-Menten kinetics (previous section) the parameters $V_{\max 1}$ and $V_{\max 2}$ are computed as the products of the velocity of product synthesis (k_3 and k_5) and the total amount of enzyme (eqs. (3.2) and (3.3)). Assuming k_3 and k_5 as constant (thus assuming a constant activation level of the DGCs and PDEs), the **relative expression level** of these enzymes w.r.t. each other is the key determinant of the c-di-GMP levels.

As discussed in section 2.3 (chapter 2), the first moment of the stationary distribution of a Markov jump process can be approximated by finding the fixed points of the equation for the corresponding deterministic trajectory $\bar{x}$, resulting from the lowest order Ω-expansion of the Chemical Master Equation. For the simple c-di-GMP module these are given as the roots of the following polynomial

$$f(\bar{x}) = \frac{V_{\max 1}}{1 + \bar{x}/\bar{K}_i} - \frac{V_{\max 2}}{1 + \bar{K}_{\mathrm{m}}/\bar{x}}, \tag{3.6}$$

where $\bar{K}_i = K_i/\Omega$ and $\bar{K}_{\mathrm{m}} = K_{\mathrm{m}}/\Omega$. The positive root of this equation is

$$\bar{x}_s = -\frac{\bar{K}_i(V_{\max 2} - V_{\max 1})}{2V_{\max 2}} + \left[\left(\frac{\bar{K}_i(V_{\max 2} - V_{\max 1})}{2V_{\max 2}}\right)^2 + \frac{V_{\max 1}\bar{K}_i\bar{K}_{\mathrm{m}}}{V_{\max 2}}\right]^{1/2}, \tag{3.7}$$

and as discussed in chapter 2.3 it is a good approximation to the stationary mean of the probability distribution (3.5).

The Chemical Master Equation (3.4) can also be generalized to c-di-GMP modules where different DGCs and/or PDEs contribute to the same c-di-GMP pool (e.g. reactions 1, 2 and 3 in fig. 3.1). We consider here the case where n different DGCs with the same K_i-constant and the m different PDEs with the same K_{m}-constant are active in the system, however the results can also be generalized to multiple DGCs and PDEs with differing K_i and K_{m}. In this case the production and degradation reactions can be summed into two reactions with the following

rates:

$$\begin{aligned} V_{\text{synthesis}} &= V_{\text{DGC 1}} + V_{\text{DGC 2}} + ... + V_{\text{DGC n}}, \\ &= \frac{V_{\max 11} + V_{\max 12} + ... + V_{\max 1n}}{1 + x/K_i}, \\ &= \frac{V_{\max 1\ \text{all}}}{1 + x/K_i}, \\ V_{\text{degradation}} &= V_{\text{PDE 1}} + V_{\text{PDE 2}} + ... + V_{\text{PDE m}}, \\ &= \frac{V_{\max 21} + V_{\max 22} + ... + V_{\max 2m}}{1 + K_{\text{m}}/x}, \\ &= \frac{V_{\max 2\ \text{all}}}{1 + K_{\text{m}}/x}, \end{aligned}$$

where $i \in \{1, .., n\}$ and $j \in \{1, .., m\}$. $V_{\text{DGC i}}$ and $V_{\text{PDE j}}$ are the synthesis and degradation rate functions of the i-th DGC and j-th PDE, respectively. Furthermore, $V_{\max 1i}$ and $V_{\max 2j}$ are the corresponding maximal catalytic velocities and $V_{\max k\ \text{all}} = V_{\max k1} + V_{\max k2} + ... + V_{\max kn}$ with $k \in \{1, 2\}$. The stationary solution of this system is obtained from equation (3.5) by substituting $V_{\max 1}$ and $V_{\max 2}$ by $V_{\max 1\ \text{all}}$ and $V_{\max 2\ \text{all}}$, respectively. In this case the ratio of expression levels of all the different DGCs versus the expression levels of all PDEs determines the level of c-di-GMP at steady state. If the K_i and K_m of the enzymes do not have the same values, then the proportionality of c-di-GMP levels and the relative expression levels of DGCs and PDEs also still holds (algebraic derivation omitted here for brevity). This implies that signaling properties deduced for a simple c-di-GMP module with one DGC-type and one PDE-type can be generalized for systems with different active DGC- and PDE-enzymes.

The role of product inhibition in signal transduction

Product inhibition (PI) of DGC enzymes is a feedback control mechanism based on allosteric binding of c-di-GMP to a secondary binding site of the enzyme [15]. In order to analyze the role of this mechanism in the c-di-GMP regulation network, the dynamical properties of c-di-GMP modules with and without product inhibition can be compared. According to the results in section 3.2.1, the production rate function $V_1(x)$ of a DGC without PI is given by the constant $V_{\max 1}$. Thus the mean of the corresponding Chemical Master Equation can be approximated by the fixed point of the following steady state equation

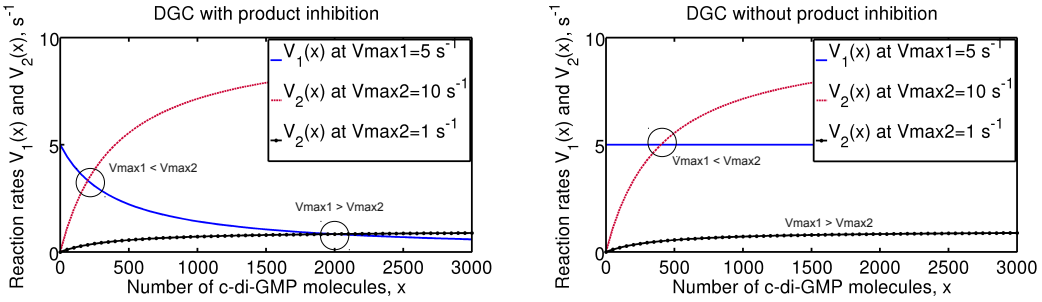
$$f(\bar{x}) = V_{\max 1} - \frac{V_{\max 2} \cdot \bar{x}}{\bar{x} + \bar{K}_{\text{m}}}, \tag{3.8}$$

which is given by

$$\bar{x}_s = \frac{V_{\max 1} \cdot \bar{K}_{\text{m}}}{V_{\max 2} - V_{\max 1}}.$$

This root is positive if $V_{\max 1} < V_{\max 2}$. In contrast, the steady state equation (3.6) has a positive root for all (non-negative) parameter configurations. This suggests

that c-di-GMP modules with product inhibition always possess a stationary state while c-di-GMP modules without product inhibition only possess a stationary state if the maximal production rate $V_{\max 1}$ is lower than the maximal degradation rate $V_{\max 2}$. This robustness property induced by the product inhibition is explained in fig. 3.5. This figure visualizes the fixed points of the approximating deterministic reaction rate equation as intersections of the production and degradation rate curves. While in a system with PI these two curves intersect for all combinations of $V_{\max 1}$ and $V_{\max 2}$ (fig. 3.5 a), in a system without PI an intersection is only possible if $V_{\max 1} < V_{\max 2}$.



(a) (b)

Figure 3.5: **Feedback inhibition ensures stationarity of c-di-GMP dynamics.** The stationary states of a stochastic c-di-GMP regulation module approximately correspond to the steady states of the deterministic reaction rate equation. The values of rate functions of c-di-GMP production and degradation (y-axis) are plotted for different numbers of c-di-GMP molecules (x-axis). The intersections between the production rate curve (blue) and the degradation rate curves represent the fixed points of the corresponding system. **(a)** In a c-di-GMP module with PI the production rate is a monotonically decreasing function of c-di-GMP numbers. As a result, the production rate curve intersects the degradation rate curve, independent of the maximal degradation value $V_{\max 2}$ (red and black curves.) **(b)** In a c-di-GMP regulation module without PI the production rate is a constant ($V_{\max 1}$) and there is only an intersection between the production and degradation rate curves if $V_{\max 1} < V_{\max 2}$. The system has no fixed point if $V_{\max 1} > V_{\max 2}$. This corresponds to a theoretical situation where c-di-GMP grows in an unbounded manner.

A second aspect related to product inhibition is the variance of the stationary c-di-GMP distribution which determines the amount of noise in signaling. A fine-tuned signal transduction requires that c-di-GMP levels are in a certain range corresponding to the binding affinities of the various effector component molecules. A possibly large noise-to-signal ratio of c-di-GMP levels indicates an unspecific signal transduction which is opposed by the significant amount of various c-di-GMP effectors with different binding affinities. As a negative feedback control mechanism, product inhibition has a potential role in reducing the signaling noise. In order to study the effect of product inhibition on the variability of the c-di-GMP levels, we aim at computing the variance of c-di-GMP levels with and without product inhibition.

The variance can be computed using the stationary distribution (3.5) by computing the difference

$$\sigma^2 = \sum_x x^2 P_s(x) - \left(\sum_x x P_s(x) \right)^2 ,$$

which leads to an equation for the noise-to-signal ratio, also referred to as the *Fano factor*:

$$\frac{\sigma^2}{\mu} = \frac{\sum_x x^2 P_s(x) - \left(\sum_x x P_s(x) \right)^2}{\sum_x x P_s(x)}. \tag{3.9}$$

Figure 3.6 shows the dependance of the Fano factor on the parameters K_i, K_m and the ratio $V_{\max 1}/V_{\max 2}$. Obviously the amount of noise relative to the signal becomes particularly large when $V_{\max 1} \approx V_{\max 2}$ (fig. 3.6 (b)) and the catalysis of degradation works near $V_{\max 2}$ (small K_m). A small K_i (high affinity of c-di-GMP for the secondary binding site) reduces this signaling noise, while a large K_i has the effect that both enzymes work at saturation, leading to a hypersensitive system, as described previously [29, 20]. The results in figure 3.6 thus indicate the noise reduction effect of product inhibition. In fig. 3.6 (a) the same computations are carried out for a smaller ratio $V_{\max 1}/V_{\max 2} = 0.96$ leading to a significantly lower amount of stochastic fluctuations. In order to explain the observed parameter

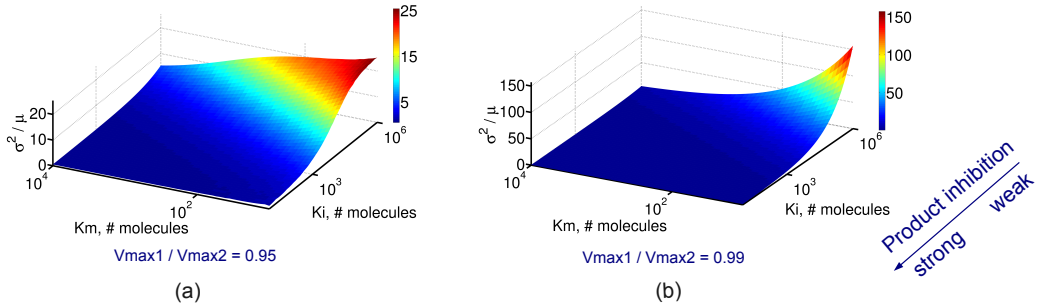


Figure 3.6: **Parameter dependence of the noise-to-signal ratio of c-di-GMP levels.**
The ratio of variance and the mean level (Fano factor) of the stationary c-di-GMP distribution increases as the catalytic rates of production and degradation approach saturation (i.e. as $V_1(x) \to V_{\max 2}$ and $V_2(x) \to V_{\max 2}$, respectively). This effect is caused by an increasing K_m and a decreasing product inhibition i.e. an increasing K_i. Furthermore the Fano factor is inversely proportional to the ratio $V_{\max 1}$ to $V_{\max 2}$, as shown by the two experiments. **(a)** If $V_{\max 1}/V_{\max 2} = 0.95$ the Fano factor is ≈ 20 at its maximum. **(b)** In the case $V_{\max 1}/V_{\max 2} = 0.99$ the maximal value of the Fano factor reaches 150. Note the different axes scaling in the figures.

dependance of the noise-to-signal ratio, an analytical expression for the variance is

needed. Since an exact analytical expression based on the probability distribution (3.5) is not available, a Linear Noise Approximation (LNA) [77] can be used (cf. section 2.3.4), as discussed in section 2.3.4 of chapter 2. In particular, a relation between the reaction rates $f(\bar{x})$ of the macroscopic equations (3.6) and (3.8) and the stationary covariance matrix $\boldsymbol{\Sigma}_s$ of the approximating Fokker-Planck equation is given by the Lyapunov equation

$$0 = \boldsymbol{\Gamma}_s \cdot \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_s + \boldsymbol{\Gamma}_s^T + \boldsymbol{D}_s, \tag{3.10}$$

where $\Gamma_{ij}(t) = \frac{\partial f_i}{\partial \bar{x}_j}|_{\bar{x}=\bar{x}_s}$ is the corresponding Jacobian at the fixed point $\bar{x}_s$, $\boldsymbol{D} = \mathbf{S} \cdot \mathrm{diag}[\mathbf{w}(\bar{x}_s)] \cdot \mathbf{S}^T$ is the diffusion matrix, $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{w}(\bar{x}_s)$ is the propensity vector.

Stochastic fluctuations in a system with product inhibition

According to equation (3.6), the Jacobian of the macroscopic equation of a c-di-GMP module with PI is given by the scalar derivative

$$\begin{aligned} \boldsymbol{\Gamma}_s &= \frac{\partial}{\partial \bar{x}} \left(\frac{V_{\max 1}}{1 + \bar{x}/\bar{K}_i} - \frac{V_{\max 2}\bar{x}}{\bar{x} + \bar{K}_\mathrm{m}} \right) \Bigg|_{\bar{x}=\bar{x}_s} \\ &= -\frac{V_{\max 1}/\bar{K}_i}{(1 + \bar{x}_s/\bar{K}_i)^2} - \frac{V_{\max 2}\bar{K}_\mathrm{m}}{(\bar{x}_s + \bar{K}_\mathrm{m})^2}. \end{aligned} \tag{3.11}$$

Furthermore, since the stoichiometric matrix of this two-reaction system is

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

the scalar diffusion coefficient is given by

$$\boldsymbol{D_s} = \frac{V_{\max 1}}{1 + \bar{x}_s/\bar{K}_i} + \frac{V_{\max 2}\bar{x}_s}{\bar{x}_s + \bar{K}_\mathrm{m}}.$$

Due to equation 3.10 the approximated variance of the stationary probability distribution results in

$$\boldsymbol{\Sigma_s} = -\frac{\boldsymbol{D}_s}{2\boldsymbol{\Gamma}_s} = \frac{1}{2} \frac{\dfrac{V_{\max 1}}{1 + \bar{x}_s/\bar{K}_i} + \dfrac{V_{\max 2}\bar{x}_s}{\bar{x}_s + \bar{K}_\mathrm{m}}}{\dfrac{V_{\max 1}/\bar{K}_i}{(1 + \bar{x}_s/\bar{K}_i)^2} + \dfrac{V_{\max 2}\bar{K}_\mathrm{m}}{(\bar{x}_s + \bar{K}_\mathrm{m})^2}}.$$

The signal-to-noise ratio can thus be computed as the ratio of the variance and the mean level (3.6) of c-di-GMP in the cell, resulting in:

$$\frac{\sigma^2}{\mu} \approx \frac{\boldsymbol{\Sigma_s}}{\bar{x}_s} = \frac{1}{2\bar{x}_s} \frac{\dfrac{V_{\max 1}}{1 + \bar{x}_s/\bar{K}_i} + \dfrac{V_{\max 2}\bar{x}_s}{\bar{x}_s + \bar{K}_\mathrm{m}}}{\dfrac{V_{\max 1}/\bar{K}_i}{(1 + \bar{x}_s/\bar{K}_i)^2} + \dfrac{V_{\max 2}\bar{K}_\mathrm{m}}{(\bar{x}_s + \bar{K}_\mathrm{m})^2}}, \tag{3.12}$$

where $\bar{x}_s$ is the positive root of the equation (3.6).

The results in figure 3.6 showed that signaling noise increases with

1. decreasing product inhibition ($\bar{K}_i >> \bar{x}_s$),
2. an action of PDE enzymes approaching saturation ($\bar{K}_\mathrm{m} << \bar{x}_s$,) and
3. with a vanishing difference of maximal catalytic velocities ($V_{\max 1} \approx V_{\max 2}$).

This has the implication for mean level equation (3.7) and the Fano factor that $\bar{K}_i + \bar{x} \approx \bar{K}_i$ and $\bar{K}_\mathrm{m} + \bar{x} \approx \bar{x}$. Substituting these parameter relations into the mean level equation (3.7) result in $\bar{x}_s \approx \sqrt{\bar{K}_i \bar{K}_\mathrm{m}}$ which is proportional to the increasing $\bar{K}_i$. Furthermore, inserting this into the equation for the Fano factor (3.12), results in

$$\frac{\boldsymbol{\Sigma_s}}{\bar{x}_s} \approx \frac{\sqrt{\bar{K}_i}}{2\sqrt{\bar{K}_\mathrm{m}}}, \tag{3.13}$$

which indicates that the Fano factor increases as the inverse square root of product inhibition (see Appendix A for derivation). A physical interpretation can be found in [20]. This effect is accounted to the independence of the derivative of the reaction rates $\boldsymbol{\Gamma}_s$ of the number of substrates (c-di-GMP) since this term determines the rate of return to the equilibrium. The described parameter regime results in $\boldsymbol{\Gamma}_s \approx -(V_{\max 1} + V_{\max 2})/K_i$ (see Appendix A), indicating that the absolute value of the restoring power $\boldsymbol{\Gamma}_s$ decreases due to a decreasing product inhibition of the DGC enzymes. Thus small perturbations lead to large deviations from the mean level and increase the signaling noise.

Stochastic fluctuations in a system without product inhibtion

For a comparison, the Fano factor in a c-di-GMP module without PI can be computed. From the macroscopic reaction rates (3.8) their derivative and the diffusion coefficient can be obtained

$$\boldsymbol{\Gamma}_s = \frac{\partial}{\partial \bar{x}} \left(V_{\max 1} - \frac{V_{\max 2} \bar{x}}{\bar{x} + \bar{K}_\mathrm{m}} \right) \Big|_{\bar{x}=\bar{x}_s} = -\frac{V_{\max 2} \bar{K}_\mathrm{m}}{(\bar{x}_s + \bar{K}_\mathrm{m})^2},$$

$$\boldsymbol{D_s} = V_{\max 1} + \frac{V_{\max 2} \bar{x}_s}{\bar{x}_s + \bar{K}_\mathrm{m}}.$$

The resulting Fano factor is

$$\frac{\sigma^2}{\mu} \approx \frac{\boldsymbol{\Sigma_s}}{\bar{x}_s} = \frac{1}{2\bar{x}_s} \frac{V_{\max 1} + \dfrac{V_{\max 2} \bar{x}_s}{\bar{K}_\mathrm{m} + \bar{x}_s}}{\dfrac{V_{\max 2} \bar{K}_\mathrm{m}}{(\bar{x}_s + \bar{K}_\mathrm{m})^2}} \tag{3.14}$$

where from equation (3.8) the macroscopic mean $\bar{x}_s$ is given by

$$\bar{x}_s = \frac{V_{\max 1} \cdot \bar{K}_\mathrm{m}}{V_{\max 2} - V_{\max 1}}.$$

As a limit $K_i \to \infty$ leading to a system without PI, equation (3.14) gives an insight into the source of the large stochastic fluctuations observed in figure 3.6. Assuming an action of the PDE enzymes at saturation, $K_{\mathrm{m}} << x$ and thus $(K_{\mathrm{m}} + x) \approx x$, results in the following noise-to-signal ratio:

$$\frac{\sigma^2}{\mu} \approx \frac{V_{\max 1} + V_{\max 2}}{V_{\max 1} - V_{\max 2}}, \tag{3.15}$$

which is inversely proportional to the difference of the maximal catalytic rates of the DGC and PDE enzymes. This results in a significantly larger amount of signaling noise in fig. 3.6 b than in a.

Response times of c-di-GMP regulation

So far only the stationary properties of c-di-GMP regulation have been analysed. As shown in figure 3.1, the expression levels of the enzymes regulating c-di-GMP levels are under the control of the master regulator RpoS. Stress-induced changes of RpoS expression levels lead to variations in the levels of DGCs and PDEs. As a result, the stationary distribution (3.5) of c-di-GMP also changes. It is thus of interest to know how fast the new stationary distribution is reached since the corresponding first passage times determine the responsiveness of signal transduction upon changes of stress levels. This can be stated as a first passage time problem e.g. from a system state corresponding to low amount of c-di-GMP x_0 to a state with a high amount of c-di-GMP $x_1, x_1 > x_0$, as shown in figure 3.7.

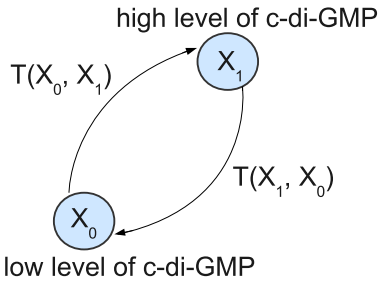


Figure 3.7: **Responsiveness of signal transduction stated as a first passage time problem.**
Low and high level states of the signaling molecule c-di-GMP are denoted by x_0 and x_1, respectively. The first passage time is a stochastic variable denoting the time that the system, started in a state x_0, requires to reach a state x_1. The probability distribution of the first passage times from x_0 to x_1 is denoted by $T(x_0, x_1)$. Accordingly. the opposite direction is described by $T(x_1, x_0)$.

Using the reaction rates (3.2) and (3.3) of the c-di-GMP regulation module, the results of section 2.4 from chapter 2 can be used to derive a recursion for the mean

first passage time

$$\mu(T_{x_0,x_1}) \begin{cases} \dfrac{1}{V_1(x_0)} + \dfrac{V_2(x_0)}{V_1(x_0)}, & \text{if } x_0 = x_1 - 1, \\ \dfrac{1}{V_1(x_0)} + \dfrac{V_2(x_0)}{V_1(x_0)} + \mu(T_{x_0+1,x_1}), & \text{if } x_0 \in \{0, x_1 - 2\}, \end{cases} \tag{3.16}$$

where T_{x_0,x_1} denotes the probability distribution of first passage times from a state x_0 to a state x_1 with $x_0 < x_1$ and $V_1(x_0), V_2(x_0)$ are the production and degradation rates, respectively.

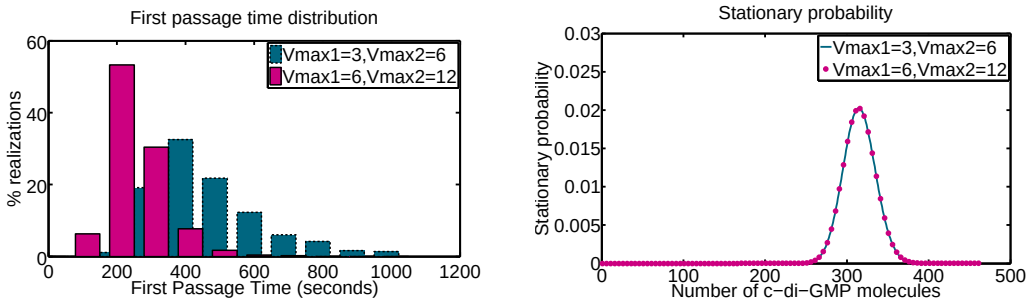
The values of the parameters K_i and K_m are usually constant in the cell, since they are based on intrinsic catalytic properties of the corresponding enzymes. The parameters which are subject to regulation in the signaling system (and are thus of interest in this analysis) are $V_{\max 1}$ and $V_{\max 2}$ due to their dependance on the expression levels of DGC and PDE enzymes, respectively. The first ratio $1/[V_1(x_0)]$ in the iteration scheme (3.16) is given by $1/[V_1(x_0)] = (x_0 + K_i)/(V_{\max 1} K_i)$. This indicates that the mean first passage time is inversely proportional to $V_{\max 1}$ and since $V_{\max 1} = ESS \cdot k_3$ (section 3.2.1), it is also inversely proportional to the expression level of DGC enzymes. Since in the second ratio $V_2(x_0)/V_1(x_0)$ it holds that $V_2(x_0) = (V_{\max 2} \cdot x_0)/(K_m + x_0)$, the first passage time is proportional to $V_{\max 2} = (E_2 + E_2 X) \cdot k_5$ (eq. (3.3)). However, similarly to the stationary distribution (3.5), the degradation rate is scaled by the production rate. Thus an increase in the expression levels of DGCs can result in shorter response times of the system, but if a simultaneous increase of expression levels of PDEs takes place, the stationary distribution of c-di-GMP levels does not change.

In an opposite situation, where the time of decrease of c-di-GMP levels from x_1 to x_0 is computed (see fig. 3.7), the recursion scheme for the mean first passage time results in

$$\mu(T_{x_1,x_0}) \begin{cases} \dfrac{1}{V_2(x_1)} + \dfrac{V_1(x_1)}{V_2(x_1)}, & \text{if } x_1 = x_0 + 1, \\ \dfrac{1}{V_2(x_1)} + \dfrac{V_1(x_1)}{V_2(x_1)} + \mu(T_{x_1-1,x_0}), & \text{if } x_1 > x_0 + 2. \end{cases} \tag{3.17}$$

In this case the first passage time depends on the degradation rate $V_2(x)$ and the ratio of $V_1(x)$ and $V_2(x)$, indicating that the response time varies with $V_{\max 2}$, although the stationary probability distribution does not change if a proportional change in $V_{\max 1}$ takes place.

Figure 3.8 shows the first passage time distribution in the situation $x_0 < x_1$, corresponding to eq. (3.16). Two different configurations of $V_{\max 1}$ and $V_{\max 2}$ with 1000 SSA simulations each where conducted. The two distinct parameter settings result in different first passage time distributions (fig. 3.8 a). However, since the ratio



(a) (b)

Figure 3.8: **First passage time distribution of c-di-GMP synthesis depends on maximal catalytic velocity $V_{\max 1}$ which is in turn determined by the expression levels of DGC enzymes.**
(a) First passage times were computed with a start state $x_0 = 30$ and end state $x_1 = 314$ (mean of the stationary probability) for two different parameter configurations: $V_{\max 1} = 3s^{-1}, V_{\max 2} = 6^{-1}$ (cyan) and $V_{\max 1} = 6s^{-1}, V_{\max 2} = 12s^{-1}$ (red) using 1000 Gillespie trajectories for each configuration. The latter configuration reduces the response times and shifts the FPT-distribution to the left. The other parameters are chosen to be $K_{\mathrm{m}} = 480$, $K_i = 1200$ (both have units # molecules). **(b)** Since the ratio of $V_{\max 1}/V_{\max 2} = 0.5$ is constant for both parameter configurations, the stationary probability of c-di-GMP numbers does not change (see also eq. (3.5)).

$V_{\max 1}/V_{\max 2}$ remains constant, the stationary distribution does not change.

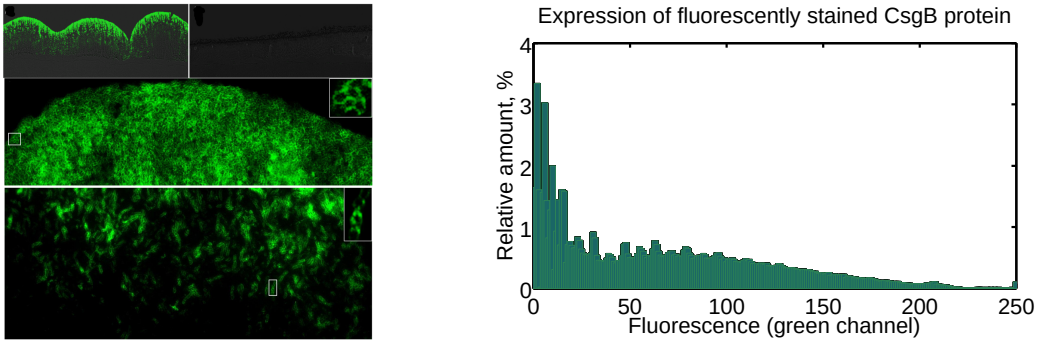
The biological implication of this insight becomes evident if the interaction network of DGCs and PDEs is considered (fig. 3.1). Since the production and degradation enzymes of c-di-GMP are induced by the same master regulator RpoS, their expression levels might vary simultaneously, depending on different stress levels or growth phases of the bacterial population. It is thus a realistic scenario, that an increase of RpoS levels does not significantly change the levels of c-di-GMP but it increases the responsivity of the system. Thus adjustments to c-di-GMP levels can take place much faster as if a stress-induced alarm mode is activated.

3.2.3 Impact of c-di-GMP dynamics on the expression of curli fimbriae

A major role of signal transduction networks is a processing of external stimuli and initiation of adequate responses. In many situations these responses are based on discrete decision making processes where single cells choose between different fates and the cell population becomes heterogeneous. Signaling systems enabling such behaviour consist of multiple stationary states and due to the intrinsic noise of signal transduction each stationary state contains different amounts of the total probability mass. **Bistability** is a special case of this behaviour and it has been shown to be involved in decision making processes of various signaling systems [70].

Initially, bistability was described in the lactose-regulation system of *E. Coli* by the lac-operon [52]. Later on, it has been found in other gene regulatory systems such as the lytic pathway regulating the response of *E. Coli* cells to an infection by the virus bacteriophage λ [3, 74] or the ability of uptake of external DNA, regulated by the competence mechanism [69]. All of these signaling systems contain network motifs with direct positive autoregulation or indirect one, based on double negative feedback, such as the interaction between YciR and the global pool of c-di-GMP in fig. 3.1 [2]. In a recent theoretical study it was suggested that bistable decision making confers a selection advantage to bacteria in environments with changing conditions and may appear naturally during evolution if the dynamics are sufficiently noisy [44].

There is strong evidence that the expression of the CsgB protein (fig. 3.1), which gives rise to curli fibers and the resulting bacterial biofilm, exhibits bistable dynamics. Discrete all-or-nothing expression of curli was qualitatively observed in *E. Coli* [67] and quantitatively measured in its genetically close relative Salmonella [31]. Indication for the population heterogeneity has been found in *E. Coli* using fluorescence microscopy imaging as shown in fig. 3.9 (a). It shows images of a 7-day old bacterial colony, where the protein CsgB was stained by a green fluorescent dye. The middle image in fig. 3.9 (a) shows the intensity of fluorescence at the top of the



(a) (b)

Figure 3.9: **Quantification of bistable expression of CsgB using fluorescence microscopy.**
(a) Electron microscopic images of a 7-day old colony of the W3110-strain of *E.Coli* K-12 with a fluorescent staining of the CsgB protein. The upper left figure shows the top of the colony characterized by high fluorescence intensity due to a majority of cells producing the curli-protein CsgB. To its right a dark image of a control CsgB-knockout mutant colony is shown, which does not produce curli. The middle image shows a zoom into the upper part of the wild type colony confirming that most of the cells express CsgB. The image at the bottom is a zoom into the central region of the same population where the relative frequency of curli-expressing cells is obviously lower than on the top of the culture. *Copyright: Diego Serra and Regine Hengge.* **(b)** Histogram of the green channel intensity of image pixels from figure (a) ranging between 0 and 255. The zooms into the upper and the central regions of the colony give rise to a bimodal distribution of fluorescence indicating a bistable expression of CsgB.

colony, which is highly illuminated, indicating a large number of CsgB expressing cells. The image at the bottom shows a zoom into the central part of the colony where CsgB expression is turned on in a relatively small amount of cells.

The histogram in fig. 3.9 (b) depicts the bimodal distribution of the green channel intensity (0 to 255) of the pixels in the two zoom images. The first mode of the distribution is concentrated around a mean value ≈ 80 and represents the high number of CsgB expressing cells at the top of the colony, while the second mode is characterized by a mean ≈ 0 and a significantly smaller spread, representing the low amount of CsgB expressing cells in the central region. This significant difference between the two fluorescence images is a qualitative indicator of the bistable distribution. For a quantitative assessment of population heterogeneity, single cell expression measurements, as conducted for Salmonella [31], are needed.

Although the mechanism behind the bistable behaviour is not fully dissected yet, it has been suggested that c-di-GMP might play a decisive role [31]. A recent study on interactions in the curli signaling network suggests that the PDE enzyme YciR is a key antagonist of curli expression [46] which, on the one hand, reduces the global level of c-di-GMP and is being subject to negative control by an inactivation through c-di-GMP, on the other. The resulting reactions 4 and 5 in fig. 3.1 give rise to a double negative feedback and are thus a potential source for bistability. The dynamical system induced by these interactions is the main focus of this section. Based on the analysis of c-di-GMP dynamics from the previous section, the reaction rates of the system will be derived and using a bifurcation analysis, parameter regimes will be assessed which induce bistability. The corresponding Master Equation enables to compute the probability distribution in the two stationary states. Finally, promoter activity measurements of the *csgB* gene are used in order to compare experimental measurements with a theoretical bistable model for the activity of the *csgB* promoter.

Chemical Master Equation and the reaction rate equations

The model of interaction between YciR and global c-di-GMP is based on the observation that the YciR molecule may be present in distinct states: active (unbound) state and an inactive state, which is bound by c-di-GMP. Thus the interaction system consists of three major molecular species: c-di-GMP, x_1, active YciR, x_2, and inactive YciR ($\mathrm{YciR}_{\mathrm{total}}$ - x_2), where $\mathrm{YciR}_{\mathrm{total}}$ denotes the total number of YciR molecules. For simplicity it is assumed that the total number of YciR protein molecules does not change and, at least on the time scale of c-di-GMP dynamics, it is at a steady state. The resulting reaction system consists of production and degradation of c-di-GMP, inactivation of YciR by c-di-GMP, and reactivation of YciR due to dissociation of c-di-GMP. This gives rise to the following rate functions:

1. Production of c-di-GMP by YegE and YedQ, combined into a single reaction rate and a lumped dissociation constant, K_i^{YegE}, for c-di-GMP binding at the

allosteric product inhibitory site (reactions 2 and 3 in figure 3.1):

$$V_1(x_1) = \frac{V_{\max 1}}{1 + x_1/K_i^{\text{YegE}}}.$$

This reaction corresponds to the production rate of c-di-GMP (3.2) derived in section 3.2.1.

2. Degradation of c-di-GMP by YhjH (reaction 1 in figure 3.1):

$$V_2(x_1) = \frac{V_{\max 2} \cdot x_1}{x_1 + K_{\text{m}}^{\text{YhjH}}},$$

corresponding to the degradation rate (3.3) derived in section 3.2.1.

3. Degradation of c-di-GMP by YciR (reaction 5 in figure 3.1):

$$V_3(x_1, x_2) = k_{\text{YciR\_act}} \cdot x_2 \cdot \frac{x_1}{x_1 + K_{\text{m}}^{\text{YciR}}},$$

also corresponding to eq. (3.3), according to which $k_{\text{YciR\_act}} \cdot x_2 = V_{\max(\text{YciR})}$ denotes the maximal catalytic velocity of YciR.

4. Deactivation of YciR due to binding of c-di-GMP, where K_d is the dissociation constant of c-di-GMP (reaction 4 in figure 3.1):

$$V_4(x_1, x_2) = k_{\text{YciR\_de}} \cdot x_2 \cdot \frac{x_1}{x_1 + K_d}.$$

5. Dissociation of c-di-GMP and reactivation of YciR:

$$V_5(x_2) = c_5 \cdot (\text{YciR}_{\text{total}} - x_2).$$

By defining $\{X(t)\}_{t \geq 0}$ as a Markov jump process on a state space S, described by the species vector $\mathbf{x} = \{x_1, x_2\}$, the corresponding probability distribution $P(x_1, x_2, t) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \text{time} = t)$, is obtained from the solution of the following Chemical Master Equation

$$\begin{aligned} \frac{\partial P(x_1, x_2, t)}{\partial t} &= V_1(x_1 - 1) \cdot P(x_1 - 1, x_2) \\ &+ [V_2(x_1 + 1) + V_3(x_1 + 1, x_2)] \cdot P(x_2 + 1, x_2) \\ &+ V_4(x_1, x_2 + 1) \cdot P(x_1, x_2 + 1) + V_5(x_2 - 1) \cdot P(x_1, x_2 - 1) \\ &- (V_1(x_1) + V_2(x_1) + V_3(x_1, x_2) \\ &+ V_4(x_1, x_2) + V_5(x_2)) \cdot P(x_1, x_2). \end{aligned} \tag{3.18}$$

In order to analyze parameter regimes inducing bistability, the approximating reaction rate equation can be used

$$\frac{d\bar{\mathbf{x}}}{dt} = \boldsymbol{f}(\bar{\mathbf{x}}), \tag{3.19}$$

where $\bar{\mathbf{x}} = \mathbf{x}/\Omega$ and Ω is the system volume and $\boldsymbol{f}(\bar{\mathbf{x}}) = \{f_1(\bar{x}), f_2(\bar{x})\}$ is a vector-valued function given by

$$\begin{aligned} f_1(\bar{\mathbf{x}}) = \frac{d\bar{x}_1}{dt} &= \frac{1}{\Omega}\left[\frac{V_{\max 1}}{1+\bar{x}_1/\bar{K}_i^{\mathrm{YegE}}} - \frac{V_{\max 2}\bar{x}_1}{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{YhjH}}}\right] \\ &- k_{\mathrm{YciR\_act}}\cdot\bar{x}_2\frac{\bar{x}_1}{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{yciR}}}, \\ f_2(\bar{\mathbf{x}}) = \frac{d\bar{x}_2}{dt} &= c_5\cdot\left[\overline{\mathrm{YciR}}_{\mathrm{total}} - \bar{x}_2\right] - k_{\mathrm{YciR\_de}}\cdot\bar{x}_2\cdot\frac{\bar{x}_1}{\bar{x}_1+\bar{K}_d}, \end{aligned}$$

where $\bar{K}_i^{\mathrm{YegE}} = K_i^{\mathrm{YegE}}/\Omega, \bar{K}_{\mathrm{m}}^{\mathrm{YhjH}} = K_{\mathrm{m}}^{\mathrm{YhjH}}/\Omega, \bar{K}_{\mathrm{m}}^{\mathrm{YciR}} = K_{\mathrm{m}}^{\mathrm{YciR}}/\Omega, \bar{K}_d = K_d/\Omega$ and $\overline{\mathrm{YciR}}_{\mathrm{total}} = \mathrm{YciR}_{\mathrm{total}}/\Omega$.

The two nullclines of this system result from setting $f_1(\bar{\mathbf{x}}) = 0$ and $f_2(\bar{\mathbf{x}}) = 0$ and the fixed points follow from their intersections, given by the roots of the following equation (for the x_1-component):

$$\begin{aligned} 0 &= \frac{1}{\Omega}\left[\frac{V_{\max 1}}{1+\bar{x}_1/\bar{K}_i^{\mathrm{YegE}}} - \frac{V_{\max 2}\bar{x}_1}{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{YhjH}}}\right]\cdot\left[\frac{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{YciR}}}{\bar{x}_1}\right] \\ &- c_5\cdot\overline{\mathrm{YciR}}_{\mathrm{total}}\cdot k_{\mathrm{YciR\_act}}\Big/\left[c_5 + k_{\mathrm{YciR\_de}}\cdot\frac{\bar{x}_1}{\bar{x}_1+\bar{K}_d}\right]. \end{aligned} \tag{3.20}$$

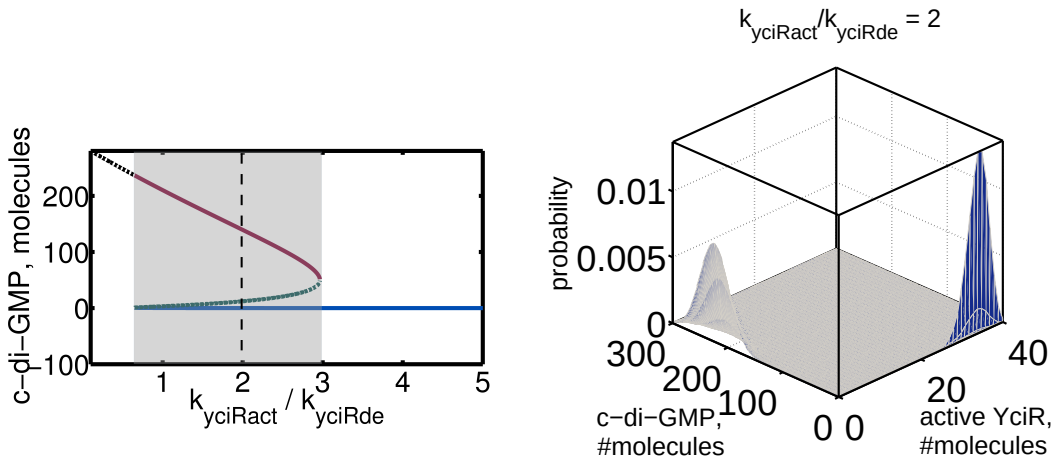
For constraining the parameter space, we focus on the situation where the affinity of c-di-GMP binding for YciR is relatively high, which is in line with the known low levels of c-di-GMP in the cell, ranging from a few molecules to a few thousand at its maximum [35]. As a result, it holds that $\bar{K}_d << \bar{x}$ and $c_5 << k_{\mathrm{YciR\_de}}$ implying that c-di-GMP binding to YciR works almost at saturation and the rate of c-di-GMP dissociation from YciR is very low. It follows that eq. (3.20) can be approximated by

$$\begin{aligned} 0 &\approx \frac{1}{\Omega}\left[\frac{V_{\max 1}}{1+\bar{x}_1/\bar{K}_i^{\mathrm{YegE}}} - \frac{V_{\max 2}\bar{x}_1}{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{YhjH}}}\right]\cdot\left[\frac{\bar{x}_1+\bar{K}_{\mathrm{m}}^{\mathrm{YciR}}}{\bar{x}_1}\right] \\ &- c_5\cdot\overline{\mathrm{YciR}}_{\mathrm{total}}\cdot\frac{k_{\mathrm{YciR\_act}}}{k_{\mathrm{YciR\_de}}}. \end{aligned}$$

This equation shows that in this parameter regime the ratio between the rate of activation of YciR, $k_{\mathrm{YciR\_act}}$, and its deactivation, $k_{\mathrm{YciR\_de}}$, determines the intersection of the nullclines and thus the location of the fixed points. For analysing the influence of the ratio $k_{\mathrm{YciR\_act}}/k_{\mathrm{YciR\_de}}$ on the fixed points of the system, the parameter $k_{\mathrm{YciR\_de}}$ was set 1 and in order to conduct a bifurcation analysis, the parameter $k_{\mathrm{YciR\_act}}$ was varied from 0.1 to 5. The other system parameters were chosen either from known *in-vitro* and *in-vivo* measurements or if corresponding data was lacking, they were estimated to be in a realistic range: $K_i^{\mathrm{YegE}} = 1200$, $K_m^{\mathrm{YhjH}} = 480$ $K_m^{\mathrm{YciR}} = 2$, $K_d = 24$, and YciRtotal = 40 (molecules), $c_5 = 0.015$ molecules^{-1} s^{-1} $V_{\max 2} = 5s^{-1}$, $V_{\max 1} = 3s^{-1}$. Prior to the analysis, the stochastic parameters were

scaled by the system volume as described for eq. 3.19.

The resulting bifurcation diagram is depicted in figure 3.10 a. It shows that in a certain range of the ratio between $k_{\text{YciR\_act}}$ and $k_{\text{YciR\_de}}$ $(0.5 - 3)$ the system exhibits three fixed points, of which two are stable due to real negative eigenvalues of the Jacobi matrix. Below and above this range there is only one fixed point with either vanishing or high c-di-GMP levels. The corresponding solution of the Chemical Master Equation 3.18 for the bistable parameter regime is depicted in figure 3.10 b, showing that the Markov process consists of two communication classes.



(a) (b)

Figure 3.10: **Analysis of parameter regions inducing bistability.**
(a) Fixed points of the c-di-GMP/YciR interaction system, depend on the ratio between the activity of YciR $\mathbf{k_{yciR\_act}}$ and its deactivation rate by c-di-GMP $\mathbf{k_{yciR\_de}}$. In the range of $k_{\text{yciR\_act}}/k_{\text{yciR\_de}}$ from 0 to 0.5 the system is monostable with a high number of c-di-GMP molecules (black dashed part of the curve). Between the values 0.5 and ~ 3 the system exhibits bistable behavior (shaded region). The continuous parts of the curve (red and blue) indicate the stable fixed points and the dashed green part indicates an unstable fixed point. Within parameter regions where the ratio is higher than 3, the system becomes monostable again with a vanishing c-di-GMP population (and high YciR levels, not shown). **(b)** Numerical solution of the Chemical Master Equation using the parameter ratio $k_{\text{yciR\_act}}/k_{\text{yciR\_de}} = 2$. (vertical black dashed line in figure a). The first stationary state corresponds to a high amount of c-di-GMP molecules and low level of active YciR molecules. The second one corresponds to a zero population level of c-di-GMP and a high amount of active YciR. The corresponding probabilities are 0.8 and 0.2, respectively.

Analysis of experimental data

The results of the previous section indicate that the interaction between c-di-GMP and YciR has the potential of generating heterogeneity in bacterial cells. Accord-

ing to this model, molecular noise stochastically drives the system into one of the two stationary states where, qualitatively, c-di-GMP levels are either significantly increased or are close to zero. Furthermore, as shown by the interaction network in fig. 3.1, the level of c-di-GMP in the cell regulates the expression of the curli protein CsgB. Thus the first stationary state of the bistable YciR/cdiGMP-model, corresponding to low c-di-GMP levels, could be responsible for generating the curli-off phenotype where the expression of the CsgB protein is turned down (dark regions in fig. 3.9 a). Also, the stationary state with high c-di-GMP levels might give rise to an induction of the CsgB expression machinery (green fluorescent regions in fig. 3.9 a)

In order to find experimental support for this hypothesis, promoter activities of the *csgB*-gene from [71] were used, which were measured during induction of the stationary phase. The corresponding measurements reflect the mean promoter activity of *csgB* in a bacterial population, resulting from the expression level of the reporter gene β-galactosidase. This mean level can be computed as the sum of probability-weighted means in each of the two stationary states

$$\mu = \sum_{x \in x_{s1}} x P_{csgB}(x) + \sum_{y \in x_{s2}} y P_{csgB}(y), \tag{3.21}$$

where $P_{csgB}(\cdot)$ denotes the bistable probability distribution of the *csgB*-promoter activity and x_{s1} and x_{s2} denote the two corresponding communication classes. As discussed above, the level of c-di-GMP determines the expression of the *csgB*-gene. Thus we assumed that the probability distribution of c-di-GMP, resulting from the solution of eq. (3.18), can be used to find an approximation to the probability distribution of the *csgB*-promoter activity, that we denote by P_{csgB}. To this end, as the most simple model, c-di-GMP levels with non-zero probability were scaled by a constant proportionality factor, to yield

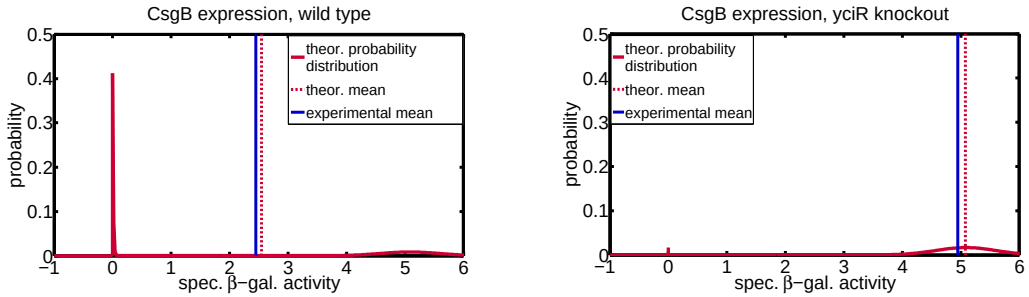
$$P_{csgB} = P(a \cdot x), \tag{3.22}$$

where $P(.)$ is the solution of the Master equation (3.18), x denotes the number of c-di-GMP molecules with non-zero probability, and a is a constant factor.

In order to fit mean level promoter activity data, the mean of the probability distribution P_{csgB} from eq. (3.21) was computed for two different genetic backgrounds: wild type and *yciR*-gene knockout mutant. In order to compute the probability $P_{\mathrm{cdiGMP}}(x)$ in the wild type, the parameters of the Master equation (3.18) were chosen as described in the bifurcation analysis (fig. 3.10). For computing the probability $P_{\mathrm{cdiGMP}}(x)$ in the *yciR*-mutant the total number of YciR molecules, $\mathrm{YciR}_{\mathrm{total}}$, was set to zero. In order to estimate the proportionality factor, the following objective function was minimized

$$a = \arg\min_{a} \sum_{i} \left(\frac{\mu^i - \mu^i_{\mathrm{exp}}}{\mu_{\mathrm{exp}}} \right)^2,$$

where i denotes the respective genetic background (here: wild type and *yciR*-knockout). μ^i is computed according to eq. (3.21) and (3.22), μ^i_{exp} is the experimentally measured promoter activity of the *csgB* gene. Based on promoter activity data from [46], the minimization of the objective function yielded an optimal scaling parameter $a = 0.021$. The experimental mean promoter activities and the fitted theoretical means are shown along with the corresponding theoretical probability distributions in fig. 3.11. The results indicate that the bistable



(a) (b)

Figure 3.11: **A bistable model of CsgB induction explains the experimentally measured mean expression level of the CsgB protein.**
The probability distribution P_{csgB} from equation (3.22), the corresponding mean value and an experimentally measured mean promoter activity are shown for **(a)** the wild type genetic background, where the CME (3.18) was parametrized, as described in the bifurcation analysis above; **(b)** *yciR*-knockout mutant genetic background, where the parameter $\text{YciR}_{\text{total}}$ was set to zero. The experimental mean promoter activities were measured in K-12 W3110 *E.Coli* strains 24 hours after the start of colony growth.[46].

model of *csgB*-induction explains experimentally measured expression levels sufficiently well. This supports the hypothesis that a bimodal distribution of c-di-GMP is a potential source for generating bistability of curli expression. Furthermore, the model enables a new qualitative interpretation of the population level promoter activities as a mean of a bimodal distribution resulting from the amount of probability mass in each of the two stationary states. This interpretation suggests that the promoter activity measured in a wild type population (fig. 3.11 a) results as the mean expression in curli-on cells with high c-di-GMP levels and curli-off cells with low c-di-GMP levels. This model implies that in a *yciR*-knockout, all cells are characterized by high c-di-GMP levels, and thus have curli expression turned on. For a better quantitative validation and a more exact parameter estimation, further experimental measurements on a single cell level are required.

3.3 Discussion and outlook

For modeling the dynamics of c-di-GMP regulation, the expression and activation levels of the catalysing enzymes were assumed to be constant (i.e. constant $V_{\max 1}$

and $V_{\max 2}$). Since gene expression, protein translation and different possible activation levels of the enzymes (e.g. due to varying stress conditions) might change these parameters, further analysis is required in order to study the resulting consequences for the dynamics. Furthermore, sufficiently accurate estimates of model parameters of the signaling system may enable a model of a better resolved transition dynamics between the two stationary states of the bistable curli expression system.

Two sets of experimental data were used for model validation: fluorescence microscopy and promoter activity measurements. The advantage of the fluorescence microscopy images is an indication of the qualitative properties of CsgB expression in the *E. Coli* population. However, their drawback is an imaging of protein expression only on the surface of the colonies and a lacking distinctness of expression levels in single cells. The measurements of promoter activities yield a basis for quantitative validation of the modeling results. Due to the inherently multistable character of the curli expression, the measurements have to be interpreted as means of a bimodal distribution. This study introduces a probability-based method for analysing such data and is suggested for verification using further experimental measurements.

3.4 Summary and conclusion

The focus of the present study was a combination of two approaches for studying the signal transduction in the curli expression network. Firstly, the dynamics of the cellular levels of the signaling molecule c-di-GMP was modeled, based on the known qualitative properties of DGC and PDE enzymes. Secondly, qualitative and quantitative measurements of curli expression were analysed and a dynamic relationship between c-di-GMP regulation and the expression of the curli gene *csgB* was established.

Modeling of c-di-GMP dynamics enabled an insight into regulatory properties of the signal transduction. Thus, the steady state of c-di-GMP in the cell was shown to be determined by the ratio of expression levels of DGC and PDE enzymes. The model also revealed that the product inhibition property of DGC enzymes contributes to an increased robustness of the steady state with respect to changes in expression levels of the enzymes and reduced signaling noise. Furthermore, the computation of the mean first passage times has shown that response times of the system are reduced when expression levels of the DGCs and PDEs are increased although the stationary distribution of molecular numbers does not necessarily exhibit a significant change. This may be an indicator for a transition to an alarm mode where the cell is ready for a fast response upon the onset of stress conditions.

The results of the c-di-GMP regulation model were used to explain the qualitatively observed bistable expression of the curli protein CsgB. To this end, it was

suggested that the recently found double negative interaction between c-di-GMP and YciR may be the source for generating bistable behaviour of this system. Corresponding parameter regimes of this dynamical interaction model were finally used to compute the probability distribution of c-di-GMP and approximate the resulting bistable probability distribution of *csgB* gene expression. The resulting theoretical mean value showed a good agreement with experimentally measured mean expression levels, yielding further support for the hypothesized source of bistable curli expression. The presented method may be considered as a general probability-based approach for evaluating population average expression data of bimodally expressed genes.

Chapter 4

Drug selection pressure and evolution of HIV

4.1 Modeling viral evolution in the presence of drug application

4.1.1 Introduction

Treatment of HIV is one of the major challenges of the modern medicine. Although it can not be completely cured, life-long treatment can significantly prolong life expectancy. Currently recommended therapies are based on a combination of nucleoside reverse transcriptase inhibitors (NRTI) and non-nucleoside reverse transcriptase inhibitors (NNRTI) or protease inhibitors (PI). A main obstacle aggravating HIV therapy and preventing its elimination is the high complexity of its evolutionary dynamics. Its genomic variability confers HIV the necessary degree of flexibility and robustness for surviving in different environmental conditions. Although in the presence of drugs the population growth can be strongly reduced, the resulting selection pressure and the high intrinsic mutability can produce genetically altered viruses which are either less vulnerable or completely resistant to the action of the drugs. An understanding of evolutionary pathways which lead to drug resistance and let HIV escape drug application is key to an implementation of improved therapeutic regimens.

We develop a mathematical approach to modeling HIV dynamics in the presence of drug treatment by integrating the key determinants of viral evolution. Furthermore, we show that intrinsic stochasticity of viral growth under drug action at low doses adds an additional complexity layer to the system and must be considered for its understanding. Prior to deriving a full stochastic model, in the following, we introduce the main theoretical concepts which give rise to viral population models. Since we are interested in the principles of resistance development of HIV under drug treatment, we will firstly clarify the notions of *mutation*, which enables the genetic flexibility of HIV and *selection*, inducing the genetic evolution and making resistance development possible.

Mutation of the viral DNA appears as a result of randomly occurring errors during the process of reverse transcription. Most of the mutant viral species are less viable than the wild type species, which prevents the new mutations from constituting

within the viral population. However, depending on the specifics of the environment, such as an application of antiviral drugs, mutations can be advantageous and cause the mutant species to outgrow the wild type.

Selection is a nonlinear function of various evolutionary factors regulating the level of (dis-)advantage of new mutations and their likelihood to establish within the population. While mutation is the random mechanism behind the evolution of HIV, selection can be considered as the driving force of evolutionary trajectories. There are two main factors contributing to the regulation of the dynamics of selection. The first one is the relative **fitness** of a viral species, which is determined by its genomic background. It is assumed that a wild type genome confers a maximal fitness of 1 and that mutations giving rise to genetically different species, induce a fitness loss. Mathematically, this can be described by a function $f(k) : \mathbb{N} \mapsto \mathbb{R} \in [0, 1]$, where k is an index referring to the specific genome composition out of a set of all possible genomes. Thus $f(k) = 1$ means that a virus carrying the genome k has a maximal possible fitness, and $f(k) = 0$ means that its viability is completely lost and the virus strain k dies out.

If a viral population is exposed to a drug treatment then the second main factor contributing to the selection pressure is the level of **resistance** of a species with respect to a particular drug. It can be formalized in the simplest way by means of a function $\eta(j, k) : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{R} \in [0, 1]$, describing the effect of a treatment, where the integer j denotes the specific condition induced by the medical treatment out of a set of all possible treatments. As before, the index k denotes the genetic background. The impact of a treatment on the viability of the virus is determined by the application of the drug, whose effect $\eta(j, k)$ is computed using the $\mathbf{E_{max}}$ **equation** (or equivalently, median-effect equation):

$$\eta(j, k) = \frac{[D_j]}{\mathrm{FR}(k) \cdot \mathrm{IC}_{50} + [D_j]}, \tag{4.1}$$

where $[D_j]$ denotes the concentration of the respective drug within the condition j and IC_{50} denotes the half-maximal inhibitory drug concentration, w.r.t. to a reference (wild type) strain. In other words, it is the concentration of the drug required to reduce the rate of viral growth by 50 %. Finally, $\mathrm{FR}(k)$ is the fold-resistance w.r.t. to the applied drug relative to the wild type strain. This parameter indicates the factor by which the IC_{50}-value, and thus the resistance, of the mutant strain k is higher than the wild type. Obviously, it follows that $\mathrm{FR}(k) = 1$ if the index k refers to the wild type. Note that in the case of HIV, the HBX2-strain from the HIV Drug Resistance Database of the Stanford University is usually considered as the standard reference viral strain [63].

The interpretation of the IC_{50}-value is explained by two hypothetical dose-response curves of a wild type and a corresponding mutant population in figure 4.1. The dose-response curve of the wild type population (reference strain) indicates that

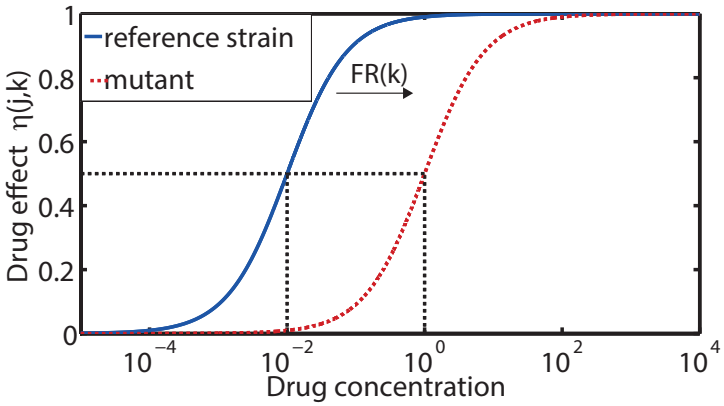


Figure 4.1: **Theoretical dose-response curves of a wild type and a mutant population with different IC_{50}-values.**
The effect $\eta(j,k)$ of the drug is depicted depending on drug concentration (x-axis, relative units) and the genomic backgrounds: wild type (blue curve) and mutant (red dashed curve). The IC_{50}-value of the wild type is 0.01 and the fold resistance conferred by the mutation is $FR(k) = 100$. The increased resistance of the mutant strain leads to a shift of the dose-response curve, implying that a higher drug concentration is needed to achieve the same effect for the mutant strain compared to the wild type.

the drug of interest reaches 50 % of its effect at a concentration of 0.01 (relative units), as shown in fig. 4.1, blue curve. This indicates that $IC_{50} = 0.01$. However, a mutation at the site of action of the drug, for instance an amino acid exchange in the binding pocket of the reverse transcriptase enzyme, might make the drug less effective. If the mutation constitutes within the population, this leads to a shift of the dose-response curve of the drug w.r.t. the mutant population, meaning that a higher drug concentration is needed to achieve the same effect for the mutant strain compared to the wild type. The dose-response curve of the mutant indicates that $FR(k) \cdot IC_{50} = 1$, increasing the 50 % inhibitory concentration by a factor of $FR(k) = 100$ relative to the wild type population.

What are the determinants for the likelihood of a randomly occurred mutation to become constituted in the population? As already mentioned, the process of selection is considered to be the result of an interplay between external conditions, such as drug treatment, the fold resistance conferred by the mutation of interest and the fitness loss that this mutation induces, relative to the wild type background. As an extreme example consider a virus species carrying an amino acid substitution at a drug binding site of a target enzyme due to a mutation event. As a result, the virus might be completely resistant to the action of the drug. However, this mutation still might not become constituted in the population during drug treatment, if it simultaneously makes the viral enzyme dysfunctional and contributes to an extreme reduction of the viral fitness. The combined independent effect of fitness and resistance on the growth rate of the viral population can be quantified by the

following product:

$$r(j,k) = r_{\varnothing} \cdot (1 - \eta(j,k)) \cdot f(k), \tag{4.2}$$

where $r(j,k)$ denotes the (time-independent) growth rate of the viral population with genomic background k under treatment j and $r_{\varnothing}$ is the growth rate of the reference (wild type) strain in the absence of drug treatment. Furthermore, $\eta(j,k) \in [0,1]$ and $f(k) \in [0,1]$ denote the drug effect and the fitness of the population, respectively.

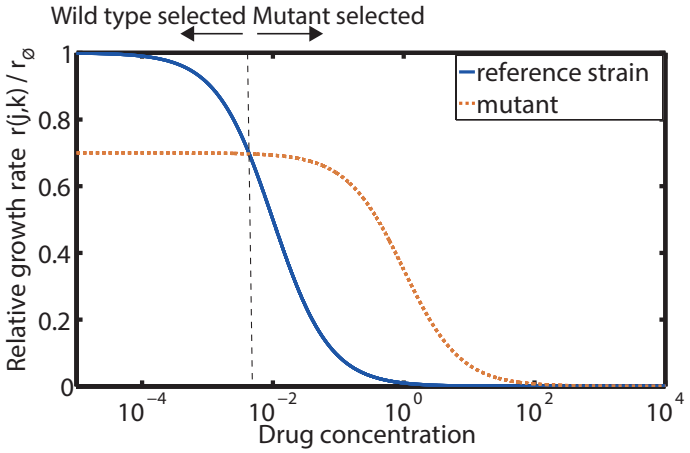


Figure 4.2: **Combined effect of fitness and resistance regulates selection of new mutations.**
Growth rates of a wild type and a mutant population are depicted, with a varying drug concentration. The IC_{50}-values of the wild type and the mutant are 0.01 and 1, respectively. The fitness of the mutant virus is $f(k) = 0.7$, relative to a fitness of 1 of the wild type. The crossing of the two curves marks a selection boundary. Drug concentrations above the selection boundary confer the mutant strain a selection advantage and lead to the constitution of the mutation in the population. Above the selection boundary the wild type virus outgrows the mutant strain and prevents a selection of the mutation.

The combined effect of resistance and fitness on the growth rate $r(j,k)$ of the virus and the resulting regulation of mutational selection is depicted in figure 4.2. The figure depicts the growth rate of the wild type and the mutant population in relative units as a function of the drug concentration. As in figure 4.1, the IC_{50}-value of the wild type and the mutant are 0.01 and 1, respectively. In addition to an increase of IC_{50}, the changed genetic background of the mutant virus introduces a fitness loss w.r.t the wild type, as described in equation (4.2). While the reference wild type strain has a fitness of 1 (blue continuous curve), the relative fitness of the mutant strain is assumed to be 0.7 (red dashed curve). The cross point of the two curves marks a selection boundary. Below the drug concentration of $\approx 0.0043 (= 10^{-2.37})$ the wild type has a higher growth rate $r(j,k)$ than the mutant due to the fitness loss induced by the mutation. With increasing drug concentration the highly susceptible wild type population becomes diminished by the action of the drug. However,

due to the (partial) resistance, induced by the mutation, the dose-response curve of the mutant population is shifted to the right, as it was discussed in figure 4.1. As a consequence, the mutant population continues growing at drug doses, which are deadly for the wild type, leading to a higher relative growth than the wild type at these high drug doses. Thus the new mutation is likely to become constituted within the population if the drug concentration exceeds the selection threshold of 0.0043 (area right to the dashed line in fig. 4.2). Below this threshold, the wild type species will outgrow the mutants and prevent the new genotype to be selected in successive virus generations (area left to the dashed line in fig. 4.2).

4.1.2 Aims, scope and the modeling strategy

In the present study we derive a stochastic model of viral growth, subject to application of NRTI and NNRTI drugs in order to dissect the phenotypic impact of different mutational events in terms of drug resistance and fitness. Incorporation of *in-vitro* measured viral passage data will enable an estimation of the resistance level and the fitness loss conferred by mutations occurring in the course of the genetic evolution. The **central question of the study** can thus be summarized as follows: given an observed time-resolved genotype of the virus as a sequence of mutational events occurred during drug treatment, what are the mechanistic principles of selection which gave rise to the particular evolutionary trajectories? We derive a mathematical framework for modeling genetic evolution of HIV under drug treatment. Stochastic modeling of viral population growth enables us to estimate biologically relevant parameters and find plausible models explaining the observed *in-vitro* genetic evolution. This yields new insights into mechanistic principles of viral infection dynamics and drug treatment.

Classical non-parametric statistics enables a deduction of basic modeling hypotheses, such as conditions that alter selection. In contrast, the stochastic dynamical model, derived here, enables a deeper insight into less evident properties of the evolutionary dynamics of the virus. These properties include the resistance level and the fitness costs conferred by individual mutations with respect to the applied drugs, the baseline (wild type) viral growth rate and the IC50 value. Furthermore, the stochastic viral growth model will enable to compare these properties between different genetic backgrounds and drug application patterns. Finally, the predictive power of the presented modeling method is enhanced by a model selection procedure which uses a **large scale estimation strategy** for selecting models and system parameters with the biggest explanatory power.

As explained in fig. 4.2 the velocity of the growth of a viral population can be taken as an indicator for its fitness and the degree of adaptation to the environment. As a consequence, the **growth rate of the viral population** is a central variable in this study. The second important variable is the **genetic composition** of the viral strain in question, mapping relevant genetic sequences and their muta-

tions on **quantitative phenotypic traits** such as the degree of resistance w.r.t. a particular drug and the level of viral viability and reproductive power influenced by mutational fitness costs. The relevant time scales of these two variables are in the order of **several days**, that a few viruses need to grow to a substantial population size and a time scale ranging from a **few days to several months** that new advantageous mutations need in order to constitute within the genome of the viral population.

The basis of the present modeling study is a one-dimensional Markov jump process describing the viral population growth subject to drug application. Also, as an outlook, we will discuss an extension of this model towards a higher-dimensional MJP, incorporating the dynamics of host-cells. A discrete stochastic population growth model is used here instead of a continuous deterministic approach since *in-vitro* viral passage experiments, underlying this model, revealed a substantial variability of population growth times [60]. Key to the present study is the notion of a **viral passage experiment** which is defined as the time that a virus population needs to grow from a predefined small size to certain maximal level. A detailed description of the corresponding *in-vitro* experiment is given in the next section. In summary, virus isolates diluted with an initial concentration V_0 are incubated with target cells and a certain amount of NRTI and/or NNRTI drugs is added. After the viral population has reached the maximal concentration $V_{\max} = 100 \cdot V_0$, the virus is extracted and subjected to consensus sequencing. This way the mutations are detected, which have occurred with respect to the baseline isolate and which have been selected during the *in-vitro* experiments. Subsequently, the extracted virus is diluted again in a medium with the concentration V_0, incubated with new target cells under drug addition and the passage experiment is repeated. For each given viral isolate and drug-treatment the described passage experiment is repeated 12 times.

The viral passage times are modeled here by the probability density of the **first passage times** of a Markov jump process describing viral growth. By assuming a sufficient structural similarity between the first passage time density induced by the model and the *in-vitro* measured viral passage times, we apply a moment-matching approach for fitting model parameters. We find that the first two moments represent the sought probability density sufficiently well and due to the low computational costs this method enables a fast and efficient large scale parameter estimation. The described moment-matching approach furthermore enables a computationally feasible parameter identifiability analysis based on a Monte Carlo sampling of the parameter space. Finally, the methodology is used to conduct a model-selection analysis, by scanning the space of all possible parameter permutations, in particular the set of relevant mutations of the reverse transcriptase enzyme, for finding model parameters with the largest impact on viral fitness and population dynamics.

4.1.3 Detailed description of experiments

Viral growth experiments conducted *in-vitro* can significantly contribute to dissecting the principles of genetic evolution and resistance development in the presence of drugs. The main advantages of this experimental methodology are the reduced system complexity due to the absence of the host immune system and a better controllability of experimental conditions. Previously, five clinical HIV isolates, cocultivated with extracts of peripheral blood mononuclear cells (PBMC), were derived from individuals who had been pre-treated with the NRTI drug Lamivudine (3TC), but have never been exposed to NNRTIs [59]. Using consensus sequencing of the primary samples up to the amino acid position 300 of the reverse transcriptase (RT) protein, a complete list of initial RT mutations was compiled, as shown in table 4.1. The amount of the viral population, reproducing by infection of the

	reverse transcriptase amino acid position													
	20	35	41	67	69	70	83	90	118	122	123	135	169	177
Hxb2	K	V	M	D	T	K	R	V	V	E	D	I	E	D
Iso 1		T	L					I		K	E			E
Iso 2	R			S	N	R			I		S	T		
Iso 3	R			S	N	R			I		S	T		
Iso 4		I	L	N			K			K			D	
Iso 5			L							K	E	T		

	reverse transcriptase amino acid position																
	184	196	201	202	208	210	211	214	215	219	272	275	277	291	293	294	297
Hxb2	M	G	K	I	H	L	R	P	T	K	S	K	R	E	I	P	E
Iso 1	V	E	V						Y			Q	R		V		Q
Iso 2	V			V				L	F	Q	P			D			T
Iso 3	V			V				L	F	Q	P			D			T
Iso 4	V				Y	W	K		Y						V		
Iso 5	V					W	K		Y		P					T	

Table 4.1: **Baseline amino acid substitutions in relation to reference (wild type) sequence (Hxb2) from the Stanford HIVDB.**
The numbers in the second row of the two tables indicate the position in the reverse transcriptase amino acid sequence. The letters correspond to amino acids detected at the respective position in the baseline isolates (rows 4-8), relative to the reference Hxb2-sequence (row 3). Empty table entries indicate equivalence to the reference Hxb2 strain. The latter is considered as the wild type sequence in this study.

cocultivated blood cells, was measured once a week using an antigen assay. At ELISA values $< 3 \times 10^4$pg/ml of the antigen, cultures were split: 2.5 Mio PBMCs were replaced by new donor PBMC. At an ELISA value $\geq 3 \times 10^4$pg/ml, indicating a predefined maximal size of the viral population, the cultures were passaged after a 2-hour incubation time. After each passage the genome of the viral population was sequenced again in order to detect newly appeared mutations with respect to the primary population. In each conducted experiment the described single viral passage was repeated 12 times.

The five viral isolates 1 to 5 were exposed to five different passage experiments A to F. Each of these experiments contained 12 single passages, as described above

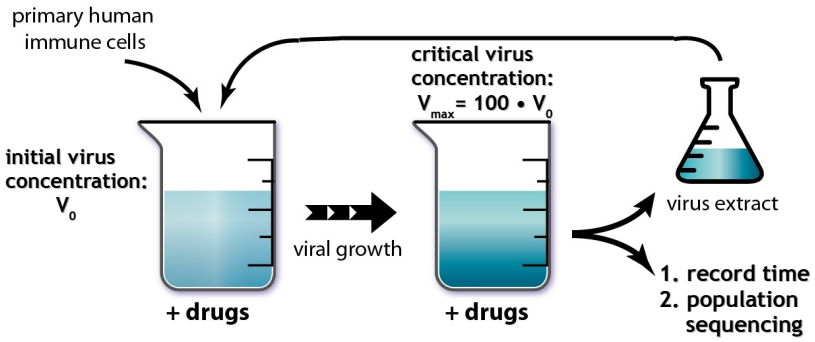


Figure 4.3: **Experimental set-up of a single passage experiment.**
Virus isolates diluted with an initial concentration V_0 are incubated with target cells and a certain amount of NRTI and/or NNRTI drugs is added. After the viral population has reached the maximal concentration $V_{\max} = 100 \cdot V_0$, the virus is extracted and subjected to consensus sequencing, enabling to detect mutations selected during the *in-vitro* experiments. Subsequently, the extracted virus is diluted again in a medium with the concentration V_0 and incubated with new target cells under drug addition.

and was determined by a different NRTI and/or NNRTI drug combination in the growth medium. The NRTI drugs consisted of Adefovir (ADV) and Lamivudine (3TC). Furthermore, the NNRTI drug Nevirapine (NVP) was used. The drug combinations in the respective experiments were as follows:

- A: No drugs were added to the medium.
- B: 1 μM 3TC and 2 μM ADV were added and maintained.
- C: NVP was added and concentrations were doubled for each passage (0.01 μM NVP during the first passage up to 20.48 μM during the last passage).
- D: 1 μM ADV and increasing concentrations of NVP were added to the medium.
- E: 2 μM 3TC and increasing concentrations of NVP were added to the medium.
- F: 1 μM 3TC and 2 μM ADV and increasing concentrations of NVP were added to the medium.

Isolates 2 and 3 were aliquots from the same baseline sample, but were run independently in experimental set-ups (C, D and E). For each experimental set-up (A-F), 12 single-passage experiments were run, in total $5 \cdot 3 \cdot 12 + 4 \cdot 3 \cdot 12 = 324$ single-passage experiments with a median duration of 21 days, respectively. In the experiments C, D, E and F the concentration of NVP was doubled with each passage. The starting dose of NVP was 0.01 μM, around the previously reported IC_{50} of the NNRTI-naïve isolates and the final concentration was 2048-fold, below previously reported cytotoxic levels [50]. On average, the cumulative time to the last passage (passage 12)

was 293 days within a range of 157-509 days. The described experimental set-up, for each of the five viral strains and the respective mutations detected during the single passages are depicted in figure 4.4.

4.2 Initial statistical analysis and conclusions for model building

4.2.1 Selection dynamics

The phenotypic background of the baseline strain is depicted in table 4.1, where relevant amino acid positions of the reverse-transcriptase enzyme are shown. None of the 55 baseline mutations are known to be in the NNRTI-binding pocket, in line with the fact that the isolates have never been treated with NNRTI drugs before. The baseline mutations of the strains 1 to 5, exhibit mutation patterns influencing the function of the reverse transcriptase enzyme and contain thymidine-analogue-associated mutations (M41L, D67N,K70R, L210W, T215F/Y, K219Q) and 3TC-related mutation M184V. Presence of 3TC prevented this mutation from reversal, while in the course of 86 % of experimental settings, where 3TC was absent, Methionine was reversed to the wild type amino acid Valine (see figure 4.4).

The viral growth experiments exhibit a characteristic pattern of mutations selected in the course of the passage experiments. Once a novel mutation with respect to the baseline strain was selected, it persisted until the last passage 12. The frequency of selected novel mutations varied between the different experiments. The number of selected mutations in experiments with NVP (set-ups C, D, E, and F) was significantly higher in passages 5-7 and 12 than in those experiments lacking NVP, as shown in figure 4.5 A. On average, 1-3 mutations occurred per passage experiment when NVP escalation was applied. Of the 43 novel mutations during NVP-escalation experiments, 38 were known to be in the NNRTI binding pocket.

4.2.2 Viral growth dynamics

In figure 4.6 the growth curves of viral populations during the individual passages are shown. The viral growth exhibits a significant variation between different experimental settings and passages. The interplay between drug concentration, mutation selection and viral growth dynamics becomes evident during NVP-escalation experiments (C-F). A doubling of the NVP concentration results in several cases in a reduced growth rate, compared to the previous passage. This is the case for instance for the strain 1 (green curve in fig. 4.6) in experiment C, passages 5 and 9. Accordingly, in several cases an accelerated population growth can be observed after a mutation occurs, as for strain 5 (yellow curve), experiment D, passages 5 and 10.

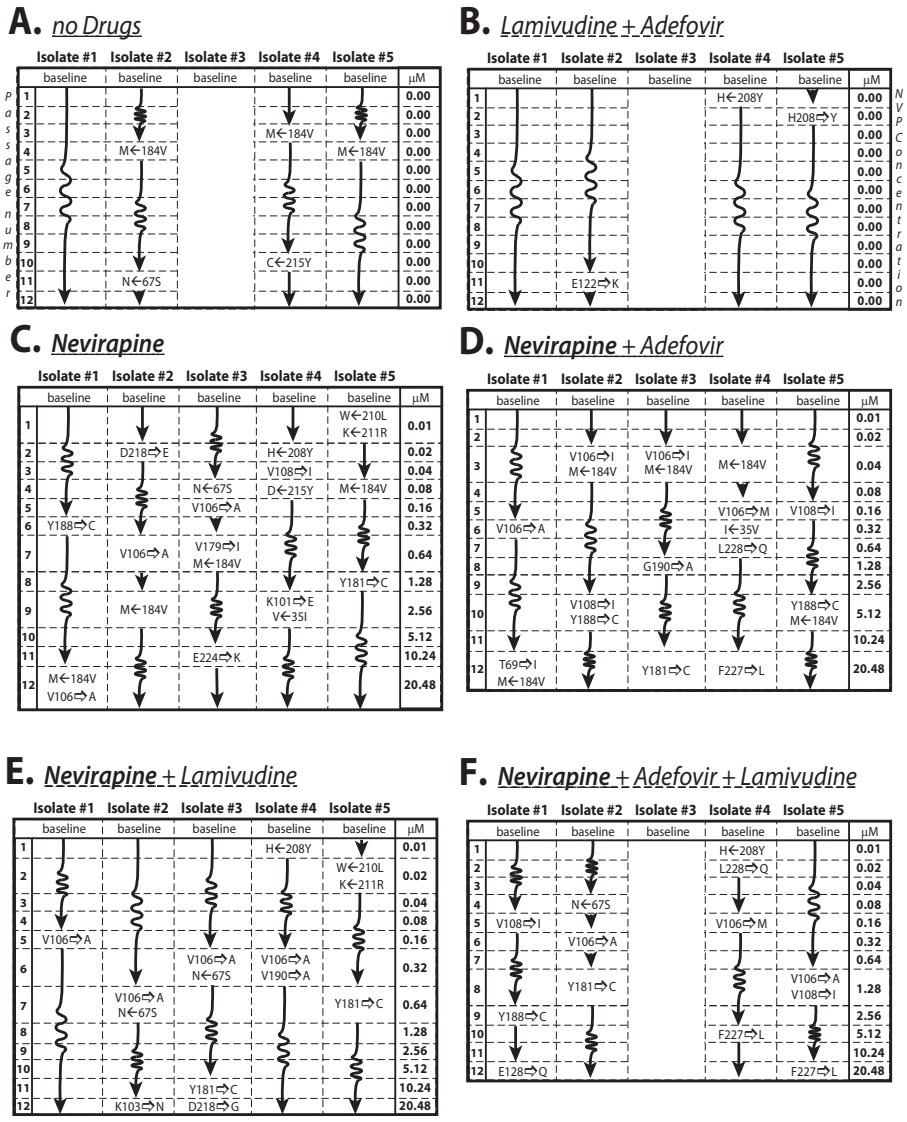


Figure 4.4: **Summary of passage experiments with sequencing data.**
The illustration provides a complete review of RT sequence changes under the following experimental set-ups: A: no drugs were added to the medium, B: 1 μM 3TC and 2 μM ADV were added and maintained, C: NVP was added and concentrations were doubled for each passage (0.01 μM NVP during the first passage up to 20.48 μM during the last passage), D: 2μM ADV and increasing concentrations of NVP were added, E: 1μM 3TC and increasing concentrations of NVP were added and F: 1μM 3TC and 2μM ADV and increasing concentrations of of NVP were added to the medium. Individual isolates 1 to 5 are indicated above the columns. Sequence changes listed are indicated in the rows that correspond to the passage number where they were first observed. NVP concentrations used in the respective passage experiment are listed on the right in units μM. Any mutation away from wild-type (Hxb2 strain) is indicated by a rightward-pointing arrow, whereas reversal to wild-type is indicated by a left-ward pointing arrow.

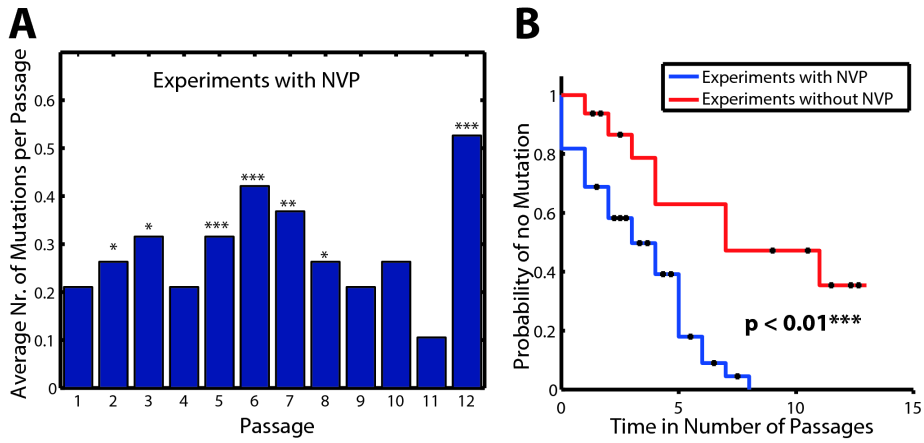


Figure 4.5: **Selection dynamics of mutations.**
A: Average number of mutations per passage in experiments with NVP (experimental set-ups C, D, E & F). Asterisks indicate whether there were significantly more mutations (wilcoxon rank-sum test) than in the NVP-free experiments (experimental set-ups A & B). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. **B**: Cumulative probability of detecting no mutation. The blue and red lines show the cumulative probability of not detecting a mutation after the indicated numbers of passages (x-axis) in experiments where NVP was added with increasing concentrations (blue line; experimental set-ups C, D, E & F) vs. experiments where no NVP was added (red line; experimental set-ups A & B).

Figure 4.7 shows the statistics of passage times over all passages of the viral growth experiments. It can be deduced that the median passage times of most experiments with NVP application are significantly higher than the medians in experiments without NVP application (experiments A, B). Possibly, the strains did not become sufficiently resistant to NVP during the passage experiments. Furthermore the median passage times of experiment B with application of ADV and 3TC, do not significantly differ from passage times of experiments with NVP-application.

In order to further elucidate the mode of action of ADV and 3TC, we compared in figure 4.8 the passage times of experiment A, where no drugs were added and experiment B, where 1μM 3TC and 2 μM ADV were added. Interestingly, for most isolates the addition of ADV and 3TC did not significantly slow the viral growth, suggesting that the concentrations of ADV and 3TC were sub-inhibitory. Only for isolate 5 the median passage times are significantly larger, when these drugs are added ($p = 0.01$). In contrast, the variance of the passage times is significantly higher in experiment B, where ADV and 3TC are added, than experiment A, without drug addition ($p < 0.05$ or $p < 0.01$). These results may indicate that addition of 1 μM 3TC and 2 μM ADV does not uniformly inhibit viral growth. In other words, viral growth is inhibited during some passage-experiments, whereas it is not inhibited in the majority of passage-experiments. The result of this mode of in-

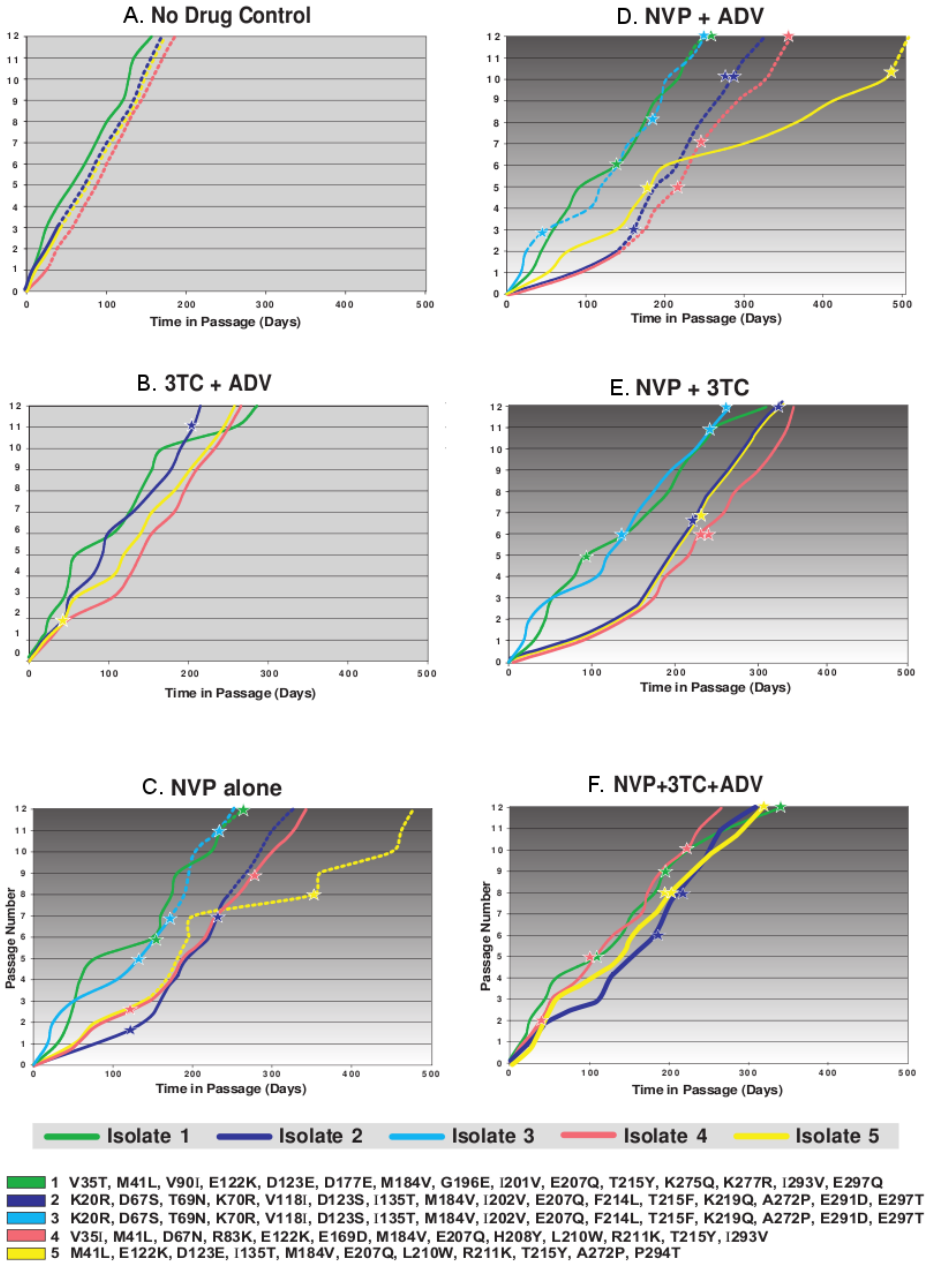


Figure 4.6: **Growth curves of the viral isolates during all passage experiments.** Each isolate is indicated by an individual color. Below the growth curves the baseline mutations for each isolate are indicated. These are the mutations detected before the start of the passage experiments.

hibition is an increase in the variance of passage times, while the central measure (median/mean passage times) may not be significantly affected.

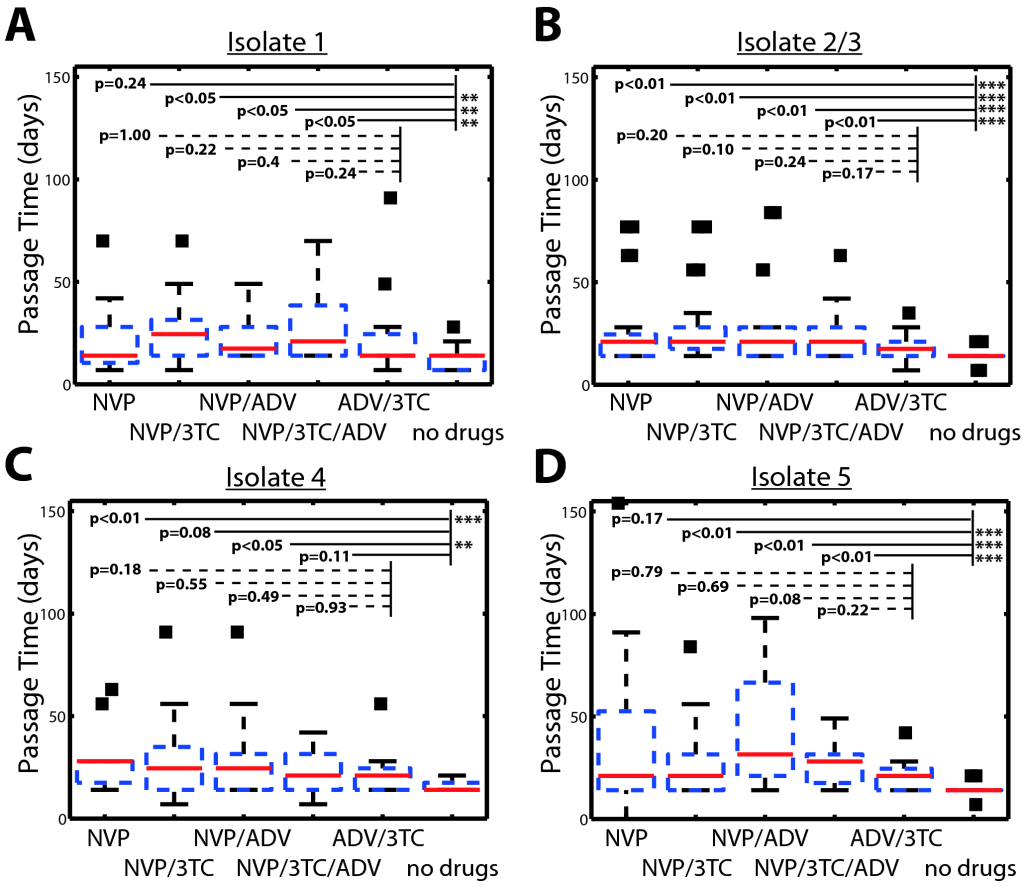


Figure 4.7: **Box plot of single passage times for all virus isolates during experimental set-ups A-F as indicated on the x-axis.**
Box plot of single passage times for virus isolate # 1, # 2/3, # 4 & # 5 during experimental set-ups A-F as indicated on the x-axis. The solid red horizontal lines indicate the respective median passage times, whereas the blue dashed boxes surrounding them indicate the range encompassed by the 25-th and 75-th percentiles. The whiskers denote the most extreme data points, which are not considered outliers and the black squares indicate outliers. A: Viral passage times for isolate # 1. B: Viral passage times for isolate # 2 & 3 (combined). C: Viral passage times for isolate # 4. D: Viral passage times for isolate # 5

From this observation we deduced a stochastic effect of action of NRTI-drugs, having a low probability of effect, due to low dosage but a strong effect if the drug succeeds in becoming integrated into the reverse-transcriptase enzyme. In order to model this effect, in the following, we will introduce a parameter describing the probability of drug integration p_{NRTI} and the intensity of the effect η_{NRTI}.

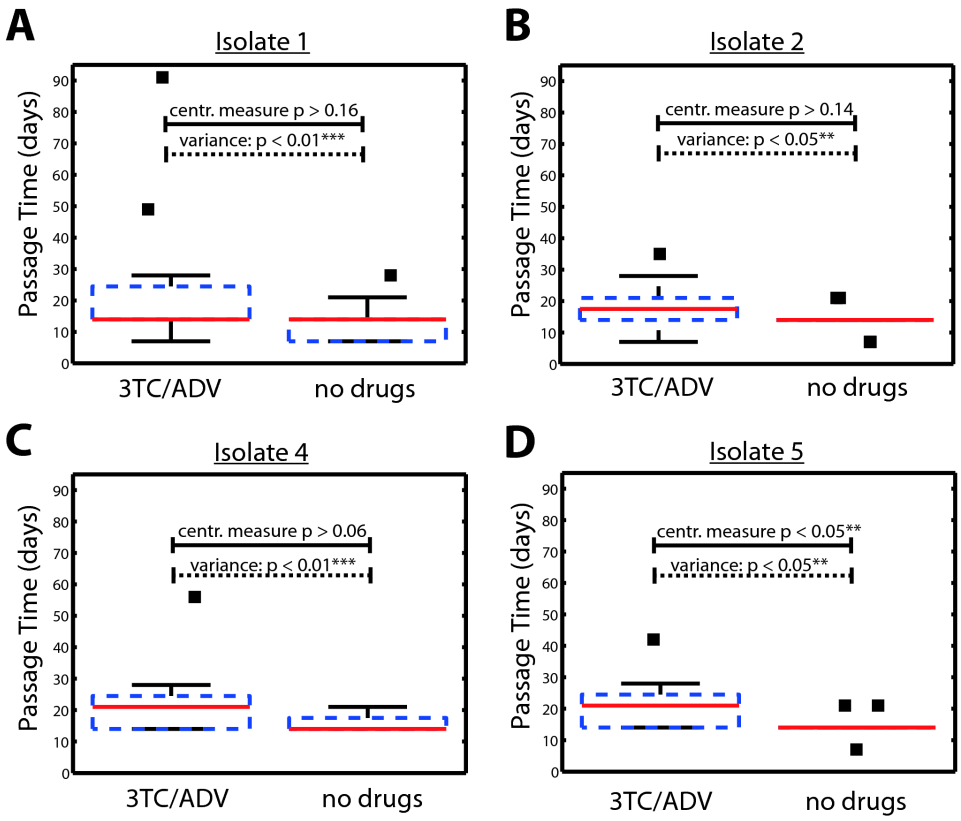


Figure 4.8: **Box plot of passage times for all virus isolates during experimental set-ups A & B (no drugs added vs. 1μM 3TC plus 2 μM ADV added) as indicated on the x-axis**
The solid red horizontal lines indicate the respective median passage times, whereas the blue dashed boxes surrounding them indicate the range encompassed by the 25th and 75th percentiles. The whiskers denote the most extreme data points, which are not considered outliers and the black squares indicate outliers. A: Viral passage times for isolate 1. B: Viral passage times for isolate 2. C: Viral passage times for isolate 4. D: Viral passage times for isolate 5.

4.3 Stochastic model of viral population growth

4.3.1 Viral growth subject to drug application

During infection, HIV enters a T-cell and reprograms it to produce viral particles. *In vitro*, the dynamics of the HIV reproduction process thus depends on the propensity of the viral integration process, the number of viral particles and the number of host cells. In *in-vitro* passage experiments, described above, the amount of host cells was kept constant. For modeling the viral reproduction in this context we thus subsumed the dynamics of host cells into the constant replication propensity of the viruses, yielding the following simple-birth reaction:

$$V \xrightarrow{r(j,p)} \mathrm{V} + \mathrm{V},$$

where $r(j,p)$ denotes the growth rate constant of the viral population in passage p under treatment, as introduced in eq. (4.2). Note that the index j refers to one of the experimental settings within the set $\{A, B, C, D, E, F\}$, described in fig. 4.4.

Drug effects and fitness

Each viral strain consists of baseline mutations and additional mutations that were derived or –lost during the course of an experiment. As previously described in eq. (4.2), the growth propensity $r(j,p)$ is affected by fitness $f(j,p)$ of viral strain i and NVP drug pressure $(1-\eta_{NVP}(j,p))$ as follows:

$$r(j,p) = r_{\varnothing} \cdot \underbrace{(1-\eta_{\mathrm{NVP}}(j,p))}_{\text{inhibition by NVP}} \cdot \underbrace{f(j,p)}_{\text{fitness}} \cdot \underbrace{[1-\eta_{\mathrm{NRTI}}(j,\rho_{\mathrm{NRTI}})]}_{\text{stoch. effect of low-dose NRTIs}} , \tag{4.3}$$

where $r_{\varnothing}$ denotes the growth rate of the baseline viral strain in the absence of any drugs. As an extension of eq. (4.2), the parameters η_{NRTI} and ρ_{NRTI} denote here the intensity of low-dose NRTI effect on viral growth and the probability of effect in experimental set-up j $\in [A..F]$, which was estimated to be $0 \leq \eta_{\mathrm{NRTI}} \leq 1$ with probability $0 \leq \rho_{\mathrm{NRTI}} \leq 1$ when NRTIs were added (experimental set-up: B, D, E, & F). In experiments without NRTI-addition (experimental set-ups A & C) the probability of effect was set to $\rho_{\mathrm{NRTI}} = 0$.

In equation (4.1) the relation between drug concentration, genetic composition and drug effect was described. In particular, it was argued that the level of resistance of genetic background k w. r. t. a drug is given by the IC_{50}-value of the wild type multiplied by the fold-resistance factor $\mathrm{FR}(k)$. In order to integrate the sequence of mutations which have been selected until the current passage into the drug effect equation, we assumed that the fold resistance conferred by each mutation contributed in an independent manner (no epistatic effects). This assumption results in a multiplicative model:

$$\mathrm{FR}(j,p) = \prod_{q \in Q(j,p)} \mathrm{FR}(q), \tag{4.4}$$

where the fold resistance $\mathrm{FR}(j,p)$ of the genetic background in passage p and experiment $j \in$ [A...F] is composed as a product of fold resistances of mutations in the set $Q(j,p)$, denoting the cumulative mutations which occurred relative to the baseline genome until passage p. See also fig. 4.4, where j corresponds to the subfigures A...F and p denotes the row number indicating the corresponding passage. The multiplicative model (4.4) of fold resistance modifies the drug-effect equation (4.1) to

$$\eta_{\mathrm{NVP}}(j,p) = \frac{[\mathrm{NVP(j)}]}{\mathrm{IC}_{50} \cdot \prod_{q \in Q(j,p)} \mathrm{FR}(q) + [\mathrm{NVP(j)}]} \tag{4.5}$$

where $\eta_{\mathrm{NVP}}(j,p)$ denotes the effect of NVP application on the viral strain. The genetic background of this strain is composed of a cumulative set of mutations $Q(j,p)$

which have been selected under treatment j until passage p.

Accordingly, the fitness of a viral strain in passage p under treatment j is given by

$$f(j,p) = \prod_q f(q), \tag{4.6}$$

where $f(q)$ denotes the relative fitness of the virus given a single mutation q w.r.t. to the wild type.

4.3.2 First passage time moment computation

The described HIV growth model implies that the viral growth dynamics in passage p under treatment j induces a simple-birth Chemical Master Equation model with a different propensity function $r(j,p)$:

$$\frac{\partial P_{jp}(k,t)}{\partial t} = (k-1)\cdot r(j,p)P_{jp}(k-1,t) - k\cdot r(j,p)P_{jp}(k,t) \tag{4.7}$$

where the probability of k viruses at time t in passage p and condition j is given by $P_{jp}(k,t) := \mathbb{P}(V = k, \text{time} = t)$, cf. section 2.2.2.

For deriving a model of passage experiments we note that in the *in-vitro* experiments described above, the virus was diluted 100-fold (100 μL supernatant was grown in 10 mL media) and the time was recorded until the initial p24 ELISA signal ($\geq 3\times 10^4$ pg/ml) was achieved. We therefore infer that the maximal number of virus particles V_1 at which the experiment was stopped is $V_1 = 100\cdot V_0$, where V_0 is the initial number of virus particles. Given the pure-birth process described by the Chemical Master Equation (4.7), the m-th moment T_m^{jp} of the PDF of the time needed to reach the state V_1 after starting in state V_0 in passage p is given by [25]:

$$T_m^{jp}\begin{cases} \frac{mT_{m-1}^{jp}}{r(j,p)}, & \text{if } V_0 = V_1 - 1, \\ \\ \frac{mT_{m-1}^{jp}}{r(j,p)} + T_m^{jp}, & \text{if } V_0 \in n-2, n-3, ..., 0. \end{cases} \tag{4.8}$$

Equation (4.8) yields the moments of the first passage times by recursion if one notes that the zero-th moment $T_0^{jp} = 1\ \forall\ i \in \{1,..,5\}, j \in \{1,..,12\}$, cf. section 2.4. For the first moment, the above recursions can be expressed in a closed form as follows

$$T_1^{jp} = \sum_{k=V_0}^{100\cdot V_0} \frac{1}{k\cdot r(j,p)}.$$

Using equation (4.2) the drug effects and the population fitness can be factored out to yield:

$$T_1^{jp} = \left[\frac{1}{(1-\eta_{\text{NVP}}(j,p))\cdot f(j,p)\cdot[1-\eta_{\text{NRTI}}(j,\rho_{\text{NRTI}})]}\right]\cdot \sum_{k=V_0}^{100\cdot V_0} \frac{1}{k\cdot r_\varnothing}.$$

Similarly, the second moment of the first passage time can be expressed in a closed form as follows:

$$T_2^{jp} = \left[\frac{2}{r(j,p)^2}\right] \cdot \sum_{k=V_0}^{100\cdot V_0} \frac{1}{k} \sum_{h=V_0}^{k} \frac{1}{h}. \tag{4.9}$$

Having computed T_1^{jp} and T_2^{jp} one obtains the mean and the standard deviation of the first passage time distribution of the viral strain in the passage j:

$$\begin{aligned} \mu(j,p) &= T_1^{jp}, \\ \sigma(j,p) &= \sqrt{T_2^{jp} - \left(T_1^{jp}\right)^2}. \end{aligned} \tag{4.10}$$

After the computation of $n = 12$ pairs of means and standard deviations $\mu(j,p), \sigma(j,p)$, $p = 1, .., 12$, the pooled mean $\tilde{\mu}(j)$ and pooled standard deviation $\tilde{\sigma}(j)$ of the strain i for all passages is obtained as follows:

$$\tilde{\mu}(j) = \frac{1}{n} \sum_{p=1}^{n} \mu(j,p), \tag{4.11}$$

$$\tilde{\sigma}(j) = \sqrt{\sum_{p=1}^{n} \left[\frac{\sigma(j,p)}{n} + \sum_{h=1}^{p-1} \frac{\sigma(j,p)^2}{n} + \frac{(\mu(j,p) - \mu(h,p))^2}{n^2}\right]}, \tag{4.12}$$

where in the second line the variance is corrected for the different means within the passages in order to obtain the variance/standard deviation for *all* passages.

4.3.3 Parameter estimation and model selection

First passage time moment fitting

The pure-birth Markov process described above gives rise to a parameter vector Θ consisting of the baseline growth propensity $r_\varnothing$, the half-maximal inhibitory drug concentration IC_{50}, fitness cost parameters $f(q)$ and fold resistances $\mathrm{FR}(q)$, $q \in Q$. Using the moments of the pure-birth process, computed above, the parameter vector Θ is estimated by minimizing the weighted least squares error $\varepsilon(\Theta)$:

$$\varepsilon(\Theta) = \min_{\Theta} \sum_{j} \left[\left(\frac{\tilde{\mu}(j) - \tilde{\mu}_{\mathrm{exp}}(j)}{\tilde{\mu}(j)}\right)^2 + \left(\frac{\tilde{\sigma}(j) - \tilde{\sigma}_{\mathrm{exp}}(j)}{\tilde{\sigma}(j)}\right)^2\right] \tag{4.13}$$

where $\tilde{\mu}(j)$ and $\tilde{\sigma}(j)$ denote the pooled mean and standard deviation, respectively computed using eqs. (4.11) and (4.12). Furthermore, $\tilde{\mu}_{\mathrm{exp}}(j)$ and $\tilde{\sigma}_{\mathrm{exp}}(j)$ denote the experimentally measured central moments, pooled over all passages, cf. fig. 4.7. Note that several unbounded parameters, e.g. $\mathrm{FR}(q)$ and IC_{50} could not be reliably estimated due to a flat residual error function resulting from eq. (4.5). In order to improve the estimation, we penalized unrealistically large parameter values in the objective function by using a constrained residual error

$$\tilde{\varepsilon}(\Theta) = \varepsilon(\Theta) + \mathrm{IC}_{50} + w \cdot \sum_{q \in Q} \log(\mathrm{FR}(q)),$$

where $w = 1/|Q|$ is a parameter proportional to the number of observed mutations, denoted by $|Q|$.

Model selection

In order to reduce the parameter search space, mutations $q \in Q$ which do not contribute to the drug resistance or fitness were *a priori* fixed at $f(q) = 1$ and $\mathrm{FR}(q) = 1$. The other parameters from the set of observed mutations Q were used to generate all possible non-empty subsets of Q, which gave rise to different sub-models. This procedure enabled to determine which observed mutations had the largest contribution to fitness costs and drug resistance. Thus, in accordance with eq. (4.3), the most basic model of viral growth without fitness and resistance effects is given by the following growth rate equation

$$r(j, p) = r_{\varnothing} \cdot (1 - \eta_{\mathrm{NRTI}}(j, \rho_{\mathrm{NRTI}})).$$

The basal growth rate $r_{\varnothing}$ is consistently included in all sub-models. In contrast, the parameter $\eta_{\mathrm{NRTI}}(j, \rho_{\mathrm{NRTI}})$ describing the stochastic NRTI-effect is only included in sub-models subject to experimental conditions with NRTI-application, i.e. B, D, E, F in fig. 4.4. For the model selection procedure we compiled for each isolate a set of all mutations selected during all experiments. First, we *a priori* assigned these mutations into two (overlapping) sets: fitness-cost-inducing mutations and resistance-conferring mutations. The first group was compiled on the basis of mutations deselected with respect to the wild type. For instance, M184V is a baseline mutation with respect to the wild type Hxb2-strain[1], while V184M denotes a reversion to wild type. This procedure gave rise to M resistance- and N fitness mutations. By successively excluding different mutations from the resistance- and the fitness model, we generated a set of sub-models based on respective parameter permutations. As a result $M + N$ fitness and resistance parameters gave rise to

$$\#\text{sub-models} = \sum_{k=0}^{M+N} \binom{M+N}{k} \tag{4.14}$$

different parameter subsets. As an illustrative example consider a set of following resistance mutations {V106A, V106I} and a set of fitness mutations {M184V, N67S}. According to eq. (4.14), this gives rise to 16 sub-models with the parameter subsets depicted in table (4.2). Each of the sub-models in this table gives rise to a different viral population growth rate of the Markov jump process, computed according to equation (4.3).

The described permutation procedure results in a different number of sub-models for each isolate, due to a different number of selected mutations. For instance, for isolate 4 there is a set of $M = 6$ resistance-conferring mutations towards NVP, selected in the course of the passage experiments: {V106M, V108I, L228Q, K101E,

[1] see table 4.1

Model Nr.	Parameters included
1	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}
2	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A)
3	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106I)
4	$r_\varnothing$, η_{NRTI}, p_{NRTI},IC50_{50}, FR(V106A), FR(V106I)
5	$r_\varnothing$, η_{NRTI}, p_{NRTI},IC50_{50}, FR(V106A), f(M184V)
6	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106I), f(M184V)
7	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A), FR(V106I), f(M184V)
8	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A) , f(N67S)
9	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106I) , f(N67S)
10	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A), FR(V106I) , f(N67S)
11	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A) , f(M184V), f(N67S)
12	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106I) , f(M184V), f(N67S)
13	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, FR(V106A)· FR(V106I) , f(M184V), f(N67S)
14	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, f(M184V)
15	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, f(N67S)
16	$r_\varnothing$, η_{NRTI}, p_{NRTI}, IC50_{50}, f((M184V), f(N67S)

Table 4.2: **Exemplary submodels resulting from two resistance inducing mutations {V106A, V106I} and two fitness-cost mutations {M184V, N67S}.**

F227L, Y181C_G190A}, where the last one corresponds to two simultaneous mutations. Furthermore $N = 4$ following potential fitness-reducing mutations were detected {M184V, H208Y, Y215C_D, I35V}. According to eq. (4.14) this gives rise to 1024 possible sub-models, which were fitted to the experimental data, using first passage time moment-matching. In order to rank the models according to their explanatory power, the Akaike information criterion (AIC) was computed

$$\text{AIC} = \log(\varepsilon) + 2 \cdot L,$$

where ε is the residual error of a model, cf. eq. (4.13). The variable L is the number of parameters of the model where $L = M + N + R$ i.e. the sum over the number of resistance-conferring mutations, fitness-loss conferring mutations and the number of basal parameters R included in each model, such as the basal growth rate. Fig. 4.9 shows the AIC-values for all sub-models of each isolate HIV strain. Each point of the curve was computed as a mean of 50 replicate numerical estimations started with random initial values. The corresponding parameter estimates for isolate strain 4 are depicted in fig. 4.10. Each sub-figure shows a box plot containing the median (red horizontal line) and a span of 25th and 75th percentiles (blue vertical line) of the k-best models, computed according to the AIC-based relative likelihood criterion [1]. Parameters, such as the fitness-cost inducing mutation I35V in isolate strain 4, did not appear in these high-ranked models. Thus they were considered as "not identifiable" due to a lack of a sufficient amount of experimental data.

In order to verify the estimation results for all isolates, the statistics of the corresponding passage times based on the models with the best parameter estimates were compared to the experimentally measured passage times. In figure 4.11 the central moments of the model predictions are plotted versus the central moments of experimental measurements. The results confirm a sufficient explanatory power

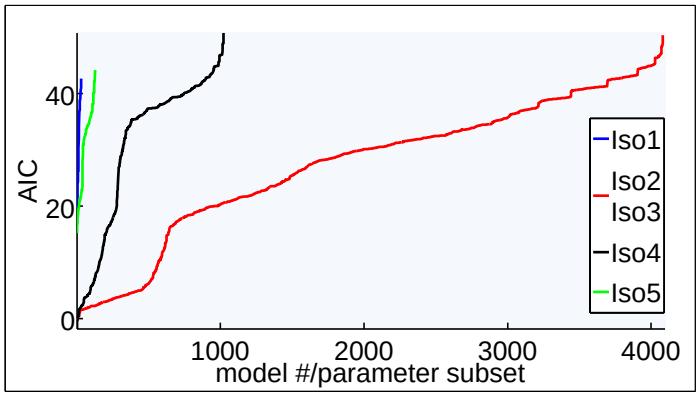


Figure 4.9: **AIC-scores resulting from parameter inference for all sub-models and viral strain isolates, as indicated.**
Different sub-models giving rise to the HIV growth rate were generated by a permutation over all parameters resulting from mutations observed during the passage experiments. The mean Akaike information criterion (AIC) of 50 estimation runs was computed for each sub-model, started with random initial parameter values. The randomly restarted replication of estimates gives an insight into the identifiability of each parameter. The first three sub-figures in the first row of fig. 4.10 show estimation results for fitness cost inducing mutation parameters. They have to be interpreted as follows: the mutations M184V and Y215C/D, H208Y and I35V have small standard deviations.

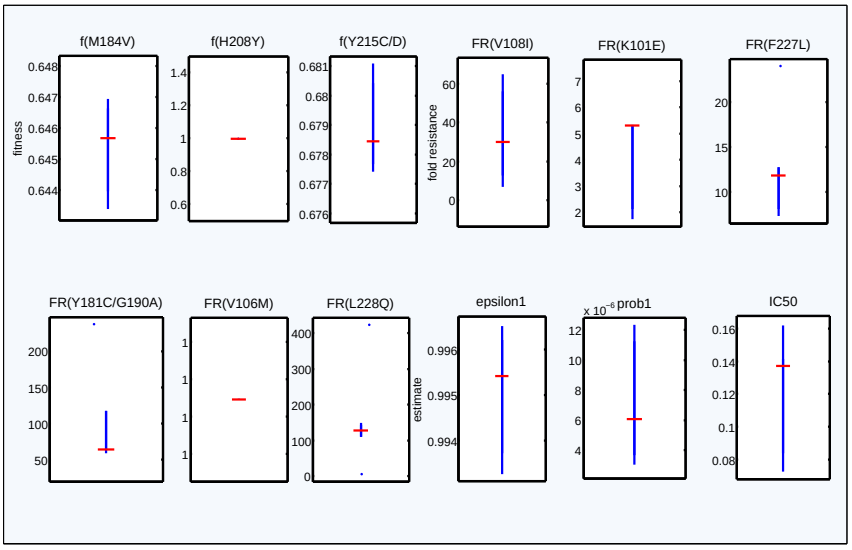
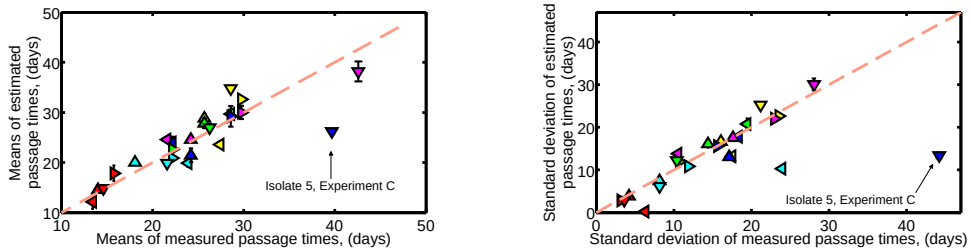


Figure 4.10: **Large scale estimation results for viral isolate 4.**
Using the AIC-score curve of isolate 4 (fig. 4.9) a ranking of the k-best sub-models (out of 1024) was compiled. Shown are error bars of the distributions of estimated values of various parameters appearing in these sub-models. The red vertical lines indicate median values and the blue vertical lines span between the 25th and 75th percentiles with outliers indicated as blue points.

of the data by the stochastic viral growth model. As an exception, the passage time distribution for isolate strain 5 subject to experimental setting C (NVP only) could not be fitted by the model (see fig. 4.11). A comparison to fig. 4.4 C yields a possible explanation for the poor fitting capability in this setting. Only one mutation (Y181C) is selected (three mutations deselected) in this strain in passage 8 although the viral growth dynamics indicates extreme variability cf. fig. 4.6 C (yellow line). This mismatch may also be a consequence of neglecting of the target-cell PBMC dynamics by the model, which was approximated here as constant, due to the described experimental setting. In the *Discussion* section we give an outlook to a possible improvement of the viral growth model in order to account for the dynamics of host cells.



(a) Means: estimates vs. means of data. (b) Standard deviation: of estimates vs. data.

Figure 4.11: **Means and standard deviations of measured vs. predicted passage times.**
Visual predictive checks of predicted (y-axis) versus observed (x-axis) data points. A: Means of first passage times $\mu_i(\Theta_i)$ and B: their standard deviations $\sigma_i(\Theta_i)$. The distinct markers indicate the different patient isolates: leftward-, upward-, rightward- and downward-pointing triangles and diamonds indicate data/predictions from/for isolates # 1, # 2, # 3, # 4 and # 5. Colors indicate the different experimental set-ups, e.g. red, cyan, blue, yellow, magenta and green denote experimental set-ups A-F respectively. Vertical bars indicate the range of predictions spanned by the 5-th and 95-th percentile of all model evaluations.

4.3.4 Biological implications of the modeling results

Efficient inhibition of viral growth and an improved medical treatment requires a dissection of the principles of viral evolution. In the previous sections we described the derivation of a stochastic model of viral growth under selection pressure of NRTI and NNRTI drugs. A central aspect for the model was a coupling of the viral growth rate to mutational dynamics. Experimental *in-vitro* data from viral passage experiments revealed specific mutational patterns of HIV under different drug combinations and doses. Using the stochastic viral growth model we were able to estimate numerical values of the fitness loss and the fold resistance induced by many of the observed mutations. Furthermore the model selection procedure, described above, enabled to determine mutational parameters which were identifiable

given the *in-vitro* viral growth data. In the following we discuss the most important mutations whose impact could be estimated. Many of the parameter values are in line with previously published results and biologically highly relevant. In some cases our estimation results suggest a high impact for mutations which have not been described before. This section is based on results published in [60].

Drug susceptibility and fitness of baseline isolates

The first passage time moment inference method enabled the estimation of key model parameters, cf. results published in [60]. The first two data columns in table 4.3 show basal growth rates $r_\varnothing$ and the IC_{50} values for the respective viral isolate strains at baseline, prior to resistance development. The growth rates of all baseline isolate strains were within the range of $0.33 - 0.42$ day^{-1}, where the growth rates of isolate 1 and isolate 4 indicated the largest and the smallest fitness, respectively. The median IC_{50} of the baseline isolates was estimated in a range from 0.07 to 0.39 μM NVP, which is consistent with published IC_{50}-values of the drug-susceptible virus (wild type: 0.1 μM) [50].

The parameter estimates of NRTI effect, i.e. intensity η_{NRTI} and probability of effect ρ_{NRTI}, are shown in table 4.3 (last two columns). In all strains except isolate 1 the estimated intensity of effect was pronounced $\eta_{NRTI} = 0.99$ while the probability of NRTI inhibition at the applied low doses (drug concentration in the range $1-2$ μM) was close to zero. These results are in line with the the drug effect model that at low doses NRTI inhibitors do not bind in the majority of instances, but if they do, their effect is strong due to the chain-terminating inhibition mechanism [79]. In contrast, for isolate 1 the estimates for the efficacy and the probability of effect were $\eta_{\text{NRTI}} = 0.65$ and $\rho_{\text{NRTI}} = 0.43$, respectively.

	$\mathbf{r}_\varnothing[(1/day)]$	$\mathbf{IC_{50}}$ $[\mu\text{M}]$	$\boldsymbol{\eta}_{\mathbf{NRTI}}$	$\boldsymbol{\rho}_{\mathbf{NRTI}}$
ISO 1	0.42 (0.42, 0.42)	0.39 (0.37, 0.48)	0.65 (0.63, 0.67)	0.43 (0.28, 0.44)
ISO 2/3	0.39 (0.39, 0.39)	0.07 (0.07, 0.1)	0.99 (0.97, 0.99)	$5.4e{-}7$ ($4.7e{-}7, 1.8e{-}4$)
ISO 4	0.33 (0.33, 0.33)	0.13 (0.07, 0.14)	0.99 (0.99, 0.99)	$1.1e{-}5$ ($3.8e{-}6, 1.2e{-}5$)
ISO 5	0.36 (0.36, 0.36)	0.39 (0.37, 0.49)	0.99 (0.99, 0.99)	$1e{-}6$ ($9.9e{-}7, 1.1e{-}6$)

Table 4.3: **Estimates of baseline parameters of viral isolate strains.**
Estimates for the viral growth rate $r_\varnothing$ in absence of drugs, lower-bound estimates for the susceptibility of baseline isolates IC_{50}, intensity η_{NRTI} and probability of NRTI effect ρ_{NRTI}. Indicated numbers are median estimates from best models according to AIC and their respective 5th and 95th percentiles.

NVP drug resistance

The fold resistance to NVP, conferred by the selected mutations, is shown in table 4.4. By computing the AIC-scores of all possible models, the mutations with the most explanatory power could be identified. Based on this selection procedure, the

parameters FR(69I), FR(101E), FR(103N), FR(108I/188C), FR(122K), FR(128Q) FR(179I), FR(208Y), FR(218G/E), FR(227L) could be excluded as "not identifiable", given the *in-vitro* data.

The estimates for the identifiable resistance parameters significantly varied between the four different baseline isolates, possibly indicating an influence of pre-existing NRTI mutations on subsequent mutations affecting NVP-susceptibility [6].
All isolates developed novel mutations at codon 106 in the presence of NVP. While mutation V106→A was estimated to induce a strong fold resistance, the mutations V106→I and V106→M conferred only an intermediate to weak resistance in isolate strains, in which they were selected (see table 4.4 and [60]).

The NVP resistance mutation V108→I arose at least once in all strains. It led to moderate resistance in isolates 4 and 5 whereas a moderate to strong fold resistance was estimated in isolate 1. Another mutation which was selected in all strains is Y181→C. Its effect, however, could only be estimated in isolates 1, 2/3 and 5 resulting in a 5- to 13-fold resistance. In isolate 4, the mutation Y181→C appeared simultaneously with G190→A, where a strong resistance to NVP (FR≥ 67) was estimated.

As shown in table 4.4 and described in [60], the estimated fold resistance of mutation L228→Q was pronounced. In all instances the selection of this mutation occurred before the selection of the mutation F227→L, with an estimated moderate effect, indicating a co-evolutionary association between the two amino acid substitutions, as described previously for NRTI-resistance mutations [62].

	ISO 1	**ISO 2/3**	**ISO 4**	**ISO 5**
FR(K101E)	n.s.	n.s.	5 (2, 5)	n.s.
FR(V106A)	80 (52, 135)	176 (22, 195)	n.s	21 (9, 47)
FR(V106I)	n.s	5 (3, 9)	n.s	n.s
FR(V106M)	n.s	n.s	1 (1,4)	n.s
FR(V108I)	25 (7,26)	1 (1, 1)	30 (7, 65)	7 (3, 7)
FR(V179I)	n.s.	1 (1, 3)	n.s.	n.s.
FR(Y181C)	5 (4, 6)	7 (6, 41)	n.i	13 (10, 13)
FR(G190A)	n.s	8 (7, 11)	n.i	n.s.
FR(Y181C/G190A)	n.s	n.s	67 (59, 300)	n.s
FR(Y188C)	23 (4, 43)	n.s	n.s	7 (2, 11)
FR(E218E)	n.s	1 (1, 1)	n.s	n.s
FR(E224K)	n.s	1 (1, 4)	n.s	n.s
FR(F227L)	n.s	n.s	12 (7, 29)	n.s
FR(L228Q)	n.s	n.s	128 (8, 423)	n.s

Table 4.4: **Estimated fold resistance against NVP exerted by single amino acid substitutions in the distinct genetic background of the baseline isolates.**
Values indicated are medians of all parameter estimates and the 5th and 95th percentile of the estimates are indicated in brackets. "n.s" means "not selected" and "n.i." means parameter "not identifiable".

Effects of baseline mutations on viral fitness

The number of distinct mutations undergoing reversal was inversely correlated with our estimates of the population growth rate of viral strains $r_\varnothing$. Thus the "fittest isolates" with the largest viral growth rate $r_\varnothing$ exhibited the fewest number of distinct mutations reversing back to wild type i.e. 1, 2, 4 and 2 distinct deselected mutations observed in isolates 1, 2/3, 4 and 5, respectively, cf. fig. 4.4 (mutations with left-ward pointing arrows) and table 4.3. Of all back mutations observed during the passage experiments only the deselection of M←184V was estimated to significantly improve viral fitness in all isolates, (table 4.5).

	ISO 1	**ISO 2/3**	**ISO 4**	**ISO 5**
f(184V)	0.79 (0.79, 0.79)	0.59 (0.59, 0.62)	0.65 (0.64, 0.65)	0.65 (0.64, 0.66)
f(215Y)	n. ds.	n. ds.	0.68 (0.68, 0.68)	n. ds.

Table 4.5: **Estimated relative fitness loss elicited on the genetic background of the baseline isolates.**
Indicated values are medians of all parameter estimates and the 5th and 95th percentile of the estimates are indicated in brackets. "n.ds" means "not deselected".

Parameter estimates depicted in table 4.5 indicate that the pre-existing mutation 184V conferred a large fitness loss due to the removal of the selective pressure by the NRTI-drug 3TC. The estimated fitness for the individual isolates ranged from 59% to 79% of the wild type fitness, which is consistent with previous *in vivo* estimates [47] and mechanistic modeling results of HIV-1 DNA polymerization process [79]. Although all baseline isolates carried resistance mutations at position 215, the reversion to wild type (C←215Y and D←215Y) was only observed in isolate 4. The estimated relative fitness in this isolate was 68%, attributable to these two pre-existing mutations (table 4.5). According to the corresponding AIC-scores, the fitness loss mutations f(67S), f(208H), f(35I) and f(210W/211K) were not included in the most informative models, and thus could not be estimated from the data.

4.4 Discussion and outlook

4.4.1 Computing first passage time moments for the 2D-model of viral growth

The model of viral growth subject to drug application enabled an estimation of major determinants of selection dynamics. These included the impact of selective disadvantage and resistance level conferred by various mutations of the reverse transcriptase enzyme. However, an aspect that could only partially be explained by the 1-D simple-birth model was the variability of passage times. We attributed this to the exclusion of target cell dynamics whose incorporation into the viral growth model leads to a two-dimensional birth-and-death system [78]. There are two major drawbacks associated with the more detailed 2D-model. Firstly, computation of

the first passage time density from the corresponding CME requires a significantly larger computational run time. In contrast to the 1D case, analytical closed-form expressions are not available. The two remaining exact strategies are based either on a direct numerical integration of the CME or an indirect sampling using Gillespie's stochastic simulation method, both of which are prohibitively slow. In particular, in the context of parameter estimation, optimization of objective functions requires a repeated computation of the first passage time density. The second drawback associated with a 2D-model is a larger number of parameters which need to be estimated, aggravating parameter identifiability. Due to these reasons we used a 1D model of viral growth, as described in the previous sections. To this end we exploited the property of the *in-vitro* experiments that the number of target cells was kept almost constant throughout the measurements.

In order to give an outlook and propose a feasible strategy for a rigorous model of host-virus interaction, in the following we derive a 2D viral growth model based on an explicit interaction of virus and target cells. We propose an alternative inexact strategy for passage time computation which approximates the Chemical Master Equation by a Fokker Planck equation with a linear diffusion term. We show that the continuous stochastic model approximates the discrete one arbitrarily well. Importantly, this method leads to a reduction of the computational run time at least by two orders of magnitude which enables its applicability for parameter estimation.

Viral growth and interaction with target cells

Infection of target cells by virus particles and the integration of viral DNA into the host genome is the basis of the viral reproduction. The most basic model of host-pathogen interaction can be described in terms of a two-dimensional birth-and-death system. The dynamics of target cells T is determined by their infection through virus particles V and their death. Viruses are in turn assembled within the target cells and are able to infect new target cells upon their release. The virus population is diminished due to inefficient host cell integration and the corresponding failure of target cell infection. Note that organism-related clearance of viral particles is not relevant in the *in-vitro* setting. The described system gives rise to the following reactions:

1. *Failed infection of target cells*:

$$\mathrm{T} + \mathrm{V}^* \xrightarrow{\mathrm{CL}} \mathrm{T}$$

2. *Successful infection*:

$$\mathrm{T} + \mathrm{V}^* \xrightarrow{\beta} \mathrm{T}^*$$

3. *Death of infected target cells*:

$$\mathrm{T}^* \xrightarrow{d} \varnothing$$

4. *Release of newly assembled viruses*:

$$\mathrm{T}^* \xrightarrow{N} \mathrm{T}^* + \mathrm{V}^*$$

Obviously, this reaction system is of second order since the first two reactions depend on target cells and virus particles. In the *in-vitro* experiments of this study (see section 4.1.3) the total number of target cells was kept nearly constant, since between and after each passage a new pool of target cells was added.

By assuming the number of uninfected target cells T as constant the second order infection model can be reduced to a first order reaction system in two dimensions. To this end, the first reaction which describes the viral death, can be reformulated as

$$\mathrm{V}^* \xrightarrow{k_1} \varnothing,$$

where k_1 is a lumped viral clearance rate adopted from [78] and given by

$$k_1 = \mathrm{CL}_\mathrm{L} \cdot \mathrm{T}.$$

Note that T is assumed to be a constant number of target cells and CL_L is computed using the probability of a successful reverse transcription $\rho_{\mathrm{rev},\phi}$ and a constant infection rate β [78]:

$$\mathrm{CL}_\mathrm{L} = \left(\frac{1}{\rho_{\mathrm{rev},\phi}} - 1\right) \cdot \beta.$$

The order of the second reaction can as well be reduced by a first-order approximation:

$$\mathrm{V}^* \xrightarrow{k_2} \mathrm{T}^*$$

where k_2 is a lumped reaction rate with $k_2 = \beta \cdot \mathrm{T}$. Similarly to the 1D model, the application of drugs can be included here by modifying the reaction rate parameters. As described in [78], the effect of drugs on the viral growth modifies the infection rate and the viral clearance. By denoting the infection rate and clearance of viral strain i under experimental setting j by $\beta(i,j)$ and $\mathrm{CL}_\mathrm{L}(i,j)$, respectively, the action of drugs can be expressed by

$$\begin{aligned} \beta(i,j) &= (1-\eta(i,j)) \cdot f(i) \cdot \beta_{\mathrm{ref},\varnothing}(i), \\ \mathrm{CL}_\mathrm{L}(i,j) &= \left(\frac{1}{\rho_{\mathrm{rev},\phi}} - (1-\eta(i,j)) \cdot f(i)\right) \cdot \beta_{\mathrm{ref},\varnothing}(i), \end{aligned} \tag{4.15}$$

where $\beta_{\mathrm{ref},\varnothing}(i)$ denotes here the infection rate of strain i in the absence of drug application. As a result, the host-pathogen dynamics can be reformulated by the following 1st-order reaction system

$$\begin{aligned} \mathrm{V}^* &\xrightarrow{k_1} \varnothing \\ \mathrm{V}^* &\xrightarrow{k_2} \mathrm{T}^* \\ \mathrm{T}^* &\xrightarrow{d} \varnothing \\ \mathrm{T}^* &\xrightarrow{N} \mathrm{T}^* + \mathrm{V}^* \end{aligned} \tag{4.16}$$

In order to study the stochastic dynamics induced by this system, the time evolution of the probability distribution of the two-dimensional system variable $X = [X_1, X_2]$ can be studied, by denoting it $P(\mathrm{T}^*, \mathrm{V}^*, t) := \mathbb{P}(X_1 = \mathrm{T}^*, X_2 = \mathrm{V}^*, \mathrm{time} = t)$. The corresponding Master equation then reads

$$\begin{aligned}\frac{\partial P(\mathrm{T}^*, \mathrm{V}^*, t)}{\partial t} &= k_1(\mathrm{V}^* + 1)\ P(\mathrm{T}^*, \mathrm{V}^* + 1, t) \\ &+ k_2(\mathrm{V}^* + 1)\ P(\mathrm{T}^* - 1, \mathrm{V}^* + 1, t) \\ &+ d(\mathrm{T}^* + 1)\ P(\mathrm{T}^* + 1, \mathrm{V}^*, t) \\ &+ N\mathrm{T}^*\ P(\mathrm{T}^*, \mathrm{V}^* - 1, t) \\ &- ([k_1 + k_2]V^* + d\mathrm{T}^* + N\mathrm{T}^*)\ P(\mathrm{T}^*, \mathrm{V}^*, t). \end{aligned} \tag{4.17}$$

As described in a previous section, we aim at fitting the moments of the first passage times $\mathrm{FPT}(\mathrm{n}_0 \to n_1)$ resulting from the Master equation (4.17). Previously, we have used this derivation for the one-dimensional model of viral growth in order to obtain closed expressions for first passage time moments (e.g. see equation (4.8)). For a two-dimensional process the computation of closed-form expressions for the moments is less straight-forward since it requires the solution of the Master equation (4.17). Alternatively, one can sample the first passage time density using the stochastic simulation algorithm [24] or the accelerated τ-leaping version [61]. The drawback of these methods is that estimation of parameters becomes prohibitively slow, since residual error optimization requires a repeated simulation of a sufficiently large number of sample paths in order to approximate the first passage time density.

4.4.2 First-passage time density computation via a Fokker-Planck-approximation

In the following we propose an alternative strategy for computing the first passage time moments. It is based on a diffusion approximation of the Master equation originating from van Kampen's Linear Noise Approximation [77]. This method was also used in the context of parameter estimation, e.g. see [30]. The main idea is a derivation of a Fokker-Planck equation which dissects the discrete stochastic dynamics induced by the Master equation into a deterministic and a stochastic part. This enables an efficient computation of the first passage time density by numerical simulation, which is significantly faster than direct or indirect CME-solution and better suited for a large scale parameter estimation.

As discussed in section 2.3.4 the propensities and the jumps of the Master equation can be expanded under the assumption that the latter are sufficiently small. This enables to express the probability distribution $\Pi(\xi, t)$ of fluctuations around the macroscopic trajectory by the following Fokker-Planck-Equation:

$$\frac{\partial \Pi}{\partial t} = -\sum_{i,j} \Gamma_{ij} \partial_i (\xi_j \Pi) + \frac{1}{2} \sum_{i,j} D_{ij} \partial_{ij} \Pi. \tag{4.18}$$

Let us denote by $\mathbf{w}(\mathbf{X})$ the vector of propensities of the Master equation:

$$\mathbf{w}(\mathbf{X}) = [k_1 \cdot X_2,\ k_2 \cdot X_2,\ d \cdot X_1,\ N \cdot X_1]^T,$$

and let $\mathbf{S}$ be the stoichiometric matrix corresponding to the reaction system (4.16)

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

Then the deterministic function $\mathbf{\Gamma}$ and the diffusion matrix $\mathbf{D}$ are respectively computed by

$$\mathbf{\Gamma} = \mathbf{S} \cdot \mathbf{w}(\mathbf{X}),$$

and

$$\mathbf{D} = \mathbf{S} \cdot \mathrm{diag}[\mathbf{w}(\mathbf{X})] \cdot \mathbf{S}^T,$$

where $\mathrm{diag}[\mathbf{w}(x)]$ is a matrix with the propensities of the Master equation on its diagonal and zero everywhere else.

The new approximating state-continuous stochastic variable $\mathbf{x}$ with fluctuations governed by the equation (4.18) is described by the following Stochastic Differential Equation

$$d\mathbf{x}(t) = f(\mathbf{x})dt + \mathbf{D}^{\frac{1}{2}} d\mathbf{B}_t, \tag{4.19}$$

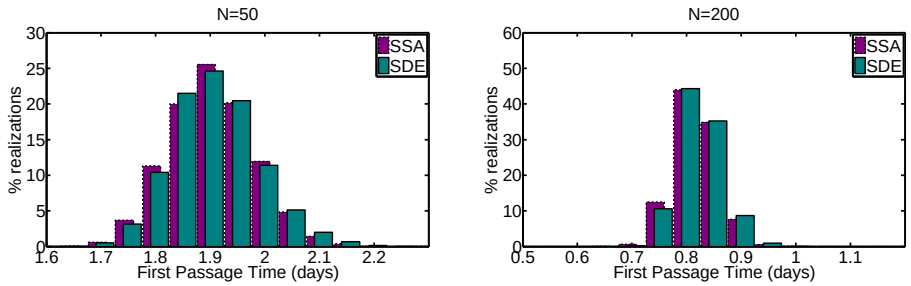
where $\mathbf{D}^{\frac{1}{2}}$ denotes a matrix root with $\mathbf{D}^{\frac{1}{2}} \left[\mathbf{D}^{\frac{1}{2}}\right]^T = \mathbf{D}$ and $d\mathbf{B}_t$ is a two-dimensional Brownian motion. In order to compute the first-passage time of a system trajectory, this SDE can be solved for instance by numerical discretization using the Euler-Maruyama method or the higher order Milstein Scheme [42]. The simulation is stopped when the number of viruses reaches the maximal level $V_{\max}$.

In figure 4.12 we compared the first passage time distributions induced by the Master equation (4.17) and the corresponding Fokker-Planck-Approximation. To this end we simulated 10^4 SSA-trajectories of the CME and SDE-trajectories from the Fokker-Planck-Equation for two different parameter settings with low and high infection rate k_2 (left column vs. right column, respectively.). The simulations were started at $X_0 = \{T_0 = 10, V_0 = 100\}$ and stopped when the number of viruses reached $V_{\max} = 10^4$ for the first time. The corresponding simulation times were recorded and used to sample the first passage time distribution. As it is shown in figure 4.12, SDE-sampling approximates the first passage time density very well for the 2D-viral growth model. We tested the goodness of the Fokker-Planck approximation for different parameters. For instance for two different viral burst rates $N = 50$ and $N = 200$ the goodness of approximation remains excellent (fig. 4.12). It is left for further investigation how good the Fokker-Planck-approximation performs for other CME-models, in particular with nonlinear propensities. Further results of this method in the context of CME-approximation can be found in [30].

Since parameter estimation is our main objective for the computation of the first passage time density, we measured the computational run times of the described simulation experiments. In table (4.6) the run times of 10^4 SSA-samples based on the original Markov jump process are compared to 10^4 SDE-samples of the corresponding Fokker-Planck approximation.

	SSA-sampling (MATLAB)	**SDE-sampling (MATLAB)**	**SDE-sampling (C++)**
$N = 50$	9.12 hours	219 s	0.71 s
$N = 200$	8.72 hours	94 s	0.34 s

Table 4.6: **Comparison of run times for the computation of the first passage time density using SSA- and SDE-sampling.**
Shown are measurements of run times for 10^4 sampled trajectories using the SSA-algorithm by Gillespie using MATLAB 7 and SDE-discretization of the Fokker-Planck-equation using MATLAB 7 and C++. The experiments were conducted for two different viral burst rates N on a Intel dual-core processor with 2 GHz and 6 MB cache on each core.



(a) **SSA**: Mean: 1.905, Std 0.079. **SDE**: mean: 1.909, std: 0.080.

(b) **SSA**: Mean: 0.819, Std 0.040. **SDE**: mean: 0.823, std: 0.040.

Figure 4.12: **Comparison of passage time statistics computed by simulation-based solution of the CME and its SDE-approximation.**
First passage times were computed for the 2D viral growth model (4.16) via SSA-simulation based on the Chemical Master Equation (4.17) and by SDE-simulation resulting from a Fokker-Planck-Approximation of the CME. Each histogram was generated using 10^4 trajectories. Model parameters were chosen as follows: $k_1 = 0.01$, $k_2 = 0.1$, $d = 0.01$. The viral burst parameter was set to **(a)** $N = 50$ and **(b)** $N = 200$.

4.5 Summary and conclusion

In this chapter we have posed a biologically motivated problem of deriving a growth model of HIV under application of NRTI and NNRTI drugs and estimating its parameters given *in-vitro* passage data. Our goal was a quantification of various factors influencing viral evolution dynamics such as the resistance gain and fitness loss, conferred by mutations of the reverse transcriptase enzyme. Initial statistical analysis of the data revealed a significant amount of stochasticity in the viral

growth dynamics and in the action of drugs. We have developed and implemented a method based on fitting the first two moments of the first passage time density of stochastic viral growth without explicit target cell interaction. The estimated parameters were biologically plausible and yielded valuable insights into the principles of viral evolution in the presence of drugs. This work was published in [60].

From a mechanistic point of view a two-dimensional model of the interaction between the viruses and the target cells would yield a more rigorous description of the stochastic viral growth dynamics. However, model inference using a two-dimensional Master equation would increase the computational run time and aggravate parameter identifiability due to an introduction of additional parameters. Using a Fokker-Planck approximation of the CME, we have shown that the run time problem can be circumvented by sampling from the approximating space-continuous density using stochastic differential equations (SDE) with a linear diffusion term. Due to a lack of a larger set of viral growth data the second problem of parameter identifiability remained unsolved. Consequently, we left the model inference of a two-dimensional viral growth dynamics as a subject for future work.

Chapter 5

Concluding remarks

The theory of Markov processes delivers an extensive framework for an analysis of the dynamics induced by complex biological systems. In the present thesis this framework was adopted for studying mechanisms involved in adaptation to environmental conditions in microorganisms. In chapter 3 qualitative properties, resulting from the interaction network of stress-induced bacterial signaling, were translated into a stochastic model. This enabled to study the signal transduction properties of the network, based on the dynamics of the key signaling molecule c-di-GMP. The solution of the Chemical Master Equation of c-di-GMP regulation demonstrated how product inhibition of DGC enzymes contributes to a reduction of stochastic fluctuations and a containment of signaling noise. Furthermore, the stochastic interaction model of c-di-GMP and YciR revealed that noise-induced bistability is potentially involved in phenotypic heterogeneity of biofilm synthesis.

A central concept of this thesis was the notion of the first passage times. In chapter 3 the mean first passage times of a regulatory c-di-GMP module were analyzed and thus the dependence of response times of c-di-GMP signaling on the parameters of the system were deduced. This enabled to understand how the expression level of the enzymes producing and degrading c-di-GMP influences the velocity of signal transduction. The results suggested that *E.Coli* cells in the stationary growth phase can be considered to be in an alarm mode since the high expression level of regulatory enzymes in this phase enables a significantly faster regulation of the c-di-GMP level and the resulting signal transduction than in other growth modes.

Throughout this work parameter estimation of Markov jump processes played an important role. As shown in chapter 2, reaction rate estimation of biochemical kinetic systems may suffer from prohibitively large state spaces. This limits the applicability of estimation methods of infinitesimal generators. Although the problem of estimating large but structured and sparse generator matrices can be reduced to estimation of a few reaction rate constants, the curse of dimensionality and prohibitive computational run times still aggravate inference for processes observed discretely in time. In chapter 4 a novel method for estimating the resistance and fitness effects of mutations from *in-vitro* passage experiments was developed. Closed analytical expressions for the moments of the first passage time distribution of the HIV growth model were used to circumvent estimation problems described above. This enabled to conduct a large-scale parameter inference and model selection using discretely observed viral growth data.

The Linear Noise Approximation of the Chemical Master Equation is a central analytical framework which complements the analysis of discrete-state Markov processes since it allows to obtain continuous-state stochastic approximations. If the jump propensities of the Master equation are linear, then the first two moments are equivalent to the moments of the approximating continuous-state system. In the nonlinear case, this equivalence is lost but the macroscopic equations are still a good approximation of the system near equilibrium states of the original process. This feature was exploited in chapter 3 in order to analyze parameter regions where the system induced bistable dynamics. Furthermore, in the outlook section of chapter 4, a Fokker-Planck approximation of the Chemical Master Equation for a two-dimensional viral growth model was suggested. A comparison of the solution of the Fokker-Planck equation via a fast numerical integration of the corresponding SDE and the SSA-sampling of the Chemical Master Equation showed a high agreement. This indicated that the approximation method is suitable for large scale parameter estimation in possible future studies.

The results of this thesis suggest that phenotypic and genotypic heterogeneity is a key mechanism behind an adaptation to perturbations of external conditions. An optimal regulation of adaptation processes is based on noise-induced dynamics with multiple equilibria. Thus, the biofilm synthesis system of *E.Coli* exhibits two phenotypic equilibria and the regulation of the corresponding stationary probabilities determines the number of biofilm expressing cells, possibly enabling a saving of resources in the isogenic curli-off cells. In HIV the genetic potential landscape is characterized by an elaborate interplay between fitness loss and resistance gain of mutations changing the amino acid sequence and eventually the 3D structure of enzymes involved in the viral life cycle. Mutational noise, induced by the erroneous process of reverse transcription, enables the viruses to efficiently search for the minima of the potential landscape where they become resistant to drug application. Here we introduced a framework for a simultaneous analysis of the impact of a large set of experimentally observed genetic mutations on viral growth dynamics under drug application. The results yield novel insights into the evolutionary dynamics of HIV and introduce a new methodology for further studies of its resistance acquisition strategies.

Summary

A key feature and a central driving force behind biological evolution is the capability of adaption to changing environmental conditions. Noise-induced transitions play a central role in these decision making processes allowing for a natural stochastic sampling between various evolutionary strategies. Mathematical analysis of such mechanisms requires experimental data, which represent the multimodal stochastic probabilities assigned to these strategies, being sampled at a sufficiently high resolution. However, in most applications the available experimental measurements are temporally and spatially too sparse for this objective. In this thesis different mathematical methods are derived for dissecting the mechanisms underlying such decision making processes despite the sparsity of data. The key idea is based on a compensation of lacking direct experimental observations using indirect inference from other, coupled system variables measured with a higher accuracy. One of the multistable systems studied here is the mutational dynamics conferring drug resistance to HIV. Since the likelihood of constitutive mutations is strongly associated with their phenotypic impact, time-discrete measurements of intrinsically stochastic viral population growth are used for inferring the principles underlying the mutational dynamical system. Furthermore, a similar idea is applied for analysing the phenotypic bistability of a stress-induced signaling network in *E. Coli*, giving rise to biofilm synthesis. Although direct single-cell measurements of *E. Coli* within the two modes of the probability distribution are not yet available, qualitative measurements of gene and protein interactions of the underlying signaling system are used for analysing dynamical properties of the bistable biofilm regulation. As a unifying framework, the theory of biochemical reactions based on Markov jump processes is adapted to the described problems and the resulting practical implications are discussed.

Zusammenfassung

Ein entscheidener Aspekt und eine zentrale Antriebskraft hinter der biologischen Evolution ist die Fähigkeit der Adaptation an sich verändernde äussere Bedingungen. Zufallsbedingte Zustandsübergänge spielen eine Schlüsselrolle in diesen Entscheidungsprozessen und ermöglichen eine natürliche stochastische Suche unter verschiedenen evolutionären Strategien. Die mathematische Analyse von solchen Mechanismen benötigt experimentelle Daten, die multimodale stochastische Wahrscheinlichkeitsverteilungen repräsentieren, gemessen mit einer hinreichend hohen Auflösung. In den meisten Anwendungen ist jedoch die zeitliche und räumliche Auflösung von experimentellen Messungen hierfür zu gering. In dieser Arbeit werden unterschiedliche mathematische Methoden hergeleitet, um Mechanismen hinter solchen Entscheidungsprozessen zu analysieren, trotz der unzureichenden Menge an Daten. Die zentrale Idee basiert auf einer Kompensation von fehlenden direkten experimentellen Messungen durch eine indirekte Schätzung, mit Hilfe von anderen gekoppelten Systemvariablen mit höherer Messhäufigkeit. Eins der multistabilen Systeme, die hier betrachtet werden, ist die Mutationsdynamik von HIV, welche Wirkstoffresistenzen verursachen kann. Da die Wahrscheinlichkeit von sich festsetzenden Mutationsereignissen eng an deren phenotypische Auswirkungen gekoppelt ist, werden zeit-diskrete Messungen von intrinsisch stochastischem viralen Populationswachstum verwendet, um Rückschlüsse auf Prinzipien der Dynamik von Mutationsereignissen zu ziehen. Weiterhin wird eine ähnliche Idee angewandt, um die phenotypische Bistabilität der durch Stress aktivierten Signalkaskade zu analysieren, die in *E. Coli* zur Biofilmbildung führt. Trotz des Fehlens von Einzelzellmessungen von *E.Coli* innherhalb der beiden Modi der Wahrscheinlichkeitverteilung, werden qualitative Messungen von Gen- und Proteininteraktionen der zugrundeliegenden Signalkaskade verwendet, um die Eigenschaften der bistabilen Biofilmregulation zu untersuchen. Als ein vereinigendes methodisches Gerüst, wird die auf Markov Sprungprozessen basierende Theorie von biochemischen Reaktionssystemen auf die beschriebenen Problemstellungen angewandt und resultierende praktische Aspekte werden diskutiert.

Eidestattliche Erklärung

Ich versichere hiermit, dass ich die von mir eingereichte Dissertation selbständig verfasst habe. Alle Hilfsmittel, wie Publikationen oder Bücher, wurden im Literaturverzeichnis angegeben und Zitate aus fremden Arbeiten sind als solche gekennzeichnet. Diese Arbeit wurde in gleicher oder ähnlicher Form bisher in keinem anderen Promotionsverfahren eingereicht und auch nicht veröffentlicht.

Berlin, 24. Januar 2014

Appendix A

Fano factor in the hypersensitive limit

In order to show that the Fano factor of a c-di-GMP module results in eq. (3.13), two main steps are applied. In the first step, the original equation for the Fano factor

$$\frac{\sigma^2}{\mu} \approx \frac{\mathbf{\Sigma_s}}{\bar{x}_s} = \frac{1}{2\bar{x}_s} \frac{\dfrac{V_{\max 1}}{1+\bar{x}_s/\bar{K}_i} + \dfrac{V_{\max 2}\bar{x}_s}{\bar{x}_s+\bar{K}_\mathrm{m}}}{\dfrac{V_{\max 1}/\bar{K}_i}{(1+\bar{x}_s/\bar{K}_i)^2} + \dfrac{V_{\max 2}\bar{K}_\mathrm{m}}{(\bar{x}_s+\bar{K}_\mathrm{m})^2}}, \tag{A.1}$$

is simplified using the follwing assumptions:

1. Vanishing product inhibition ($\bar{K}_i >> \bar{x}_s$),
2. an action of PDE enzymes approaching saturation ($\bar{K}_\mathrm{m} << \bar{x}_s$,) and
3. vanishing difference of maximal catalytic velocities ($V_{\max 1} \approx V_{\max 2}$).

Inserting the three assumptions into equation (A.1) results in

$$\frac{\mathbf{\Sigma_s}}{\bar{x}_s} = \frac{1}{2\bar{x}_s} \frac{V_{\max 1} + V_{\max 2}}{V_{\max 1}/\bar{K}_i + \dfrac{V_{\max 2}\bar{K}_\mathrm{m}}{\bar{x}_s^2}}. \tag{A.2}$$

In the second step the above assumption 3 ($V_{\max 1} - V_{\max 2} \approx 0$) is applied to the mean level equation

$$\bar{x}_s = -\frac{\bar{K}_i(V_{\max 2} - V_{\max 1})}{2V_{\max 2}} + \left[\left(\frac{\bar{K}_i(V_{\max 2} - V_{\max 1})}{2V_{\max 2}}\right)^2 + \frac{V_{\max 1}\bar{K}_i\bar{K}_\mathrm{m}}{V_{\max 2}}\right]^{1/2},$$

which results in

$$\bar{x}_s \approx \sqrt{\bar{K}_i\bar{K}_\mathrm{m}}. \tag{A.3}$$

A substitution of eq. (A.3) into eq. (A.2) yields the final result

$$\frac{\mathbf{\Sigma_s}}{\bar{x}_s} = \frac{\sqrt{\bar{K}_i}}{2\sqrt{\bar{K}_\mathrm{m}}}. \tag{A.4}$$

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec. 1974. (Cited on page 89.)

[2] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, June 2007. (Cited on pages 48 and 62.)

[3] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, 149(4):1633–1648, Aug. 1998. (Cited on page 62.)

[4] C. Beloin and J.-M. M. Ghigo. Finding gene-expression patterns in bacterial biofilms. *Trends in Microbiology*, 13(1):16–19, Jan. 2005. (Cited on page 45.)

[5] C. Beloin, J. Valle, P. Latour-Lambert, P. Faure, M. Kzreminski, D. Balestrino, J. A. J. Haagensen, S. Molin, G. Prensier, B. Arbeille, and J.-M. Ghigo. Global impact of mature biofilm lifestyle on *Escherichia coli* K-12 gene expression. *Molecular Microbiology*, 51(3):659–674, Feb. 2004. (Cited on page 45.)

[6] D. E. Bennett, R. J. Camacho, D. Otelea, D. R. Kuritzkes, H. Fleury, M. Kiuchi, W. Heneine, R. Kantor, M. R. Jordan, J. M. Schapiro, A.-M. M. Vandamme, P. Sandstrom, C. A. Boucher, D. van de Vijver, S.-Y. Y. Rhee, T. F. Liu, D. Pillay, and R. W. Shafer. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PloS one*, 4(3), 2009. (Cited on page 93.)

[7] R. Berg. The indigenous gastrointestinal microflora. *Trends in Microbiology*, 4(11):430–435, Nov. 1996. (Cited on page 45.)

[8] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, 67, 2005. (Cited on page 43.)

[9] P. Bremaud. *Markov Chains*. Springer, corrected edition, Mar. 2008. (Cited on pages 19 and 20.)

[10] G. E. Briggs and J. B. Haldane. A note on the kinetics of enzyme action. *The Biochemical journal*, 19(2):338–339, 1925. (Cited on page 51.)

[11] N. L. Brown, J. V. Stoyanov, S. P. Kidd, and J. L. Hobman. The MerR family of transcriptional regulators. *FEMS Microbiology Reviews*, 27(2-3):145–163, June 2003. (Cited on page 47.)

[12] C. Chan, R. Paul, D. Samoray, N. C. Amiot, B. Giese, U. Jenal, and T. Schirmer. Structural basis of activity and allosteric control of diguanylate cyclase. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17084–17089, Dec. 2004. (Cited on page 50.)

[13] A. L. Chang, J. R. Tuckerman, G. Gonzalez, R. Mayer, H. Weinhouse, G. Volman, D. Amikam, M. Benziman, and M. A. Gilles-Gonzalez. Phosphodiesterase A1, a regulator of cellulose synthesis in Acetobacter xylinum, is a heme-based sensor. *Biochemistry*, 40(12):3420–3426, Mar. 2001. (Cited on page 47.)

[14] E. Cheynubrata. Modellierung von chemischen Prozessen mit Markov-Sprungprozessen und Schätzung von Reaktionskonstanten. *FU Berlin, Diplomarbeit*, 2011. (Cited on page 43.)

[15] B. Christen, M. Christen, R. Paul, F. Schmid, M. Folcher, P. Jenoe, M. Meuwly, and U. Jenal. Allosteric Control of Cyclic di-GMP Signaling. *Journal of Biological Chemistry*, 281(42):32015–32024, Oct. 2006. (Cited on pages 48 and 54.)

[16] M. Christen, B. Christen, M. Folcher, A. Schauerte, and U. Jenal. Identification and characterization of a cyclic di-GMP-specific phosphodiesterase and its allosteric control by GTP. *The Journal of biological chemistry*, 280(35):30829–30837, Sept. 2005. (Cited on page 47.)

[17] A. Collet, P. Cosette, C. Beloin, J.-M. M. Ghigo, C. Rihouey, P. Lerouge, G.-A. A. Junter, and T. Jouenne. Impact of RpoS deletion on the proteome of *Escherichia coli* grown planktonically and as biofilm. *Journal of proteome research*, 7(11):4659–4669, Nov. 2008. (Cited on page 45.)

[18] S. H. Dandach and M. Khammash. Analysis of stochastic strategies in bacterial competence: a Master equation approach. *PLoS computational biology*, 6(11), Nov. 2010. (Cited on page 6.)

[19] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Springer, 2002 edition, July 2002. (Cited on page 28.)

[20] J. Elf and M. Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the Linear Noise Approximation. *Genome Research*, 13(11):2475–2484, Nov. 2003. (Cited on pages 56 and 58.)

[21] S. Engblom. Computing the moments of high dimensional solutions of the Master equation. *Applied Mathematics and Computation*, 180(2):498–515, Sept. 2006. (Cited on pages 26 and 27.)

[22] L. Ferm, P. Lötstedt, and A. Hellander. A hierarchy of approximations of the Master equation scaled by a size parameter. In *J. Ssi. Comput.*, 2008. (Cited on page 35.)

[23] C. Gardiner. *Handbook of stochastic methods*. Springer Verlag, 2004. (Cited on page 14.)

[24] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81:2340–2381, 1977. (Cited on pages 21, 24, 41, 52 and 97.)

[25] D. Gillespie. *Markov processes: an introduction for physical scientists.* Academic press, Inc., 1992. (Cited on pages 20, 24, 35, 36, 37 and 86.)

[26] D. T. Gillespie. A rigorous derivation of the Chemical Master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, Sept. 1992. (Cited on pages 22, 23 and 24.)

[27] D. T. Gillespie. The Chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000. (Cited on page 30.)

[28] D. T. Gillespie. Deterministic limit of stochastic chemical kinetics. *The journal of Physical Chemistry. B*, 113(6):1640–1644, Feb. 2009. (Cited on page 30.)

[29] A. Goldbeter and D. E. Koshland. An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 78(11):6840–6844, Nov. 1981. (Cited on page 56.)

[30] A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005. (Cited on pages 97 and 98.)

[31] N. Grantcharova, V. Peters, C. Monteiro, K. Zakikhany, and U. Römling. Bistable expression of CsgD in biofilm development of *Salmonella enterica* serovar typhimurium. *Journal of Bacteriology*, 192(2):456–466, Jan. 2010. (Cited on pages 45, 49, 62 and 63.)

[32] C. Guldberg and P. Waage. Über die chemische Affinität. *J. Prakt. Chem.*, 19:69, 1879. (Cited on pages 23 and 27.)

[33] L. Hall-Stoodley, J. W. Costerton, and P. Stoodley. Bacterial biofilms: from the natural environment to infectious diseases. *Nature Reviews Microbiology*, 2(2):95–108, Feb. 2004. (Cited on page 45.)

[34] J. Helmann. Regulation by Alternative sigma Factors. In G. Storz and R. Hengge, editors, *Bacterial Stress Responses*. ASM Press, Washington DC, 2 edition, 2011. (Cited on page 46.)

[35] R. Hengge. Principles of c-di-GMP signalling in bacteria. *Nat. Rev. Microbiology*, 7(4):263–273, Apr. 2009. (Cited on pages 7, 47, 48, 50, 51 and 65.)

[36] R. Hengge. Proteolysis of σ^S (RpoS) and the general stress response in *Escherichia coli*. *Research in Microbiology*, 160(9):667–676, Nov. 2009. (Cited on page 46.)

[37] R. Hengge. The general stress response in gram-negative bacteria. In G. Storz and R. Hengge, editors, *Bacterial Stress Responses*. ASM Press, Washington DC, 2 edition, 2011. (Cited on page 46.)

[38] A. Hobolth and E. Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Ann. Appl. Stat.*, 3(3), 2009. (Cited on page 43.)

[39] K. A. Johnson and R. S. Goody. The original Michaelis constant: translation of the 1913 Michaelis–Menten paper. *Biochemistry*, 50(39):8264–8269, Sept. 2011. (Cited on page 51.)

[40] E. Klauck and R. Hengge. σ^{S}-controlling netwoks in *Escherichia coli*. In A. Filloux, editor, *Bacterial Regulatory Networks*. Caister Academic Press, 2011. (Cited on page 46.)

[41] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice: concepts, implementation and application*. Wiley-Blackwell, 1. edition, May 2005. (Cited on page 51.)

[42] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations (Stochastic modelling and applied probability)*. Springer, corrected edition, June 2011. (Cited on pages 32, 34 and 98.)

[43] T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344+, June 1971. (Cited on page 29.)

[44] H. Kuwahara and O. S. Soyer. Bistability in feedback circuits as a byproduct of evolution of evolvability. *Molecular Systems Biology*, 8(1), Jan. 2012. (Cited on page 62.)

[45] P. Landini. Cross-talk mechanisms in biofilm formation and responses to environmental and physiological stress in *Escherichia coli*. *Research in Microbiology*, 160(4):259–266, May 2009. (Cited on page 46.)

[46] S. Lindenberg, G. Klauck, C. Pesavento, E. Klauck, and R. Hengge. The EAL domain protein YciR acts as a trigger enzyme in a c-di-GMP signalling cascade in E. coli biofilm control. *The EMBO Journal*, advance online publication, May 2013. (Cited on pages 47, 48, 49, 63 and 68.)

[47] J. Martinez-Picado, K. Morales-Lopetegi, T. Wrin, J. G. Prado, S. D. Frost, C. J. Petropoulos, B. Clotet, and L. Ruiz. Selection of drug-resistant HIV-1 mutants in response to repeated structured treatment interruptions. *AIDS (London, England)*, 16(6):895–899, Apr. 2002. (Cited on page 94.)

[48] S. Menz. Hybrid stochastic–deterministic approaches for simulation and analysis of biochemical reaction networks. *Doktorarbeit, Freie Universität Berlin*, July 2012. (Cited on pages 28 and 29.)

[49] S. Menz, J. C. Latorre, C. Schütte, and W. Huisinga. Hybrid stochastic-deterministic solution of the Chemical Master equation. *Multiscale Model. Simul.*, 10(4):1232–1262, 2012. (Cited on page 30.)

[50] V. Merluzzi et al. Inhibition of HIV-1 replication by a nonnucleoside reverse transcriptase inhibitor. *Science*, 250:1411–1413, 1990. (Cited on pages 78 and 92.)

[51] P. Metzner, I. Horenko, and C. Schütte. Generator estimation of Markov jump processes based on incomplete observations non-equidistant in time. *Phys. Rev*, 76, 2007. (Cited on page 43.)

[52] J. Monod and F. Jacob. General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:389–401, Jan. 1961. (Cited on page 62.)

[53] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the Chemical Master equation. *The Journal of Chemical Physics*, 124(4):044104+, Jan. 2006. (Cited on page 19.)

[54] B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 5th edition, Nov. 2002. (Cited on page 32.)

[55] J. A. Papin, T. Hunter, B. O. Palsson, and S. Subramaniam. Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.*, 6(2):99–111, Feb. 2005. (Cited on page 46.)

[56] R. Paul, S. Weiser, N. C. Amiot, C. Chan, T. Schirmer, B. Giese, and U. Jenal. Cell cycle-dependent dynamic localization of a bacterial response regulator with a novel di-guanylate cyclase output domain. *Genes & Development*, 18(6):715–727, Mar. 2004. (Cited on page 47.)

[57] P. Pesavento, G. Becker, N. Sommerfeldt, A. Possling, N. Tschowri, A. Mehlis, and R. Hengge. Inverse regulatory coordination of motility and curli-mediated adhesion in *Escherichia coli*. *Genes & Development*, 22:2434–2446, 2008. (Cited on page 45.)

[58] C. Prigent-Combaret, E. Brombacher, O. Vidal, A. Ambert, P. Lejeune, P. Landini, and C. Dorel. Complex regulatory network controls initial adhesion and biofilm formation in *Escherichia coli* via regulation of the csgD gene. *Journal of bacteriology*, 183(24):7213–7223, Dec. 2001. (Cited on page 46.)

[59] B. A. Rath, R. A. Olshen, J. Halpern, and T. C. Merigan. Persistence versus Reversion of 3TC Resistance in HIV-1 Determine the Rate of Emergence of NVP Resistance. *Viruses*, 4(8):1212–1234, Aug. 2012. (Cited on page 77.)

[60] B. A. Rath, K. Pouran Yousef, D. K. Katzenstein, R. W. Shafer, C. Schütte, M. von Kleist, and T. C. Merigan. *In Vitro* HIV-1 Evolution in Response to Triple Reverse Transcriptase Inhibitors & *In Silico* Phenotypic Analysis. *PLoS ONE*, 8(4):e61102+, Apr. 2013. (Cited on pages 8, 76, 92, 93 and 100.)

[61] M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119(24):12784–12794, 2003. (Cited on page 97.)

[62] S.-Y. Rhee, T. F. Liu, S. P. Holmes, and R. W. Shafer. HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation. *PLoS Comput Biol*, 3(5):e87+, May 2007. (Cited on page 93.)

[63] S.-Y. Y. Rhee, M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31(1):298–303, Jan. 2003. (Cited on page 72.)

[64] E. P. C. Rocha. The organization of the bacterial genome. *Annual Review of Genetics*, 42(1):211–233, 2008. (Cited on page 49.)

[65] U. Römling, W. D. Sierralta, K. Eriksson, and S. Normark. Multicellular and aggregative behaviour of Salmonella typhimurium strains is controlled by mutations in the agfD promoter. *Molecular microbiology*, 28(2):249–264, Apr. 1998. (Cited on page 46.)

[66] D. A. Ryjenkov, M. Tarutina, O. V. Moskvin, and M. Gomelsky. Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF protein domain. *Journal of Bacteriology*, 187(5):1792–1798, Mar. 2005. (Cited on page 50.)

[67] D. O. Serra, A. M. Richter, G. Klauck, F. Mika, and R. Hengge. Microanatomy at Cellular Resolution and Spatial Order of Physiological Differentiation in a Bacterial Biofilm. *mBio*, 4(2), May 2013. (Cited on pages 45, 49 and 62.)

[68] U. K. Sharma and D. Chatterji. Transcriptional switching in *Escherichia coli* during stress and starvation by modulation of sigma activity. *FEMS microbiology reviews*, 34(5):646–657, Sept. 2010. (Cited on page 46.)

[69] W. K. Smits, C. C. Eschevins, K. A. Susanna, S. Bron, O. P. Kuipers, and L. W. Hamoen. Stripping Bacillus: ComK auto-stimulation is responsible for the bistable response in competence development. *Molecular Microbiology*, 56(3):604–614, May 2005. (Cited on page 62.)

[70] W. K. Smits, O. P. Kuipers, and J.-W. W. Veening. Phenotypic variation in bacteria: the role of feedback regulation. *Nat. Rev. Microbiol.*, 4(4):259–271, Apr. 2006. (Cited on pages 49 and 61.)

[71] N. Sommerfeldt, A. Possling, G. Becker, C. Pesavento, N. Tschowri, and R. Hengge. Gene expression patterns and differential input into curli fimbriae regulation of all GGDEF/EAL domain proteins in *Escherichia coli*. *Microbiology*, 155(4):1318–1331, Apr. 2009. (Cited on pages 45 and 67.)

[72] N. R. Stanley and B. A. Lazazzera. Environmental signals and regulatory pathways that influence biofilm formation. *Molecular microbiology*, 52(4):917–924, May 2004. (Cited on page 45.)

[73] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800, Oct. 2002. (Cited on page 6.)

[74] R. Thomas, A.-M. Gathoye, and L. Lambert. A Complex Control Circuit. *European Journal of Biochemistry*, 71(1):211–227, Dec. 1976. (Cited on page 62.)

[75] R. Tomioka, H. Kimura, Kobayashi, and K. Aihara. Multivariate analysis of noise in genetic regulatory networks. *Journal of Theoretical Biology*, 229(4):501–521, Aug. 2004. (Cited on pages 27 and 32.)

[76] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, Feb. 2009. (Cited on page 42.)

[77] N. G. Van Kampen. *Stochastic processes in physics and chemistry*. North Holland, 3. edition, May 2007. (Cited on pages 30, 57 and 97.)

[78] M. von Kleist, S. Menz, and W. Huisinga. Drug-class specific impact of antivirals on the reproductive capacity of HIV. *PLoS Comput Biol*, 6(3), Mar. 2010. (Cited on pages 94 and 96.)

[79] M. von Kleist, P. Metzner, R. Marquet, and C. Schütte. HIV-1 polymerase inhibition by nucleoside analogs: cellular- and kinetic parameters of efficacy, susceptibility and resistance selection. *PLoS Comput Biol*, 8(1), Jan. 2012. (Cited on pages 92 and 94.)

[80] H. Weber, C. Pesavento, A. Possling, G. Tischendorf, and R. Hengge. Cyclic-di-GMP-mediated signalling within the σ^S network of *Escherichia coli*. *Molecular Microbiology*, 2006. (Cited on pages 45, 46 and 48.)